A computational model to predict the organisational performance of startups in South African incubators

A research report submitted to the Faculty of Commerce, Law and Management, University of the Witwatersrand, in partial fulfilment of the requirements for the degree of Master of Management in Entrepreneurship and New Venture Creation

Jarryd Jermaine Chengalroyen
0408374p

Dr Diran Soumonni

Table of Contents

TABLES

IMAGES

ABSTRACT

There have been several changes to the global economy in recent history. These are due to numerous factors such as globalisation, advancement in technology, accelerated innovations, and changing trends in demographics. These changes have resulted in the need to improve levels of entrepreneurship. Entrepreneurship plays a crucial role in the improvement of economic growth and development. It also plays a vital role in facilitating poverty reduction, creating employment and structural changes. Entrepreneurship is a tool which can be utilised to improve living standards and general well-being.

Failure rates for new businesses, however, are extremely high. The success of new businesses is a necessary factor to grow the economy. Business failures, particularly for new businesses, are a waste of valuable resources which could be used to grow the economy. Business incubators have been created in order to solve this problem. Incubators add value by combining the entrepreneurial drive of a startup with a plethora of resources usually not available to these under resourced startups.

There have been several models developed to predict the success of startups. For this research, rather than measuring only success or failure, organisation performance was measured. This study creates a computational model, using machine learning, which will be able to predict the organisational performance of start-ups within incubators, based on specific factors. The organisational performance has been defined as a composite of both turnover and number of staff employed.

In order to create the model, a literature review was performed, in which 15 factors were determined as being significant in terms of predicting organisational performance. This was used to create a survey, which was distributed to incubators. There were 103 respondents to this survey. When doing statistical analysis on the results of the 103 respondents, only five factors were found to have statistical significance - age, number of founders, capital rating, professional advisors and education level.

Statistically, the predictability of the initial statistical model proved to be low at 23,8% for turnover and 25,4% for number of staff employed. Using the random forest machine learning algorithm, the predictability was improved to 35,92% for turnover predictability and 40,78% for number of staff employed.

DECLARATION

I, _____,
declare that this research report is my own work except as indicated in the references and acknowledgements. It is submitted in partial fulfilment of the requirements for the degree of Master of Management in the Field of Entrepreneurship at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in this or any other university.

_____

Jarryd Jermaine Chengalroyen

Signed at ……

On the …….. day of …….. …… 2018.

ACKNOWLEDGEMENTS

Performing this research has been a challenging experience, however, this is an exciting area of research which I hope gains momentum. It has grown my interest in entrepreneurship and has ignited a spark within me which I hope to use to bring positive and sustainable changes to the society around me.

My deepest gratitude and appreciate goes out to the following people:

- To my friends, family and business partner for the understanding.
- To my supervisor, Dr. Diran Soumonni, for his guidance.
- To my MMENVC syndicate and classmates at Wits Business School for their support and assistance.
- To my colleagues and managers for their patience.

If you don't build your own dream, someone will hire you to help build theirs.
Tony Gaskins

CHAPTER 1

1.1 Introduction

There have been several changes to the global economy in recent history. These are due to numerous factors such as globalisation, advancement in technology, accelerated innovations, and changing trends in demographics. These changes have resulted in the need to improve levels of entrepreneurship which also plays a crucial role in their improvement of growth and development. The importance of entrepreneurship - the creation of small businesses - has been recognised by governments all over world, acknowledging their contribution to economic growth and job creation (Marom & Lussier, 2014).

Entrepreneurship plays a vital role in facilitating poverty reduction, creating employment and structural changes. Entrepreneurship is a tool which can be utilised to improve living standards and the general well-being of people. Additionally, new venture creation is a highly utilised strategy by government for attaining sustainable national economic growth and development (Salem, 2014).

Failure rates for new businesses, however, are extremely high. The success of new businesses is a necessary factor to grow the economy. The failure of businesses are a waste of valuable resources which could be used to grow the economy. It has proven to be difficult to predict whether business will succeed or fail and even more difficult to understand to what extent they will be successful (Marom & Lussier, 2014).

Since the recent economic recession and the debt crisis, there has been an increased effort to boost entrepreneurship and small business. Business incubators are perceived to be a cornerstone of economic development programs. Incubators add value by combining the entrepreneurial drive of a startup with a plethora of resources usually not available to these under resourced businesses. Among the primary

objectives of business incubators are creation of job opportunities which in turn is expected to facilitate economic growth (Salem, 2014).

This research will examine which factors are most pertinent to increasing organisational performance within startups based in South African incubators. This will be done by creating a computational model examining specific factors based on literature.

1.2 Theoretical background to study

There have been several models developed to predict the success of startups. Lussier developed a 15-variable model to predict failure or success within startups. Mwangi et al. (2013) attempted to find the constructs of successful and sustainable SMEs in East African economies by doing a qualitative study. Benzing et al. (2009) examined the characteristic of Turkish entrepreneurs in order to determine the motivation and factors contributing to their success. Lussier's 15-variable model is among the most popular of these models, which has been utilised in a number of studies in many different countries in order to predict the success of startups. Lussier’s 15-variable model has been extended in a number of studies such as the exploratory model developed by Siow Song Teng et al. (2011).

Human capital factors are a primary constituent of all models which attempt to predict entrepreneurial success. The relation between human capital and entrepreneurship is well documented. Theodore W. Shultz developed a theory relating human capital to entrepreneurship. He famously stated “The man without skills and knowledge is leaning terrifically against nothing” (Shultz, 1961, pg. 16). Human capital is generally shown to have a positive relation to entrepreneurial success. A meta-analytical review of data performed by Unger et al. (2009) of 70 samples over the past 30 years, showed a small but positive relationship between human capital and entrepreneurial success. There are exceptions - Davidsson et al. (2003) showed that none of the human capital variables were associated with obtaining a first sale or being profitable during the

study which they conducted. In this literature, human capital theory will be examined as one of the major factors influencing organisational performance.

1.3 Purpose of study

This research aims to create a computational model, using machine learning, which will be able to predict the organisational performance of startups within incubators, based on specific factors. The primary objective is to determine which factors are most pertinent to increasing organisational performance of startups based in South African incubators. Various factors such as human capital, operational and environmental factors will be analysed to determine their impact on organisational performance. The results of this research could potentially assist both incubator managers and entrepreneurs in understanding which factors lead to improved organisational performance in this specialised startup environment, and assist incubator managers in selecting entrepreneurs to enter the incubator. This research could also allow investors to add a new dimension of evaluation in terms of risks based on past organisational performance. Furthermore, it could also allow investors to suggest changes to be made in the organisation before investment or further investment is considered. Additionally, it may allow entrepreneurs to understand what changes could be made to an organisation in order to increase their performance. Lastly, suppliers could also use it to evaluate risk and limit credit facilities to these clients.

1.4 Context of study

In South Africa, there is a view that incubators have a significant impact on economic growth and development (Lose, 2016). The purpose of an incubator is to accelerate the successful development of entrepreneurial ventures. This is done by providing business support services and resources (Lose and Tengeh, 2015). It is suggested that incubators, as well as the incubatees, each face their own challenges (Ndabeni, 2008). Ndabeni (2008) identified several factors used to determine the success of incubators. This indicates that factors within the incubators themselves

have an influence on the incubated entrepreneur. This research, however, will not consider that. The only thing considered is the particular factors which affect the performance of startups which have entered incubators in South Africa.

Performance within an organisation is a multi-dimensional construct (Unger et al., 2011). The research leans on the organisational performance measures listed by Combs et. al (2006). There are many ways of measuring operational and financial performance. There are two measures upon which organisational performance will be measured in the context of this research - revenue and employment. The manner in which organisational performance is measured will be two fold. Firstly, statistical regression will be used and secondly, machine learning techniques will be explored in order to determine if they lead to improved predictability.

Machine learning was born from pattern recognition. It is based on the theory that computers can learn without being programmed to perform specific tasks. Machine learning algorithms learn from previous computations. This allows them to produce reliable and repeatable decisions and results. Machine learning algorithms have many applications (Alpaydin, 2014). In this research, machine learning will be used within the field of entrepreneurship to test if an algorithm can learn to identify the attributes which lead to positive organisational performance.

The aim of this research is to marry the field of computer science and entrepreneurship theory. The research requires a definition of theoretical relationships between particular sets of variables and organisational performance obtained from examining previous work. The theory will be used to create a set of hypotheses which will be used as predictors. The predictors serve to form the learning data for the machine learning algorithm. Machine learning algorithms can be used in conjunction with the human capital and environmental theory to examine the factors that lead to positive organisational performance of startups within incubators. The information obtained in the research could potentially assist both

technopreneurs and incubator managers in determining which startups are likely to succeed. It could also assist incubation managers in their selection processes.

1.5 Problem statement

The influx of incubators and startups within Africa continent and more specifically, South Africa, proves that this business model is viewed as an exciting and innovative model for entrepreneurs (Kelly & Firestone, 2016). While the adoption of the incubator model is increasing, little is understood as to which factors will lead to successful entrepreneurship upon entering this environment. Models to predict success have been explored previously, but there are limitations to these models. Firstly, the models only predict success or failure while this research will utilise the construct of organisational performance, which will give a wider range of relationships. While other models also typically use regression, none have attempted to use machine learning in order to predict organisational performance. This research attempts to understand which factors, within incubators in South Africa, lead to increased organisational performance by creating a predictive computational model using machine learning techniques.

1.6 Main Problem

There are models which are able to accurately predict business success and failure. This research attempts to explore the possibility of a more accurate computation model formed using machine learning techniques. This would be contrasted to the traditional linear regression model. This computational model would not only attempt to measure whether the business has succeeded or failed but would measure the organisational performance - composed of the turnover and number of employees.

Research Question one - which factors are likely to lead to the success of startups within incubators?

Research Question two - how much more accurate would a machine learning algorithm be at predicting organisational performance than a traditional regression model which measures business success or failure?

1.7 Significance of study

Incubators are increasing exponentially on the African continent. Understanding which factors are most likely to lead to entrepreneurial success is imperative. Although the area of incubation has been explored and the area of predictive models has also been widely examined in different contexts, little research has been done to investigate which factors are pertinent to incubator based startups within sub-saharan Africa. The aim of this paper is to contribute to the understanding of predictive performance within this unique sector. This information could be used by incubator managers and entrepreneurs. Incubator managers could use this information to evaluate whether a startup should be accepted within their hub or whether further refinement or skills are required in order to increase their likelihood of succeeding. Entrepreneurs wishing to be based in incubators could use this information to understand which elements of human capital they need to up-skill on. They could also, in conjunction with the incubator, consider injecting the necessary skills into the startup.

In summary, work in this area has been done before with many models attempting to measure success or failure ratios of businesses. This work is unique in a few regards: firstly, it attempts to measure organisational performance which gives a larger range of values as opposed to measuring success or failure. Secondly, it create a unique model combining elements from the literature. Lastly, it creates a computational model which can be compared to a regression model in order to see which leads to the best predictor.

1.8 Definition of terms

Incubator - an organisation that develops and promotes an entrepreneurial idea from initiation to commercialisation

Machine Learning - a field of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. It gives computers the ability to learn without being explicitly programmed.

Regression - a statistical measure used in many disciplines that attempts to determine the strength of the relationship between one dependent variable and a series of other changing variables (known as independent variables).

1.9 Assumptions

- Participants will answer the interview questions in an honest and candid manner.

- The inclusion criteria of the sample are appropriate and therefore assures a well-balanced sample is chosen.

1.10 Conclusion

In summary, this research will examine which factors are most pertinent to increasing organisational performance within startups based in South African incubators. While research has been done in this area before, this work is unique in a three regards: its attempt to measure organisational performance differs to previous research; it creates a unique model; lastly, it creates a computational model.

CHAPTER 2

2.1 Introduction

The literature review will begin by establishing the context of incubators in South Africa. The human capital element of entrepreneurship theory is then examined. This is done to form theoretical relationships which could be used in the predictability model and creating hypotheses. Other models which have been created in order to predict success or failure of business will then be explored. These studies will be contrasted. General machine learning techniques are then explored, determining which technique is appropriate for this research. This will be followed by a more detailed explanation of the ten machine learning algorithm techniques which have been used in this study. Lastly, the measures of organisational performance used in this research will be discussed.

The hypotheses formed are stated based on previous research on human capital and organisational performance. There are ten machine learning techniques chosen for this research. These are explained in detail. Some of these techniques have been found to be the appropriate technique to use for business data mining by Bose and Mahapatra (2001). Cooper et al. (1994) found that by using a model based on human and financial capital, it was possible to predict the performance of startups with a high degree of confidence. Using this machine learning technique with the learning data obtained from the hypotheses formed, it is then determined if a computational model can be formed.

2.2 Business Incubators

The first generation of incubation models was introduced in the early nineties. The primary focus of these incubators was to provide both physical and financial resource support. New incubation models began to rapidly emerge. Incubators began to provide a wider range of basic services, moving away from the traditional financial and office space support. Services such as market evaluation, knowledge access

services, product development support, networks of entrepreneurs and provision of entrepreneurial finance were offered. Furthermore, there has been a greater shift recently toward knowledge intensive business services, moving almost entirely away from the primary services for which the incubation models (Pauwels et al., 2016). Most models, however, are hybrid models, offering a number of different services.

Business incubators were designed to create self-reliant entrepreneurs. This is achieved by offering a broad spectrum of tailored services. The aim of business incubators is to raise the success rate of small and medium size enterprises (Lose and Tengeh, 2015). Business incubation is fairly recent concept in developing countries including South Africa (Ndabeni, 2008).

According to the Word Development Report done by the World Bank in 2016, more than 117 incubators are currently active in Africa (Kelly and Firestone, 2016). In late 2016, the GSMA Ecosystem Accelerator reported a total of 314 technology specific incubators on the African continent - a higher estimate than the World Bank report. Incubators are becoming increasingly relevant on the African continent and this suggests that they are clearly seen as an dynamic and useful model for supporting technology entrepreneurs.

2.3 South African economy

South Africa is an efficiency-driven economy with a population of 55 million. The current GDP is $313 billion with a GDP per capita of $5,695. The SME contribution to GDP is estimated to be 36% (Harrington. & Kew, 2016). Over the past few years South Africa has experienced a number of socioeconomic setbacks due to political and economic instability.

South Africa slipped into a technical recession after its gross domestic product (GDP) declined 0.7% during the first quarter of 2017 after contracting by 0.3% in the fourth quarter of 2016. However, in the second quarter of 2017, GDP growth recovered and grew 2.5%. This is

the second recession experienced in post democratic South Africa. Negative credit rating decisions from credit rating agencies, including Moody's and S&P have exerted further pressure on the economy (Stats S.A, 2016).

Recent figures from the Quarterly Employment Statistics (QES) indicated that South Africa's formal non-agricultural sector shed 31 000 jobs in the third quarter of 2017. Previously, losses of 41 000 were recorded in March 2017 and 31 000 in June 2017. This shows a total lost of 83 000 formal jobs lost from September 2016 to September 2017 (Stats S.A, 2016).

According to Harrington & Kew (2016) in the GEMS 2016 report, early-stage entrepreneurial activity has dropped considerably. Comparing the 2016 results to the 2015 results, the nascent entrepreneurial rate is down by 30%, while the TEA (total entrepreneurship activity) rate has dropped by 25%. South Africa has substantially lower TEA rates compared to their African counterparts. On average, countries in the African region demonstrate a TEA rate 2.5 times higher than South Africa. South Africa also has one of the lowest TEA rates among the efficiency-driven economies. It is ranked 28th out of 32 countries.

The country's National Development Plan (NDP), endorsed by Cabinet in 2012, is the product of interactions with and inputs from many South Africans. accompanied by extensive research. One of the major aims of the NDP is to target the country's high unemployment rate. The NDP targets a decrease in the unemployment rate from 24.9% in 2012 to 6% by 2030, which requires an additional 11-million jobs to be created. Total employment should rise from 13 million to 24 million (National Planning Commission, 2011). Creating an entrepreneurial economy is key to driving this vision.

2.4 Existing models predicting failure or success of startups

Lussier (1995) produced a model to predict the success or failure of businesses. This is commonly known as Lussier's 15-variable success vs failure prediction model. It is a statistical model. For the purposes of this research, a computational model will be used as opposed to a statistical model.

The model was shown to be statistically significant. The parameters and statistical significance of each variable are shown in the table below:

Table 1. Lussier's Variables

	Variable	Significance	Type
1	Capital	.1189	Positive
2	Record Keeping and Financial Control	.2672	Positive
3	Industry Experience	.3170	Positive
4	Management Experience	.8301	Positive
5	Planning	.0052	Negative
6	Professional Advisors	.0005	Positive
7	Education	.0210	Positive
8	Staffing	.0045	Positive
9	Product/Service Timing	.6616	Positive
10	Economic Timing	.8149	Positive
11	Age of Owner	.9834	Positive
12	Partners	.3851	Positive
13	Parents Owned Business	.0550	Positive
14	Minority	.2123	Negative
15	Marketing	.4085	Positive

The business success versus failure prediction model developed by Lussier had proven to be statistically significant and was able to reliably outperform random classification of a group of businesses (as successful

or failed) over 99 percent of the time. The model has an approximately 70% success rate for predicting success or failure.

Lussier's model is largely based on the work done by Cooper (1979), who arguably created the first prediction model. Many of the constructs chosen as predictors by Cooper (1979) are included in Lussier's 15-variable model. Therefore, in this research, more focus will be given to Lussier's model.

Mwangi (2009) demonstrated constructs that emerged from their qualitative research on already successful SME's in Kenya and Uganda. These results were obtained through interviews and focus group discussions. The study sought to construct an account of leadership practices and ascriptions of success for SME's that had succeeded by interviewing. The analysis of data resulted in sixteen constructs, eight of which were closely linked to the SMEs' success. These constructs could be grouped into the following categories: visioning, building commitment, social capital, personal values, anticipation and resilience, resourcefulness, responsiveness, and entrepreneurial orientation.

Table 2. Mwangi's Variables

Constructs related to SMEs' success	**Constructs related to SMEs' success**	**No. of Open Codes**
Organization	16	54
Experience with relatives	12	44
Employee qualities	13	37
Family support	16	36
Locating self before success	10	33

Constructs related to SMEs' success	Constructs related to SMEs' success	No. of Open Codes
Source of funds	17	31
Reason for starting business	14	28
Challenges in business	8	26

Siow Song Teng et al. (2011) created an exploratory model based on the Lussier prediction model. Using logistic regression analysis, it was found that both the Lussier model and the exploratory model were significant predictors of business success and failure. The exploratory model was slightly more accurate, however, it had 26 variables as opposed to the 15-variable model of Lussier. In this study, the Lussier model accurately predicted success versus failure in 85,6 percent of the surveyed firms while the exploratory model predicted 86.3 percent.

Table 3. Siow Song Teng et al. (2011) Exploratory Model

	Variable	Significance	Type
1	Capital	0,757	Positive
2	Record Keeping and Financial Control	0,382	Positive
3	Industry Experience	0,563	Positive
4	Management Experience	0,477	Positive
5	Planning	0,176	Negative
6	Professional Advisors	0,632	Positive
7	Education	0,113	Positive
8	Staffing	0,074	Positive
9	Product/Service Timing	0,043	Positive
10	Economic Timing	0,349	Positive
11	Age of Owner	0,728	Positive
12	Partners	0,864	Positive

	Variable	Significance	Type
13	Parents Owned Business	0,114	Positive
14	Minority	0,189	Negative
15	Marketing	0,281	Positive
16	Relationship with customers	0,065	Positive
17	Niche product/service area	0,594	Positive
18	Cost of running business	0,771	Negative
19	Technology edge	0,954	Positive
20	Competition from rivals	0,954	Negative
21	Leadership of senior management	0,023	Positive
22	People bonding in firm	0,452	Positive
23	Organizational capability	0,272	Positive
24	Broad access to resources	0,912	Positive
25	Local knowledge of market	0,048	Positive
26	Good government policy	0,191	Positive

Lussier and Corman (1996) did a comparison of variables between the variable included in Cooper et al. (1990 + 1991), and Reynolds and Miller (1987 + 1989) - which were the only two models which were non-financial empirically-tested predictive models. The two models were tested on the same data set. They created a list of 15 observations comparing these variables which give further insights:

1. **Capital** - the only tested significant variable in the four models mentioned about

2. **Record** keeping and financial control - not tested in the Cooper models. Was significant in the Reynolds model.

3. **Industry experience** - significant in the Cooper 1991 model and 0-10 employee model. Not significant in the Cooper 1990 model. Not a

tested variable in the Reynolds models.

4. **Management experience** - not significant variable in the Cooper models. Not tested with the Reynolds models.

5. **Planning** - significant variable in Cooper 1990 and Reynolds models. Not a tested variable in the Cooper 1991 model.

6. **Professional advisors** - significant in the Cooper models. Not tested variable in the Reynolds models.

7. **Education** - significant in the Cooper 1991. Not significant in the Cooper 1990 model or the Reynolds 1989 mode. Not a tested variable in the Reynolds 1987 model.

8. **Staffing** - Not significant in the Reynolds models. Not tested variable in the Cooper models.

9. **Product timing** - significant variable in the Cooper 1990 model and the Reynolds model. Not significant in the Cooper 1991 model.

10. **Economic timing** - significant variable in the Cooper 1990 model. Not in the Cooper 1991 model, and the Reynolds models.

11. **Age of owner** - significant variable in the Cooper 1990 model. Not significant in Cooper 1991, Reynolds 1989 model. Not tested in the Reynolds 1987 model.

12. **Partners** - significant variable in the Cooper 1990 and Reynolds and Miller 1989 model. Not significant in the Cooper 1991 model. Not included in the Reynolds 1987 model.

13. **Parents owned a business** - significant variable in the Cooper 1991 model. Not tested in the Reynolds models or Cooper 1990 model.

14. **Minority ownership** - significant variable in both Cooper models. Not a tested variable in either Reynolds models.

15. **Marketing skills** - not significant variable in any of the models. Rejected as significant in the Reynolds 1987 model and the Cooper models. Reynolds 1989 model did not include it.

The table below has been extracted verbatim from da Silva (2016). It demonstrates which variables have been considered a contributor to failure of startups, across 16 pieces of literature. As can be seen the top 5 factors with regards to failure are: capital, record keeping and financial transactions, industry experience, management experience and planning.

Table 4. Comparison of factors in literature contributing to business

	capt	rkft	inex	maex	plan	prad	deuc	staf	psti	ecti	age	part	pent	miot	mrkt
Bruno	F	F	-	F	F			F	F	F					F
Cooper90	F	-	N	N	F	F	N	-	F	F	F	F	-	F	-
Cooper91	F	-	F	N	-	F	F	-	N	N	N	N	F	F	-
Crawford	-	-	F	-	-	F	F	-	-	N	N	-	-	-	-
D+BSt.	F	F	F	F	-	-	-	-	-	F	-	-	-	-	-
Flahvin	F	F	F	F	-	F	-	F	-	-	-	-	-	-	-
Hoad	-	-	F	N	N	F	F	-	-	-	-	-	-	-	-
Kennedy	F	-	-	F	F	-	-	-	-	F	-	-	-	-	-
Lauzen	F	F	-	F	F	-	-	F	-	-	-	-	-	-	-
McQueen	F	-	F	F	-	-	-	-	-	-	-	-	-	-	F
Reynolds87	F	F	-	-	F	-	-	N	F	-	-	-	-	-	N
Reynolds89	F	F	-	-	F	-	N	N	F	-	N	F	-	-	-

	capt	rkft	inex	maex	plan	prad	educ	staf	psti	ecti	age	part	pent	miot	mrkt
Sommers	-	-	-	F	F	-	-	F	-	-	-	-	-	-	-
Thompson	N	-	-	F	F	-	-	F	F	-	-	-	-	-	F
Vesper	F	F	F	F	N	F	F		F	F		F			F
Wight	F	F	-	F	-	F	-	-	-	-	-	-	-	-	-
Wood	-	F	F	F	F	-	F	-	-	-	-	-	-	-	-
TotalF	12	9	9	11	9	7	5	5	6	5	1	3	1	2	4
TotalN	1	0	1	3	2	0	2	2	1	2	3	1	0	0	1
Total	4	8	8	3	6	10	10	10	10	10	13	13	16	15	12

Independent variables: capt: capital, tkft:record keeping and financial controls; inex: industry experience; maex: management experience; plan: planning; prad: professional advisors; educ: education; staff : staffing ; psti: product or service timing; ecti: economic timing; age: founder age; part: partners; pent: parents; mior: minority; mrkt: marketing.

F supports variables as a factor contributing to failure; N does not mention variable as a contributing factor

2.4 Factors affecting organisational performance

The following discussion will review previous research to understand the theoretical relationships between particular sets of variables and organisational performance.

Human capital theory supports the view that characteristics like managerial experience influence strategic choice and firm performance. Human capital theory also suggests that an experienced entrepreneurial team should be more productive than a less experienced one. Experience is considered to be a valuable asset which has a positive impact on productivity, thus increasing both the economic

value of the firm and the managerial compensation. Management experience allows managers to make more informed strategic choices (Shrader and Siegel, 2007). Baptista et al. (2007) examine the role played by the backgrounds and pre-entry experiences of founders in influencing new firm success, where pre-entry knowledge and capabilities are considered part of entrepreneurial human capital. A number of hypotheses were formed in this research, pertaining to both general human capital and specific human capital. The results in this research have indicated a distinction between opportunity and necessity driven entrepreneurship. It was found that general human capital has a more significant impact on opportunity driven entrepreneurship. Baptista et al. (2007) concluded that people who have worked in the industry where they are founding a new firm, and who have had at least some managerial experience, embody the entrepreneurial human capital characteristics more likely to contribute to the early success of startups (Baptista et al., 2007). Entrepreneurs who have not worked in industry or non-profit organisations may have less opportunity to develop skills relevant to managing a business, implying that their ventures are more likely to not perform well. Level of management experience could prove to be useful in a business. Entrepreneurs who have previous experience in supervising managers or managing a business should have a higher chance of being successful. The breadth of experience would leave those entrepreneurs better prepared to tackle the wide range of challenges confronting new ventures (Cooper et al., 1991). However, Cooper et al. (1994) found that management level did not have significant effect on performance. Nonetheless, this leads to the first hypothesis:

H1: Managerial experience has a positive effect on organisational performance.

Strategies embraced by new ventures are often paralleled with specialised experience in functional areas. Functional experience enhances the firm's implement specific and innovative strategies. A venture pursuing a strategy of differentiation based on innovative new products would likely

benefit from the specific type of human capital developed though years of experience in technical jobs. This experience would likely be accompanied by very specific insights and expertise which would enhance the ventures technical skills which could positively impact the product features (Shrader and Siegel, 2007). Shrader and Siegel (2007) have assessed the role of human capital in the growth of new technology ventures. Their results suggested that strategy and team fit were strongly linked to long term performance of high-tech entrepreneurial endeavours. The study, based on human capital theory, indicates that managers with more knowledge and experience have an advantage in terms of helping firms successfully adapt to new technologies. This was found to be particularly useful to high-technology entrepreneurial firms. The study also found that for smaller firms, the technological experience of the team appeared to be the most contributing factor (Shrader and Siegel, 2007). This leads to the second hypothesis:

H2: Technical expertise is positively associated with organisational performance.

Human capital basically comprises of a stock of knowledge and skills that resides within individuals. It can be developed and/or transferred between individuals. This is a differentiating characteristic between human capital and characteristics such as personality traits (Wright et al., 2007). Formal education is a component of human capital which plays a role in assisting in the accumulation of explicit knowledge - a skill which may provide skills useful to entrepreneurs. It indicates that individuals have a learning aptitude and organisation skills which better allow them to exploit opportunities. It enhances problems skills and is also likely to lead to an increased social network (Baptista et al., 2014). Empirical research has demonstrated many associations between education and entrepreneurial success. Education demonstrates a non-linear effect in supporting the probability of becoming an entrepreneur. It has also been shown to have a similar relationship in terms of achieving entrepreneurial success. Studies have indicated that returns are dependent on two factors - the level of education as well as the type of

industry (Davidsson and Honig, 2003). Davidsson and Honig (2003) examined nascent entrepreneurship by comparing individuals engaged in nascent activities with a control group. Their findings indicated that formal education played a role in predicting which type of individuals would take up entrepreneurship, however, they also found that formal education did not play a factor in determining the success of the exploitation of this process.

Marvel and Lumpkin (2007) investigated the effect of human capital on innovation radicalness within incubators. For human capital attributes such as education and depth and breadth of experience, it was found that the greatest difference between the innovation groups was among formal education and depth of experience - both of which were positively associated with innovation radicalness. Prior technology knowledge alone played the largest role in innovation radicalness. The empirical results of this research indicate that for small technology based startups, the technological experience of the team appears to be the most important determinant of the success of a differentiation strategy regardless of the strategy employed (Marvel and Lumpkin, 2007). Studies done by Van der Sluis et al. (2005) have concluded too, that entrepreneurial performance, irrespective of what measure is being used has a positive association to formal education. One can therefore, expect a founder with formal education to have a higher probability of increased performance (Baptista et al., 2014). This leads to the third hypothesis:

H3: Higher levels of education have a positive effect on organisational performance.

Brüderl et al. (1992) defined a proponent of human capita, specific human capital, to include work experience and industry-specific experience (Baptista et al. 2017).

Cassar (2014) performed an empirical investigation into the role of both industry and startup experience in forecasting the performance of 2304

startup entrepreneurs. Cassar (2014) demonstrated that that industry experience is associated with more accurate entrepreneurial expectations. It was also found that high technology industries showed a greater benefit.

Bosma et al. (2004) attempted to investigate “To what extent does investment in human and social capital, besides the effect of “talent”, enhance entrepreneurial performance?” (Bosma et al, 2004, pg. 1). Using Dutch longitudinal data set of firm founders, their findings were that investment in industry-specific human enhances performance, irrespective of the performance measure used.

Industry-specific professional knowledge, managerial and entrepreneurial experience were found to positively affect firm size, when the start up size of 391 Italians firms were investigated by Colombo et al. (2004). The specific component of human capital which had the most positive effect on initial firm size is industry-specific professional knowledge and managerial and entrepreneurial experiences. This was proxied by education and general working experience (Colombo et al., 2004).This leads to the fourth hypothesis:

H4: Previous work in the same industry has a positive effect on organisational performance

Cooper et al. (1994) examined predictors that could be used to predict performance of startups by examining whether ventures fail, marginally survive or grow. Human and financial capital were divided into four categories - general human capital, financial capital, management know-how and industry specific know-how. Cooper et al. (1994) found that education generally had a positive relation to organisational performance as discussed above. They also found that racial minorities and gender roles played a role in the probabilities of survival. Female entrepreneurship has risen over the past 10 years, however, it is still the case that most entrepreneurial activity is generated by male entrepreneurs. There has been research performed into observing the gender gaps in performance. Findings have varied, with certain scholars finding no relationship while

others have demonstrated mixed findings (Johnsen and McMahon 2005; Watson 2012). Many of the studies, however, have indicated that female owned firms underperform compared to male owned firms (Lee and Marvel, 2014). Two perspectives can be adopted when examining male and female perspectives with regards to entrepreneurship - liberal feminist theory and social feminist theory. Liberal feminist theory takes the view that female entrepreneurs demonstrate lower performance due to systematic discrimination and limitations which are pressed upon females. Social feminist theory, on the other hand, takes the view that males and females are inherently different (Lee and Marvel, 2014). The approach taken may be fundamentally different and this may or may not be more effective (Fischer et al. 1993).

Rosa et al. (1996) surveyed 300 women and 300 men who owned Scottish and English small businesses as part of a three year study to investigate the impact of gender on small business management. Their results suggest a complex relationship between business performance and gender. They determined that gender appears to be a significant determinant even with other key factors being controlled for. One of the arguments put forth by Rosa et al. (1996) is that women may show lower performance in many financially related indicators such as turnover, jobs created and profitability. This may be due women entering businesses for fundamentally different reasons than men. Women often enter business in order to increase flexibility and independence, normally interfacing family and work commitments (Rosa et al., 1996). Cooper et al. (1994) also noted that startups headed by female entrepreneurs have had mixed findings - Sexton and Robinson (1989) reporting poorer performance while Kalleberg and Ieicht (1991) found that businesses headed by women performed no differently to those of men.

Another factor found to be a determinant to performance was having parents who own a business. This was found to have a positive impact on survival but not necessarily growth (Cooper et al., 1994). As parents are often seen as role models, entrepreneurs may also be more

attracted to entrepreneurial activities if their parents had previously owned businesses (Evans and Leighton,1989; Ziegler,1992). Growing up in a family where entrepreneurship is seen as a viable career path, increases the likelihood of that path being pursued. Therefore, entrepreneurs who have an intrinsic motivation to be involved in entrepreneurial activities, are more likely to obtain a higher psychic income from entrepreneurship than those who do not have those backgrounds and motivations (Gimeno et al., 1997).

The number of partners in a venture also contributed significantly in terms of growth of these ventures. Generally, it was found that management plays an insignificant role when measuring failures and marginal survival but play a more insignificant role with regards to growth (Cooper et al., 1994). According to Cooper et. al (1991) an increased number of partners leads to a greater breadth and depth of expertise. According to resource-dependence theory, an increased number of partners can also be viewed as a means of adding more resources into the venture. The credibility and legitimacy of the business, as viewed by other constituents, is enhanced as a result of more partners. (Teach, et al., 1986). Furthermore, an incorporation requires a considerably higher initial investment than which generally shows a higher level of commitment. This larger financial investment implies larger risk which in turn should require additional planning and more consideration before launching the firm.

According to Cooper et al. (1994), Sexton and Robinson (1989) had reported that ventures started by minority groups were more likely to generate lower earnings. Given history, one might assume that minority entrepreneurs have had fewer opportunities to develop relevant experience, have fewer contacts who can provide assistance, and have greater difficulty in assembling resources (Sexton and Robinson, 1989).

For this purposes of this research, we hypothesise that firms started by women and by minority groups would do less well. This leads to the following four hypotheses:

H5: Businesses with a male founder display higher levels of organisational performance over female entrepreneurs.

H6: Businesses founded by a minority group individual display a higher level of organisational performance than majority group entrepreneurs.

H7: Businesses with founders whose parents have their own business display a higher level of organisational performance than those who do not.

H8: Businesses founded by entrepreneurial teams display a higher level of organisational performance than businesses with a single founder.

The amount of capital raised should be positively associated with venture survival. A greater pool of capital permits more ambitious strategies as well as flexibility for surviving mistakes or buying time. Ventures with higher initial capital may also reflect greater planning and the approval of investors and lenders (Bruno and Tyebjee, 1984).

Davila, Foster, and Gupta (2003) examined the association between the venture capital and the growth of startups - whether venture capital leads to revenue growth and growth in terms of employees. The finding was that an increase in venture capital led to an increase in growth. The study found support for past growth being a predictor of future growth. Lastly, the study found that the increase in venture capital led to an increase in employees. This leads to the ninth hypothesis:

H9: An increase in capital leads to an increase in organisational performance.

Marketing skills prove to be a much needed skill for startups. Benedetto (1999) found that effective product launches are a key driver of top performance. He found that successful launches were strongly related to perceived superior skills in marketing research, distribution, R&D,

promotion and engineering and sales force. Lussier and Corman (1996) state that Cooper et al. (1990 + 1991) and Reynolds and Miller (1987 + 1989) have included marketing skills in their models. However, they were found to have no significant impact. Boag (1987) created a framework to examine the issue of marketing control from a theoretical as well as an empirical perspective. The findings indicated that a greater control of marketing operations correlated with a stronger market performance. This leads to the tenth hypothesis:

H10: Business owners with strong marketing skills show an increase in organisational performance.

Huck and McEwen (1991) found customer relations as well as technical knowledge to be the competencies most pertinent to the success of small business. Ghosh et al. (2001) attempted to determine the strategy dynamics and key success factors (KSFs) for excellence in performance. They found that many businesses believe that customer relationships are more important than a good product/service. They also listed a good customer relationship as being one of the key factors to performance excellence in businesses. This leads to the eleventh hypothesis:

H11: Businesses with good customer relations have an improved level of organisational performance.

Previous studies of new firm performance have found substantial differences by industry, with retail firms having the highest discontinuance rates (Reynolds, 1987). As such, we would expect new ventures in retail and personal services, both of which are characterised by low entry barriers and high turnover rates, to do less well.

H12: Businesses that receive a higher level of professional advice have an improved level of organisational performance.

They hypothesis as developed by Lussier (1995) stated that businesses that do not keep updated and accurate records and do not use adequate financial controls have a greater chance of failure than firms which do. da Silva (2016) found that the relationship between financial controls and startup success were not in line with the expectations - this was contradictory to the hypothesis made by Lussier (1995). da Silva (2016) showed, looking at past research that out of 16 studies, recording keeping and financial control were found to be a contributing factor to failure in 9 of the studies. While

H13: Businesses that demonstrate high level of financial control have an improved level of organisational performance.

Staffing problems are often faced by small businesses. Staffing was found to be a significant factor in the majority of business success and failure studies (Hyder and Lussier, 2016). Small businesses usually experience a high turnover in staff. Amongst the many reasons, a relatively lower wage and higher workload are considered to be the main factors. Informal sectors are normally not subject to labour laws and a large number of workers are paid a daily wage rate. In the study performed by Saha (2006), rigid labor laws were found to have a negative effective on the attempt by India to deregulate the industrial sector. The recommendation was that flexible laws should be implemented in order to reduce high employee turnover, thus creating a more employee friendly environment. In the study performed by Hyder and Lussier (2016), anecdotal evidence suggested that lower wages were the main cause of staffing issues. This could be found to be strongly related to the amount of capital which the startup has raised. This leads to the fourteenth hypothesis:

H14: Businesses that have less difficulty obtaining staff have an improved level of organisational performance.

Lussier and Pfeifer (2000) hypothesised that younger people who start a business have a greater chance of failure than older people. Evidence supporting this factor was found in Cooper (1990) and Reynolds (1987). However, many studies have also found that age is not a factor which affects the performance of a business. Bosma et al. (2004) found that age appears to have no effect on the performance measures. Gimeno et al. (1997) stated that an in-depth study by Mayer and Goldstein (1961) found that job seeking at an older age is what encourage older entrepreneurs to continue on a path of entrepreneurship despite the business having very marginal economic performance. Also noted was that if age were to serve as a proxy for general human capital, then it also may be linked to the monetary performance of a venture. Regarding all of the above mentioned considerations, one may expect older entrepreneurs to perform differently from younger ones. This leads to our 15th and final hypothesis:

H15: An increase in the average age of founders of the startup has a positive effect on organisational performance.

2.6 Machine Learning

Lussier (1995) and Cooper et al. (1994) have used statistical models aiming to predict either organisational growth or the success or failure of organisations. The purpose of this research is to use a computational model as opposed to a statistical model. Machine learning will form the foundation of that computation model.

Machine learning is a branch of artificial intelligence. It is concerned with the construction and study of systems that can learn from data (Alpaydin, 2014). Machine learning employs a variety of statistical, probabilistic and optimisation techniques that allows computers to “learn” from past examples and to detect hard-to-discern patterns from large, complex data sets (Cruz and Wishart, 2006).

There are many techniques for machine learning such as artificial neural networks, case-based reasoning, decision trees, rule induction, genetic algorithms and induction logic programming. Each of these techniques are unique in terms of accuracy, explanatory value and configurability (Mair et al., 2000).

Data mining applications in particular areas have been examined by Bose and Mahapatra (2001), where it was found that rule induction is the most suitable machine learning technique with regard to business data.

Statistical models have certain distributional hypothesis that financial statement data do not always fit. Hence some non-parametric techniques have been developed to overcome the constraints of traditional statistical models. Most of them belong to the data mining domain such as artificial intelligence. Most researchers dealt with the issue of comparing data mining methods with traditional statistical models.

A decision has been taken in this research to test a number of machine learning techniques in order to test which gives the most accurate results. Two forms of machine learning will tested: decision tree learning and support vector machines. Decision tree learning is a method commonly used in data mining whereby a model that predicts the value of a target variable based on several input variables
Looking further at decision trees, rule induction employs a decision tree or a set of decision rules. These are formed from training examples with a known classification. If these examples belong to two or more classes, then the most discriminating attribute is selected and the set is split into multiple classes. This process of attribute selection and splitting is continued until each terminal node represents a different class of examples. The decision tree formed as a result of this is then applied to a test data set to evaluate its accuracy in classifying new examples. Overfitting a decision tree to a training data set may cause its classification accuracy with new data to diminish. The tree must then be

pruned to eliminate overfitting before it is deployed in a real life application (Bose and Mahapatra, 2001).

Rule induction has characteristics which make it an appropriate technique for developing data mining applications in business. It is able to process large data sets with high predictive accuracy and is well suited for classification and prediction tasks. There are also many tools which can be used to implement this technique (Bose and Mahapatra, 2001).

2.7 Machine Learning Usage

Many scholars have researched the prediction of bankruptcy. The application of data computational algorithms such as neural networks, support vector machines may often fit data well, however, there is a lack of clarity of comprehensibility which make these methods less prominent.

Zhao et al. (2009) used these various techniques to mine bankruptcy data. One of the aims was to compare the accuracy fo the data. Their findings indicated that decision trees were relatively more accurate compared to neural networks and support vector machines, however, had more rule nodes. Adjustment of minimum support yielded more tractable rule sets.

The vast majority of studies in this domain have focused on neural networks, comparing them to statistical methods such as regression. There is, however, an issue with neural networks - they are largely seen as being blackboxes. Therefore, they lack transparency, comprehensibility and transportability. Decision trees do not suffer from those issues (Zhao et al., 2009).

Sun and Wu (2010) used classification and regression trees in order to predict business failure - a critical task for investors, government officials, stock holders, managers, employees, investors and researchers. They stated that in comparison with the other classification mining methods, there are many advantages in the use of classification and regression tree

methods. Amongst these are simplicity of results, a high level of accuracy and the implementation can be quite simple and it is non parametric. Sun and Wu (2010) concluded that the predictive accuracy and significance of classification and regression test generally outperformed statistical method by at least 5%. They deduced that this method was applicable to the area of business failure prediction.

2.8 Machine Learning Algorithms

Weka is the tool of choice to run machine learning algorithms. Weka workbench offers a variety of algorithms and tools for data analysis, clustering and predictive modelling. This is wrapped in a simple-to-use user interface, making it easy to access functionality. Weka is freely available and well documented. It is also available across a range of platforms. Weka offers several standard data mining tasks - classification, association, clustering , feature selection and preprocessing. The focus for the purposes of this paper will be classification (Aher and Lobo, 2011).

Classification, also known as supervised learning, is a data mining task that maps the data into predefined groups and classes. It is used to model continuous-valued functions, i.e. it predicts unknown or missing values. In this case the model is meant to predict the organisational performance of the startup (as defined as turnover and number of staff) based on 15 dependent variables. Classification consists of two steps:

1. **Model construction**: The data is given and the classifier rules are determined based on the data. It consists of set of predetermined classes. The data given is used for model construction as a training set. The model is then represented as a set of classification rules, decision trees or formulae.

2. **Model usage**: Here there is an existing model - that is an existing set of classification rules, decision trees or formulae. The model is then used to classify future or unknown objects. The known label of test sample is compared with the classified result from the model.

The model then has a percentage of correct and incorrect classifications as performed by the model. In these cases, the test set is normally an independent training set so as to avoid over-fitting.

The following algorithms will be used for testing the predictability of the model:

2.8.1 Zero R

This is a rule based classifier. ZeroR is the simplest classification method which is available on Weka. ZeroR classifier predicts the majority category in the training data. For numerical prediction problems it will choose the average. There is no predictability power in ZeroR, however, it used as a baseline measure to which the other algorithms or classifiers may be compared to (Witten et al., 1999).

2.8.2 Decision Trees

A decision tree is a predictive machine-learning model. A decision tree consists of a root, decision nodes and terminal leaves. The nodes of a decision tree denote attributes while the branches between each of the nodes denotes the possible values that these attributes could contain. The terminal nodes, i.e. the leaves, will determine the final value (classification) of the dependent variable. The process of creating a decision tree starts at the root and is repeated recursively until a leaf node is reached. This is the final or emergent classification. There are many different types of decision trees such as J48, random trees and random forests (Aktan, 2011). Decision trees are a nonlinear architecture. They are able to discriminate nonlinear patterns and do not have any distributions constraints. The results are very easy to determine and analyse. There is not much preparation required on the initial data set and it performs well with large data (Aktan, 2011).

J48 Decision tree - The J48 Decision tree classifier follows a simple algorithm. The basic idea behind yr algorithm is to divide the data into range based sets on the attribute values for that set found in the training

sample. The classification works as follows: if a new item requires classification, a decision is created based on the attribute values of the training data. When another item requires classification, the attribute which discriminates the instance is identified. This feature that is able to tell us most about the data instances so that it can be best classified, is said to have the highest information gain. While recursing through the values, if there is any value for which there is no ambiguity – meaning there is no data instances for which its category has the same value as the target variable – then that branch is terminated and the branch is assigned to the new value. J48 creates a binary tree which is very useful in classification problems (Patil and Sherekar, 2013).

Decision Stump - A decision stump is a machine learning model which uses a decision tree of one level (Iba and Langley, 1992). The tree is very simple - one root which is connected to every terminal nodel - hence the word stump. The split from the root is based on the attribute value pair. A decision stump makes a prediction based on the value of a single input feature (Zhao and Zhang, 2008).

Random Forest – This is part of the ensemble learning algorithm set, and is considered to be more accurate and robust than single noise classifiers (Breiman, 1996; Dietterich, 2000). The premise of ensemble learning algorithms is that a set of classifiers outperforms a single classifier. The random forest classifier was introduced by Breiman. It was considered to be very promising due to its high level of efficiency and its ability to handle a high number of input variables. A random forest operates by constructing a combination of decision trees classifiers and outputting the most common classification or average classification of the individual trees (Rodriguez-Galiano et al., 2012). Each classifier is generated using a random vector which has been sampled independently. Random forest trees are able to avoid overfitting of the data set.

2.8.3 Decision Table

Decision tables, like decision trees, are a set of machine learning algorithms used as classification models for prediction. A decision table works as follows: there exists a hierarchical table. Each entry in the hierarchal table is broken down by the values of a pair of additional attributes to form another table. The structure is similar to dimensional stacking (Becker, 1998).

2.8.4 Adaptive Boosting

Adaptive Boosting, more commonly known as AdaBoost, is a machine learning algorithm which can be classified as a meta algorithm. It was formulated by Freund and Schapire (1997). In basic terms, AdaBoost can be used in conjunction with other machine learning algorithms in order to improve performance. The output of the other learning algorithm may be used in conjunction with this algorithm in order to create a weight sum. That weighted sum represents the final output of the classier (known as the boosted classifier). This will improve the accuracy of the single model in general. AdaBoost is not known to overfit with very large models (Hastie et al.,2002).

2.8.5 Bagging

Bagging is an acronym from Bootstrap AGGregatING. Bagging was formulated by Beirman in 1996. It is an ensemble machine learning algorithm. Bagging trains each classifier on a random redistribution of the training data set. Each sample consists of many patterns. Each training sample trains one of the classifiers. However all classifiers in the ensemble would be used, as they are combined using averages or weighted majority. Training samples are chosen at random; this selection process means that some patterns may be chosen multiple times while other are never chosen at all. Bagging is therefore considered to be a selected or stable learning algorithm - similar to decision trees or neural networks (Tharwat et al., 2016).

2.8.6 Support Vector Machines

Support Vector Machines (SVM) was formulated by Vapnik (1995). It takes a blended approach - using linear modelling and instance based learning. A number of critical boundary points, known as ‘critical boundary instances’ are selected from each class. These are used to form a linear discriminant function which separate each class as widely as possible. SVM is intended to use the linear model in order to represent non-linear class boundaries. Like neural networks, SVM suffers from the constraint of statistical distributions. SVM is not prone to overfitting and can produce very accurate classifiers (Aktan, 2011).

2.8.7 Naive Bayes

The Naive Bayes is the simplest implementation of the Bayesian network. It is based on the so-called Bayesian theorem. All attributes are independent of the value of the class variable. The conditional independence assumption is rarely met in real world examples. The Naive Bayes classifier does not really produce predications but rather produces probability estimates. A probability estimate can be considered to be the conditional probability distribution of the values of the class attribute. This class attribute is based on the values of the other attributes. Naive Bayes can be seen to be an alternative manner in representing conditional probability distribution (Witten and Frank, 2005). The Bayesian Network is a well-founded manner in which probability distributions can be represented concisely in a graphical manner (Aktan, 2011).

2.9 Organisational Performance

Performance within an organisation is a multi-dimensional construct (Unger et al., 2011). This research is largely based on the organisational performance measures listed by Combs et. al (2006). There are many ways of measuring operations and financial performance (Unger et al., 2011). The measures used in the context of this research will be revenue and employment.

In South Africa, the unemployment rate in Q4 2016 is listed at 26,5% (Statistics South Africa, 2016). Creating employment is therefore arguably the most important factor when considering entrepreneurial success. However, employment without positive revenue and cash flow is unsustainable. The measure of profitability was not opted for. Due to the early capital investments required in a startup environment, profits may not be a good indication of entrepreneurial success. The study will evaluate positive revenue and cash flows instead.

According to GEMS (2015), the percentage of early-stage entrepreneurs who expect to generate no jobs within the next five years has decreased considerably from 30% to 14%. The average over the African region is roughly 35% while the average for efficiency driven economies is 46%. This is a positive metric for South Africa.

The percentage of early stage-entrepreneurs who expect to create between one and five new jobs within the next five years is around 60%, a percentage which has grown steadily over the past few years. The reason behind this increase in expected job growth may be linked to a marked increase in opportunity-driven entrepreneurship and activities. South Africa has a low established business rate and job creation must be viewed in that context (GEMS, 2015)

In this study, turnover and employment will also be used in indicating organisational performance in terms of marginal survival and growth (Read et al., 2009).

Marginal survival will be based on two factors measured over the life span of the venture - turnover and employment. Where both employment and turnover are < 5%, this will be defined as marginal growth. Similarly, where employment and turnover > 5%, this will be considered growth.

2.9.1 Predictors for organisational performance

Given the literature pertaining to human capital the following hypotheses can be drawn indicating an impact on organisational performance. These hypotheses will be used to generate surveys, the responses of which will form the training data set for the machine learning algorithm. To summarise, the following hypotheses are considered when gathering input data for our machine learning algorithm.

H1: Higher levels of education have a positive effect on organisational performance.

H2: Technical expertise are positively associated with organisational performance.

H3: Managerial experience has a positive effect on organisational performance.

H4: Previous work in the same industry has a positive effect on organisational performance

H5: Businesses with a male founder display higher levels of organisational performance over female entrepreneurs.

H6: Businesses founded by a minority group individual display a higher level of organisational performance than majority group entrepreneurs.

H7: Businesses with founders whose parents have their own business display a higher level of organisational performance than those who do not.

H8: Businesses founded by entrepreneurial teams display a higher level of organisational performance than businesses with a single founder.

H9: An increase in capital leads to an increase in organisational performance.

H10: Business owners with strong marketing skills show an increase in organisational performance.

H11: Businesses with good customer relations have an improved level of organisational performance.

H12: Businesses that receive a higher level of professional advice have an improved level of organisational performance.

H13: Businesses that demonstrate high level of financial control have an improved level of organisational performance.

H14: Businesses that have less difficulty obtaining staff have an improved level of organisational performance.

H15: An increase in the average age of founders of the startup has a positive effect on organisational performance.

2.10 Conclusion of literature review

The literature review established the context of the incubator space in South Africa. The human capital element of entrepreneurship theory was examined with the purpose of testing theoretical relationships which will be used to form surveys. The results of the surveys will be used as machine learning data input for the machine learning algorithm. General machine learning and machine learning techniques were explored. The machine learning technique chosen for this research was rule induction. This was chosen found to be the appropriate technique to use for business data mining by Bose and Mahapatra (2001) due its ease of use and ability to have high predictive accuracy. Lastly, the measures of organisational performance which also will impact the learning data were discussed. The measures in this case would be revenue and

employment numbers. The venture outcomes will be also be measured to indicate organisational performance.

CHAPTER 3

3.1 Introduction

This chapter presents the research design and methodology that was followed for the study. The research design includes the research approach and paradigm. The methodology of the study examines the population, the sampling procedures, procedures for data collection and data analysis.

3.2 Research Paradigm

The position of this study, based on the beliefs, assumptions and perceptions of the world which determine the choices made in the research, follows the research paradigm of positivism. Positivism is the dominant paradigm used in research (Chell, 2013). The aim of this study is to focus on the theoretical element of indicators of organisational performance by establishing the relationship between performance and predictors. The relationships established are based on existing research as covered in the literature review. The nature of this relationship is further examined in the context of this study by using quantitative means. The information retrieved by those means are then plugged into a structured research instrument - a machine learning algorithm.

3.3 Research Design

This research will use the quantitative method of research. A survey will be used to capture information. The survey will be used to capture a cross-sectional time for a particular business as a particular time. Similar studies related to measuring organisational performance have been conducted by Carlos et al. (2014), Ejere and Abasilim (2013) and Birasnav (2014).

The rationale behind using a survey was that through using this approach, the research is able to quantify the extent to which the factors under consideration affect organisational performance as well as show if

the factors are related. Using this research design, questions were pre-set, allowing the content of the research to be controlled and easily measured.

While the survey is cross-sectional, there will be questions which require data over a period of time. This survey data will be used to test the relationship as defined by the literature review. The results of the survey will be prepared as input data for the machine learning algorithm.

The research seeks to identify the factors pertinent to organisational performance. Looking at the nature of the questions posed, one is able to see that this is a perception study. The research does not seek to show inference or causality. Future longitudinal studies could be performed in order to measure the organisational performance of incubatees.

3.4 Population

The population of this research was South African incubatees. An incubatee refers to any company which was in incubation at the time of responding to the survey. In order to be eligible for this study, the

Table 9. Respondents by Incubator

Incubator	Number of respondents	Percentage
Innovation Hub	18	17,48
Shanduka Black Umbrellas	56	54,37
Riversands	20	19,42
LaunchLab	5	4,85
JoziHub	3	2,91
SoftStart	1	0,97

startups must be founded in South Africa and presently be in an

incubator. The founders, however, are not required to be South African. The sample does not discriminate between any age, gender, race groups or status. The sample also does not discriminate on the incubators themselves or the length of time the startup has been in the incubator.

The incubators used in this study may fall within any region (province) of South Africa. Most incubators in the sample reside in the Gauteng region. Gauteng, is the most highly populated province in South Africa. It hosts 24,1% of the South African population with an estimated 13 498 200 residents, according to the mid-year population estimates of 2016 (Stats S.A, 2016). Gauteng's municipalities contribute to about 30% of the national GVA (Gross Value Add) and have residents with much higher income on average than the rest of the country (National Treasury, 2016).

Table 5. Incubators in Barrett (2016)

	Innovation Hub	**Shanduka Black Umbrellas**	**Riversands**	**LaunchLab**	**JoziHub**	**SoftStart**
Type of incubator	Gauteng provincial government	Non-profit enterprise programme	Non-profit enterprise programme	Non-profit enterprise programme	Gauteng provincial government	Non-profit enterprise programme
Mandate	Advance innovation	Advance enterprises	Blacked owned businesses	Hi tech incubator	Hi tech incubator	Hi tech incubator
Location	Science and technology park, Pretoria	Johannesburg, South Africa	Johannesburg, South Africa	Cape Town South Africa	Johannesburg, South Africa	Cape Town South Africa

	Innovation Hub	Shanduka Black Umbrellas	Riversands	LaunchLab	JoziHub	SoftStart
Entry criteria	Must be innovative. Must have future pipeline for expansion. Entrepreneur must understand industry, have networks. Requires viable business case. Once first product is sold may be incubated.	Intense training and mentoring must take place for three months. Founder must then pitch to panel and panel will decide if they may successfully enter full incubation program.	Preferably black owned. Must contribute to economic growth.	Help build a viable business through coaching and mentorship programmes. Partner with universities.	Must be innovative and sustainable. Contribute to economic growth.	Must be innovative. Must have future pipeline for expansion.

The population of this sample will include South African ventures and startups within identified incubators. Due to the limited time available for this project, the project will make use of incubators who list their clients publicly. All of these have been listed in Table 4.

3.5 Sample and sampling method

All participants currently in the incubation phase or pre-incubation phase in any of the identified incubators, will be eligible to be a part of the sample. A venture which is able to meet all criteria in order to join these incubators will meet all criteria required for this research.

A simple sampling method will be utilised in this research. In this case each business within an incubator is chosen entirely by chance and

each member of the population has an equal chance, or probability, of being selected.

3.6 The Research Instrument

Information will be captured from the incubatees via a questionnaire. The data from the questionnaire will serve as input to the machine learning algorithm. Machine learning requires as a large a data set as possible. Questionnaires are an efficient and easy way to collect and administer a large amount of information from a large number of respondents. The main advantage is that responses are gathered in a standardised manner. This allows the data to be analysed and in this case prepared for the machine learning algorithm. Questionnaires are also a relatively quick way to collect information (Kwong, E. and Lee, W.B., 2009).

The disadvantages of using a questionnaire is that the return rate may be low. It is also not possible to probe for information and you are limited to the responses given without further insight. Conducting interviews is a more suitable format for eliciting information regarding knowledge and opinions (Kwong, E. and Lee, W.B., 2009). Analysing this data, however, is difficult and time consuming for a large number of subjects.

Based on previous research findings related to quantitative research, using questionnaires to create input data to create computational model using machine learning, has proven to be an effective method.

3.7 Measuring tool

The purpose of this research it to create a computation model to predict performance in organisations. This will be achieved by utilising a machine learning algorithm. Machine learning is being widely explored for the creation of computation models such as (Yu et al., 2014.) created a model for the prediction of bankruptcy. Delen et al. (2013) used machine learning tools to examine knowledge management practices in terms of both financial and non-financial performance. Geng et al.

(2015) used data mining techniques to build models for predicting financial distress of 107 Chinese listed companies using 31 predictors (classifiers). This research will perform this by utilising an existing implementation of an algorithm which is deployed in a cloud solution. There are a number of solutions available as of May 2017:

Table 6. Machine Learning Services

	Cloud Based	Open Source	Paid for Service
Google Cloud Machine Learning	Yes	No	Yes
BigML	No	No	Yes
TensorFlow	Yes	Yes	Yes
Amazon Machine Learning	Yes	No	Yes
Weka	No	Yes	No

The platform selected from the above list is the Weka Platform. This was on the basis that it was well documented, simple and was easily accessible. The advantages of using machine learning as a modelling technique is that it is flexible enough to handle complex problems with multiple interacting elements and outcompete other statistical models. This makes it ideal for research in complex ecological systems. The disadvantages of using machine learning to create a model is that it may require a large amount of data and the amount of data required in order to increase its accuracy is unclear. The input data is also required to be representative of the population, otherwise the model will be inaccurate. A further disadvantage is that it is not as transparent as other statistical models and the ability to understand how each variable is weighted is complex (Olden et al., 2008).

The Waikato Environment for Knowledge Analysis (WEKA) is a set of Java libraries which implement state-of-the-art machine learning and

data mining algorithms. It is accompanied by a full document with the details of all of the algorithms it contains.

The objectives of WEKA are to:

- make ML techniques more highly available
- apply ML to practical problems
- create new machine learning algorithms
- contribute to a theoretical framework

WEKA is a collection of machine learning algorithms for data mining in Java. It was created in Waikato University of New Zealand. WEKA is capable of performing data preprocessing, classification, regression transformations and clustering. It demonstrates certain tree algorithms visually which can be useful. It allows one to switch between different machine learning algorithms on the same data set, allowing for the comparison between algorithms (Loukeris and Matsatsinis, 2006).

For machine learning, the input required can be divided into the following:

- Training set: The input data used for learning
- Validation set: The input data used for fine tuning
- Test set: The data used to classify performance of the ML algorithm

With regard to machine learning, it is understood that the more data available, the more effective the computational model. It has been found that data size is more effective than adding more predictors (or classifiers). This was supported by Brants et al. (2007), where a new method dubbed ‘Stupid Backoff’ was introduced. This indicated that it is inexpensive to train on large data sets and approaches higher quality as the training data increases.

In this study, the minimum data set size for training will be deemed to be 100. Cross-validation will be used for the entire sample set. Given the large number of predictors (classifiers) used in this case, there is the risk

of overfitting. Overfitting could occur, amongst other reasons, when a model has too many parameters relative to the number of observations.

3.8 Questionnaire Design

The following process was used to convert the qualitative data into a survey instrument:

Figure 1. Questionnaire Design Process

Steps 1 and 2 have been done through preceding chapters of this paper. Step 3 and 4 are defined below, where the dependent variables are defined based on the research hypotheses. A pilot test will be conducted on 10 members in order to test the validity of the survey as the research instrument.

To increase reliability, the questionnaire was carefully developed based on literature as well as previous questionnaires – largely that of Lussier. The sampled respondents are not aware of how organisational performance was measured nor would they need to be. Organisational performance was merely inferred from employment and turnover. The questionnaires were mailed to each owner/CEO repeatedly. The questionnaire was distributed via email five times in total. The questionnaire had 18 questions designed to measure organisational performance.

3.9 Independent Variables and Codes

Table 7. Dependent variables

	Variable Name	Measurement	Type
1	Education Level	Matric/Bachelors/ Diploma/Honours/ Masters/PhD/ Certificates	Integer (mapped to NQF level)
2	Technical Expertise	Number of Years and Likert Scale	Integer
3	Management Experience	Number of Years	Integer
4	Founders/Partners	Number of People	Integer
5	Experience in same industry	Number of Years	Integer
6	Gender	Male/Female/Other	String
7	Race	Black/White/Indian/ Coloured/Foreigner	String
8	Capital raised	Currency in ZAR	Integer
9	Age	Number of Years	Integer
10	Parents owned business	True/False	Boolean
11	Marketing Skills	Little/High use on Likert Scale	Integer
12	Customer Relationships	Highly important/Not important on Likert Scale	Integer

	Variable Name	Measurement	Type
13	Industry	- Agriculture, Forestry and Fishing - Mining - Construction - Manufacturing - Transportation, Communications, Electric, Gas and Sanitary service - Wholesale Trade - Retail Trade - Finance, Insurance and Real Estate - Services - Public Administration - Information Technology	String

3.10 Coded Variables

Table 8. Dependent variables Coded

	Variable Name	Measurement	Coded to	Grouped as
1	Education Level	Matric/Bachelors/ Diploma/ Honours/ Masters/PhD/ Certificates	Sum of NQF Level	Sum of NQF Level
2	Technical Expertise	Number of Years and Likert Scale	Sum of Years	Sum of Years
3	Management Experience	Number of Years	Sum of Years	Sum of Years
5	Experience in same industry	Number of Years	Sum of Years	Sum of Years
6	Gender	Male/Female/ Other	All males, all females, mixed	All males, all females, mixed

	Variable Name	Measurement	Coded to	Grouped as
7	Race	Black/White/ Indian/Coloured/ Foreigner	All White, all black, all indian, all coloured, mixed	All White, all black, all indian, all coloured, mixed
9	Age	Number of Years	Integer	Integer
10	Parents owned business	True/False	All parents, no parents, some parents	All parents, no parents, some parents

Previous work done by Lussier (1995) and Cooper (1990) only considered a single founder. This work is significantly different as it requires coding of data for multiple founders.

3.11 Procedure for Data Collection

The first element of data collection was to find incubators which were willing to participate. South Africa has the largest number of incubators on the continent. The relevant incubators were identified based on a) their willingness to assist and b) whether their incubatees' information

Table 9. Respondents by Incubator

Incubator	Number of respondents	Percentage
Innovation Hub	18	17,48
Shanduka Black Umbrellas	56	54,37
Riversands	20	19,42
LaunchLab	5	4,85
JoziHub	3	2,91
SoftStart	1	0,97

was easily available. Several incubators refused to assist and preferred to keep information confidential. Questionnaires were administered to them by means of either email or in person. The questionnaires were only be administered in person if the response rate on the emails proved to be low. This methodology ensured a higher response rate and more comprehensive completion. Questionnaires were a fast, effective and inexpensive means of collecting data. The questionnaire used a closed format, allowing respondents to select responses from a delineated group of answers. The respondents of the survey will remain anonymous. The questionnaires administered in person were collected as soon as they were complete. The emailed questionnaires were stored on a remote server once submitted.

The data from the emailed surveys and in-person respondents were then combined. Once combined, the data was prepared into a format suitable as input into the machine learning algorithm and data analysis.

3.12 Data analysis

The data was be downloaded in a granular format from the survey tool used. Data cleanup, transformations and mappings were performed. The data was coded and mapped as specified in Table 3. This included converting the Likert scale questions from text to numerical data. Other ordinal data was also be captured on a scale. The data was then transformed into a format appropriate for the consumption of WEKA, the machine learning algorithm tool. Multiple algorithms were run on WEKA. The output of these were then captured, analysed and compared using descriptive statistics in order to determine the accuracy of the machine learning algorithm. Descriptive statistics, means, medians, standard deviations and variances were generated to describe the cohort of respondents on measures of central tendency and measures of spread. Descriptive statistics allowed for evaluating the viability and assumptions of multivariate statistical tests.

3.13 Validity and reliability of research design

A pilot study was performed in order to determine whether the questionnaire was asking the right questions. It also ensured there was no ambiguity or errors. The length of the questionnaire was examined and found to be short enough not to cause fatigue which could impact the results. Cronbach's alpha, as discussed below, was used to perform the reliability test.

Internal Validity measures how well the questionnaires measure what they were intended to. It can be broken up in to:

3.13.1 Content Validity

This measures the extent to which the questions provides adequate coverage of the investigative questions. The literature review provided an in-depth analysis of the content

3.13.2 Predictive Validity

This measures the ability of and the extent to which the questions can make accurate predictions. This was shown by looking at previous research which utilised computational models to accurately predict.

3.13.3 Reliablity

The test of reliability used was Cronbach's alpha. A Cronbach alpha of 0.77 and above was perceived to be satisfactory. An alpha value of above 0.90 may reflect redundancy and would have to be interpreted cautiously.

Golafshani (2003) identify three types of reliability referred to in quantitative research, which relates to:

- The degree to which a measurement, given repeatedly, remains the same;
- The stability of a measurement over time; and
- The similarity of measurements within a given time period.

The reliabilities of scales were analysed using Cronbach's alpha.

3.14 Uses of the model

Technology startups can benefit from this study. It allows investors to add a new dimension of evaluation in terms of risks based on past organisational performance. It also allows them to control suggest changes to me made in the organisation before investment/further investment is considered. It allows entrepreneurs to understand what changes could be made to an organisation to increase performance. Suppliers can use it to evaluate risk and limit credit facilities to these clients. Incubators and accelerators could use the model to assist and improve businesses.

3.15 Limitations of the study

- It only applies to the South African startups
- The model does not deal with probabilities and required judgement and analysis when applying to organisations
- The model is based on the information which it is fed. It is possibly subject to biased depending on data collection. Every attempt has been made to ensure this is not the case.
- The model does not indicate the strength and weakness of any particular variable. Therefore, we do not know which is the most influential factor or variable in the model.
- Recall bias due to respondents having to recall what sustained the business.
- The study intentionally did not consider failed SMEs. This is mostly due to unavailability of that information.
- Cross sectional studies have an inherent limitation as the phenomena are viewed as a snapshot in time whereas in reality that phenomena may be dynamic.
- There is always possible self-selection bias due to the sampling method employed.

CHAPTER 4

4.1 Introduction

The results section of this study presents the data analysis of the results by examining the following:

- The sample respondents
- The missing data and how it has been dealt with
- The regression
- The assumptions of regression
- The reliability of the data

Included in this chapter are various statistical tests demonstrated using data tables and images. The results of the data analysis are conducted with regards to the data populated from online surveys and consider the various research hypotheses, constructs and variables described earlier.

4.2 Demographic Profile of respondents

In total there were 103 respondents in a total of 6 incubators. 54,37% were from Shanduka Black Umbrellas, which was by far the largest incubator surveyed. Shandukah Black Umbrellas is a national incubator hence the responses were from different regions. Riversands and Innovation hub, which are both located in the Gauteng region brought in a combined 36,9% of respondents. The remainder of the respondents came in smaller numbers from the other three incubators - a combined value of 8,73%.

4.2.1 Gender

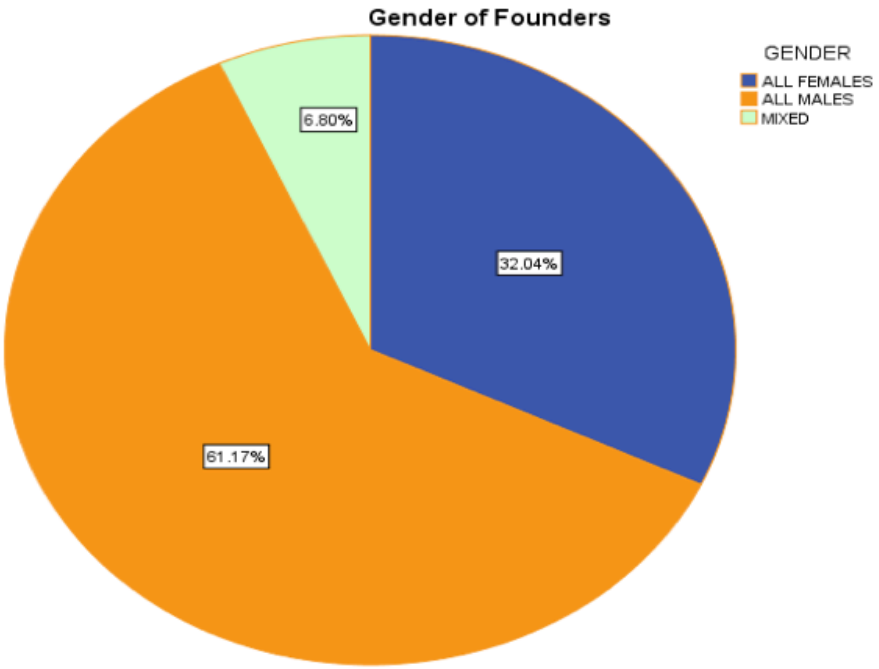

Figure 2. Gender of founders

It was found that 61.17% of the businesses were founded by men, 32.04% were founded by women while 6.8% of the businesses were founded by both male and female. This leads us to conclude there are more male more entrepreneurs than women.

4.2.2 Race

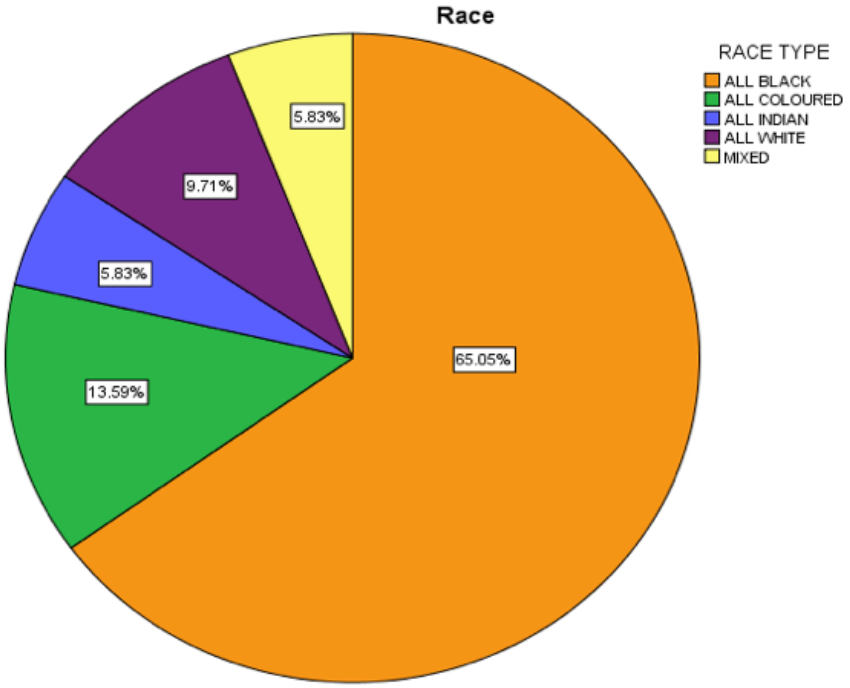

Figure 3. Race of founders

Most of the sampled businesses (65.05%) were found to be owned by blacks followed by coloured people, who own 13.59% of the businesses. Indians and whites own 5.83% and 9.71% of the sampled businesses respectively while 5.83% of the businesses had founders of mixed races.

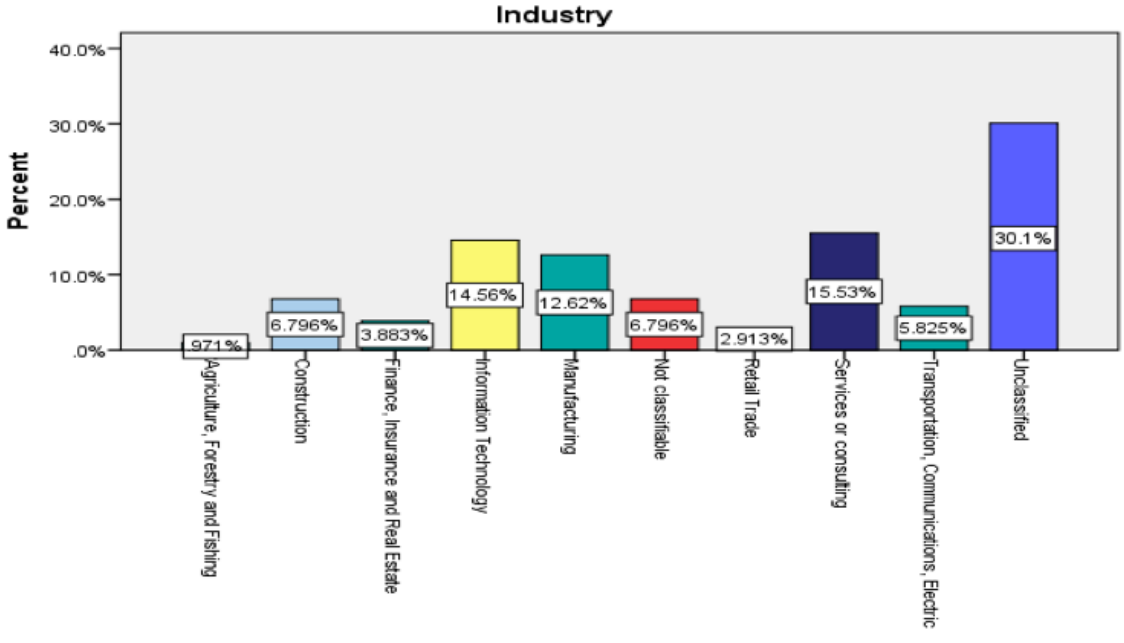

Figure 5. Breakdown by Industry

The distribution of races might be a reflection of the real population proportions on the sampled location, or an unrepresentative sample.

4.2.3 Number of Founders

70.87% of the businesses were found to be sole proprietorship, 19.42% were founded by two founders, 5.825% are founded by 3 people while 3.883% are owned by 4 people. Four was found to be the maximum number of founders in the sample.

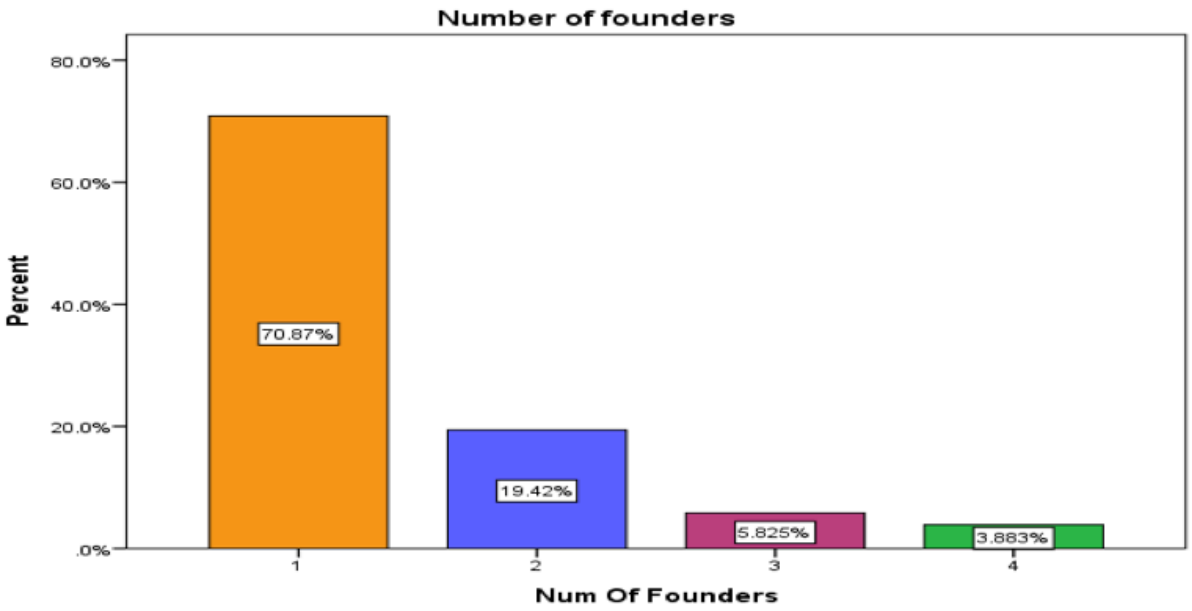

Figure 4. Number of founders

4.2.4 Industry

30.1% of the businesses are from unclassified fields: Services and consulting, information technology and manufacturing follows with 15.53%, 14.56% and 12.62% respectively, the least involved industry is Agriculture with .971% .

4.2.5 Turnover

Most businesses (27.18%) were found to make a turnover less than R1000 while only a few make between R500 to R1000000, since the classes(turnover classes) are ordered this graph can be used to describe distribution of turnover, which is a not a normal distribution.

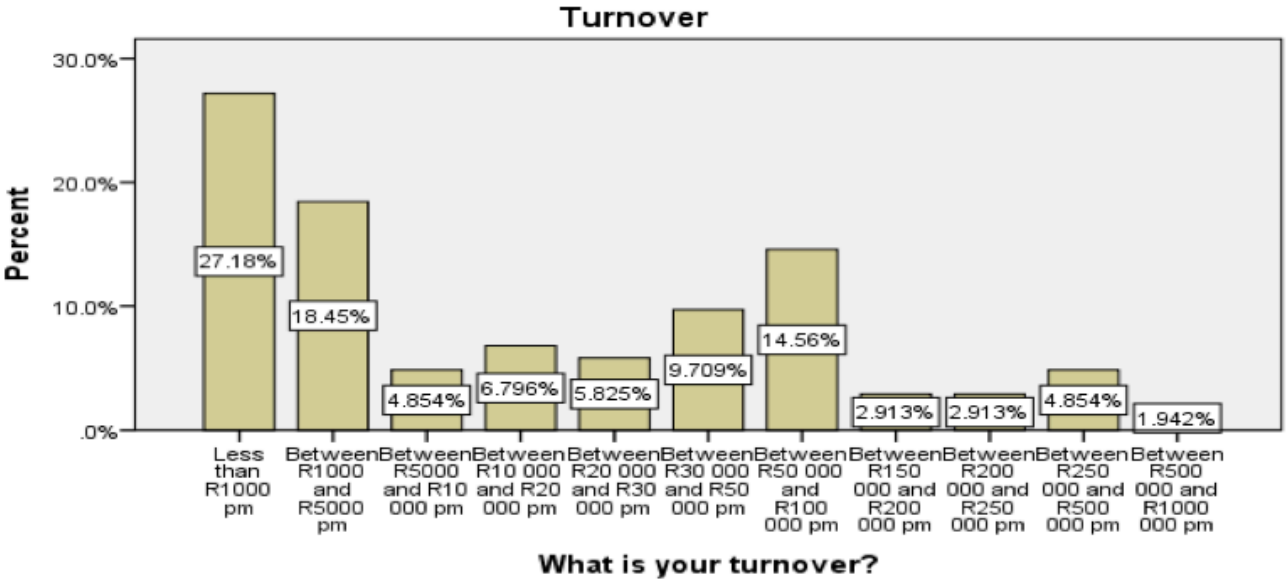

Figure 6. Turnover

Table 10. Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
AVERAGE AGE	103	22	67	39.37	9,986	99,713	0,877	0,238
SUM EDUCATION	103	0	36	8.38	5,655	31,983	2,276	0,238
SUM TECH EXPERTISE	103	1	20	5.11	3,945	15,567	1,911	0,238

SUM MANAGEMENT EXPERIENCE	103	0	84	8.16	11,230	126,11	3,605	0,238
SUM SAME INDUSTRY	103	0	105	11.29	15,978	255,30	3,122	0,238
Valid N (listwise)	103							

For all the sampled businesses the average age of the founders was found to be 39.37 with standard deviation 9.986 Average years of expertise is 8.38 with a standard deviation 3.945 years. Mean sum Management experience, and same industry is 8.16 and 11.29 respectively, with standard deviations 11.230 and 15.978 respectively.

4.3 Missing Data

Due the structure of the survey, and given multiple founders, it was not possible to make all fields mandatory without making a much larger survey and increasing the chances of survey fatigue. The decision was made to structure the survey in an easy manner and encourage more participants. This has led to a fair amount of missing data. For each variable below, the numbers of missing values are discussed and the resolution for these missing values is given.

Table 11. Missing Data

Attribute	Missing Values	Resolution
Name - Founder 1	0	N/A

Attribute	Missing Values	Resolution
Surname - Founder 1	1	None Required. This field will not have an impact on statistical analysis.
Gender - Founder 1	1	The gender was assumed from the name.
Race - Founder 1	2	The race was assumed from the name.
Highest Qualification - Founder 1	3	
Technical Expertise (Level 1 to 5) - Founder 1	13	Assumed to have no technical expertise. The lowest figure of 1 was given.
Business experience in same industry (Years) - Founder 1 - Years	17	Assumed to have no business experience. The lowest figure of 1 was given.
Schooling - Founder 1	3	None Required. This field will not have an impact on statistical analysis.
Management experience in same industry (Years) - Founder 1 - Years	20	Assumed to have no management experience. The lowest figure of 1 was given.
Parents Owned Business - Founder 1	20	Assumed that parents do not own business. The value of false was given.
Age - Founder 1 - Years	3	The average age across the board of all founders was given.
Name - Founder 2	0	N/A
Surname - Founder 2	0	N/A
Gender - Founder 2	2	The gender was assumed from the name.
Race - Founder 2	1	The race was assumed from the name.
Highest Qualification - Founder 2	2	

Attribute	Missing Values	Resolution
Technical Expertise (Level 1 to 5) - Founder 2	2	Assumed to have no technical expertise. The lowest figure of 1 was given.
Business experience in same industry (Years) - Founder 2 - Years	5	Assumed to have no business experience. The lowest figure of 1 was given.
Schooling - Founder 2	1	None Required. This field will not have an impact on statistical analysis.
Management experience in same industry (Years) - Founder 2 - Years	7	Assumed to have no management experience. The lowest figure of 1 was given.
Parents Owned Business - Founder 2	8	Assumed that parents do not own business. The value of false was given.
Age - Founder 2 - Years	1	The average age across the board of all founders was given.
Name - Founder 3	0	N/A
Surname - Founder 3	0	N/A
Gender - Founder 3	0	N/A
Race - Founder 3	0	N/A
Highest Qualification - Founder 3	0	N/A
Technical Expertise (Level 1 to 5) - Founder 3	0	N/A
Business experience in same industry (Years) - Founder 3 - Years	1	Assumed to have no business experience. The lowest figure of 1 was given.

Attribute	Missing Values	Resolution
Schooling - Founder 3	0	N/A
Management experience in same industry (Years) - Founder 3 - Years	3	Assumed to have no management experience. The lowest figure of 1 was given.
Parents Owned Business - Founder 3	4	Assumed that parents do not own business. The value of false was given.
Age - Founder 3 - Years	9	The average age across the board of all founders was given.
Name - Founder 4	0	N/A
Surname - Founder 4	0	N/A
Gender - Founder 4	0	N/A
Race - Founder 4	0	N/A
Highest Qualification - Founder 4	0	N/A
Technical Expertise (Level 1 to 5) - Founder 4	0	N/A
Business experience in same industry (Years) - Founder 4 - Years	1	Assumed to have no business experience. The lowest figure of 1 was given.
Schooling - Founder 4	1	None Required. This field will not have an impact on statistical analysis.
Management experience in same industry (Years) - Founder 4 - Years	1	Assumed to have no management experience. The lowest figure of 1 was given.

Attribute	Missing Values	Resolution
Parents Owned Business - Founder 4	1	Assumed that parents do not own business. The value of false was given.
Age - Founder 4 - Years	0	N/A
How much capital has been placed into the business?	0	N/A
How would you rate the capital in your business? - Inadequate Capital:Adequate Capital	0	N/A
What was the sources of capital?	3	None Required. This field will not have an impact on statistical analysis.
How long have you been in incubation for?	0	N/A
What was your turnover at the point of incubation?	0	N/A
What is your current staff compliment? - Staff size	0	N/A
What was the size of your staff at the point of incubation? - Staff size	0	N/A
Do you believe that incubation has led to an increase in turnover? - Definitely not:Definitely yes	0	N/A
Do you believe that incubation has led to an increase in your number of employees? - Definitely not:Definitely yes	0	N/A

Attribute	Missing Values	Resolution
Do you believe that incubation has contributed to increasing your chances of success? - Definitely not:Definitely yes	0	N/A
How would you rate your record keeping and financial control? - Very Poor:Very Professional	0	N/A
How often do you professional advice? - Never used:Used often	0	N/A
How difficult did you find it to obtain staff? - Easy:Difficult	0	N/A
How would you rate your marketing skills? - Little Marketing:Great use of marketing	0	N/A
How would you rate the importance of customer relations to your business? - Not Important:Highly Important	30	Average taken.

4.4 Examining Independent variable - Turnover

4.4.1 Normality tests

From the above histogram the condition of normality of residuals was found to be poorly met as the shape is not bell shaped. The probability plot is also not a straight line.

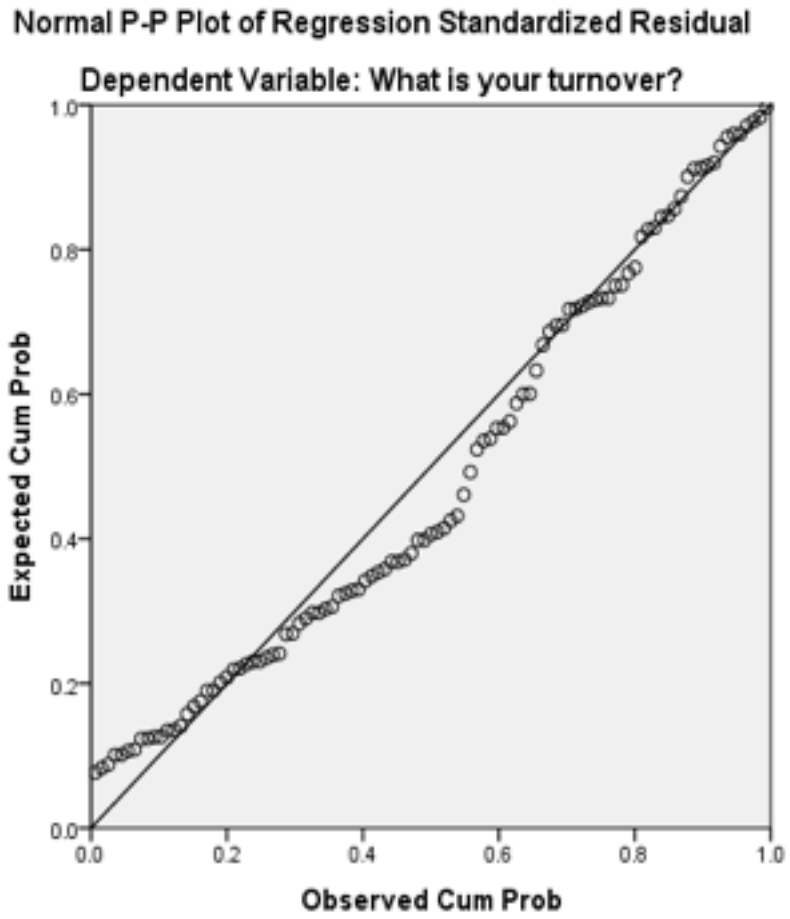

Figure 7. Normal Regression Standardised Residual

Classical test of normality of variance:

- H_0: errors are normally distributed
- H_1: errors are not normally distributed

Table 12. Normality Test - Turnover

Tests of Normality						
	Kolmogorov-Smirnov[a]			*Shapiro-Wilk*		
	Statistic	*df*	*Sig0,*	*Statistic*	*df*	*Sig0,*
Unstandardized Residual	0,098	103	0,016	0,964	103	0,007

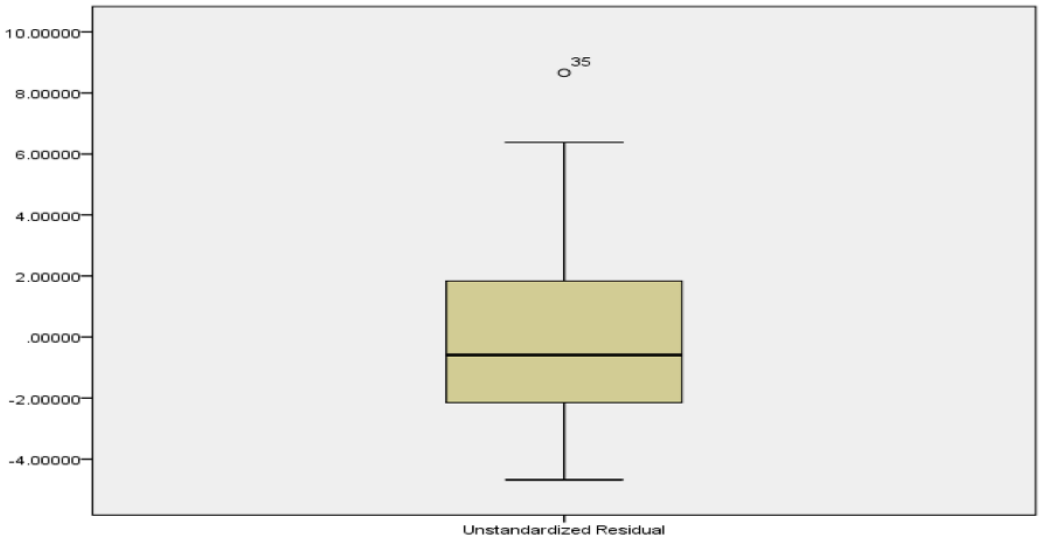

Figure 8. Outliers for turnover

The Kolmogorov-Smirnov test is used to test for 'goodness of fit' between a sample distribution and another distribution, which often is the Normal (bell-shaped) distribution. The test compares the set of scores in the sample to a normally distribute set of scores with the same mean and standard deviation. The two-sample Kolmogorov-Smirnov test is a nonparametric test that compares the cumulative distributions of two data sets. The test is nonparametric. It does not assume that data are sampled from Gaussian distributions.

We reject the null hypothesis at both .05 as the value is 0,016. The conclusion can be made that the errors are not normal hence the condition of normality is not made.

4.4.2 Outliers

The outliers for the independent variable turnover, is now examined:

Unstandardized Residual Stem-and-Leaf Plot

```
Frequency Stem & Leaf

 3.00 -4 . 234
 12.00 -3 . 134455568899
 14.00 -2 . 12222334567799
 19.00 -1 . 0001122334456667799
 10.00 -0 . 0355677889
 9.00 0 . 122444677
```

```
12.00 1 . 035557788999
7.00 2 . 0123899
5.00 3 . 11359
5.00 4 . 12239
4.00 5 . 2448
2.00 6 . 25
1.00 Extremes (>=8.7)

Stem width: 1.00000
Each leaf: 1 case(s)
```

A box plot and a stem diagram of the same shows that only one point is outlying. There is no much violation of this condition hence it is approximately met

4.4.3 Homoscedacity

The Points of the residual plot are randomly distributed signifying constant variance. The Breush-Pagan test creates a statistic that is chi-squared distributed and for data that statistic = 7.18. The p-value is the

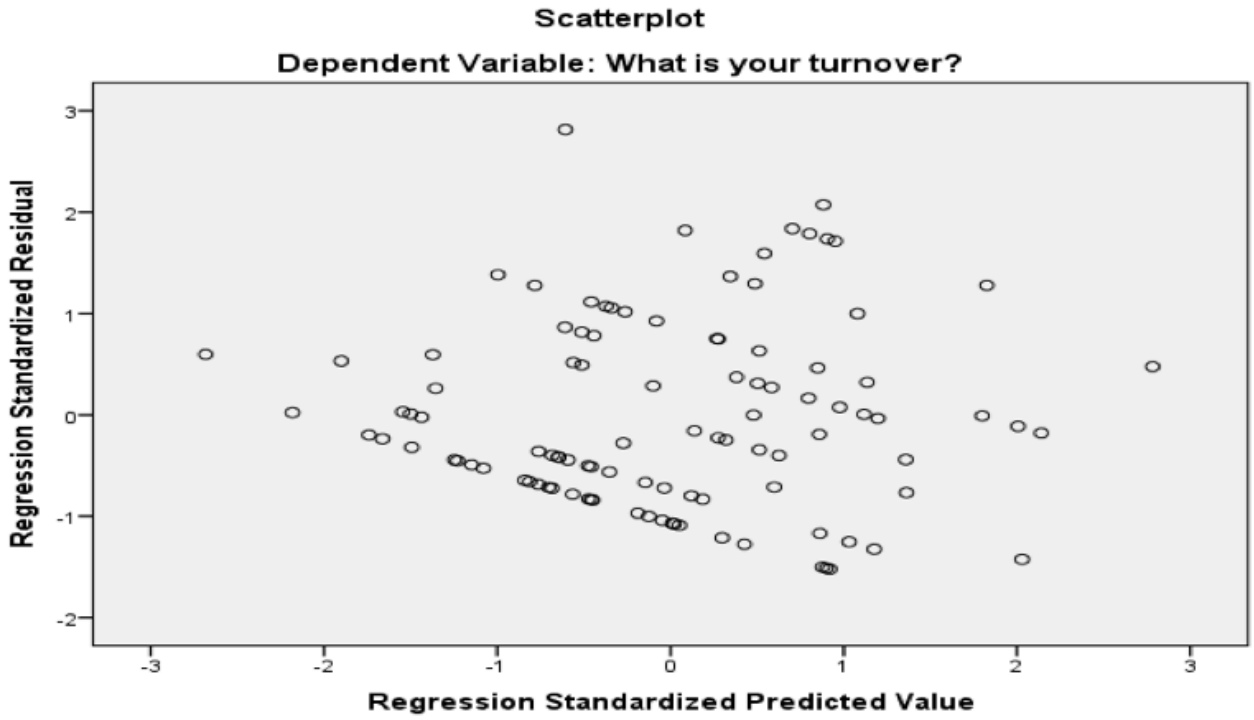

Figure 9. Scatterplot for turnover

result of the chi-squared test and (normally) the null hypothesis is rejected for p-value < 0.05.

Classical test Breusch-Pagan

H_0 : Error variance is homoscedastic

H_1: Error variance is not homoscedastic

Breusch-Pagan and Koenker test statistics and sig-values

```
BP      15.022 .450
Koenker 16.444 .353
```

The p value for both BP and Koenker is greater 0.05 we therefore fail to reject the null hypothesis and conclude that error variance is constant for this model. The seen homoscedasticity is not significant.

4.4.4 Collinearity

The variance inflation factor does not show significant variance inflation although most variables has a high variance inflation

Table 13. Collinearity - Turnover

Coefficients[a]							
	Unstandardized Coefficients		*Standardized Coefficients*	*t*	*Sig0,*	*Collinearity Statistics*	
	B	*Std0, Error*	*Beta*			*Tolerance*	*VIF*
(Constant)	-0,241	30,467		-0,07	0,945		
How would you rate the capital in your business? - Inadequate Capital:Adequate Capital	0,351	0,189	0,188	1,858	0,067	0,869	1,15
How would you rate your record keeping and financial control? - Very Poor:Very Professional	-0,383	0,283	-0,167	-1,357	0,178	0,589	1,697
How often do you professional advice? - Never used:Used often	0,423	0,214	0,208	1,973	0,052	0,799	1,252
How difficult did you find it to obtain staff? - Easy:Difficult	-0,09	0,195	-0,05	-0,461	0,646	0,754	1,326

How would you rate your marketing skills? - Little Marketing:Great use of marketing	-0,158	0,186	-0,085	-0,849	0,398	0,885	1,13
How would you rate the importance of customer relations to your business? - Not Important:Highly Important	-0,055	0,499	-0,011	-0,109	0,913	0,822	1,217
Num Of Founders	0,274	10,095	0,066	0,251	0,803	0,129	7,762
SUM EDUCATION	-0,188	0,145	-0,33	-1,299	0,197	0,138	7,25
SUM TECH EXPERTISE	0,339	0,208	0,414	10,626	0,108	0,137	7,297
SUM MANAGEMENT EXPERIENCE	-0,044	0,052	-0,152	-0,844	0,401	0,275	3,633
SUM SAME INDUSTRY	-0,016	0,043	-0,082	-0,379	0,705	0,192	5,206
RACE TYPE	-0,234	0,296	-0,101	-0,789	0,432	0,545	1,834
AVERAGE AGE	0,122	0,045	0,376	20,719	0,008	0,465	2,149
PARENTS OWNED BUSINESS - SOME/ALL/NONE	-0,036	0,491	-0,008	-0,072	0,942	0,747	1,339
GENDER	0,337	0,649	0,065	0,519	0,605	0,568	1,761
a0, Dependent Variable: What is your turnover?							

4.5 Linear Regression - Turnover

Table 14. Model Summary - Turnover

Model Summary[b]						
Model	*R*	*R Square*	*Adjusted R Square*	*Std0, Error of the Estimate*	*Durbin-Watson*	
1	0,487a	0,238	0,106	0,74143	1,883	

The r square of the model was found to be .238 meaning that only 23.8 % of the variation in square root of turnover can be explained by the independent variables.

Table 15. Anova - Turnover

ANOVA[a]						
Model		*Sum of Squares*	*df*	*Mean Square*	*F*	*Sig0,*
1	Regression	14,906	15	0,994	1,808	0,046b
	Residual	47,826	87	0,55		
	Total	62,731	102			
a0, Dependent Variable: newturn						

The p value of the ANOVA statistic was found to be .046. The model is therefore significant at 0.05 level of significance, hence can be considered predictive. Now that the hypothesis adopted is that at least one regression coefficient is different from zero, we go on to interpret the significant coefficients.

Table 16. Coefficients - Turnover

Coefficients[a]								
Model		*Unstandardized Coefficients*		*Standardized Coefficients*	*t*	*Sig 0,*	*Collinearity Statistics*	
		B	*Std0, Error*	*Beta*			*Tolerance*	*VIF*
1	(Constant)	0,994	0,836		1,189	0,238		
	How would you rate the capital in your business? - Inadequate Capital:Adequate Capital	0,103	0,046	0,226	2,25	0,027	0,869	1,15

	How would you rate your record keeping and financial control? - Very Poor:Very Professional	-0,104	0,068	-0,185	-10,52	0,132	0,589	1,697
	How often do you professional advice? - Never used:Used often	0,108	0,052	0,22	2,097	0,039	0,799	1,252
	How difficult did you find it to obtain staff? - Easy:Difficult	-0,02	0,047	-0,045	-0,418	0,677	0,754	1,326
	How would you rate your marketing skills? - Little Marketing: Great use of marketing	-0,046	0,045	-0,101	-10,017	0,312	0,885	1,13
	How would you rate the importance of customer relations to your business? - Not Important: Highly Important	-0,022	0,12	-0,019	-0,183	0,855	0,822	1,217
	Num Of Founders	0,117	0,264	0,115	0,442	0,659	0,129	7,762

	SUM EDUCATION	-0,053	0,035	-0,383	-10,519	0,132	0,138	7,25
	SUM TECH EXPERTISE	0,079	0,05	0,399	10,579	0,118	0,137	7,297
	SUM MANAGEMENT EXPERIENCE	-0,01	0,012	-0,143	-0,801	0,425	0,275	3,633
	SUM SAME INDUSTRY	-0,005	0,01	-0,097	-0,454	0,651	0,192	5,206
	RACE TYPE	-0,066	0,071	-0,116	-0,919	0,361	0,545	1,834
	AVERAGE AGE	0,027	0,011	0,345	20,516	0,014	0,465	2,149
	PARENTS OWNED BUSINESS - SOME/ ALL/ NONE	-0,012	0,118	-0,011	-0,098	0,923	0,747	1,339
	GENDER	0,08	0,156	0,063	0,511	0,611	0,568	1,761
a0, Dependent Variable: newturn								

The following variables were found to significantly affect turnover at 0.05 levels:

- Capital
- Professional advice
- Average age

The coefficient of the capital rating is .103 which means that turnover increases significantly with the amount of capital pumped to business. A single unit increase in startup capital leads to a .103 unit increase in square root turnover and . 0110609 units in turnover.

The coefficient of the variable *professional advice* is .108 which means that turnover increases significantly with the amount of time one seeks professional advice for; for every increased unit of professional advice (on the liberty scale), the square root of turnover increases by .108 units.

The coefficient of the variable *average age* is .027 which means that turnover increases significantly with the average age. For every one year increase in *average age,* there is a square root of turnover increase by .027 units. For every unit increase in turnover, there is an average of .000729 units in age.

4.6 Examining Independent variable - Number of staff

4.6.1 Normality tests

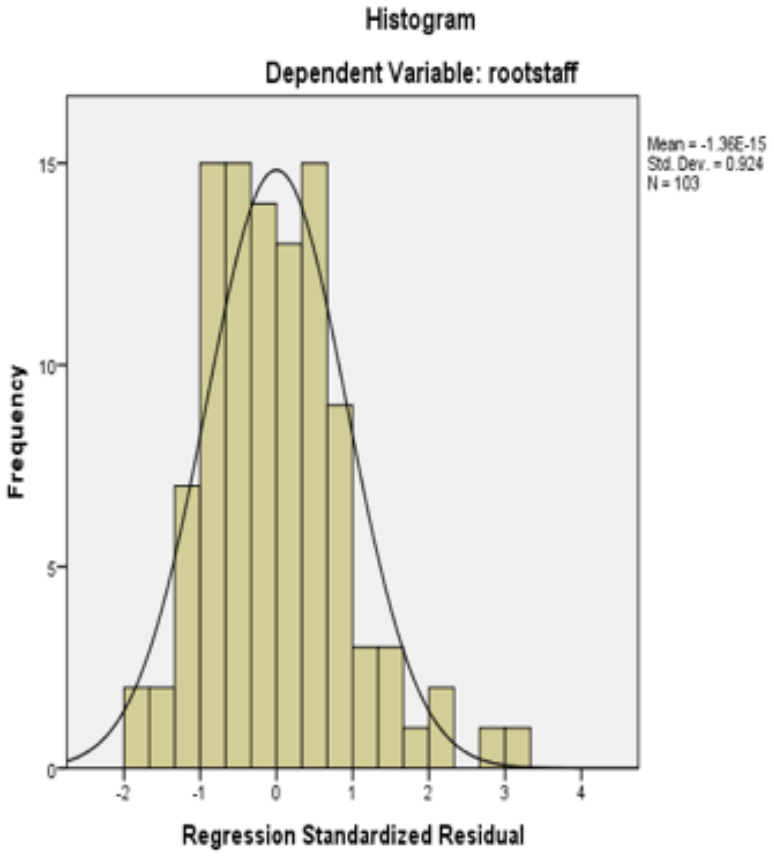

Figure 10. Histogram - Staff

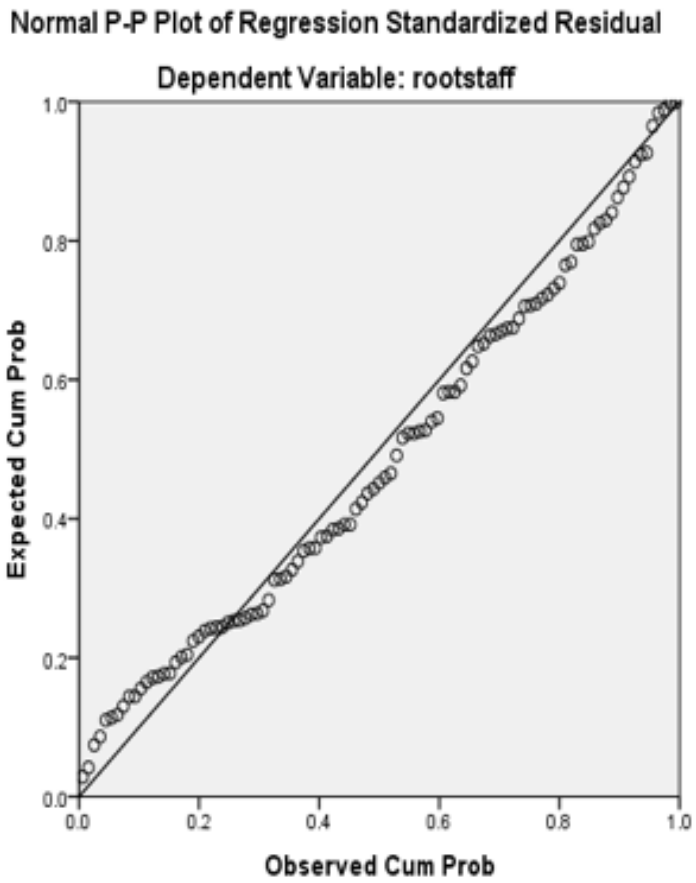

Figure 11. Normal Regression Standardised Residual

From the above histogram the condition of normality of residuals was found to be met as the shape is close to a bell shaped. The probability plot is approximately a straight line

H_0: errors are normally distributed

H_1: errors are not normally distributed

Table 17. Normality test - number of staff

Tests of Normality						
	Kolmogorov-Smirnov[a]			*Shapiro-Wilk*		
	Statistic	df	Sig.	Statistic	df	Sig.
Studentized Residual	0,076	103	0,157	0,963	103	0,006
a. Lilliefors Significance Correction						

We fail to reject the null hypothesis at.05 for Kolmogorov test and conclude that errors are normal. The Shapiro test is significant, and it is therefore difficult to reach a conclusion regarding normality, though most of the test indicates that normality is met. Regression is also a robust test, hence it is less affected by such minor violation of conditions.

4.6.2 Outliers

Table 18. Descriptive Statistics - Number of Staff

Descriptive Statistics					
	N	*Minimum*	*Maximum*	*Mean*	*Std. Deviation*
Studentized Residual	103	-2,06596	3,37235	0,0030973	1,00266675
Valid N (listwise)	103				

Taking the most extreme values of the standardised value found some influential points. (greater than 3)

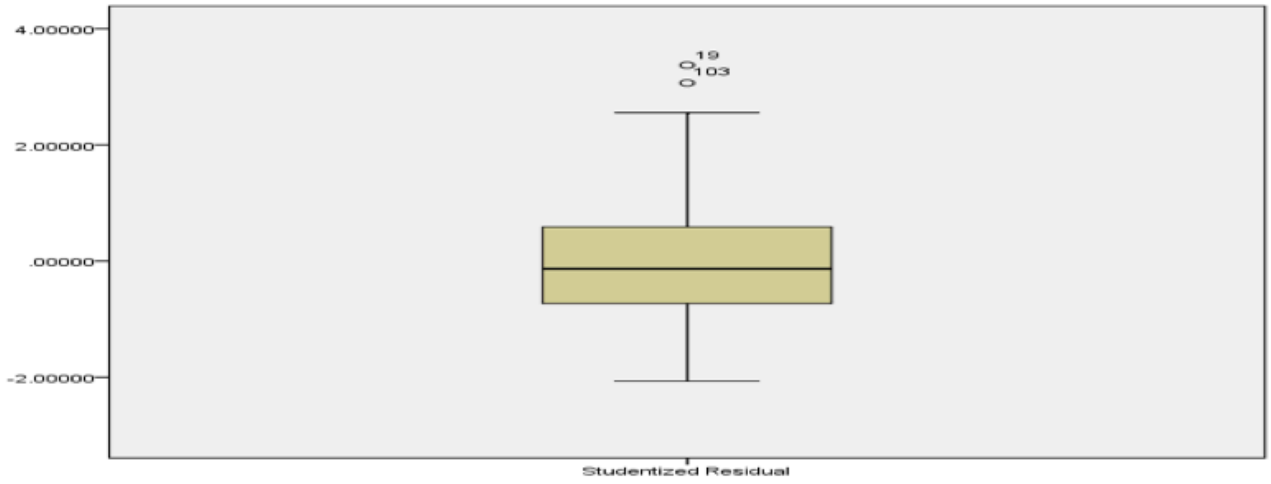

Figure 12. Boxplot - Staff

Table 19.Casewise Diagnostics

Casewise Diagnostics[a]				
Case Number	*Std. Residual*	*rootstaff*	*Predicted Value*	*Residual*
19	3,136	3,16	1,7812	1,38107
a. Dependent Variable: rootstaff				

Studentized Residual Stem-and-Leaf Plot

```
Frequency Stem & Leaf

 1.00  -2 . 0
 2.00  -1 . 59
13.00  -1 . 0000001122334
21.00  -0 . 555556666777777778888
18.00  -0 . 011111223333333344
19.00  0 . 0000011222233444444
13.00  0 . 5555666677889
 9.00  1 . 001111234
 3.00  1 . 559
 1.00  2 . 2
 1.00  2 . 5
 2.00 Extremes (>=3.1)

Stem width: 1.00000
Each leaf:  1 case(s)
```

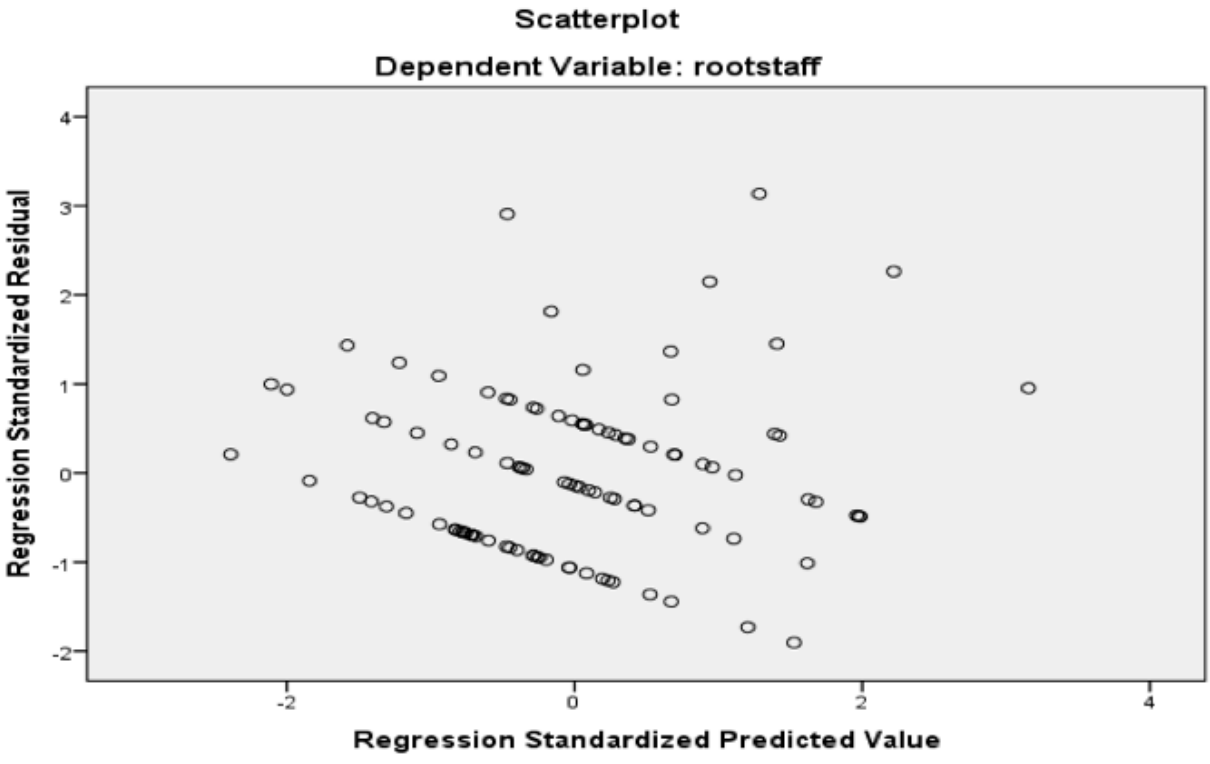

Figure 13. Scatterplot for number of staff

A box plot and a stem diagram of the same shows that two points is outlying. There is no excessive violation of this condition hence it is approximately met.

4.6.3 Homoscedacity

The points of the residual plot are randomly distribution signifying constant variance.

Classical test Breusch pagan

H_0 : Error variance is homoscedastic
H_1: Error variance is not homoscedastic

Breusch-Pagan and Koenker test statistics and sig-values

```
BP        24.181   .062
Koenker   15.633   .407
```

The p value for both BP and Koenker is greater 0.05 we therefore fail to reject the null hypothesis and conclude that error variance is constant for this model. The seen heteroskedasticity is not significant.

4.6.4 Independence of error terms

H_0: = 0
H_1: > 0

Table 20. Model Summary - Number of staff

Model Summary[b]					
Model	*R*	*R Square*	*Adjusted R Square*	*Std. Error of the Estimate*	*Durbin-Watson*
1	0,504[a]	0,254	0,126	0,44045	1,690

From the model summary the Durbin Watson test statistic is 1.690, which is between the critical values 1.414 and 1.847 the test is inconclusive

Table 21. Coefficients - Number of staff

Coefficients[a]								
Model		*Unstandardized Coefficients*		*Std. Coeff,*	*t*	*Sig0,*	*Collinearity Statistics*	
		B	*Std0, Error*	*Beta*			*Tolerance*	*VIF*
1	(Constant)	1,699	0,497		3,422	0,001		
	How would you rate the capital in your business? - Inadequate Capital:Adequate Capital	-0,018	0,027	-0,068	-0,682	0,497	0,869	1,15
	How would you rate your record keeping and financial control? - Very Poor:Very Professional	0,039	0,04	0,116	0,96	0,34	0,589	1,697
	How often do you professional advice? - Never used:Used often	0,055	0,031	0,186	1,791	0,077	0,799	1,252
	How difficult did you find it to obtain staff? - Easy:Difficult	-0,036	0,028	-0,136	-1,272	0,207	0,754	1,326
	How would you rate your marketing skills? - Little Marketing:Great use of marketing	-0,041	0,027	-0,151	-1,535	0,128	0,885	1,13
	How would you rate the importance of customer relations to your business? - Not Important:Highly Important	-0,01	0,072	-0,015	-0,146	0,884	0,822	1,217
	Num Of Founders	0,565	0,157	0,93	30,604	0,001	0,129	7,762
	SUM EDUCATION	-0,058	0,021	-0,694	-20,786	0,007	0,138	7,25
	SUM TECH EXPERTISE	-0,008	0,03	-0,065	-0,26	0,795	0,137	7,297

	SUM MANAGEMENT EXPERIENCE	0,003	0,007	0,074	0,422	0,674	0,275	3,633
	SUM SAME INDUSTRY	-0,009	0,006	-0,31	-1,466	0,146	0,192	5,206
	RACE TYPE	0,007	0,042	0,02	0,158	0,875	0,545	1,834
	AVERAGE AGE	-0,009	0,006	-0,197	-1,453	0,15	0,465	2,149
	PARENTS OWNED BUSINESS - SOME/ALL/ NONE	0,019	0,07	0,029	0,269	0,789	0,747	1,339
	GENDER	-0,063	0,093	-0,083	-0,678	0,5	0,568	1,761
a0, Dependent Variable: rootstaff								

The variance inflation factor doesn't show significant variance inflation although most variables have a high variance inflation.

4.7 Linear Regression - Number of staff

Table 22. Model Summary - Number of staff

Model Summary[b]						
Model	*R*	*R Square*	*Adjusted R Square*	*Std, Error of the Estimate*	*Durbin-Watson*	
1	,504a	0,254	0,126	0,44045	1,69	

The r square of the model was found to be .254 meaning that only 25.4 % of the variation in square root of staff complement can be explained by the independent variables. Square root to improve the adherence to the model assumption and significance of the model.

Table 23. Anova - Number of staff

ANOVA[a]						
Model		*Sum of Squares*	*df*	*Mean Square*	*F*	*Sig,*
1	Regression	5,752	15	0,383	1,977	,026b
	Residual	16,878	87	0,194		

	Total	22,63	102			

The p value of the ANOVA statistic was found to be .026. The model is therefore significant at 0.05 level of significance hence can be considered predictive. Given that we adopted the hypothesis that at least one regression coefficient is different from zero, we go on to interpret the significant coefficients.

Table 24. Coefficients - Number of staff

Coefficients[a]								
Model		Unstandardized Coefficients		Std. Coeff.	t	Sig,	Collinearity Statistics	
		B	Std, Error	Beta			Tolerance	VIF
1	(Constant)	1,699	0,497		3,422	0,001		
	How would you rate the capital in your business? - Inadequate Capital:Adequate Capital	-0,018	0,027	-0,068	-0,68	0,497	0,869	1,15
	How would you rate your record keeping and financial control? - Very Poor:Very Professional	0,039	0,04	0,116	0,96	0,34	0,589	1,697
	How often do you professional advice? - Never used:Used often	0,055	0,031	0,186	1,791	0,077	0,799	1,252
	How difficult did you find it to obtain staff? - Easy:Difficult	-0,036	0,028	-0,136	-1,27	0,207	0,754	1,326

	How would you rate your marketing skills? - Little Marketing:Great use of marketing	-0,041	0,027	-0,151	-1,53	0,128	0,885	1,13
	How would you rate the importance of customer relations to your business? - Not Important:Highly Important	-0,01	0,072	-0,015	-0,14	0,884	0,822	1,217
	Num Of Founders	0,565	0,157	0,93	3,604	0,001	0,129	7,762
	SUM EDUCATION	-0,058	0,021	-0,694	-2,78	0,007	0,138	7,25
	SUM TECH EXPERTISE	-0,008	0,03	-0,065	-0,26	0,795	0,137	7,297
	SUM MANAGEMENT EXPERIENCE	0,003	0,007	0,074	0,422	0,674	0,275	3,633
	SUM SAME INDUSTRY	-0,009	0,006	-0,31	-1,46	0,146	0,192	5,206
	RACE TYPE	0,007	0,042	0,02	0,158	0,875	0,545	1,834
	AVERAGE AGE	-0,009	0,006	-0,197	-1,45	0,15	0,465	2,149
	PARENTS OWNED BUSINESS - SOME/ALL/ NONE	0,019	0,07	0,029	0,269	0,789	0,747	1,339
	GENDER	-0,063	0,093	-0,083	-0,67	0,5	0,568	1,761
a, Dependent Variable: rootstaff								

Two variables were found to be significant:

- Number of founders
- Level of education

At .05 level of significance the number of founders a business has was found to significantly affect the staff complement. It has a p value =.001, which is very small, indicating strong evidence against the null hypothesis that the number of founders a business has doesn't affect the magnitude of the staff group. The regression coefficient is .565

meaning that a unit increase in the number of founders would lead to an average of .585 increase in square root of staff group and .342222 in staff group.

The level of education also affects the staff complement significantly. It has a p value =.007, which. The regression coefficient is -.058, meaning that a unit increase in level of education leads to -.058 in the square root of staff compliment rating. This might be due to increased commitment to multiple jobs as education increases.

4.8 Reliability

Table 25. Cronbach's Alpha criteria

Cronbach's Alpha Value	Reliability
>0.9	Excellent
>0.8	Good
>0.7	Acceptable
>0.6	Questionable
>0.5	Poor
<0.5	Unacceptable

The Cronbach coefficient is .603 which is a moderate figure, which indicates a moderate level of internal consistency for our scale with this specific sample.

Table 26. Cronbach's Alpha

Reliability Statistics		
Cronbach's Alpha	*Cronbach's Alpha Based on Standardized Items*	*N of Items*
603	669	19

4.9 Machine Learning Results

This research aimed to present a machine learning model which is able to accurately classify the organisational performance of startups within incubators. The research highlights different machine learning algorithms and examines how they work as well as some of the advantages and disadvantages of each. Classification accuracies along with misclassification rates are examined in order to see which machine learning algorithm proves to be the most fitting. Weka has been chosen as the machine learning tool of choice. Weka is able to easily interchange between multiple algorithms and give a detailed output as to the performance of the algorithm.

Below are some of the guidelines to the results of the machine learning algorithms which have largely been adapted from the MTech website (2018):

Total number instances - this refers to the number of entries which have been analysed. In this case, there were 103 surveys processed.

Confusion matrix - The confusion matrix gives information regarding the actual and predicted classifications performed by the algorithm. It is largely based on true positive and false positive values. The values across the matrix horizontally represent the predicted values while the values across the horizontal represent the actual values. The diagonal from top left to bottom right represents the values which have been correctly classified. The comparison is made on accuracy, sensitivity and specificity using true positive and false positive in confusion matrix generated by the respective algorithms. There are four outcomes to an entry in the confusion matrix:

- true positives (TP): Correctly classified instances where value is true.
- true negatives (TN): Correctly classified instances where value is false.

- false positives (FP): Known as a Type I error where prediction is yes and actual is false.
- false negatives (FN): Known as Type II error where prediction is no and actual is yes.

Correctly classified instances - this can be seen as all the true positives in the confusion matrix.
These are the correct predictions of the classifier.

Incorrectly classified instances - this can be seen as everything besides the true positives in the confusion matrix. It is every case which has led to an incorrect prediction from the classifier.

Kappa statistic – basically refers to the chance-corrected measure of agreement between the classifications and the true classes. It is calculated by taking the difference between the agreement expected by change and the observed agreement. This is then divided by the maximum possible agreement. If the value of the kappa statistic is greater than 0, then the classifier is doing better than chance.

Mean absolute error – measures the proximity of the forecasts or predictions as compared to the eventual outcomes. The calculation is similar to that of variance, however, rather than using the square difference, the absolute value is used. The reason this is done, is that taking the absolute value assigns equal weight to spread of data whereas using a square may over-emphasise the extremes.

Root mean square error – measures the differences between values predicted by a model or an estimator and the values actually observed. Represents sample standard deviation of the differences between predicted values and observed values. The magnitude of errors are then aggregated into a single measure - the value indicates the predictability. The measure is scale-dependent and while it is a good measure of accuracy, it is only suitable when comparing the forecasting of the difference between a particular variable over multiple models and not multiple variables.

Root relative squared error - this is computed by taking the mean absolute error of the classifier and dividing it by the mean absolute error which was obtained when predicting class probabilities in the training data (and not non-class probabilities). A value of close to 100% means that the classifier has not learnt anything useful from the data.

4.10 Machine Learning Results - Turnover

4.10.1 Zero R

Table 27. Zero R - Turnover Summary

Correctly Classified Instances	28	27.1845 %
Incorrectly Classified Instances	75	72.8155 %
Kappa statistic		0
Mean absolute error		0.1435
Root mean squared error		0.2672
Relative absolute error		100 %
Root relative squared error		100 %
Total Number of Instances		103

Table 28. Zero R Turnover Confusion Matrix

a	b	c	d	e	f	g	h	i	j	k	l	<-- Classified as
28	0	0	0	0	0	0	0	0	0	0	0	a = Less than R1000 pm
17	0	0	0	0	0	0	0	0	0	0	0	b = Between R1000 and R5000 pm
7	0	0	0	0	0	0	0	0	0	0	0	c = Between R5000 and R10 000 pm
7	0	0	0	0	0	0	0	0	0	0	0	d = Between R10 000 and R20 000 pm
6	0	0	0	0	0	0	0	0	0	0	0	e = Between R20 000 and R30 000 pm
10	0	0	0	0	0	0	0	0	0	0	0	f = Between R30 000 and R50 000 pm
15	0	0	0	0	0	0	0	0	0	0	0	g = Between R50 000 and R100 000 pm
1	0	0	0	0	0	0	0	0	0	0	0	h = Between R100 000 and R150 000 pm

a	b	c	d	e	f	g	h	i	j	k	l	<-- Classified as
2	0	0	0	0	0	0	0	0	0	0	0	i = Between R150 000 and R200 000 pm
2	0	0	0	0	0	0	0	0	0	0	0	j = Between R200 000 and R250 000 pm
5	0	0	0	0	0	0	0	0	0	0	0	k = Between R250 000 and R500 000 pm
3	0	0	0	0	0	0	0	0	0	0	0	l = Between R500 000 and R1 000 000 pm

The predictability of the model using this algorithm is low but higher than that off the statistical regression - which is 23,8%. The model demonstrates a Kappa statistic of 0 which shows that predictions may be due to chance. The root relative square error also indicates that nothing useful may have been learnt from the data. Looking at the confusion matrix, all items were classified as ‘Less than R1000 pm’. The Zero R algorithm will serve as a baseline for all other algorithms implemented.

4.10.2 J48

Table 29. J48 Turnover Summary

Correctly Classified Instances	23	22.3301 %
Incorrectly Classified Instances	80	77.6699 %
Kappa statistic		0.0813
Mean absolute error		0.1321
Root mean squared error		0.3255
Relative absolute error		92.0345 %
Root relative squared error		121.8351 %
Total Number of Instances		103

Table 30. J48 Turnover Confusion Matrix

a	b	c	d	e	f	g	h	i	j	k	l	<-- Classified as
16	4	3	1	0	3	1	0	0	0	0	0	a = Less than R1000 pm

a	b	c	d	e	f	g	h	i	j	k	l	<-- Classified as
3	2	0	0	0	0	1	0	0	0	1	0	c = Between R5000 and R10 000 pm
1	1	0	1	0	1	2	0	0	0	0	1	d = Between R10 000 and R20 000 pm
2	2	0	0	0	2	0	0	0	0	0	0	e = Between R20 000 and R30 000 pm
1	3	2	0	2	1	0	0	0	0	0	1	f = Between R30 000 and R50 000 pm
0	3	1	3	1	2	4	0	0	1	0	0	g = Between R50 000 and R100 000 pm
0	0	0	0	1	0	0	0	0	0	0	0	h = Between R100 000 and R150 000 pm
1	0	0	0	0	1	0	0	0	0	0	0	i = Between R150 000 and R200 000 pm
1	0	0	0	1	0	0	0	0	0	0	0	j = Between R200 000 and R250 000 pm
3	0	2	0	0	0	0	0	0	0	0	0	k = Between R250 000 and R500 000 pm
1	0	0	0	1	0	1	0	0	0	0	0	l = Between R500 000 and R1 000 000 pm

The J48 predictability is lower than that of traditional statistical regression. It does show a positive Kappa statistic and a root absolute error of less than 100%. These are good indicators that the algorithm has learnt something from the data. The confusion matrix indicates that most positive matches came from those classified as ‘Less than R1000 pm’ followed by those classified as ‘Between R50 000 and R100 000 pm’.

4.10.3 Decision Stump

Table 31. Decision Stump Turnover Summary

Correctly Classified Instances	22	21.3592 %
Incorrectly Classified Instances	81	78.6408 %
Kappa statistic		-0.0301
Mean absolute error		0.1429
Root mean squared error		0.2716
Relative absolute error		99,58 %
Root relative squared error		101,67 %

Total Number of Instances	103

Table 32. Decision Stump Turnover Confusion Matrix

a	b	c	d	e	f	g	h	i	j	k	l	<-- Classified as
19	6	0	0	0	0	3	0	0	0	0	0	a = Less than R1000 pm
12	2	0	0	0	0	3	0	0	0	0	0	b = Between R1000 and R5000 pm
5	1	0	0	0	0	1	0	0	0	0	0	c = Between R5000 and R10 000 pm
5	1	0	0	0	0	1	0	0	0	0	0	d = Between R10 000 and R20 000 pm
4	2	0	0	0	0	0	0	0	0	0	0	e = Between R20 000 and R30 000 pm
5	3	0	0	0	0	2	0	0	0	0	0	f = Between R30 000 and R50 000 pm
11	3	0	0	0	0	1	0	0	0	0	0	g = Between R50 000 and R100 000 pm
1	0	0	0	0	0	0	0	0	0	0	0	h = Between R100 000 and R150 000 pm
1	1	0	0	0	0	0	0	0	0	0	0	i = Between R150 000 and R200 000 pm
2	0	0	0	0	0	0	0	0	0	0	0	j = Between R200 000 and R250 000 pm
4	1	0	0	0	0	0	0	0	0	0	0	k = Between R250 000 and R500 000 pm
2	1	0	0	0	0	0	0	0	0	0	0	l = Between R500 000 and R1 000 000 pm

The decision stump predictability is lower than that of traditional statistical regression. It does not show a positive Kappa statistic and the root absolute error is very close to 100%. This indicates that the algorithm has not been able to learn much from the data. The confusion matrix demonstrates that most positive matches came from those classified as ‘Less than R1000 pm’.

4.10.4 Random Tree

Table 33. Random Tree Turnover Summary

Correctly Classified Instances	25	24.2718 %
Incorrectly Classified Instances	78	75.7282 %
Kappa statistic		0.1235

Mean absolute error	0.1262
Root mean squared error	0.3546
Relative absolute error	87,96 %
Root relative squared error	132,71 %
Total Number of Instances	103

Table 34. Random Tree Turnover Confusion Matrix

a	b	c	d	e	f	g	h	i	j	k	l	<-- Classified as
13	2	0	0	1	1	3	0	1	0	4	3	a = Less than R1000 pm
3	2	1	1	2	0	1	0	0	1	1	5	b = Between R1000 and R5000 pm
2	1	1	0	0	0	1	0	0	1	1	0	c = Between R5000 and R10 000 pm
1	0	1	0	0	1	2	0	1	0	1	0	d = Between R10 000 and R20 000 pm
1	2	0	0	0	1	1	0	0	0	0	1	e = Between R20 000 and R30 000 pm
2	0	0	1	1	1	3	0	1	1	0	0	f = Between R30 000 and R50 000 pm
1	5	0	1	0	0	6	0	0	0	0	2	g = Between R50 000 and R100 000 pm
0	0	0	0	1	0	0	0	0	0	0	0	h = Between R100 000 and R150 000 pm
1	0	0	0	1	0	0	0	0	0	0	0	i = Between R150 000 and R200 000 pm
0	0	0	0	0	1	1	0	0	0	0	0	j = Between R200 000 and R250 000 pm
0	0	1	0	0	1	0	0	0	1	2	0	k = Between R250 000 and R500 000 pm
1	1	0	0	0	0	1	0	0	0	0	0	l = Between R500 000 and R1 000 000 pm

The random tree predictability is higher than that of traditional statistical regression. It does show a positive Kappa statistic and a root absolute error much less than 100%. These are good indicators that the algorithm has learnt something from the data. The confusion matrix indicates that most positive matches came from those classified as ‘Less than R1000 pm’ followed by those classified as ‘Between R50 000 and R100 000 pm’.

4.10.5 Random Forest

Table 35. Random Forest Turnover Summary

Correctly Classified Instances	37	35.9223 %
Incorrectly Classified Instances	66	64.0777 %
Kappa statistic		0.2252
Mean absolute error		0.1323
Root mean squared error		0.2595
Relative absolute error		92,20 %
Root relative squared error		97,13 %
Total Number of Instances		103

Table 36. Random Forest Turnover Confusion Matrix

a	b	c	d	e	f	g	h	i	j	k	l	<-- Classified as
20	3	0	3	0	0	1	0	0	0	1	0	a = Less than R1000 pm
5	5	1	0	1	0	3	0	0	0	1	1	b = Between R1000 and R5000 pm
0	3	2	0	0	0	1	0	0	0	1	0	c = Between R5000 and R10 000 pm
5	1	0	0	1	0	0	0	0	0	0	0	d = Between R10 000 and R20 000 pm
0	2	0	1	0	0	2	0	1	0	0	0	e = Between R20 000 and R30 000 pm
3	2	0	0	1	3	0	0	0	0	1	0	f = Between R30 000 and R50 000 pm
4	2	0	2	1	1	5	0	0	0	0	0	g = Between R50 000 and R100 000 pm
0	0	0	0	1	0	0	0	0	0	0	0	h = Between R100 000 and R150 000 pm
2	0	0	0	0	0	0	0	0	0	0	0	i = Between R150 000 and R200 000 pm
0	0	0	0	0	0	1	0	0	0	1	0	j = Between R200 000 and R250 000 pm
0	1	0	0	0	1	0	0	0	1	2	0	k = Between R250 000 and R500 000 pm
0	2	0	0	1	0	0	0	0	0	0	0	l = Between R500 000 and R1 000 000 pm

The random forest predictability is substantially higher than that of traditional statistical regression - 35,9% vs 23,8%. It does show a positive Kappa statistic and a root absolute error of less than 100%. These are good indicators that the algorithm has learnt something from the data. The confusion matrix indicates that most positive matches came from those classified as ‘Less than R1000 pm’ followed by those classified as ‘Between R1000 and R5000 pm’ and ‘Between R50 000 and R100 000 pm. This algorithm demonstrates the highest predictability.

4.10.6 Decision table

Table 37. Decision Table Turnover Summary

Correctly Classified Instances	28	27.1845 %
Incorrectly Classified Instances	75	72.8155 %
Kappa statistic		0.0034
Mean absolute error		0.1438
Root mean squared error		0.2675
Relative absolute error		100,23 %
Root relative squared error		100,13 %
Total Number of Instances		103

Table 38. Decision Table Turnover Confusion Matrix

a	b	c	d	e	f	g	h	i	j	k	l	<-- Classified as
28	0	0	0	0	0	0	0	0	0	0	0	a = Less than R1000 pm
16	0	0	0	0	0	1	0	0	0	0	0	b = Between R1000 and R5000 pm
6	0	0	0	0	0	1	0	0	0	0	0	c = Between R5000 and R10 000 pm
7	0	0	0	0	0	0	0	0	0	0	0	d = Between R10 000 and R20 000 pm
6	0	0	0	0	0	0	0	0	0	0	0	e = Between R20 000 and R30 000 pm
10	0	0	0	0	0	0	0	0	0	0	0	f = Between R30 000 and R50 000 pm
15	0	0	0	0	0	0	0	0	0	0	0	g = Between R50 000 and R100 000 pm

a	b	c	d	e	f	g	h	i	j	k	l	<-- Classified as
1	0	0	0	0	0	0	0	0	0	0	0	h = Between R100 000 and R150 000 pm
2	0	0	0	0	0	0	0	0	0	0	0	i = Between R150 000 and R200 000 pm
2	0	0	0	0	0	0	0	0	0	0	0	j = Between R200 000 and R250 000 pm
5	0	0	0	0	0	0	0	0	0	0	0	k = Between R250 000 and R500 000 pm
3	0	0	0	0	0	0	0	0	0	0	0	l = Between R500 000 and R1 000 000 pm

The decision table predictability is higher than that of traditional statistical regression - 27,18% vs 23,8%. It does show a positive but very low Kappa statistic, however, the root absolute error is close to 100%. These indicate that the algorithm has not learnt much from the data. The confusion matrix indicates that all positive matches came from those classified as ‘Less than R1000 pm’.

4.10.7 Adaptive Boosting

Table 39. Adaptive Boosting Turnover Summary

Correctly Classified Instances	22	21.3592 %
Incorrectly Classified Instances	81	78.6408 %
Kappa statistic		-0.0301
Mean absolute error		0.1429
Root mean squared error		0.2716
Relative absolute error		99,58 %
Root relative squared error		101,67 %
Total Number of Instances		103

Table 40. Adaptive Boosting Turnover Confusion Matrix

a	b	c	d	e	f	g	h	i	j	k	l	<-- Classified as
19	6	0	0	0	0	3	0	0	0	0	0	a = Less than R1000 pm
12	2	0	0	0	0	3	0	0	0	0	0	b = Between R1000 and R5000 pm
5	1	0	0	0	0	1	0	0	0	0	0	c = Between R5000 and R10 000 pm

a	b	c	d	e	f	g	h	i	j	k	l	<-- Classified as
5	1	0	0	0	0	1	0	0	0	0	0	d = Between R10 000 and R20 000 pm
4	2	0	0	0	0	0	0	0	0	0	0	e = Between R20 000 and R30 000 pm
5	3	0	0	0	0	2	0	0	0	0	0	f = Between R30 000 and R50 000 pm
11	3	0	0	0	0	1	0	0	0	0	0	g = Between R50 000 and R100 000 pm
1	0	0	0	0	0	0	0	0	0	0	0	h = Between R100 000 and R150 000 pm
1	1	0	0	0	0	0	0	0	0	0	0	i = Between R150 000 and R200 000 pm
2	0	0	0	0	0	0	0	0	0	0	0	j = Between R200 000 and R250 000 pm
4	1	0	0	0	0	0	0	0	0	0	0	k = Between R250 000 and R500 000 pm
2	1	0	0	0	0	0	0	0	0	0	0	l = Between R500 000 and R1 000 000 pm

The adaptive boosting predictability is lower than that of traditional statistical regression. It does not show a positive Kappa statistic and has a root absolute error close to 100%. These are good indicators that the algorithm has not learnt something from the data. The confusion matrix indicates that most positive matches came from those classified as ‘Less than R1000 pm’.

4.10.8 Bagging

Table 41. Bagging Turnover Summary

Correctly Classified Instances	31	30.0971 %
Incorrectly Classified Instances	72	69.9029 %
Kappa statistic		0.1323
Mean absolute error		0.1342
Root mean squared error		0.2693
Relative absolute error		93,55 %
Root relative squared error		100791,00 %
Total Number of Instances		103

Table 42. Bagging Turnover Confusion Matrix

a	b	c	d	e	f	g	h	i	j	k	l	<-- Classified as
21	1	0	3	0	0	3	0	0	0	0	0	a = Less than R1000 pm
5	6	0	0	3	2	1	0	0	0	0	0	b = Between R1000 and R5000 pm
3	2	0	0	0	1	1	0	0	0	0	0	c = Between R5000 and R10 000 pm
2	1	1	0	0	0	1	0	0	1	1	0	d = Between R10 000 and R20 000 pm
2	3	0	0	0	0	1	0	0	0	0	0	e = Between R20 000 and R30 000 pm
4	4	0	0	0	2	0	0	0	0	0	0	f = Between R30 000 and R50 000 pm
5	6	0	0	0	1	2	0	0	1	0	0	g = Between R50 000 and R100 000 pm
1	0	0	0	0	0	0	0	0	0	0	0	h = Between R100 000 and R150 000 pm
1	0	0	0	0	0	1	0	0	0	0	0	i = Between R150 000 and R200 000 pm
0	0	0	1	0	0	1	0	0	0	0	0	j = Between R200 000 and R250 000 pm
2	2	0	1	0	0	0	0	0	0	0	0	k = Between R250 000 and R500 000 pm
1	1	0	0	0	1	0	0	0	0	0	0	l = Between R500 000 and R1 000 000 pm

The bagging predictability is substantially higher than that of traditional statistical regression - 30% vs 23,8%. It does show a positive Kappa statistic and a root absolute error of less than 100%. These are good indicators that the algorithm has learnt something from the data. The confusion matrix indicates that most positive matches came from those classified as ‘Less than R1000 pm’ followed by those classified as ‘Between R1000 and R5000 pm’.

4.10.9 Bayes Net

Table 43. Bayes Net Turnover Summary

Correctly Classified Instances	19	18.4466 %
Incorrectly Classified Instances	84	81.5534 %
Kappa statistic		-0.0383
Mean absolute error		0.1411
Root mean squared error		0.2708
Relative absolute error		98,36 %
Root relative squared error		101,35 %
Total Number of Instances		103

Table 44. Bayes Net Turnover Confusion Matrix

a	b	c	d	e	f	g	h	i	j	k	l	<-- Classified as
15	4	1	1	0	2	4	0	0	1	0	0	a = Less than R1000 pm
10	3	0	0	0	1	3	0	0	0	0	0	b = Between R1000 and R5000 pm
4	1	0	0	0	0	2	0	0	0	0	0	c = Between R5000 and R10 000 pm
5	0	0	0	0	0	2	0	0	0	0	0	d = Between R10 000 and R20 000 pm
5	1	0	0	0	0	0	0	0	0	0	0	e = Between R20 000 and R30 000 pm
7	0	0	0	0	0	3	0	0	0	0	0	f = Between R30 000 and R50 000 pm
9	3	0	0	0	1	1	0	0	0	1	0	g = Between R50 000 and R100 000 pm
1	0	0	0	0	0	0	0	0	0	0	0	h = Between R100 000 and R150 000 pm
1	0	0	0	0	0	1	0	0	0	0	0	i = Between R150 000 and R200 000 pm
0	0	0	0	0	0	1	0	0	0	1	0	j = Between R200 000 and R250 000 pm
2	1	0	1	0	0	1	0	0	0	0	0	k = Between R250 000 and R500 000 pm
1	1	0	0	0	0	1	0	0	0	0	0	l = Between R500 000 and R1 000 000 pm

The Bayes Net predictability is substantially lower than that of traditional statistical regression. It does not demonstrate a positive Kappa statistic

and the root absolute error is close to 100%. These are good indicators that the algorithm has not learnt something from the data. The confusion matrix indicates that most positive matches came from those classified as ‘Less than R1000 pm’.

4.10.10 SMO

Table 45. SMO Turnover Summary

Correctly Classified Instances	20	19.4175 %
Incorrectly Classified Instances	83	80.5825 %
Kappa statistic		-0.0335
Mean absolute error		0.1465
Root mean squared error		0.2697
Relative absolute error		102,07 %
Root relative squared error		100,93 %
Total Number of Instances		103

Table 46. SMO Turnover Confusion Matrix

a	b	c	d	e	f	g	h	i	j	k	l	<-- Classified as
14	4	2	0	0	2	4	0	0	0	2	0	a = Less than R1000 pm
12	4	0	0	0	0	0	0	0	0	1	0	b = Between R1000 and R5000 pm
5	0	0	0	0	0	2	0	0	0	0	0	c = Between R5000 and R10 000 pm
6	0	0	0	0	0	1	0	0	0	0	0	d = Between R10 000 and R20 000 pm
5	1	0	0	0	0	0	0	0	0	0	0	e = Between R20 000 and R30 000 pm
8	0	0	0	0	0	2	0	0	0	0	0	f = Between R30 000 and R50 000 pm
9	3	0	1	0	0	2	0	0	0	0	0	g = Between R50 000 and R100 000 pm
1	0	0	0	0	0	0	0	0	0	0	0	h = Between R100 000 and R150 000 pm
1	1	0	0	0	0	0	0	0	0	0	0	i = Between R150 000 and R200 000 pm
1	0	0	1	0	0	0	0	0	0	0	0	j = Between R200 000 and R250 000 pm

a	b	c	d	e	f	g	h	i	j	k	l	<-- Classified as
3	1	0	0	0	0	0	0	0	1	0	0	k = Between R250 000 and R500 000 pm
0	2	0	0	0	0	1	0	0	0	0	0	l = Between R500 000 and R1 000 000 pm

The SMO predictability is substantially lower than that of traditional statistical regression. It does not demonstrate a positive Kappa statistic and the root absolute error is over 100%. These are good indicators that the algorithm has not learnt something from the data. The confusion matrix indicates that most positive matches came from those classified as ‘Less than R1000 pm’.

4.11 Machine Learning output for number of staff

4.11.1 Zero R

Table 47. Zero R Staff Summary

Correctly Classified Instances	36	34.9515 %
Incorrectly Classified Instances	67	65.0485 %
Kappa statistic		0
Mean absolute error		0.1499
Root mean squared error		0.2713
Relative absolute error		100,00 %
Root relative squared error		100,00 %
Total Number of Instances		103

Table 48. Zero R Staff Confusion Matrix

a	b	c	d	e	f	g	h	i	j	<-- Classified as
0	26	0	0	0	0	0	0	0	0	a = Staff < 2
0	36	0	0	0	0	0	0	0	0	b = Staff < 5
0	29	0	0	0	0	0	0	0	0	c = Staff < 10
0	4	0	0	0	0	0	0	0	0	d = Staff < 15
0	2	0	0	0	0	0	0	0	0	e = Staff < 20
0	1	0	0	0	0	0	0	0	0	f = Staff < 30

a	b	c	d	e	f	g	h	i	j	<-- Classified as
0	3	0	0	0	0	0	0	0	0	g = Staff < 40
0	0	0	0	0	0	0	0	0	0	h = Staff < 50
0	1	0	0	0	0	0	0	0	0	i = Staff < 75
0	1	0	0	0	0	0	0	0	0	j = Staff < 100

The Zero R predictability is substantially higher than that of traditional statistical regression - 34,95% vs 25,4%. It does not demonstrate a positive Kappa statistic. The root absolute error is close to 100%. These are good indicators that the algorithm has not learnt something from the data. The confusion matrix indicates that all positive matches came from those classified as ‘Staff < 5’.

4.11.2 J48

Table 49. J48 Staff Summary

Correctly Classified Instances	36	34.9515 %
Incorrectly Classified Instances	67	65.0485 %
Kappa statistic		0.0889
Mean absolute error		0.1306
Root mean squared error		0.3201
Relative absolute error		87,14 %
Root relative squared error		117,99 %
Total Number of Instances		103

Table 50. J48 Staff Confusion Matrix

a	b	c	d	e	f	g	h	i	j	<-- Classified as
8	8	8	2	0	0	0	0	0	0	a = Staff < 2
7	17	10	1	0	0	0	0	0	1	b = Staff < 5
4	12	11	2	0	0	0	0	0	0	c = Staff < 10
2	1	1	0	0	0	0	0	0	0	d = Staff < 15
2	0	0	0	0	0	0	0	0	0	e = Staff < 20
0	0	1	0	0	0	0	0	0	0	f = Staff < 30
1	0	2	0	0	0	0	0	0	0	g = Staff < 40
0	0	0	0	0	0	0	0	0	0	h = Staff < 50
0	0	1	0	0	0	0	0	0	0	i = Staff < 75
0	1	0	0	0	0	0	0	0	0	j = Staff < 100

The Zero R predictability is substantially higher than that of traditional statistical regression - 34,95% vs 25,4%. It does not demonstrate a positive Kappa statistic. The root absolute error is close to 100%. These are good indicators that the algorithm has not learnt something from the data. The confusion matrix indicates that all positive matches came from those classified as ‘Staff < 5’.

4.11.3 Decision Stump

Table 51. Decision Stump Staff Summary

Correctly Classified Instances	35	33.9806 %
Incorrectly Classified Instances	68	66.0194 %
Kappa statistic		0.0156
Mean absolute error		0.1461
Root mean squared error		0.2752
Relative absolute error		97,46 %
Root relative squared error		101,42 %
Total Number of Instances		103

Table 52. Decision Stump Staff Confusion Matrix

a	b	c	d	e	f	g	h	i	j	<-- Classified as
2	20	4	0	0	0	0	0	0	0	a = Staff < 2
4	30	2	0	0	0	0	0	0	0	b = Staff < 5
4	22	3	0	0	0	0	0	0	0	c = Staff < 10
0	3	1	0	0	0	0	0	0	0	d = Staff < 15
0	0	2	0	0	0	0	0	0	0	e = Staff < 20
0	1	0	0	0	0	0	0	0	0	f = Staff < 30
1	2	0	0	0	0	0	0	0	0	g = Staff < 40
0	0	0	0	0	0	0	0	0	0	h = Staff < 50
1	0	0	0	0	0	0	0	0	0	i = Staff < 75
1	0	0	0	0	0	0	0	0	0	j = Staff < 100

The decision stump predictability is slightly higher than that of traditional statistical regression. It does not demonstrate a positive Kappa statistic.

The root absolute error is lower than 100%. It is unclear as to whether much was learnt from the data. The confusion matrix indicates that most positive matches came from those classified as ‘Staff < 5’.

4.11.4 Random Tree

Table 53. Random Tree Staff Summary

Correctly Classified Instances	28	27.1845 %
Incorrectly Classified Instances	75	72.8155 %
Kappa statistic		-6
Mean absolute error		0.1446
Root mean squared error		0.3786
Relative absolute error		96,44 %
Root relative squared error		139,54 %
Total Number of Instances		103

Table 54. Random Tree Staff Confusion Matrix

a	b	c	d	e	f	g	h	i	j	<-- Classified as
1	11	9	0	0	1	3	0	0	1	a = Staff < 2
6	17	10	1	0	0	2	0	0	0	b = Staff < 5
5	12	10	0	0	2	0	0	0	0	c = Staff < 10
0	0	4	0	0	0	0	0	0	0	d = Staff < 15
1	0	0	0	0	0	1	0	0	0	e = Staff < 20
0	1	0	0	0	0	0	0	0	0	f = Staff < 30
3	0	0	0	0	0	0	0	0	0	g = Staff < 40
0	0	0	0	0	0	0	0	0	0	h = Staff < 50
0	0	1	0	0	0	0	0	0	0	i = Staff < 75
1	0	0	0	0	0	0	0	0	0	j = Staff < 100

The random tree predictability is slightly higher than that of traditional statistical regression. It does not demonstrate a positive Kappa statistic. The root absolute error is lower than 100%. It is unclear as to whether much was learnt from the data. The confusion matrix indicates that most positive matches came from those classified as ‘Staff < 5’.

4.11.5 Random Forest

Table 55. Random Forest Staff Summary

Correctly Classified Instances	42	40.7767 %
Incorrectly Classified Instances	61	59.2233 %
Kappa statistic		0.1481
Mean absolute error		0.1402
Root mean squared error		0.2692
Relative absolute error		93,56 %
Root relative squared error		99,22 %
Total Number of Instances		103

Table 56. Random Forest Staff Confusion Matrix

a	b	c	d	e	f	g	h	i	j	<-- Classified as
6	12	8	0	0	0	0	0	0	0	a = Staff < 2
4	24	8	0	0	0	0	0	0	0	b = Staff < 5
4	14	10	1	0	0	0	0	0	0	c = Staff < 10
1	1	2	0	0	0	0	0	0	0	d = Staff < 15
0	0	0	0	2	0	0	0	0	0	e = Staff < 20
0	0	1	0	0	0	0	0	0	0	f = Staff < 30
0	1	2	0	0	0	0	0	0	0	g = Staff < 40
0	0	0	0	0	0	0	0	0	0	h = Staff < 50
0	0	1	0	0	0	0	0	0	0	i = Staff < 75
0	1	0	0	0	0	0	0	0	0	j = Staff < 100

The random forest predictability is substantially higher than that of traditional statistical regression - 40,7% vs 25,4%. It does demonstrate a positive but low Kappa statistic. The root absolute error is much lower than 100%. These are good indicators that the algorithm has learnt something from the data. The confusion matrix indicates that most positive matches

came from those classified as ‘Staff < 5’, followed by ‘Staff < 10’ and ‘Staff < 2’. This algorithm demonstrates the highest predictability.

4.11.6 Decision table

Table 57. Decision Table Staff Summary

Correctly Classified Instances	31	30.0971 %
Incorrectly Classified Instances	72	69.9029 %
Kappa statistic		-0.0522
Mean absolute error		0.1558
Root mean squared error		0.2776
Relative absolute error		103,96 %
Root relative squared error		102,30 %
Total Number of Instances		103

Table 58. Decision Table Staff Confusion Matrix

a	b	c	d	e	f	g	h	i	j	<-- Classified as
1	22	3	0	0	0	0	0	0	0	a = Staff < 2
5	29	1	1	0	0	0	0	0	0	b = Staff < 5
2	26	1	0	0	0	0	0	0	0	c = Staff < 10
0	4	0	0	0	0	0	0	0	0	d = Staff < 15
0	2	0	0	0	0	0	0	0	0	e = Staff < 20
0	1	0	0	0	0	0	0	0	0	f = Staff < 30
0	3	0	0	0	0	0	0	0	0	g = Staff < 40
0	0	0	0	0	0	0	0	0	0	h = Staff < 50
0	1	0	0	0	0	0	0	0	0	i = Staff < 75
0	1	0	0	0	0	0	0	0	0	j = Staff < 100

The decision table predictability is higher than that of traditional statistical regression - 30,09% vs 25,4%. It does not demonstrate a positive Kappa statistic. The root absolute error is more than 100%. These are good indicators that the algorithm has not learnt something from the data. The confusion matrix indicates that most positive matches came from those classified as ‘Staff < 5’.

4.11.7 Adaptive Boosting

Table 59. Adaptive Boosting Staff Summary

Incorrectly Classified Instances	68 66.0194 %
Kappa statistic	0.0156
Mean absolute error	0.1461
Root mean squared error	0.2752
Relative absolute error	97,46 %
Root relative squared error	101,42 %
Total Number of Instances	103

Table 60. Adaptive Boosting Staff Confusion Matrix

a	b	c	d	e	f	g	h	i	j	<-- Classified as
2	20	4	0	0	0	0	0	0	0	a = Staff < 2
4	30	2	0	0	0	0	0	0	0	b = Staff < 5
4	22	3	0	0	0	0	0	0	0	c = Staff < 10
0	3	1	0	0	0	0	0	0	0	d = Staff < 15
0	0	2	0	0	0	0	0	0	0	e = Staff < 20
0	1	0	0	0	0	0	0	0	0	f = Staff < 30
1	2	0	0	0	0	0	0	0	0	g = Staff < 40
0	0	0	0	0	0	0	0	0	0	h = Staff < 50
1	0	0	0	0	0	0	0	0	0	i = Staff < 75
1	0	0	0	0	0	0	0	0	0	j = Staff < 100

The adaptive boosting predictability is much higher than that of traditional statistical regression - 33,98% vs 25,4%. It does demonstrate a positive Kappa statistic. The root absolute error is lower than 100%. These are good indicators that the algorithm has learnt something from the data. The confusion matrix indicates that most positive matches came from those classified as ‘Staff < 5’.

Table 61. Bagging Staff Summary

Correctly Classified Instances	28	27.1845 %
Incorrectly Classified Instances	75	72.8155 %
Kappa statistic		-54
Mean absolute error		0.1461
Root mean squared error		0.2799
Relative absolute error		97,46 %
Root relative squared error		103,16 %
Total Number of Instances		103

Table 62. Bagging Staff Confusion Matrix

a	b	c	d	e	f	g	h	i	j	<-- Classified as
3	12	11	0	0	0	0	0	0	0	a = Staff < 2
6	19	11	0	0	0	0	0	0	0	b = Staff < 5
9	14	6	0	0	0	0	0	0	0	c = Staff < 10
0	2	2	0	0	0	0	0	0	0	d = Staff < 15
0	0	2	0	0	0	0	0	0	0	e = Staff < 20
0	1	0	0	0	0	0	0	0	0	f = Staff < 30
1	0	2	0	0	0	0	0	0	0	g = Staff < 40
0	0	0	0	0	0	0	0	0	0	h = Staff < 50
0	1	0	0	0	0	0	0	0	0	i = Staff < 75
0	1	0	0	0	0	0	0	0	0	j = Staff < 100

The Bagging predictability is slightly higher than that of traditional statistical regression. It does not demonstrate a positive Kappa statistic. The root absolute error is lower than 100%. It is unclear as to whether the algorithm has learnt much from the training data. The confusion matrix indicates that most positive matches came from those classified as ‘Staff < 5’ followed by ‘Staff < 10’ and ‘Staff < 5’ respectively.

4.11.9 Bayes Net

Table 63. Bayes Net Staff Summary

Correctly Classified Instances	30	29.1262 %
Incorrectly Classified Instances	73	70.8738 %
Kappa statistic		-43

Mean absolute error	0.1478
Root mean squared error	0.2782
Relative absolute error	98,58 %
Root relative squared error	102,52 %
Total Number of Instances	103

Table 64. Bayes Net Staff Confusion Matrix

a	b	c	d	e	f	g	h	i	j	<-- Classified as
3	19	4	0	0	0	0	0	0	0	a = Staff < 2
7	22	6	1	0	0	0	0	0	0	b = Staff < 5
4	20	5	0	0	0	0	0	0	0	c = Staff < 10
0	2	2	0	0	0	0	0	0	0	d = Staff < 15
0	2	0	0	0	0	0	0	0	0	e = Staff < 20
0	1	0	0	0	0	0	0	0	0	f = Staff < 30
1	1	1	0	0	0	0	0	0	0	g = Staff < 40
0	0	0	0	0	0	0	0	0	0	h = Staff < 50
0	1	0	0	0	0	0	0	0	0	i = Staff < 75
0	1	0	0	0	0	0	0	0	0	j = Staff < 100

The Bayes Net predictability is slightly higher than that of traditional statistical regression. It does not demonstrate a positive Kappa statistic. The root absolute error is lower than 100%. It is unclear as to whether the algorithm has learnt much from the training data. The confusion matrix indicates that most positive matches came from those classified as ‘Staff < 5’.

4.11.10 SMO

Table 65. SMO Staff Summary

Correctly Classified Instances	29	28.1553 %
Incorrectly Classified Instances	74	71.8447 %
Kappa statistic		-0.0366
Mean absolute error		0.1667
Root mean squared error		0.2846
Relative absolute error		111,21 %

Root relative squared error	104,89 %
Total Number of Instances	103

Table 66. SMO Staff Confusion Matrix

a	b	c	d	e	f	g	h	i	j	<-- Classified as
9	14	3	0	0	0	0	0	0	0	a = Staff < 2
10	17	9	0	0	0	0	0	0	0	b = Staff < 5
5	20	3	1	0	0	0	0	0	0	c = Staff < 10
0	1	3	0	0	0	0	0	0	0	d = Staff < 15
1	0	1	0	0	0	0	0	0	0	e = Staff < 20
1	0	0	0	0	0	0	0	0	0	f = Staff < 30
1	1	1	0	0	0	0	0	0	0	g = Staff < 40
0	0	0	0	0	0	0	0	0	0	h = Staff < 50
1	0	0	0	0	0	0	0	0	0	i = Staff < 75
0	1	0	0	0	0	0	0	0	0	j = Staff < 100

The SMO predictability is slightly higher than that of traditional statistical regression. It does not demonstrate a positive Kappa statistic. The root absolute error is more than 100%. This demonstrate that the has not learnt much from the training data. The confusion matrix indicates that most positive matches came from those classified as ‘Staff < 5’f followed by ‘Staff < 2’ and ‘Staff < 5’ respectively.

4.12 Results

Here is a quick visual breakdown of the results:

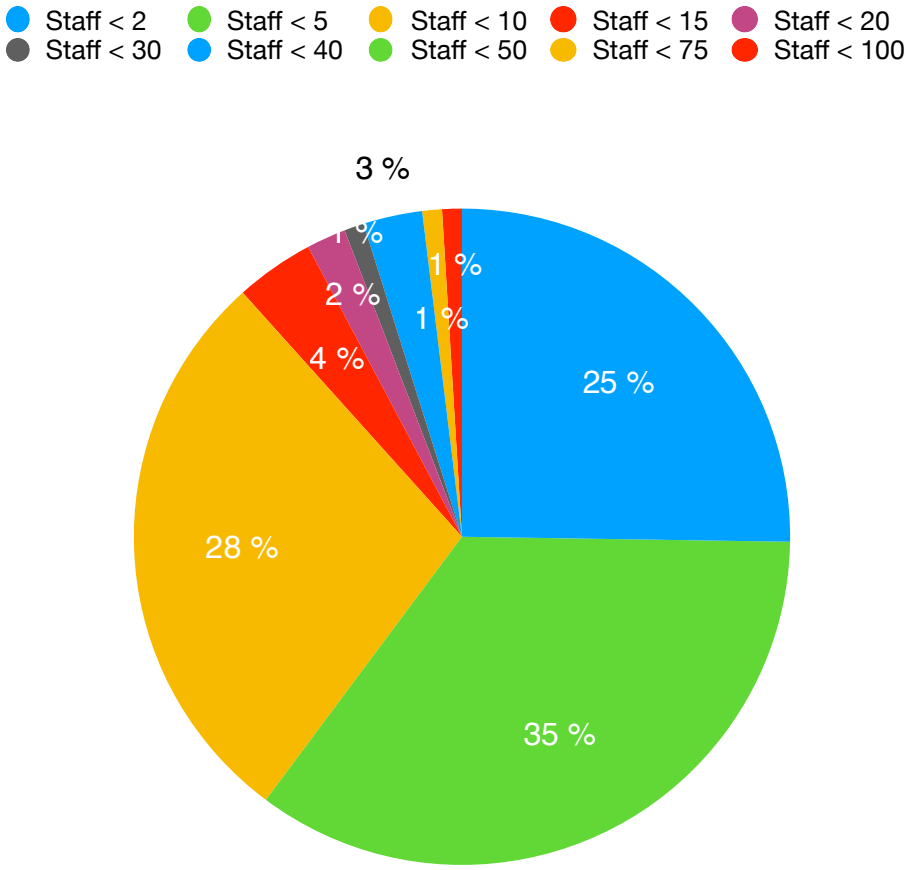

Figure 14. Staff Pie chart

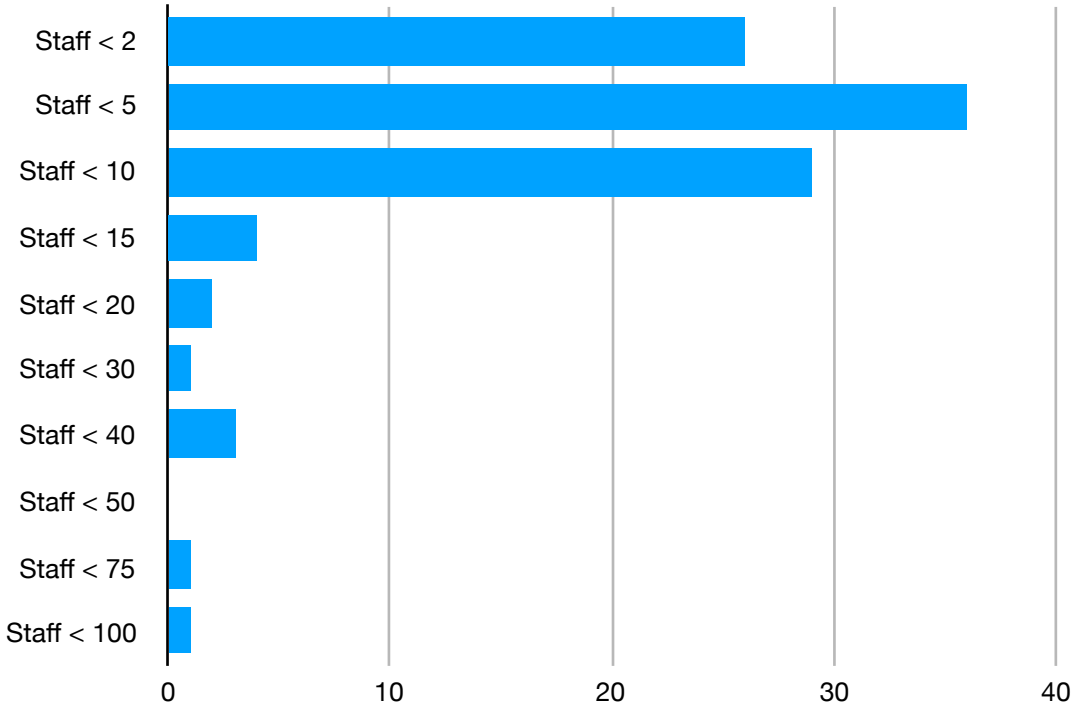

Figure 15. Staff Bar chart

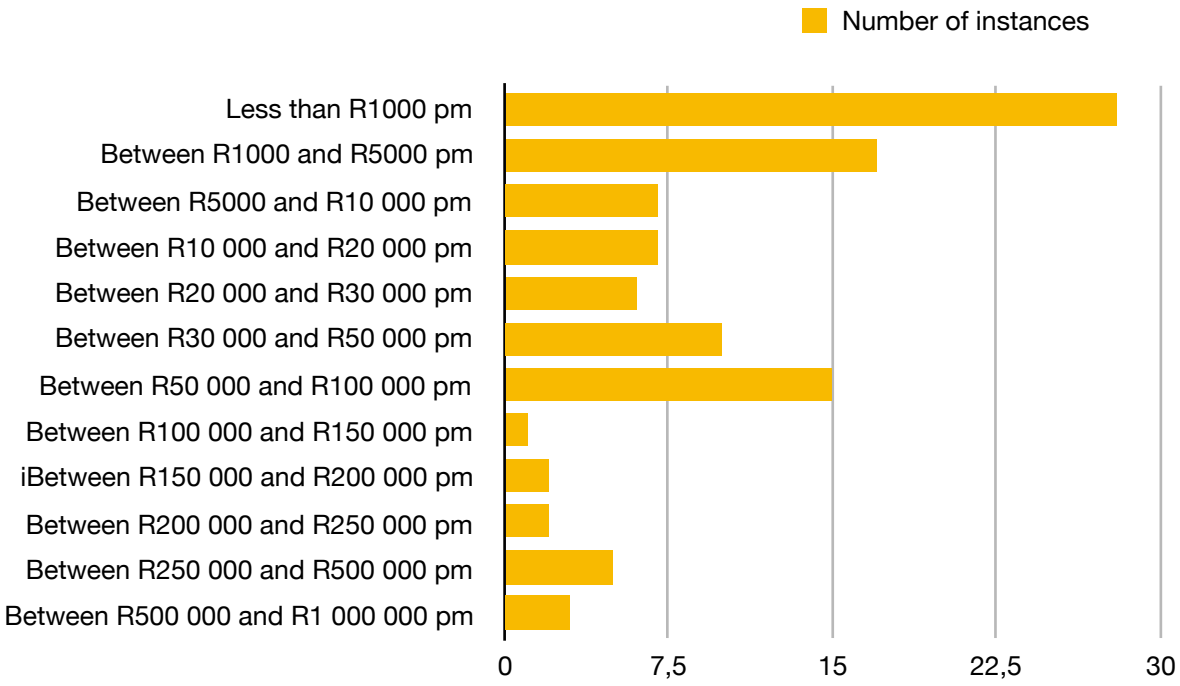

Figure 16. Turnover Bar chart

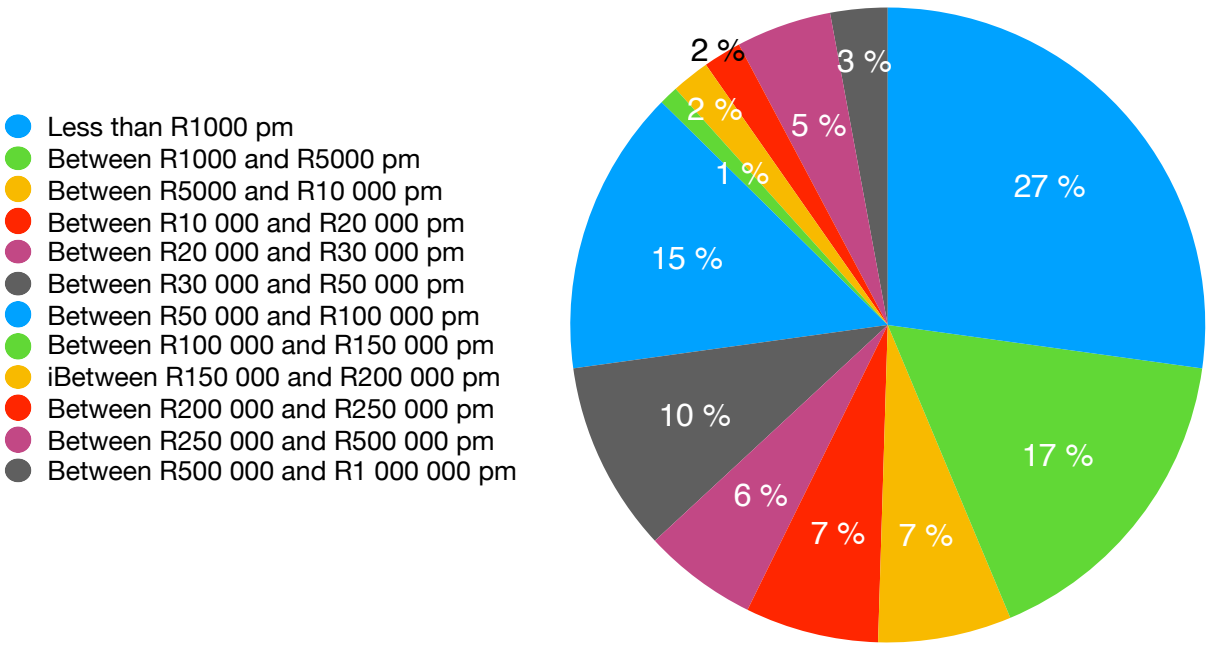

Figure 17. Turnover Pie chart

Table 67. Breakdown of staff

Value	Number of instances
Staff < 2	26
Staff < 5	36
Staff < 10	29
Staff < 15	4
Staff < 20	2
Staff < 30	1
Staff < 40	3
Staff < 50	0
Staff < 75	1
Staff < 100	1

Table 68. Breakdown of turnover

Value	Number of instances
Less than R1000 pm	28
Between R1000 and R5000 pm	17
Between R5000 and R10 000 pm	7
Between R10 000 and R20 000 pm	7
Between R20 000 and R30 000 pm	6
Between R30 000 and R50 000 pm	10
Between R50 000 and R100 000 pm	15
Between R100 000 and R150 000 pm	1
iBetween R150 000 and R200 000 pm	2
Between R200 000 and R250 000 pm	2
Between R250 000 and R500 000 pm	5
Between R500 000 and R1 000 000 pm	3

4.13 Summary

The purpose of this research was to investigate if a computational model could be formed which could accurately determine organisational performance of a startup. This research differs from other research in a few regards:

- It is conducted in South Africa where very limited studies have been performed with regards to startups. Much research has been performed in developing countries, however, there have been very few in African countries.

- It measures organisational performance - which is defined as turnover and the number of staff employed. Most studies perform business failure or success research. While understanding whether a business will fail or not is very useful, it was thought to be more useful if the actual level of success of that business could be measured.

- This study created a computational model based on previous research and literature. The literature review had led to the utilisation of 15 dependent variables and 2 independent variables. A survey was created based upon the factors discussed in the literature review. The survey was distributed to incubators via email and in person. The results of the survey were collected online and statistical analysis was performed on the results.

The traditional statistical regression demonstrates a classification rate of 23,8% for the independent variable ‘turnover’ and 25,4% for the independent variable ‘Number of staff’.

It compared traditional statistical methods to machine learning algorithms in order to see if the classification rate would be better. This

research has proved that a computational model performed using machine learning algorithms could indeed lead to an improved classification rate as compared to traditional regression.

Ten machine learning algorithms were then run on the data in order to determine if this could lead to an improved classification rate. For both independent variables, ‘turnover’ and ‘number of staff’, the Random Forest machine learning algorithm lead to an improved classification rate. For turnover, a rate of 35,93% was calculated - 7,52% higher than the regression figure. For number of staff, a rate of 40,78% was calculated - 15,38% higher than the regression figure.

Despite the positive results, there are some limitations to this research which would require additional research involving creating a more robust theoretical framework as well as more qualitative work in this area of research. In summary, machine learning algorithms could be used to create an alternative computational model. However, the results of the model have proved to be poor.

CHAPTER 5

5.1 Introduction

As previously stated, the purpose of this research was to create a computational model which could accurately predict the organisational performance of startups - where organisational performance was defined as both the turnover and number of staff employed. In this chapter, the findings pertaining to this research will be discussed.

In Chapter 4, the details of the results of the finding were presented. It was found that model displayed a low to moderate level of accuracy (35,92% for turnover predictability and 40,78% for number of staff). It was, however, demonstrated that by using a machine learning algorithm, the predicability was improved. In this chapter the details of those findings and a comparison of those algorithms will be discussed. The most accurate algorithm - the random forest algorithm - will be discussed in more detail.

SPSS was used to perform the statistical analysis on this data. WEKA was used to perform the machine learning algorithms. The details of the findings using both these tools will be discussed in more detail.

5.2 Demographic Profile of respondents

Six incubators were included in this study. In total there were 103 respondents. 54,37% were from the national incubator, Shanduka Black Umbrellas. Riversands and Innovation hub, which are both located in the Gauteng region, had a combined 36,9% of respondents. The rest of the incubators accounted for 8,73% of respondents.

The respondent sizes were representative of the hubs themselves. Shandukah Black Umbrellas was by far the biggest incubator, in terms of the incubatees held, followed by Innovation Hub and then Riversands.

The sample seems to be reasonably in line with the demographics of South Africa. The last reported census results of 2017 were:

Table 69. Breakdown by race

Race	Percentage
Black	80,8 %
Coloured	8,8 %
Indian/Asian	2,5 %
White	8,0 %

The purpose of many of these incubators are to empower previously disadvantaged people and the data does demonstrate that incubators are certainly doing this.

The gender breakdown, however, seems to favour males considerably. There were very few founding partners consisting of males and females. Females still need to be empowered more than they currently are. The department of Trade and Industry in South Africa spends considerable effort on the empowerment of women by creating various programmes. There is perhaps good reason to create incubators which focus solely on the empowerment of previously disadvantage women.

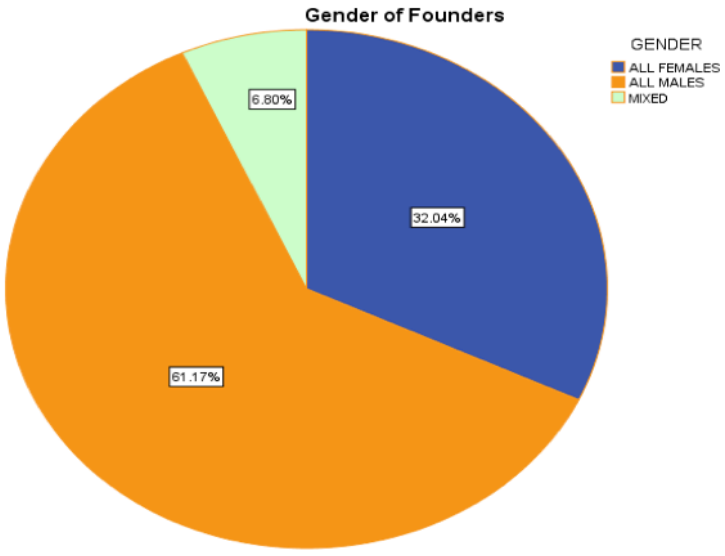

Figure 18. Gender of founders

Table 70. Turnover breakdown by gender

Row Labels	ALL FEMALES	ALL MALES	MIXED
Less than R1000 pm	8	20	
Between R1000 and R5000 pm	6	10	1
Between R5000 and R10 000 pm	2	4	1
Between R10 000 and R20 000 pm	2	4	1
Between R20 000 and R30 000 pm	1	5	
Between R30 000 and R50 000 pm	4	6	
Between R50 000 and R100 000 pm	6	7	2
Between R100 000 and R150 000 pm		1	
Between R150 000 and R200 000 pm	1	1	
Between R200 000 and R250 000 pm		1	1
Between R250 000 and R500 000 pm	2	2	1
Between R500 000 and R1 000 000 pm	1	2	
Grand Total	**33**	**63**	**7**

Table 71. Turnover breakdown by race

Row Labels	ALL BLACK	ALL COLOURED	ALL INDIAN	ALL WHITE	MIXED
Less than R1000 pm	14	4	1	6	3
Between R1000 and R5000 pm	14	1		1	1
Between R5000 and R10 000 pm	5			2	
Between R10 000 and R20 000 pm	3	2	2		
Between R20 000 and R30 000 pm	5	1			
Between R30 000 and R50 000 pm	9	1			
Between R50 000 and R100 000 pm	11	3	1		
Between R100 000 and R150 000 pm		1			
Between R150 000 and R200 000 pm	1			1	
Between R200 000 and R250 000 pm		1	1		

Row Labels	ALL BLACK	ALL COLOURED	ALL INDIAN	ALL WHITE	MIXED
Between R250 000 and R500 000 pm	2		1		2
Between R500 000 and R1 000 000 pm	3				
Grand Total	**67**	**14**	**6**	**10**	**6**

Table 72. Staff breakdown by gender

Row Labels	ALL FEMALES	ALL MALES	MIXED
Staff < 2	9	17	
Staff < 5	11	23	2
Staff < 10	9	17	3
Staff < 15	1	2	1
Staff < 20	2		
Staff < 30		1	
Staff < 40	1	1	1
Staff < 75		1	
Staff < 100		1	
Grand Total'	**33**	**63**	**7**

Table 73. Staff breakdown by race

Row Labels	ALL BLACK	ALL COLOURED	ALL INDIAN	ALL WHITE	MIXED
Staff < 2	16	6	1	3	
Staff < 5	23	4	3	2	4
Staff < 10	18	3	2	5	1
Staff < 15	4				
Staff < 20	2				
Staff < 30	1				
Staff < 40	2	1			
Staff < 75	1				

Row Labels	ALL BLACK	ALL COLOURED	ALL INDIAN	ALL WHITE	MIXED
Staff < 100					1
Grand Total'	**67**	**14**	**6**	**10**	**6**

5.3 Traditional Regression vs Machine Learning

Table 74. Regression Comparison

	Variable	Turnover	Staff
1	Capital	**0,027**	0,497
2	Record Keeping and Financial Control	0,132	0,340
3	Industry Experience	0,651	0,146
4	Management Experience	0,425	0,674
5	Technical Expertise	0,118	0,795
6	Professional Advisors	**0,039**	0,077
7	Education	0,132	**0,007**
8	Staffing	0,677	0,207
9	Customer relations	0,855	0,884
10	Gender	0,611	0,500
11	Age	**0,014**	0,150
12	Number of Partners	0,659	**0,001**
13	Parents Owned Business	0,923	0,789
14	Race	0,61	0,875
15	Marketing	0,312	0,128

As can be seen in the table above only 5 variables played a significant role in predicting organisational performance. Those are:

- Capital
- Professional Advisors
- Education
- Age
- Number of partners

These finding will be discussed in more detail when the findings of each hypothesis are discussed in later sections.

Table 75. Machine Learning Algorithm Comparison

	Turnover	Number of Staff
Traditional regression	23,80 %	25,40 %
Zero R	27,18 %	34,95 %
J48	24,27 %	34,95 %
Decision Stump	21,36 %	33,98 %
Random Tree	24,27 %	27,18 %
Random Forest	**35,92 %**	**40,78 %**
Decision Table	27,18 %	30,10 %
Ada Boost	21,36 %	33,98 %
Bagging	30,10 %	27,18 %

5.4 Discussion of Hypotheses

H1: Higher levels of education have a positive effect on organisational performance.

The significance value reveals a weak relationship between level of education and turnover, however, a strong relationship is revealed between number of staff employed and level of education. It is good to see this positive relationship and it is an indicator that education is definitely a multi-faceted driving factor in solving employment issues in South Africa. In order to alleviate the problem of unemployment, incubators alone may not solve the issue.

One of the questions included in the survey which has not been exposed in the study was the growth in staff size. Prior to incubation, the 103 startups employed a total of 482 employees. At the point of taking the survey (while in incubation) that increased to 677 employees. This is an increase of just over 28%, which is quite significant. This demonstrates that incubation certainly positively affects employment.

H2: Technical expertise are positively associated with organisational performance.

In the cases of both turnover and number of staff employed, technical expertise has shown to have a weak relationship. One of the reasons may be that technical expertise on average was fairly high. There were 147 founders in a total of 103 startups. The average technical expertise of each individual was approximately 3,6 (on a scale of 1 to 5). If we look at the average per startup, it worked out to 5,1 - basically higher than the Likert scale. While the intention here was to capture the sum of all expertise, perhaps the average should have been taken rather than sum. This could be investigated in future research.

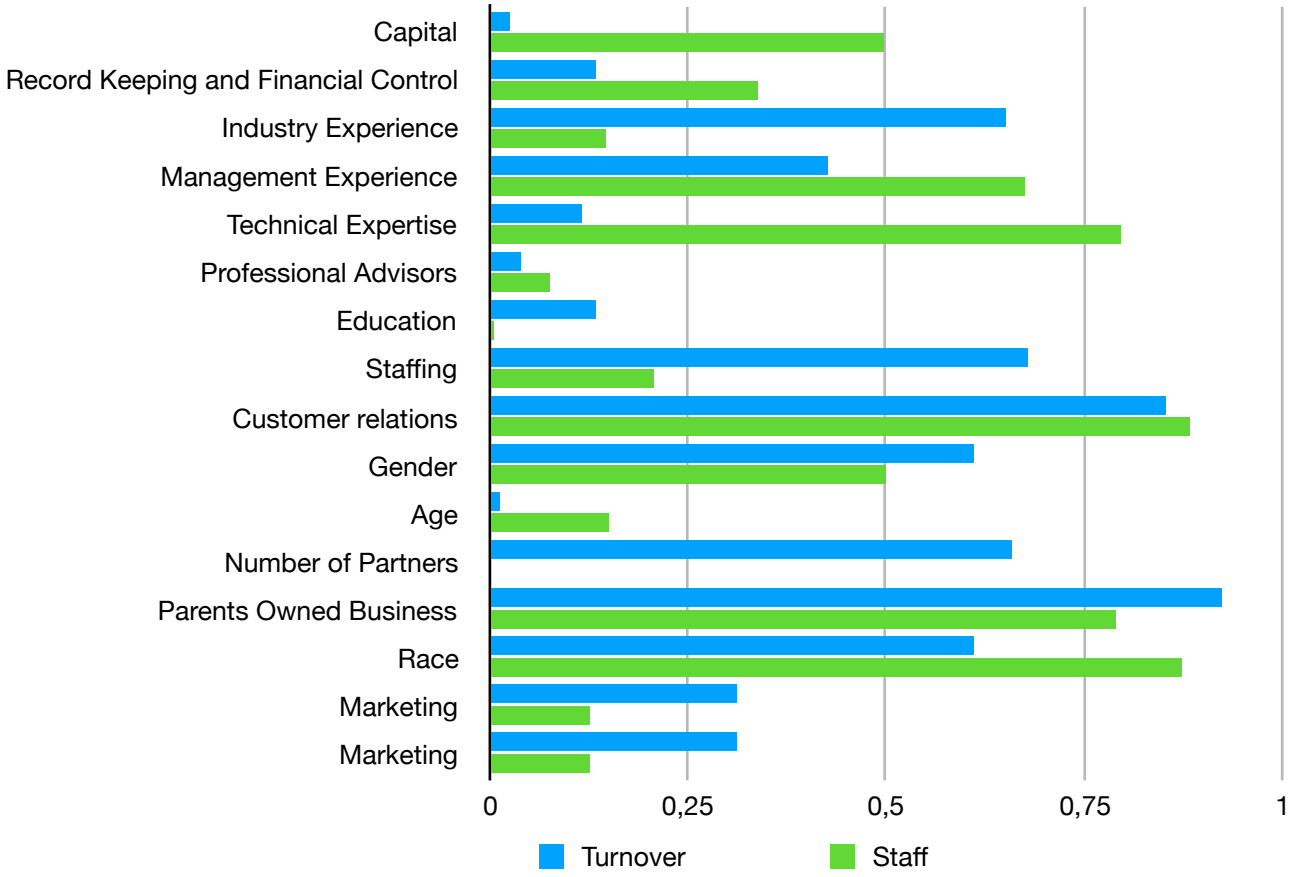

Figure 19. Comparison staff vs turnover (Regression)

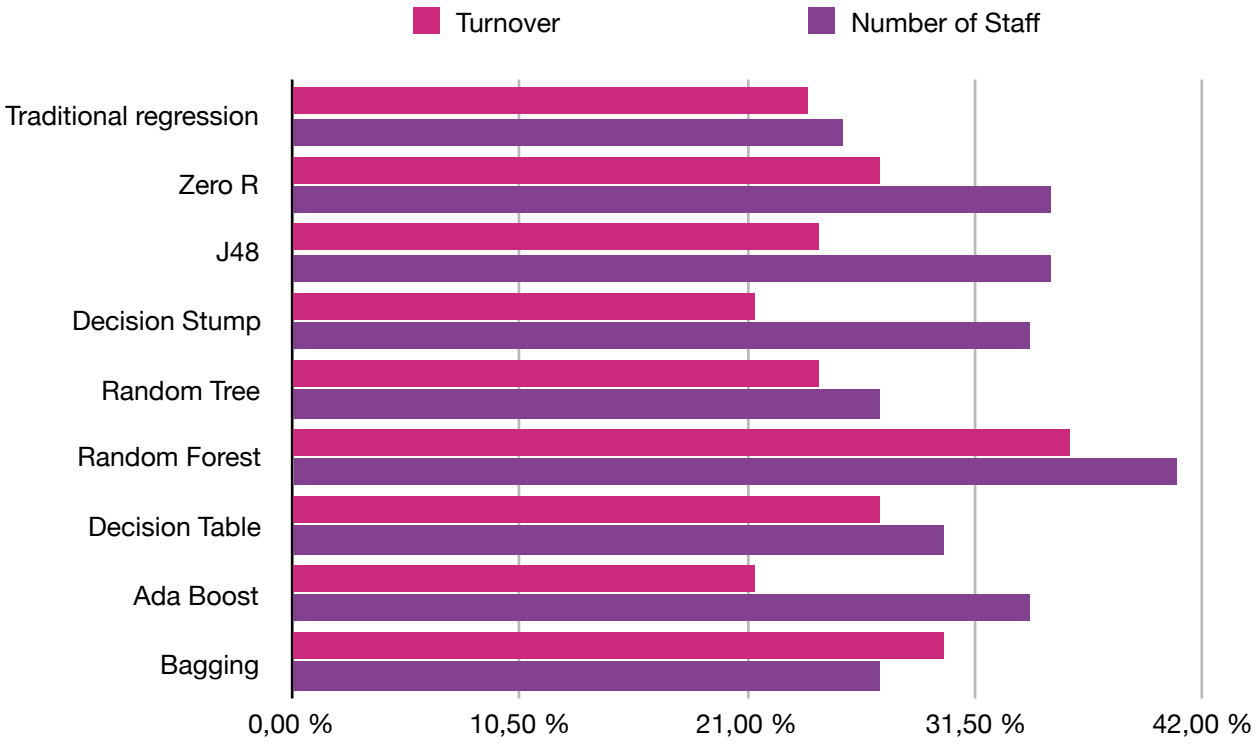

Figure 20. Comparison of machine learning algorithms

H3: Managerial experience has a positive effect on organisational performance.

Managerial experience has shown to have a weak correlation to both turnover and number of staff employed. In total, amongst the 147 founders, there was a combined 840 years of management experience. There were, however, edge cases where certain startups had extremely high values such as 84 years of combined management experience while many had no management experience at all. The graph below demonstrates the scatter of the 103 startups.

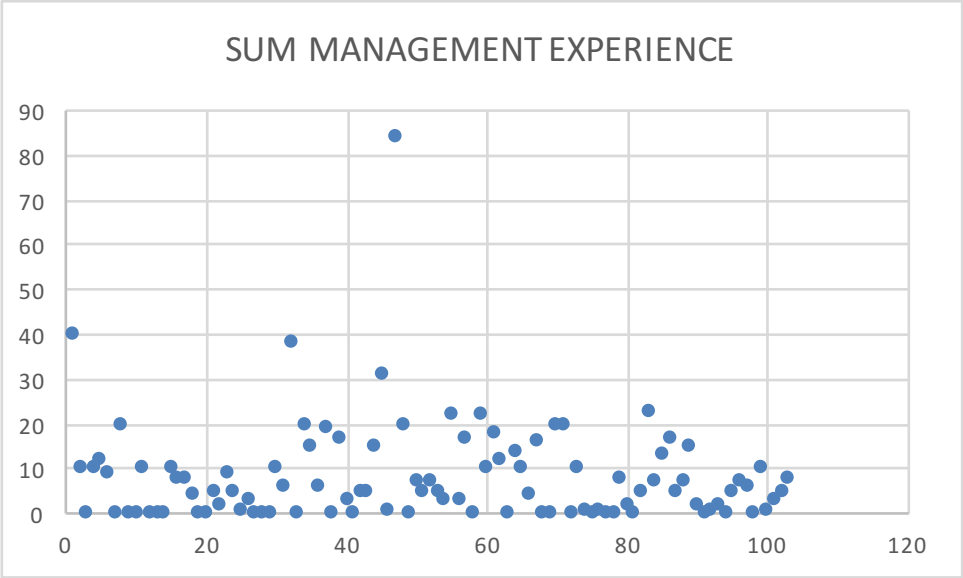

Figure 21. Sum of management experience scatter

H4: Previous work in the same industry has a positive effect on organisational performance

Previous experience in the same industry was found to not have a strong correlation to both turnover and number of staff employed. Amongst the 147 founders, from 103 startups, there was a total of 1163 years of industry experience. On average, a startup had 11,3 years of industry experience. The data, however, was found to have extreme values as demonstrated in the graph below. Looking at the data, 22 startups displayed no previous industry experience, 6 startups displayed over 40 years' worth of industry experience and 28 startups had industry experience between 10 and 40 years.

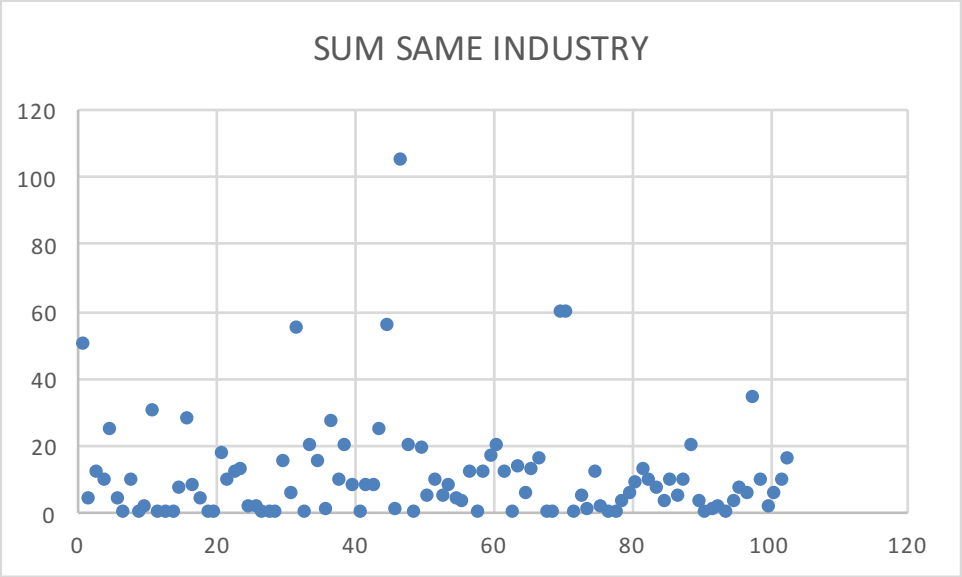

Figure 22. Sum of same industry experience scatter

H5: Businesses with a male founder display higher levels of organisational performance over female entrepreneurs.

Gender was shown to have no correlation to organisational performance - both turnover and number of staff. Out of the 147 founders, 49 were female and 98 were male. There is still a much larger number of male founders than female founders.

H6: Businesses founded by a minority group individual display a higher level of organisational performance than majority group entrepreneurs.

There was no strong correlation between race and organisational performance. It is a very positive indicator given the history of South Africa. It is a good indication of multiracial empowerment, however, it must be said that the majority of surveys were obtained from Shanduka Black Umbrellas, whose sole purpose as an incubator is the empowerment of previously disadvantaged people. Therefore, there may be sampling bias present in the data. Further research may need to be conducted in order to determine if race still plays a vital role in startups.

H7: Businesses with founders whose parents have their own business display a higher level of organisational performance than those who do not.

Whether the founders parents owned a business or not does not play a positive role in organisational performance. It was found that only 22 out of the 147 founders had parents who owned businesses. These founders were spread over 18 startups.

H8: Businesses founded by entrepreneurial teams display a higher level of organisational performance than businesses with a single founder.

Number of founders was found to have a very weak correlation to turnover, however, it was found to have a very strong correlation to number of staff employed. With regards to number of staff employed, this was found to be the variable with the strongest correlation to number of staff employed. There were 73 businesses with a single founder and 20 with two founders. This could possibly be related to the increase in network when having multiple founders and/or access to capital. When examining the number of founders in relation to access to capital the following may be noted. On a Likert scale from 1 to 7 (rating the capital from Inadequate to Adequate):

- Out of the 73 single founder startups, only six have rated the capital in the business as being six and above on the Likert scale

- Out of 20 startups founded by two members, four startups indicate their capital as being six and above on the Likert scale

- Out of 10 startups with more than 2 founders, only one startups rated their capital as being six and above

It is not possible to assume anything from the results, however, further research could be done into understanding the effects of multiple founders on a startup.

H9: An increase in capital leads to an increase in organisational performance.

An increase in capital was found to have a strong correlation to turnover, however, did not have a strong correlation to number of staff employed. In most circumstances, one could expect that a high turnover would result in a high number of staff, however, there are a number of factors which need to be examined. These would revolve around the nature of the business and model of the business itself and whether it lends itself to employment opportunities - perhaps the businesses do not require many employees. The table below demonstrates the relationship between capital and the rate of capital in the business (Likert scale from 1 to 7):

Table 75. Capital vs Turnover

Count of What is your turnover?	1	2	3	4	5	6	7	Grand Total
Less than R1000 pm	11	6	2	7		1	1	28
Between R1000 and R5000 pm	7	3	4	3				17
Between R5000 and R10 000 pm	1	1	1	3		1		7
Between R10 000 and R20 000 pm		3	1	1	1		1	7
Between R30 000 and R50 000 pm	2	1	3	4				10
Between R20 000 and R30 000 pm	1		2	2	1			6
Between R50 000 and R100 000 pm	5		1	4		2	3	15
Between R100 000 and R150 000 pm				1				1
Between R150 000 and R200 000 pm			1	1				2
Between R200 000 and R250 000 pm		2						2
Between R250 000 and R500 000 pm		1	2			2		5

Count of What is your turnover?	1	2	3	4	5	6	7	Grand Total
Between R500 000 and R1 000 000 pm	3							3
Grand Total	**30**	**17**	**17**	**26**	**2**	**6**	**5**	**103**

In many respects the above measures perceptions, and one can notice that there are very few people who rate the capital in their business as being adequate. Only 11 out of the 103 startups rate their capital as being 5 and above in terms of adequacy (on a scale of 1 to 7).

Focusing again on the findings, this research shows that while increasing the capital to entrepreneurs does positively affect the turnover in a business, it does not necessarily lead to more employment. This could prove to be an important finding in terms of policy and regulations.

H10: Business owners with strong marketing skills show an increase in organisational performance.

Marketing does not show a strong correlation to both turnover and number of staff employed. The average marketing skills per startup was 3,84 (on the Likert scale from 1 to 7). 11 out of the 103 startups rated their marketing skills at the highest, 7. Also, from the 103 startups, 8 gave themselves the lowest rating of 1. This finding shows that incubators need not place emphasis on marketing skills as other qualities would play a more vital role in organisational performance.

H11: Businesses with good customer relations have an improved level of organisational performance.

Customer relations too, proved to not have a strong correlation to organisational performance - both turnover and number of staff employed. The average customer relationship was extremely high: 6,58

when taking the average. However, there was a large number of missing data regarding this variable as it was only introduced into the survey later. The actual average of inputed data from surveys was even higher at 6,65. Again, it is important to emphasise that this is based solely on perception and that there is no actual manner of measuring how strong the customer relations of the business are.

H12: Businesses that receive a higher level of professional advice have an improved level of organisational performance.

Professional advice was found to have a positive correlation with regards to turnover. It did not have a strong correlation with regards to the number of staff employed, however, there was a moderate correlation to (p value was 0,077). Professional advice was measured on a Likert scale from 1 to 7. The average across the 103 respondents was 4,53. This is a positive finding for incubators as it does show that involving professional to guide and mentor incubatees does have a positive impact on the startup. This could be considered when incubators look at the services which they offer.

H13: Businesses that demonstrate high level of financial control have an improved level of organisational performance.

Financial control was found to not have strong correlation to organisational performance - both turnover and number of staff employed. The average level of financial control, as measured on a Likert scale from 1 to 7, was found to be 4.99. This is fairly high. It is worthwhile reiterating that the survey is largely based on perception and that “level of financial control” is a somewhat abstract concept. There is no instrument defined which could accurately standardise a measure of financial control. A variable like this may require further research and work in order to understand the impact it has on organisational performance.

H14: Businesses that have less difficulty obtaining staff have an improved level of organisational performance.

Difficulty obtaining staff was found to not to have a strong correlation with regards to organisational performance - both turnover and number of staff employed. One may have the expectation that there should be a strong correlation between the difficulty obtaining staff and the number of staff employed. However, this may not be the case as the question was not raised as to whether the startup is seeking to employ more people. Furthermore, the startup may have had difficulty obtaining new staff but have overcome that difficulty. This too, is an area which could require more research in order to fully understand the impact it has on organisational performance. Qualitative research could play a vital role in truly understanding how this factor affects startups.

H15: An increase in the average age of founders of the startup has a positive effect on organisational performance.

Age was found to have a strong correlation to turnover, however, it did not have a strong correlation to the number of staff employed. For startups with multiple founders, the average age was taken. Across the 103 startups, the average age was found to be 39,36. There were 18 startups with an average age of over 50 and only 5 startups with an average age of 25 or less. This may have implications for incubators, for example, teaming up younger teams with older members to create a more well-rounded team. As mentioned when analysing the literature, younger people may have easier access to the job market - which means they are likely to leave entrepreneurship for a more lucrative career if they are only marginally successful. Incubators could manage the expectations of younger entrepreneurs and prevent this from happening.

5.5 Comparison to other findings

As mentioned previously, this study is unique in many ways. This study aimed to create a computational model in order to predict the

organisational performance of startups within incubators. The model itself did not demonstrate a high level of predictability when compared to other models such as Lussier (1995). Lussier demonstrated a 75% level of predictability. However, there were key difference between the Lussier study and this research.

- Lussier attempted only to predict business or success - a boolean value. The aim of this research was not simply to predict failure or success but to predict whether a business would thrive. This was done by measuring organisation performance which was defined as a composite between turnover and number of staff employed.

- Lussier used statistical regression while this research attempted to create a computation model.

- This research was performed in South Africa where studies of this nature are limited.

Other studies have posted similar results to that off Lussier (1995). Siow Song Teng et al. (2011) created an exploratory model based on the Lussier prediction model. The Lussier model accurately predicted the failure or success off 85.6 percent of the surveyed firms while the exploratory by Siow Song Teng et al. (2011) predicted 86.3 percent.

Results on the prediction of success or failure of businesses has vastly differed. Lussier and Corman (1995) found 4 variables to be significant. Successful business used professional advisors more often and had more parents whom owned business. Failed business owners had a higher level of education and seemed to have less difficulty obtaining staff. The only overlap in this study was that the use of professional advisors was also found to be significant. Lussier and Corman (1996) performed a study in which they created a success vs failure mode for business with 0 to 10 employees. The results demonstrated a prediction rate of 75%.

Marom and Lussier (2014) created a model which predicted the success or failure of businesses in Israel. This resulted in an overall prediction accuracy of 85,4%. The model had an 86% accuracy for business failures and 84% for business success. The results are varied across many counties. Lussier and Pfeifer (2001) had a 72% accuracy in their Croatian study, while Lussier and Halabi (2010) only had a 63% accuracy in their Chilean study.

While these studies demonstrated very positive findings overall, they are not attempting to measure organisation performance. This makes it difficult to make a comparison between this study and the above mentioned success vs failure prediction models. The aim of this study is to highlight the importance of not merely understanding whether a company will succeed or fail, but to what extent would the succeed. This gives far richer information to stakeholders such as investors, incubators and policy makers.

5.6 Summary

This chapter began by analysing some of the demographic and attempting to understand the impact it had on the research. Thereafter, a comparison was made between the traditional statistical regression and the machine learning algorithms.

While the predictability of this model was lower than desired, it was demonstrated that using a machine learning algorithm, the predicability of the model was increased. The model was attempting to measure organisational performance, which was defined a composite of both turnover and number of staff employed. The machine learning algorithm used(the random forest algorithm) resulted in an increase in classification of 12,12% for turnover and 15,4% for number of staff employed. These are both significant increases.

A statistical analysis is performed on each variable. Lastly, a comparison is done between other models predicting success or failure vs this model which measured organisational performance..

CHAPTER 6

6.1 Introduction

This aim of this study was to create a computational model which would be able to predict organisational performance of startups within incubators. The literature review began by attempting to define which factors could lead to increased organisational performance. This was done by looking at past literature and models. Some of the models examined were Lussier (1995), Mwangi (2009), Cooper (1990) and Siow Song Teng et al. (2011). These finding have had various results. Besides looking at other models, various source of literature were used to create a 15-variable model. This essentially created 15 hypotheses.The aim was to measure organisational performance which was defined as a composite between turnover and number of staff employed.

These 15 hypotheses were used to create a survey. The survey was distributed via email to various incubators. In total, 103 responses were received. The 103 response were analysed and coded. The coding was given in Chapter 3. The results then went through stringent statistical analysis in order to determine the normality, reliability and demographics of the data. The statistical analysis demonstrated a 23,8% classification percentage for turnover and a 25,4% classification rate for number of staff employed. These findings were lower than desired, however, this work is substantially different to previous work done.

Weka was then used as the instrument of choice in order to determine whether a computational model could result in an improved classification rate. Weka was used to run 10 difference machine learning algorithms. The results indicated that the highest predictability was generated by using the random forest algorithm. The random forest algorithm resulted in a classification rate of 35,92% for turnover and 40,78% for number of staff employed.

Lastly, the implications of the results were discussed. The demographics were analysed, followed by a comparison between the statistical regression and machine learning algorithms. Lastly, each hypothesis was discussed in detail.

The findings were that although the model did not have the predictability as high as desired, this result could be improved substantially using machine learning algorithms - in this case the random forest algorithm. There are many areas of future work and further areas which can be explored.

6.2 Conclusions of study

This study has attempted to create a computational model in order to predict organisational performance of startups within incubators. The results of the model, however have been much lower than desired. Using the random forest algorithm has resulted in a significantly improved model in terms of predictability.

This study is unique in what it has attempted to measure and cannot truly be compared to previous studies which only test business success or failure. While knowing whether a business succeeds or fails is extremely important, the level of success of a business is really what would drive further investments into that business. It would be highly useful information to entrepreneurs, incubators, investors and suppliers. If the model proved to be highly accurate, it would allow investors and incubators to add a new dimension of evaluation in terms of risk. The information could also be used by incubators in order to determine which services need to be acquired by the startup prior to entering incubation, and what should be offered once in incubation

The study has made a few important findings when examining the hypotheses. Only four factors were found to be of any statistical significance in this study. These were capital, professional advice, education level, age and number of partners.

Looking at the machine learning algorithms, the best performance was obtained using the random forest classifier. It lead to a vast improvement of predictability for both turnover and number of staff employed.

One could conclude the findings of this study by stating that the study is unique and the results are not comparable to the business success or failure results. The aim of this study is to measure organisational performance and not business success or failure. This study also has a different population - startups within incubators. Previous studies have focused on any business. This study has also attempted to create a computation model rather than focus on the traditional statistical regression. Lastly, very few studies focus on South Africa or any country on the African continent. As stated earlier, the results of the model are much lower than desired. Despite that, there are numerous useful finding in this work as well as many areas for future research.

6.3 Implications and recommendations

The benefits of this study would be to:

- To improve the state of incubators
- To alleviate the issue of unemployment in South Africa
- To improve the economic margins of startups within incubators
- To add to the entrepreneurship literature and knowledge, specifically relating to startups

This study has many implications for the stakeholders involved. These include startup founders, incubator managers, investors and policy makers as well as others involved in researching the performance of startups. Some of the implications of this study are:

- Machine learning algorithm mays be used as opposed to traditional statistical regression in order to obtain more accurate classifiers

- Incubator managers need to focus on having professional advice easily and inexpensively available to startups

- Education of entrepreneurs within incubators is vitally important to solving the unemployment issue and is something that should also be encouraged within incubators or serve as a prerequisite prior to entering incubation

- Startups should be encourage to have multiple founders or perhaps incubators could look at having joint ventures between startups.

- If a startup founder is young it may be beneficial to pair them up with an older, experienced entrepreneur

- For policymakers in organisation such as the Department of trade and industry and other private venture capitalist and funders, the aim of the policy needs to be well understood. If the aim is to alleviate unemployment, then there are certain policies which may be put into place, however, if the issue is startups are not generating enough turnover, then perhaps a different policy may need to be applied. Although the two have been examined as being a composite of organisational performance, the study reveals them to be vastly different.

6.4 Limitations of study and further research avenues

This study forms a solid academic and practical ground for future research. There is limited research regarding the organisational performance of startups, and research pertaining to the African continent is even more limited. Retrospectively, there are a number of limitations to this research. There is also much room for future research and improvements in terms of model accuracy.

One of the major shortcomings of the manner in which this research was done, was that the research only employed quantitative research method. A mixed method research, consisting of both qualitative and quantitative research would definitely result in a much richer understanding. It allows for the scrutiny of the quality of survey responses. The manner in which responses were gained here was by emailing the startup founders directly. The email, in some cases, were sent as many as 8 times. Sending out the email too many times, increases the risk of irritating the survey respondent which could in turn affect the quality of the responses given. Furthermore, an increased survey sample would have increased the number of results and eliminated any potential survey bias. This should be considered as an avenue for future research.

Another shortcoming of this survey was that it was based on perceptions. Using qualitative measures could allow for the creation of instruments as well as more accurate measurements of the variables. Examples of these may be defining financial control or actually giving the number of hours or budget used for professional advice. These indicators could be more accurate and lead to more solid findings. This too, should be considered as an avenue for future research.

One of the emergent predecessors to this study was determining that business success or failure, although valuable, may not be considered a good indicator to investors or incubator managers. The understanding of how much a business thrives could be far more useful. Future research could explore this avenue further as opposed to simply measuring success or failure.

This study has attempted to explore a model which is able to determine the performance of startups within incubators. Entrepreneurship is the cornerstone of many modern economies. Entrepreneurship can change the way society lives and works. Successful entrepreneurship can create innovation, improve general standards of living, create employment and ultimately create a prosperous society. Being able to understand the underlying factors which create an entrepreneurial economy and create successful startups are of imperative importance. This study hopes to add to this body of knowledge which, hopefully, if further explored will result a positive impact for the South African startup environment.

Bibliography

Aher, S. B., & Lobo, L. M. R. J. (2011). Data mining in educational system using weka. In International Conference on Emerging Technology Trends (ICETT) (Vol. 3, pp. 20-25).

Aktan, S. (2011). Application of machine learning algorithms for business failure prediction. Innovations, 8, 2.

Alpaydin, E., 2014. Introduction to machine learning. MIT press.

Baptista, R., Karaöz, M. and Mendonça, J. (2007), Entrepreneurial backgrounds, human capital and start-up success. Jena economic research papers, 2007(045)

Baptista, R., Karaöz, M., & Mendonça, J. (2014). The impact of human capital on the early success of necessity versus opportunity-based entrepreneurs. Small Business Economics, 42(4), 831-847.

Barrett, T., 2016. Business model innovation in South African startups (Doctoral dissertation, University of Pretoria).

Becker, B. G. (1998, October). Visualizing decision table classifiers. In Information Visualization, 1998. Proceedings. IEEE Symposium on (pp. 102-105). IEEE.

Benedetto, C. A. (1999). Identifying the key success factors in new product launch. Journal of product innovation management, 16(6), 530-544.

Benzing, C., Chu, H. M., & Kara, O. (2009). Entrepreneurs in Turkey: A factor analysis of motivations, success factors, and problems. Journal of small business management, 47(1), 58-91.

Boag, D. A. (1987). Marketing control and performance in early-growth companies. Journal of Business Venturing, 2(4), 365-379.

Bose, I. and Mahapatra, R.K. (2001). Business data mining—a machine learning perspective. Information & management, 39(3), pp.211-225.

Bosma, N., Van Praag, M., Thurik, R., & De Wit, G. (2004). The value of human and social capital investments for the business performance of startups. Small Business Economics, 23(3), 227-236.

Brants, T., Popat, A.C., Xu, P., Och, F.J. and Dean, J. (2007). Large language models in machine translation. In In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

Birasnav, M. (2014). Knowledge management and organizational performance in the service industry: The role of transformational leadership beyond the effects of transactional leadership. Journal of Business Research, 67(8), pp.1622-1629.

Carlos Pinho, J., Paula Rodrigues, A. and Dibb, S. (2014). The role of corporate culture, market orientation and organisational commitment in organisational performance: the case of non-profit organisations. Journal of Management Development, 33(4), pp.374-398.

Cassar, G. (2014). Industry and startup experience on entrepreneur forecast performance in new firms. Journal of Business Venturing, 29(1), 137-151.

Colomboa, M. G., Delmastrob, M. and Grillia, L. (2004), Entrepreneurs' human capital and the start-up size of new technology-based firms. International Journal of Industrial Organization, 22(8-9): pp. 1183–1211. http://dx.doi.org/10.1016/j.ijindorg.2004.06.006

Combs J.G., Crook T.R. and Shook C.L. (2005), The Dimensionality of Organisational Performance and its Implications for Strategic Management Research. Research Methodology in Strategy and Management. 2. pp.259 - 286

Cooper, R. G. (1979). The dimensions of industrial new product success and failure. The Journal of Marketing, 93-103.

Cooper, A. C., Gimeno-Gascon, F. J., & Woo, C. Y. (1991, August). A RESOURCE-BASED PREDICTION OF NEW VENTURE SURVIVAL AND GROWTH. In Academy of Management Proceedings (Vol. 1991, No. 1, pp. 68-72). Academy of Management.

Cooper, A.C., Gimeno-Gascon, F.J. and Woo, C.Y. (1994). Initial human and financial capital as predictors of new venture performance. Journal of business venturing, 9(5), pp.371-395.

Cruz, J.A. and Wishart, D.S. (2006). Applications of machine learning in cancer prediction and prognosis. Cancer informatics, 2.

Crotty, M. (1998). Foundations of social research: Meaning and Perspective in the Research Process. p.256.

Davidsson, P. and Honig, B. (2003) The role of social and human capital among nascent entrepreneurs. Journal of Business Venturing 18(3):pp. 301-331.

Davila, A., Foster, G., & Gupta, M. (2003). Venture capital financing and the growth of startup firms. Journal of business venturing, 18(6), 689-708.

Delen, D., Zaim, H., Kuzey, C. and Zaim, S. (2013). A comparative analysis of machine learning systems for measuring the impact of knowledge management practices. Decision Support Systems, 54(2), pp. 1150-1160.

Ejere, E.I. and Abasilim, U.D. (2013). Impact of transactional and transformational leadership styles on organisational performance: empirical evidence from Nigeria. The Journal of Commerce, 5(1), pp. 30-41.

Geng, R., Bose, I. and Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. European Journal of Operational Research, 241(1), pp.236-247.

Ghosh, B. C., Liang, T. W., Meng, T. T., & Chan, B. (2001). The key success factors, distinctive capabilities, and strategic thrusts of top SMEs in Singapore. Journal of Business Research, 51(3), 209-221.

Gimeno, J., Folta, T. B., Cooper, A. C., & Woo, C. Y. (1997). Survival of the fittest? Entrepreneurial human capital and the persistence of underperforming firms. Administrative science quarterly, 750-783.

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. The qualitative report, 8(4), 597-606.

Harrington, M. & Kew, P.(2016) 2015/16 GEM South Africa Report

Hyder, S., & Lussier, R. N. (2016). Why businesses succeed or fail: a study on small businesses in Pakistan. Journal of Entrepreneurship in Emerging Economies, 8(1), 82-100.

Iba, W., & Langley, P. (1992). Induction of one-level decision trees. In Machine Learning Proceedings 1992 (pp. 233-240).

Kelly, T. and Firestone, R. (2016). How Tech Hubs are helping to Drive Economic Growth in Africa. Background Paper for the World Development Report 2016: Digital Dividends.[online] Available at: http://documents.worldbank.org/curated/en/626981468195850883/pdf/102957-WP-Box394845B-PUBLIC-WDR16-BP-How-Tech-Hubs-are-helping-to-Drive-Economic-Growth-in-Africa-Kelly-Firestone.pdf

Kwong, E. and Lee, W.B. (2009). Knowledge elicitation in reliability management in the airline industry. Journal of Knowledge Management, 13(2), pp.35-48.

Chell, E., 2013. Review of skill and the entrepreneurial process. International Journal of Entrepreneurial Behavior & Research, 19(1), pp. 6-31.

Lee, I. H., & Marvel, M. R. (2014). Revisiting the entrepreneur gender–performance relationship: a firm perspective. Small Business Economics, 42(4), 769-786.

Li, H., Sun, J., & Wu, J. (2010). Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods. Expert Systems with Applications, 37(8), 5895-5904.

Lose, T. (2016). The role of business incubators in facilitating the entrepreneurial skills requirements of small and medium size enterprises in the Cape metropolitan area, South Africa. MTech Thesis, Cape Town. Cape Peninisula University of Technology).

Lose, T., and Tengeh, R. K. (2016). An evaluation of the effectiveness of business incubation
programs: a user satisfaction approach.

Loukeris, N., & Matsatsinis, N. (2006). Corporate financial evaluation and bankruptcy prediction implementing artificial intelligence methods. WSEAS Transactions on Business and Economics, 3(4), 343.

Lussier, R.N. (1995). A nonfinancial business success versus failure prediction mo. Journal of Small Business Management, 33(1), p.8.

Lussier, R. N., & Corman, J. (1996). A business success versus failure prediction model for entrepreneurs with 0-10 employees. Journal of Small Business Strategy, 7(1), 21-36.

Lussier, R. N., & Corman, J. (1995). There are few differences between successful and failed small businesses. Journal of Small Business Strategy, 6(1), 21-34.

Mair, C., Kadoda, G., Lefley, M., Phalp, K., Schofield, C., Shepperd, M. and Webster, S. (2000). An investigation of machine learning based prediction systems. Journal of Systems and Software, 53(1), pp.23-29.

Marom, S., & Lussier, R. N. (2014). A business success versus failure prediction model for small businesses in Israel. Business and Economic Research, 4(2), 63.

Marvel, M. R. and Lumpkin, G.T. (2007), Technology Entrepreneurs' Human Capital and Its Effects on Innovation Radicalness. Entrepreneurship Theory and Practice, 31: 807–828. doi:10.1111/j. 1540-6520.2007.00209.x

Mothibi, G. (2014). The Influence of Business Incubation Services on the Performance of Small Medium Enterprises in the South African Tourism Industry. In European Conference on Management, Leadership & Governance (p. 569). Academic Conferences International Limited.

Mtech website, https://katie.mtech.edu/classes/csci347/Resources/Weka_error_measurements.pdf

Mwangi, R.M., Sejjaaka, S., Owino, E.O., Canney, S., Maina, R., Kairo, D., Rotich, A., Nsereko, I. and Mindra, R., (2013). Constructs of successful and sustainable SME leadership in east Africa.

National Planning Commission. 2011. Diagnostic Overview of National Development Plan 2030. [Online]. Available at: www.info.gov.za/views [Accessed 21 October. 2017].

National Treasury, 2016. Socio-economic review and outlook (2016). [Online]. Available at: http://www.treasury.gpg.gov.za/Documents/SERO%202016.pdf [Accessed 23 October. 2017].

Olden, J.D., Lawler, J.J. and Poff, N.L. (2008). Machine learning methods without tears: a primer for ecologists. The Quarterly review of biology, 83(2), pp.171-193.

Patil, T. R., & Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International Journal of Computer Science and Applications, 6(2), 256-261.

Read, S., Song, M. and Smit, W. (2009). A meta-analytic review of effectuation and venture performance. Journal of Business Venturing, 24(6), pp.573-587.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 67, 93-104.

Rosa, P., Carter, S., & Hamilton, D. (1996). Gender as a determinant of small business performance: Insights from a British study. Small business economics, 8(6), 463-478.

Salem, M. I. (2014). The role of business incubators in the economic development of Saudi Arabia. The International Business & Economics Research Journal (Online), 13(4), 853.

Schultz, T. W. (1961), Investment in Human Capital. The American Economic Review, 51(1), pp 1-17

Shrader, R. and Siegel, D. S. (2007), Assessing the Relationship between Human Capital and Firm Performance: Evidence from Technology-Based

New Ventures. Entrepreneurship Theory and Practice, 31: 893–908. doi: 10.1111/j.1540-6520.2007.00206.x

Siow Song Teng, H., Singh Bhatia, G., & Anwar, S. (2011). A success versus failure prediction model for small businesses in Singapore. American Journal of Business, 26(1), 50-64.

Statistics South Africa (2016) “Quarterly Labour Force Survey Quarter 4: 2016.” Statistical release P0211. [online] Available at: http://www.statssa.gov.za/publications/P0211/P02114thQuarter2016.pdf . [Accessed 07 July. 2017].

Statistics South Africa Midyear (2016) Midyear Population Estimates, 2016. [online] Available at: https://www.statssa.gov.za/publications/P0302/P03022016.pdf. [Accessed 01 July. 2017].

Tharwat, A., Gaber, T., Awad, Y. M., Dey, N., & Hassanien, A. E. (2016). Plants identification using feature fusion technique and bagging classifier. In The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Beni Suef, Egypt (pp. 461-471). Springer, Cham.

Unger, J., Rauch, A.,Frese, M. and Rosenbusch, N. (2011) Human capital and entrepreneurial success : a meta-analytical review. Journal of Business Venturing, 26 (3). pp. 341-358. http://dx.doi.org/10.1016/j.jbusvent.2009.09.004

Witten, I. H., Frank, E., Trigg, L. E., Hall, M. A., Holmes, G., & Cunningham, S. J. (1999). Weka: Practical machine learning tools and techniques with Java implementations.

Wright, M., Hmieleski, K. M., Siegel, D. S., & Ensley, M. D. (2007). The role of human capital in technological entrepreneurship. Entrepreneurship Theory and Practice, 31(6), 791-806.

Yu, Q., Miche, Y., Séverin, E. and Lendasse, A. (2014). Bankruptcy prediction using extreme learning machine and financial expertise. Neurocomputing, 128, pp.296-302

Zhao, H., Sinha, A. P., & Ge, W. (2009). Effects of feature construction on classification performance: An empirical study in bank failure prediction. Expert Systems with Applications, 36(2), 2633-2644.

Zhao, Y., & Zhang, Y. (2008). Comparison of decision tree methods for finding active objects. Advances in Space Research, 41(12), 1955-1959.

Appendices

Appendix 1 - Draft Research Instrument

1) How many founders are there in the organisation?

   _____

2) Please give the following information relating to the founders of the organisation:

	Founder 1	Founder 2	Founder 3	Founder 4
Name				
Surname				
Age				
Race				
Gender				
Highest Qualification				
Level of technical expertise (1 to 5)				
Business experience in same industry				
Management experience				
Years of technical experience				
Parents owned business				
Marital Status				
Born in				
Private/Public school				

3) How much capital has been placed into the business?

- None at all

- Less than R10 000
- Between R10 000 and R100 000
- Between R100 000 and R250 000
- Between R250 000 and R500 000
- Between R500 000 and R1 000 000
- Between R1 000 000 and R2 500 000
- Between R2 500 000 and R5 000 000
- Between R1 000 000 and R2 500 000
- Between R2 500 000 and R5 000 000
- Between R5 000 000 and R7 500 000
- Between R7 500 000 and R10 000 000
- Over R10 000 000

3) How long have you been in incubation for?

- Less than 6 months
- 6 months to one year
- One year to two years
- Two to three years
- Three to five years
- Five to ten years
- More than five ten years

4) What is your turnover?

- Less than R1000 pm
- Between R1000 and R5000 pm
- Between R5000 and R10 000 pm
- Between R10 000 and R20 000 pm
- Between R20 000 and R30 000 pm
- Between R30 000 and R50 000 pm
- Between R50 000 and R100 000 pm
- Between R100 000 and R150 000 pm
- Between R150 000 and R200 000 pm
- Between R200 000 and R250 000 pm
- Between R250 000 and R500 000 pm
- Between R500 000 and R1 000 000 pm

5) What is your current staff compliment? _____

6) What was the size of your staff at the point of incubation?

_____

7) How often do you professional advice?

Never used										Used often

8) How would you rate your record keeping and financial control?

Very Poor	Very Professional

9) How difficult did you find it to obtain staff?

Easy	Difficult

10) How would you rate your marketing skills?

Little Marketing	Great use of marketing

11) Which industry are you involved in?

Little Marketing	Great use of marketing

12) How would you rate the importance of customer relations to your business?

Not important	Highly important

13) How would you rate the importance of customer relations to your business?

- Agriculture, Forestry and Fishing
- Mining
- Construction
- Manufacturing
- Transportation, Communications, Electric, Gas and Sanitary service
- Wholesale Trade
- Retail Trade
- Finance, Insurance and Real Estate
- Services or consulting
- Information Technology
- Public Administration
- Not classifiable

Appendix 2 - Draft Cover Letter

A computational model to predict the organisational performance of startups in South African incubators

Dear Sir/Madam,

My name is Jarryd Jermaine Chengalroyen. I am currently studying a Masters in Management at WBS (Wits Business School). The title of my research topic is ‘A computational model to predict the organisational performance of startups in South African incubators’. The objective of this is to evaluate which predictors are likely to lead to increased organisational performance. The research may result in a number of benefits. Technology startups can benefit from this study. It allows investors to add a new dimension of evaluation in terms of risks based on past organisational performance. It also allows them to suggest changes to be made in the organisation before investment/further investment is considered. It allows entrepreneurs to understand what changes could be made to an organisation to increase performance. Suppliers can use it to evaluate risk and limit credit facilities to these clients. Incubators and accelerators could use the model to assist and improve businesses.

As a startup founder entrepreneur, you are invited to participate in this research by completing a questionnaire. The questionnaire consists of 25 questions and will take about 10 to 15 minutes to complete.

What will happen if you choose to participate in the research?

1. There should be no harm or risk if you participate.

2. You are requested to sign a consent form indicating that you are voluntarily agreeing to participate in the research.

3. You may stop participating in this research at any stage.

4. Your information will remain as confidential as to the extent the law allows.

5. The data will be kept for a maximum of 5 years. Published data will still keep you anonymous.

What will happen if you choose not to participate in the research?

1. You may choose to not participate without any harm or prejudice.

The research study was approved unconditionally by the Wits Business School research panel. Should you have queries related to the research, please feel free to contact my supervisor: Dr Diran Soumonni on 011 717 3646 or Email: Diran.Soumonni@wits.ac.za. You may directly request copies of the results of the research to me on 083 235 3893 or 0408374p@students.wits.ac.za.

Jarryd Jermaine Chengalroyen

Appendix 3 - Draft Consent Form

Study name

A computational model to predict the organisational performance of startups in South African incubators

Researchers

Name: Jarryd Jermaine Chengalroyen
Institution: Wits Business School
Course: Masters Management in Entrepreneurship and New Venture Capital
Contact Number: +27 83 235 3893
email:0408374p@students.wits.ac.za

Purpose of the research

The purpose of this research is to conduct a quantitative study to understand which factors or attributes lead to an increase in organisational performance for technical startups within the Gauteng Province. The research will require to fill out a questionnaire. It should require about 15 to 20 minutes of your time.

Risks and discomforts

There will be no additional risk or discomfort anticipated from participating in this research.

Benefits of the research and benefit to you

There are no immediate benefits to you for participating in this study. The study will be extremely useful in finding the factors which lead to improved organisational performance and if you like I would forward my finding to you.

Your participation

Your participation in this research is completely voluntary. You are by no means being forced to participate in this study. Whether you choose to

participate or not, you will not be affected in any way whatsoever. If you initially agree to participate, you may at any stage, stop participating in the research. If this is the case, all associated data collected will be destroyed. Participation in this research will not prejudice you any way. Your relationship with the interviewer or Wits Business School will not be affected in any manner.

Confidentiality

Any records in which you are identified you will be kept confidential to the extent possible by law. The records from your participation may be reviewed by people responsible for making sure that research is done properly, including my academic supervisor/s. (All of these people are required to keep your identity confidential.)

All records will be destroyed after the completion and marking of my research assignment. I will refer to you by a code number or pseudonym in the research.

Questions about the research

If you have any questions or concerns regarding this research or the manner in which it is being conducted please feel free to contact either:

1) the Research Administrator at the Wits Business School, Kedibone Tyeda. Email: tyeda@wits.ac.za

2) my academic research supervisor Dr Diran Soumonni on 011 717 3646 or Email: Diran.Soumonni@wits.ac.za.

CONSENT

I hereby agree to participate in research on '*A computational model to predict the organisational performance of startups in South African incubators'.* I understand that I am that I am participating at my own will and not being forced to participate in any manner.

I understand I may cease to participate at any moment in time without repercussion.

I understand that I will not directly benefit from this research in the short term.

I understand that my participation and data will remain confidential as to the extent of the law.

Signature of participant _____ **Date**: _____

Appendix 4 - Consistency Matrix

Table 77. Consistency Matrix

Aim of research	Literature review	Hypotheses or propositions or research questions	Source of data	Type of data	Analysis
Create a computation model to predict organisational performance within tech startups within South Africa	Colomboa, M. G., Delmastrob, M. and Grillia, L. (2004), Combs J.G., Crook T.R. and Shook C.L. (2005), Cooper, A.C., Gimeno-Gascon, F.J. and Woo, C.Y., 1994, Davidsson, P. and Honig, B. (2003)	* Hypotheses listed below	Survey data gathered using research instrument	Cross sectional	Machine learning output with factor analysis
Evaluate whether certain incubators generally outperform others	Mothibi, G. (2014),Kelly, T. and Firestone, R. (2016)	Which factors of incubators generally allow one to perform better than others?	Survey data gathered using research instrument	Cross sectional	Correlation analysis

Table 78. Consistency Matrix of Hypotheses

H1	Managerial experience has a positive effect on organisational performance.	Baptista et al. (2007)
H2	Technical expertise are positively associated with organisational performance.	Shrader and Siegel (2007)
H3	Higher levels of education have a positive effect on organisational performance.	Marvel and Lumpkin (2007)
H4	A larger number business partners has a positive effect on organisational performance.	Colombo et al. (2004)
H5	Businesses with a male founder display higher levels of organisational performance over female entrepreneurs.	Cooper et al. (1994)
H6	Businesses founded by a minority group individual display a higher level of organisational performance than majority group entrepreneurs.	Cooper et al. (1994)
H7	Businesses with founders whose parents have their own business display a higher level of organisational performance than those who do not.	Cooper et al. (1994)
H8	Businesses founded by entrepreneurial teams display a higher level of organisational performance than businesses with a single founder.	Cooper et al. (1994)
H9	An increase in venture capital leads to an increase in organisational performance.	Davila, Foster, and Gupta (2003)
H10	Business owners with strong marketing skills show an increase in organisational performance.	Benedetto (1999) Boag (1987) Boag (1987)
H11	Businesses with good customer relations have an improved level of organisational performance.	Huck and McEwen (1991) Ghosh et al. (2001)
H12	Businesses that receive a higher level of professional advice have an improved level of organisational performance.	Reynolds (1987)

H13	Businesses that demonstrate high level of financial control have an improved level of organisational performance.	da Silva (2016) Lussier (1995)
H14	Businesses that have less difficulty obtaining staff have an improved level of organisational performance.	(Hyder and Lussier, 2016) Saha (2006)
H15	An increase in the average age of founders of the startup has a positive effect on organisational performance.	Gimeno et al. (1997) Bosma et al. (2004)