

https://dspace.org/

Using Supervised Machine Learning Algorithms to Predict Prostate Cancer Cases: A Case of Men of African Descent Carcinoma of the Prostate Consortium (MADCaP), South Africa

Mhone, Phale

2025
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/49755

Mhone, Phale . (2025). Using Supervised Machine Learning Algorithms to Predict Prostate Cancer Cases: A Case of Men of African Descent Carcinoma of the Prostate Consortium (MADCaP), South Africa [Master's dissertation, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/49755

Downloaded from a DSpace repository. https://dspace.org/

Using Supervised Machine Learning Algorithms to Predict Prostate Cancer Cases:

A Case of Men of African Descent Carcinoma of the Prostate Consortium (MADCaP), South Africa

By

Friday Phale Mhone(2603349)

UNIVERSITY OF THE WITWATERSRAND



Faculty of Health Sciences

School of Public Health

Supervisor: Dr Michael Mapundu

Co-Supervisor: Dr Wenlong Carl Chen

September 2025

Declaration of Authorship

I, Friday Mhone, declare that this research report titled, *"Using Supervised Machine Learning Algorithms to Predict Prostate Cancer Cases"*, and the work presented in it are my own, except where specific references are cited to the work of others. This work has not been submitted elsewhere for any other degree or award except for this current fulfillment of the Master of Science degree at The University of Witwatersrand. This research work reflects my independent work except in instances where explicit references have acknowledged other sources.

Signed:



Date: 2nd September 2025

Dedication

This thesis is dedicated to myself for the perseverance and resilience I have demonstrated throughout this journey. Being a self-sponsored student in a foreign country was an immense financial challenge, but I am proud of myself for overcoming it and making it to this point.

To my brother, Peter Duncan this one is for you. Your generosity, unwavering support, and willingness to host me made all the difference. Without your help, I might not have been able to complete this journey. Your kindness knows no bounds, and I am forever grateful.

To my mentors and supervisors, Dr. Michael Mapundu and Dr. Wenlong Carl Chen your support has been invaluable. Dr. Mapundu, your prompt responses to my emails and phone calls felt like a father-son mentorship, and I deeply appreciate your dedication. Dr. Chen, your wisdom and guidance have profoundly shaped this work. Your commitment to knowledge and research has inspired me to strive for excellence.

Finally, I dedicate this work to all individuals who have battled prostate cancer. Your experiences have fueled my passion for research in public health and epidemiology. May this work contribute, in some way, to the advancement of knowledge and better health outcomes for future generations.

Abstract

Prostate cancer is a major public health concern globally, particularly affecting men of African descent who experience higher incidence and mortality rates. This study aimed to develop predictive models for prostate cancer risk using machine learning algorithms applied to data from the Men of African Descent Carcinoma of the Prostate (MADCaP) consortium at Chris Hani Baragwanath Academic Hospital (CHBAH).

The study population included 1096 prostate cancer cases enrolled between January 2017 and December 2021. A retrospective cross-sectional analysis of secondary data was conducted. Using Python in Google Colaboratory, we implemented Naïve Bayes, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Support Vector Machine (SVM) models. Data preprocessing followed the Knowledge Discovery in Databases (KDD) methodology, including cleaning, feature selection, and addressing class imbalance with a hybrid oversampling technique. Model performance was evaluated using accuracy, precision, recall, and F1 score.

Results showed that KNN and Decision Tree models achieved the best predictive performance, with KNN providing high accuracy and Decision Tree offering interpretability. Random Forest performed well but

indicated minor overfitting. The study population analysis revealed significant differences in age across risk categories ($H = 2631.90$, $p < 0.001$) and a statistically significant change in PSA levels from referral to diagnosis ($t = 3.20$, $p = 0.001$). Other factors, such as smoking status, showed no significant association with risk classification.

In conclusion, machine learning models, particularly KNN, show promise for predicting prostate cancer risk in men of African descent. These findings support targeted screening and risk stratification in high-risk populations. Future studies should validate these models across diverse datasets and consider integrating additional genetic and environmental factors to improve predictive accuracy and address health disparities.

Contents

1 Introduction **1**

1.1 Background 2

1.2 Problem Statement 5

1.3 Justification 6

1.4 Aim 7

1.5 Objectives 7

2 Literature review **9**

2.1 Overview of the South African PC screening guidelines 10

2.2 Genetic risk and aggressive disease manifestation 12

2.3 Case identification 14

2.4 Conceptual framework 15

2.5 Big data and ML in PC research 16

2.6 ML model classification 17

2.7 Rule based Classifiers: a background 18

2.7.1 Why ML is better 19

2.8 ML classifiers 19

2.8.1 Unsupervised learning 20

2.8.2 Semi supervised learning 21

2.8.3 Supervised learning 21

2.9 Model evaluation . . . 22
2.10 Optimization . . . 23
2.10.1 Cross validation and grid search . . . 24
2.10.2 Model selection . . . 24
2.11 Summary . . . 24

3 Materials and methods **26**
3.1 Introduction . . . 26
3.2 Study setting and population . . . 26
3.3 Methods . . . 27
3.3.1 Computational environment . . . 28
3.4 Overview of proposed solutions . . . 29
3.5 Data definition . . . 29
3.6 Feature selection and engineering . . . 30
3.6.1 Handling Class Imbalance and Data Augmentation . . . 32
3.7 Analysis . . . 34
3.7.1 Association and correlation . . . 34
3.7.2 ML models . . . 34
3.8 Performance evaluation of the algorithms . . . 42
3.8.1 Accuracy . . . 43
3.8.2 Sensitivity . . . 44
3.8.3 Specificity . . . 44
3.8.4 Area under the curve (AUC) of the receiver operating characteristic (ROC) curve . . . 45
3.8.5 Precision . . . 45
3.8.6 F1 score . . . 46
3.9 Ethical considerations . . . 47
3.10 Summary . . . 47

4 Results **48**

4.1 Descriptive statistics 48

4.1.1 PC analytical data 48

4.2 Risk factors and disparities 54

4.3 Predicting PC progressiveness and cancer severity 64

4.4 Model performance 67

4.4.1 Confusion matrices 67

4.4.2 ROC curves 71

4.4.3 Comparing Model Performance: Original, Resampled, and Augmented Data 75

4.4.4 Conclusion 77

5 Discussion **78**

5.1 Descriptive discussion 79

5.1.1 Ethnicity and PC risk 79

5.1.2 Smoking and PC risk 79

5.1.3 PSA Levels and risk classification 80

5.1.4 Education and PC risk 80

5.1.5 Age and PC risk 81

5.1.6 Cigarette consumption and PC severity 81

5.1.7 Implications of findings 82

5.2 Model evaluation summary 82

5.3 Interpretation of model-specific performance 83

5.3.1 GNB: simplicity at a cost 83

5.3.2 KNN: robustness and flexibility 83

5.3.3 DT: interpretability and consistency 84

5.3.4 RF: High accuracy with potential overfitting 84

5.3.5 SVM: Consistency with limited predictive power 85

5.3.6 GB: Balanced and competitive performance 85
5.4 Comparison and implications for model selection 86
5.4.1 Selection of optimal model: KNN and DT 86
5.5 Limitations of the Study 87
5.6 Summary 87

6 Conclusions and future directions **89**
6.1 missing value report 107
6.2 Correlation matrix 108
6.3 Plagiarism declaration 109
6.4 TurnItIn Report 110
6.5 HREC research clearance certificate 111
6.6 Research ethics training certificate 112
6.7 MADCaP ethics certificate 2022 113
6.8 Python Code for Data Analysis 114

List of Figures

2.1 Recommended age-related screening intervals for prostate-specific antigen screening (John et al., 2024) 10
2.2 Summary of genome-wide association studies (GWAS) in Africa and ancestral fractions.(Janivara, R., Chen, W.C., Hazra, U. et al., 2024) 13
2.3 Conceptual framework showing factors associated with PC case. . . 16

3.1 Typical supervised learning algorithm 27
3.2 KNN classifier by Rajvi Shah 37
3.3 DT classifier branch 38
3.4 RF classifier branch by Ankit Chauhan 39
3.5 Comparison of SVM classifiers with different hyperplanes by James Le. 42
3.6 Confusion matrix 43

4.1 Before processing 49
4.2 After processing 49
4.3 PC risk categories by smoking status 55
4.4 Comparison and change in PSA levels from referral to diagnosis . . 56
4.5 Heatmap of risk category by education codes 58
4.6 Average gleason total score by cigarette consumption (HRF cigarettes count) 59

4.7 Age distribution by PC risk category 61
4.8 Stacked bar graph showing the distribution of PC risk categories across income groups. 63
4.9 Confusion matrices for all models 70
4.10 Area under the receiver operating characteristic curve (AUROC) for all models 72

6.1 Correlation matrix 108

List of Tables

3.1 Computational environment 28
3.2 Variable description 32
3.3 Dataset sizes of initial data, after balancing, and after augmentation 34

4.1 Distribution of categorical variables across prostate cancer risk categories 50
4.2 Summary statistics for risk factors across PC risk categories 52
4.3 Test performance metrics for different ML models on PC prediction (in percentages) 66
4.4 Test performance metrics using original data. 75
4.5 Test performance metrics after resampling using ROSE. 76

List of Acronyms

DT	Decision Tree
GB	Gradient Boost Model
GNB	Gaussian Naive Bayes
KNN	K-Nearest Neighbors
MADCaP	Men of African Descent Carcinoma of the Prostate
ML	Machine Learning
PC	Prostate Cancer
RF	Random Forest
SVC	Support Vector Classifier

Chapter 1

Introduction

This chapter sets the stage for the research by introducing two important topics: Prostate Cancer (PC) and the role of supervised machine learning (ML) in healthcare. PC is a pressing health issue, particularly in regions where medical resources are stretched thin, such as Africa. On the other hand, ML, specifically supervised ML, is an emerging tool that has the potential to transform the way we diagnose and treat diseases. By bringing these two areas together, this research aims to explore how ML can help improve PC diagnosis and treatment.

While there have been significant advances in PC care in some parts of the world, the same cannot be said for every region. Disparities in healthcare access and outcomes still exist, especially in countries with fewer resources. In this context, the combination of medical expertise and advanced technology like ML becomes even more crucial [1]. This chapter outlines the gaps in current research, explains why filling these gaps is important, and introduces the research question, aim, and objectives of the study. The chapters that follow will explore the development and evaluation of a ML algorithm designed specifically for PC diagnosis and treatment.

1.1 Background

PC is one of the most common cancers in men around the world, and the situation is particularly concerning in Africa [1]. In this region, PC is the leading cancer diagnosis among men and the second most common cause of cancer-related deaths [2]. In 2020, over 103,000 new and 375,000 deaths cases were reported in Africa, making up 6.5% of all cancer cases in men [3]. More alarmingly, about 67,000 men lost their lives to PC that year, representing 8.3% of all cancer deaths among African men [3]. These numbers reveal the heavy burden of PC on the continent, where access to early diagnosis and modern treatments is often limited [4].

Projections indicate that the global burden of prostate cancer will increase substantially in the coming decades. Annual cases are expected to rise from 1.4 million in 2020 to 2.9 million by 2040, with annual deaths projected to nearly double to almost 700,000 during the same period[5]. This surge is anticipated to affect low- and middle-income countries disproportionately, primarily due to aging populations and limited access to early detection and treatment services [5].

Worldwide, cancer remains one of the biggest causes of death, with nearly 19.3 million new cases and close to 10 million deaths in 2020 [6]. PC is a major part of this global health challenge. However, the impact of the disease is felt most strongly in regions where healthcare systems are unable to provide widespread screening or early intervention [7]. This situation makes it crucial to develop more effective diagnostic tools and treatment options, particularly in Africa.

The MADCaP Consortium, which focuses on PC in men of African descent, is working hard to address the high rates of the disease through collaborative research [8]. This group connects researchers from across Africa and beyond, allowing them to share insights and resources. The collaboration is important because men of African ancestry are at a higher risk of developing aggressive forms of PC and often

have worse outcomes. By studying the genetic, environmental, and social factors that contribute to these patterns, the MADCaP Consortium aims to improve how PC is understood, diagnosed, and treated in African populations.

In this study, the concept of *progressiveness* refers to the degree to which prostate cancer advances from lower-risk to higher-risk stages. Prostate cancer progression is typically evaluated using clinical indicators such as prostate specific antigen (PSA) levels, Gleason score from biopsy, tumor volume, and imaging or staging assessments. A patient is considered to exhibit disease progressiveness when there is a measurable escalation in any of these markers, for example, a faster PSA doubling time, an increase in Gleason grade on repeat biopsy, or evidence of tumor growth on imaging. Understanding progressiveness is crucial, as it helps identify patients who may require more aggressive interventions and informs predictive modelling efforts aimed at anticipating shifts to higher-risk categories[9]. By operationalizing progressiveness in this manner, this study aims to evaluate risk stratification and improve the accuracy of predictive models for prostate cancer outcomes.

In recent years, supervised ML has emerged as a powerful tool in healthcare. These algorithms are trained using datasets where the inputs and outputs are known, allowing them to identify patterns and make predictions on new, unseen data [10]. For PC, ML has been used to group patients based on their risk levels, taking into account their genetic information [11]. This approach is critical for personalized treatment, where decisions are tailored to an individual's unique profile. ML has also been used to predict whether cancer will come back after surgery by analyzing MRI scans [12]. While these early applications are promising, more research is needed to prove their usefulness in everyday clinical settings [13].

The way these algorithms work is by learning from a dataset that contains both input information (like a patient's medical history, lab results, or imaging scans)

and corresponding outcomes (such as whether the patient has cancer or how they responded to treatment) [14]. Over time, the algorithm learns how to connect certain patterns in the data to specific outcomes. Once trained, it can predict the results for new patients, helping with tasks like diagnosing PC, estimating the risk of disease progression, or planning treatments [15].

Despite its potential, using ML in healthcare especially for PC faces several challenges. First, there's a need for large, diverse datasets to train these algorithms effectively. In regions like Africa, collecting such data is difficult due to limited healthcare infrastructure and resources [16]. Second, the models themselves need to be understandable by doctors. Healthcare professionals are unlikely to rely on a ML system unless they can trust its recommendations and see how it arrived at its conclusions [17].

Even with these challenges, the future looks bright for ML in PC research. As technology continues to advance and more data becomes available, these algorithms will likely become a regular part of healthcare, leading to more accurate diagnoses, personalized treatments, and better outcomes for patients [18]. This research aims to contribute to that future by developing a supervised ML algorithm that addresses the unique challenges of diagnosing and treating PC, particularly in African populations.

In summary, PC is a serious global health issue, and the situation is particularly dire in Africa due to a lack of resources and access to timely medical care. Supervised ML offers a potential solution, with its ability to analyze large amounts of data and provide predictions that can guide diagnosis and treatment. However, more work needs to be done to overcome current obstacles and make these technologies widely accessible. This thesis seeks to advance that goal by exploring the development of a ML model tailored for PC care in Africa.

1.2 Problem Statement

Prostate cancer (PC) is a pressing global health issue and has emerged as a particularly severe public health concern in Africa. Men of African descent are disproportionately affected, exhibiting higher incidence rates, earlier onset, and greater mortality compared to other populations worldwide [19]. In Sub-Saharan Africa, the burden of PC is compounded by health system challenges, including a critical shortage of trained specialists. For example, the region has approximately one pathologist per one million people, in stark contrast to ratios of about one per 25,000 in high resource countries such as the United States and United Kingdom [19]. This shortage leads to delays in diagnosis, frequent misdiagnoses, and suboptimal treatment decisions, ultimately increasing morbidity and mortality [20]. The lack of infrastructure and specialized human resources therefore exacerbates the already high disease burden.

Conventional screening approaches, such as prostate-specific antigen (PSA) testing and digital rectal examination (DRE), remain the most widely used methods for PC detection. However, these tools are limited by poor specificity and sensitivity, with PSA testing in particular yielding high false positive rates and failing to consistently detect aggressive disease [21]. Cultural, social, and economic barriers such as limited awareness of PC, stigma, and reluctance to undergo screening further undermine early detection among African men, contributing to the high proportion of late stage diagnoses [22]. In addition, there is a near absence of validated, population specific risk prediction models that reflect the unique genetic, environmental, and socio-economic factors shaping PC outcomes in African settings [23]. These gaps limit the potential of existing diagnostic strategies and highlight an urgent need for innovative, context specific solutions.

Emerging computational approaches, particularly machine learning (ML), offer

a promising pathway to addressing these challenges. ML methods are capable of integrating heterogeneous data including clinical, demographic, lifestyle, and socio economic information to improve risk prediction and stratify patients for targeted screening and early intervention. Beyond structured data, ML techniques such as deep learning have also shown promise in histopathological image analysis, potentially compensating for the shortage of pathologists and reducing diagnostic delays [24]. The application of ML in African contexts could therefore facilitate earlier detection, improve prognostic accuracy, and inform resource efficient allocation of limited healthcare services. By leveraging these advances, this study seeks to enhance PC risk prediction and early detection among African men, directly addressing the limitations of current screening methods and the unmet need for population specific models [25].

1.3 Justification

Despite significant advances in prostate cancer (PC) research, African men remain severely underrepresented in global predictive modelling studies, resulting in a lack of validated, population specific risk prediction tools [23]. This gap is critical, as African populations face disproportionately high PC incidence and mortality, yet current screening tools such as PSA and DRE are poorly specific and often miss aggressive disease [21]. Moreover, the severe shortage of pathologists across the continent contributes to delayed diagnoses and inconsistent treatment decisions, underscoring the urgent need for scalable, data-driven solutions [19].

The MADCaP dataset provides a unique opportunity to address these gaps. It contains clinical, socio-demographic, and lifestyle data from African men a population rarely reflected in large scale predictive models allowing for the development of contextually relevant risk stratification tools. By applying machine learning

(ML) algorithms to this dataset, the study can explore associations that traditional methods may overlook and generate models specifically tailored to African men, thereby improving prediction accuracy and early identification of high-risk cases [14, 26].

This study is justified on three main grounds. First, it directly addresses the underrepresentation of African data in global PC research by leveraging the MADCaP cohort. Second, it seeks to develop risk prediction models that go beyond PSA and DRE, thereby filling a critical diagnostic gap. Third, its findings have the potential to inform targeted screening, risk stratification, and clinical decision-making in resource-limited African settings, where cost-effective and scalable tools are urgently needed [27, 25]. By aligning methodological innovation with population-specific needs, this research advances both scientific understanding and public health equity in PC management.

1.4 Aim

The study aims to apply supervised learning models to predict the progressiveness of PC and hidden correlations and patterns.

1.5 Objectives

1. To apply ML models to predict PC progressiveness and severity.

2. To investigate hidden correlations and risk factors that lead to various cancer stages and explore potential disparities or associations.

3. To evaluate ML algorithm performance compared to expert diagnosis.

Thesis structure

The remainder of this report is structured as follows: Chapter 2 delves into key concepts related to PC, covering areas such as case identification, clinical presentations, and treatment options. It also reviews the current approaches to PC surveillance. Chapter 3 outlines the materials and methods used in this research, with the empirical data and findings presented in Chapter 4. Chapter 5 offers a discussion of the research outcomes, while Chapter 6 presents the conclusions drawn from the study.

Chapter 2

Literature review

This section begins with a brief overview of PC, focusing on its manifestation and management in men of African descent, particularly within the context of Men of African Descent Carcinoma of the Prostate (MADCaP) in South Africa, highlighting key diagnostic parameters that aid in early detection. Following this, we provide an overview of the South African PC Screening Guidelines, which inform best practices for early screening and intervention in high-risk populations. We then delve into the supervised ML algorithms employed to predict PC cases in this study, including decision trees (DTs), random forest (RFs),Support Vector Classifiers (SVCs), and gradient boost (GB) methods. The selection of these classifiers is based on their capacity to handle complex clinical and demographic data. Additionally, we discuss key factors such as feature selection, model training, and validation techniques to optimize prediction accuracy. Finally, we examine performance metrics like accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC), which are crucial for assessing the effectiveness of the models and selecting the most suitable approach for predicting PC cases within the MADCaP. This section lays the foundation for understanding the research methodology and its significance in addressing PC in South Africa.

2.1 Overview of the South African PC screening guidelines

Figure 2.1, extracted from The South African PC Screening Guidelines [28], outlines the recommended Prostate-Specific Antigen (PSA) screening protocols for PC in South Africa, considering factors such as age, risk factors, and clinical conditions. It highlights the key guidelines for initiating and discontinuing PSA screening, emphasizing the importance of tailoring screening approaches based on patient-specific factors.

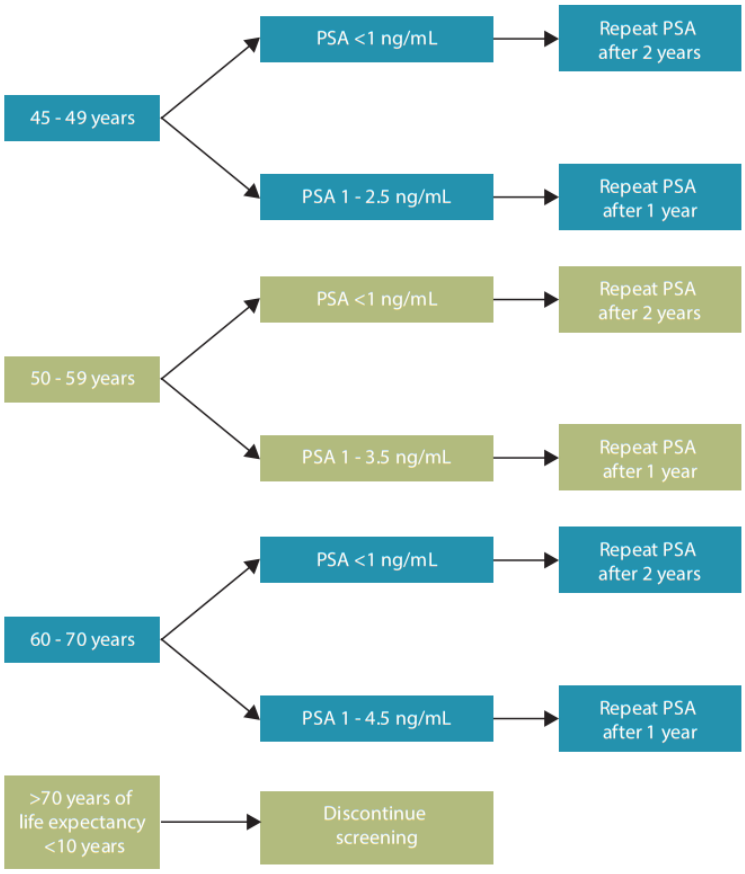


Figure 2.1: Recommended age-related screening intervals for prostate-specific antigen screening (John et al., 2024)

PSA-based screening for prostate cancer has several limitations that impact its effectiveness and reliability. One major issue is the lack of specificity, as elevated PSA levels can be caused by benign conditions such as prostatitis or benign prostatic hyperplasia (BPH), leading to false positives and unnecessary biopsies [29]. Conversely, PSA screening can also produce false negatives, where aggressive prostate cancers go undetected due to low PSA levels [30]. Moreover, there is no universally agreed-upon PSA threshold for recommending further testing, leading to variability in clinical decision-making [31]. These challenges highlight the need for improved screening strategies, such as incorporating additional biomarkers or risk-based approaches, to enhance the accuracy and effectiveness of prostate cancer detection [32].

For average-risk men, routine PSA screening is typically recommended to begin at age 50 [33]. However, the guidelines advocate for earlier screening in high-risk populations, including black African men and individuals with a family history of prostate or breast cancer, who should begin screening at age 45 [28].Men with known genetic predispositions, such as carriers of *BRCA1*, *BRCA2*, *HOXB13*, *ATM*, or *CHEK2* genes, are advised to initiate screening as early as age 40, or 10 years before the youngest affected family member was diagnosed if that occurred before age 40. [34].

In contrast, for men over 70 years or those with a life expectancy of less than 10 years, the guidelines recommend against routine PSA screening, as studies have shown no significant benefit in reducing PC mortality in this group [35]. Data from the European Randomised Study of Screening for PC and other clinical trials have demonstrated that the risks of curative treatments, such as surgery and radiation, often outweigh the benefits as age increases [36]. In such cases, the competing mortality risks from other causes are likely to reduce the advantages of PC screening.

The guidelines also advocate for individualized screening intervals, particularly for men who present with normal initial PSA levels. While international guidelines suggest that PSA testing can be spaced out to as long as eight years, more frequent testing is recommended for black African men due to their higher incidence of and mortality from PC [37]. Additionally, the guidelines provide detailed instructions on deferring PSA tests under circumstances that could lead to temporary PSA elevations, such as bacterial infections, recent prostate procedures, or acute urinary retention.

The diagram visually consolidates these recommendations, illustrating the decision-making process for determining optimal PSA screening intervals based on patient age, health status, and risk factors. It serves as a crucial tool for guiding evidence-based screening practices, ensuring that patients receive appropriate and timely interventions while minimizing unnecessary risks.

2.2 Genetic risk and aggressive disease manifestation

Men of African descent have a higher incidence and mortality rate of PC compared to other ethnic groups [38]. In South Africa, a study conducted in the Free State province found that prostate cancer is the most frequently diagnosed cancer among men, significantly affecting rural communities with limited access to healthcare [39]. Research suggests genetic, environmental, and lifestyle factors contribute to the increased risk, particularly in regions like Southern and Northern Africa [40]. PC risk rises significantly after the age of 50, and other factors such as dietary habits (e.g., high red meat consumption and low intake of fruits and vegetables) also play a role in exacerbating the disease's impact on this population [41].

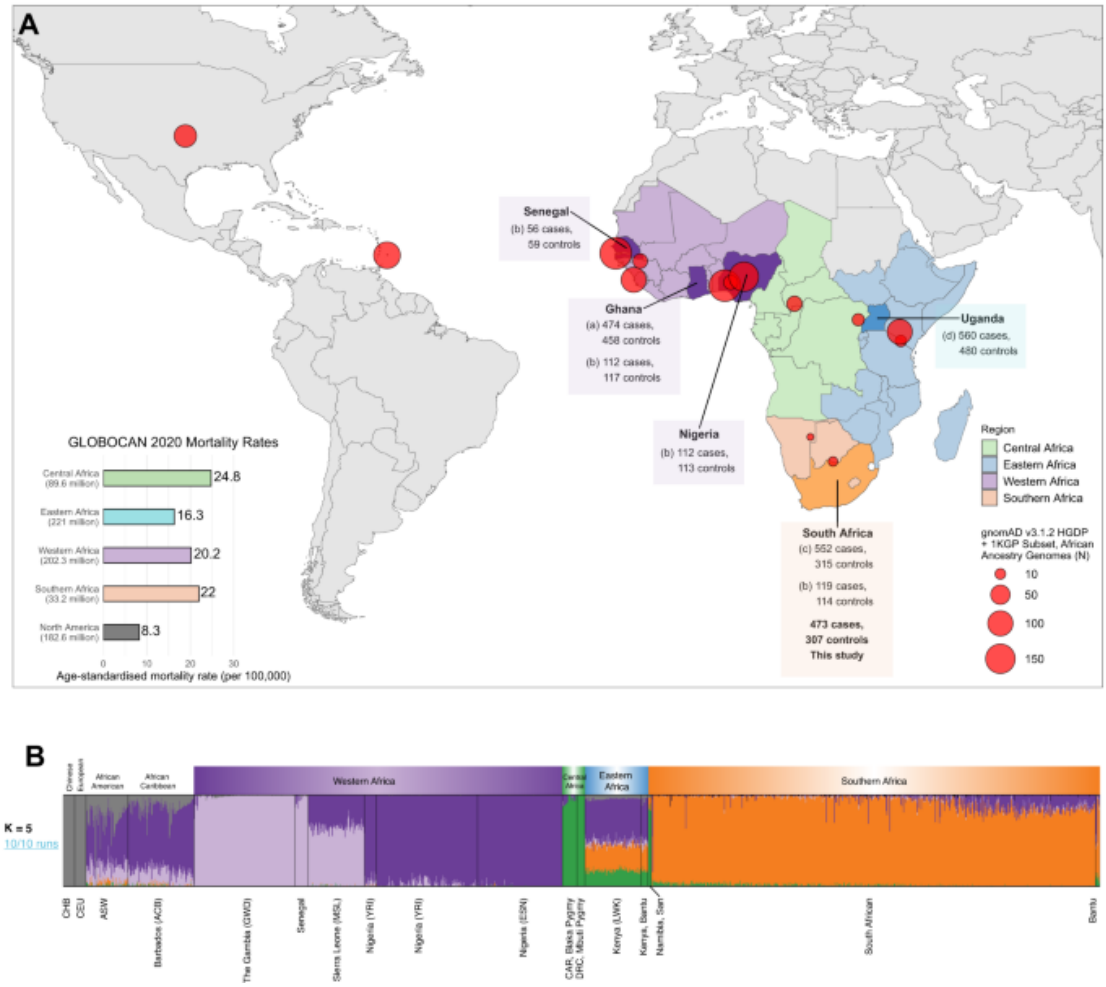


Figure 2.2: Summary of genome-wide association studies (GWAS) in Africa and ancestral fractions.(Janivara, R., Chen, W.C., Hazra, U. et al., 2024)

The map in Figure 2.2, based on the work of Janivara et al. (2024), shows where Genome-Wide Association Studies (GWAS) on PC have been conducted in African populations [42]. The red circles on the map mark the locations and sizes of samples collected from people of African ancestry, using data from the Human Genome Diversity Project (HGDP) and the 1000 Genomes Project (1KGP). The sizes of the circles represent how many samples were taken from each area, indicating which regions had more extensive genetic research.

Additionally, the map includes a bar plot in the bottom left corner that shows PC mortality rates from GLOBOCAN 2020 [6]. This plot gives important information about the impact of PC in different areas, with the number of men in each region noted in brackets below the region names.

The study sought to investigate the genetic risks linked to prostate cancer and its aggressive forms in men of African ancestry using their own generated data. However, the researchers noted in their paper that they were unable to assess polygenic risk scores for prostate cancer [43]. The research found significant genetic variants that are linked to both the likelihood of developing PC and the severity of the disease. By looking at these genetic factors, the study sheds light on the unique biological pathways that might affect PC risk in southern Africa. It also stresses the importance of further research focused on African populations, which could help develop better methods for early detection and treatment of PC, ultimately reducing health disparities in these communities.

2.3 Case identification

The primary risk factors for PC include advanced age, family history, and genetic predispositions, with certain populations, such as those of African ancestry, facing an increased likelihood of developing aggressive forms of the disease [44]. Early detection is crucial for improving outcomes, yet it remains challenging due to the typically asymptomatic nature of PC in its early stages [30]. Many men may only present with symptoms when the disease has advanced, making treatment more difficult and often less effective. Conventional diagnostic tools, such as prostate-specific antigen (PSA) testing and digital rectal exams (DRE), are commonly used for early screening but are often inadequate in distinguishing between indolent and aggressive forms of PC [45]. PSA testing, for instance, can lead to false

positives or miss aggressive cancers, resulting in over-diagnosis and over-treatment, respectively [35].

To address these limitations, there has been a growing focus on integrating more sophisticated diagnostic techniques. Genome-wide association studies (GWAS) have contributed to identifying specific genetic markers associated with higher PC risk, particularly in high-risk populations [46]. This genetic data, combined with demographic and clinical information, is at the core of initiatives like the MADCaP network, which aims to improve predictive accuracy and tailor screening approaches based on individual risk profiles. By incorporating genetic screening alongside traditional methods, the hope is to enhance early detection, allowing for timely interventions and more personalized treatment strategies, especially for populations at higher risk [47].

Advances in genetic research, coupled with better diagnostic tools and large-scale data integration, offer promising opportunities to refine PC screening and improve outcomes, particularly in regions where disparities in healthcare access and outcomes are most pronounced.

2.4 Conceptual framework

This study adopts a framework based on Chow et al. (2019), integrating demographic, genetic, clinical, and lifestyle factors to predict PC in men of African descent. The framework focuses on the following factors [48]:

- Demographic factors: Age, nationality, and place of birth.
- Clinical risk factors: PSA levels and biopsy results.
- Lifestyle factors: Diet and smoking habits.
- Genetic factors: Family history, race, and specific genetic variants.

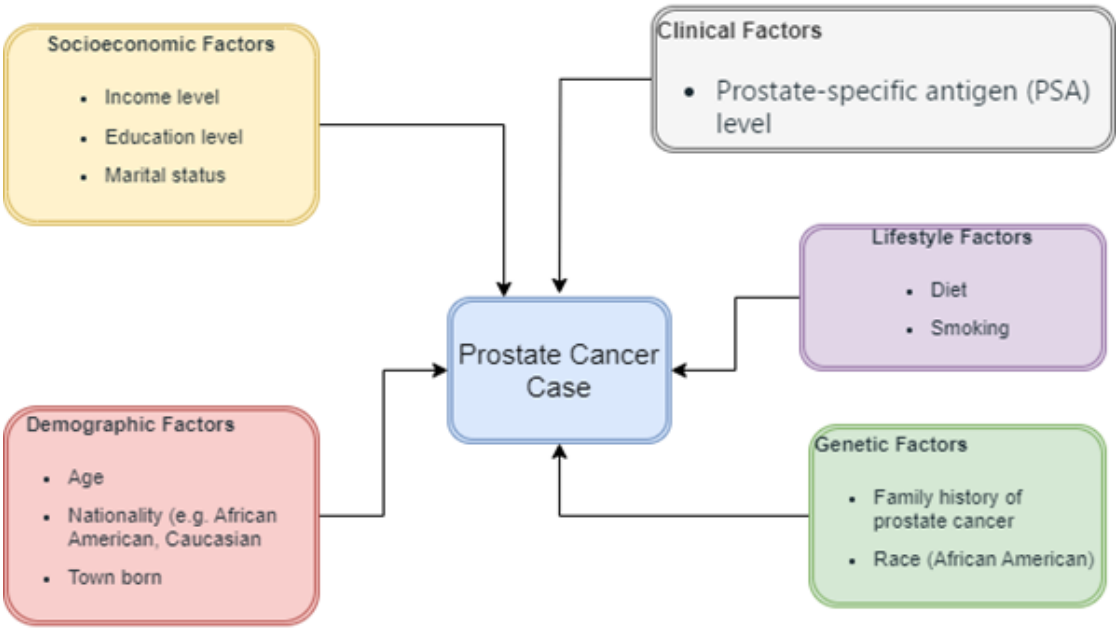


Figure 2.3: Conceptual framework showing factors associated with PC case.

This holistic framework provides a nuanced approach to understanding PC risk, enabling the development of more tailored predictive models for men of African descent .

2.5 Big data and ML in PC research

The availability of large datasets in healthcare, including genomic, clinical, and imaging data, has transformed PC research. These big data sources allow for the development of personalized treatment strategies using ML (ML). For example, Negpal et al. (2020) created a model with over 85% accuracy using genomic data from large PC cohorts, highlighting the power of ML in advancing cancer diagnostics [49]. ML algorithms are also enhancing diagnostic accuracy in imaging, particularly in magnetic resonance imaging (MRI). Deep learning models, such as those developed by Gulshan et al. (2016), have shown human-comparable accuracy

in detecting PC from histopathology slides, while Anderson et al. (2019) achieved 97% accuracy using an SVM model to predict PC based on demographic and clinical data [50, 51]. Convolutional neural networks (CNNs) have further demonstrated up to 98% accuracy in PC detection, particularly in high-risk populations like men of African descent [52].

2.6 ML model classification

In this study, classification refers to the process of assigning cases of PC to distinct categories based on clinical and genetic features, specifically focusing on the MADCaP Consortium in South Africa. The goal is to leverage supervised ML algorithms to predict PC cases and classify them according to the Gleason score, which measures the aggressiveness of PC tumors. These categories include "Low Risk", "Intermediate Risk", and "High Risk", as defined by Gleason score thresholds. [53].

Traditionally, classification is understood as the assignment of observations to predefined categories in a finite space. However, our interpretation of classification differs from that of the WHO International Classification of Disease (ICD), which organizes causes of death and informs morbidity and mortality statistics [54]. In the context of ML, classification relates to predicting a qualitative or categorical outcome based on data inputs, with the objective of grouping observations into specific classes [55]. For example, given PC patient data, a classification algorithm might predict whether an individual is at low, intermediate, or high risk of aggressive disease.

Supervised ML classifiers work by identifying patterns within a dataset and generating predictive models. These models are designed to segregate cases into distinct categories, such as the three risk levels of PC aggressiveness. For example,

in PC diagnosis, algorithms analyze various features such as age, PSA levels, and family history alongside genetic markers to determine whether a patient falls into the low, intermediate, or high-risk group, based on their Gleason score.

In this research, the focus is on multi-class classification, where a classifier predicts three possible outcomes: Low Risk, Intermediate Risk, and High Risk based on their Gleason score, rather than binary classification (presence or absence of disease). This is crucial for tailoring clinical treatment plans based on the aggressiveness of the tumor. By utilizing supervised ML algorithms, such as DTs, RFs, and SVMs, we can more accurately predict the category a patient belongs to, leading to earlier intervention and more personalized care.

Additionally, supervised classifiers in this study are trained on the MADCaP dataset, which includes rich demographic, clinical, and genetic data from men of African descent, a population that is often underrepresented in genetic studies but exhibits a disproportionately high burden of PC [56]. ML models trained on this data aim to improve the predictive accuracy of PC diagnosis and prognosis, potentially reducing disparities in healthcare outcomes for this high-risk group.

2.7 Rule based Classifiers: a background

Rule-based classifiers, also called inferential classifiers, work by using a set of predefined rules to classify data into categories. These rules are based on features such as Gleason scores, age, family history, and prostate-specific antigen (PSA) levels. For example, in PC prediction, a rule might look like this:

$$\text{IF Gleason score} \geq 8 \text{ THEN classify as High Risk.} \tag{2.1}$$

Each rule has two parts: a condition (the 'IF' part) and a result (the 'THEN' part).The classifier applies these rules to predict whether a patient falls into a

specific risk category, such as *Low Risk*, *Intermediate Risk*, or *High Risk*.

Although rule-based classifiers were once widely used, they have some problems. They can struggle to handle complex data and are hard to maintain when many rules are added. If two rules conflict, the predictions may become less accurate. This makes them less effective for large datasets with many features.

2.7.1 Why ML is better

While rule-based classifiers were useful in the past, we now have more advanced methods like ML classifiers. These methods do not rely on predefined rules. Instead, they learn patterns directly from the data, making them more flexible and accurate.

ML algorithms, such as RFs, SVMs, and GNB, are better at handling complex and large datasets. They can find patterns and relationships that are too difficult for rule-based systems to capture [57].

In this thesis, we did not use rule-based classifiers. Instead, we focused on advanced ML techniques. These modern methods are more powerful and reliable for predicting PC risk. While rule-based approaches are included in the background section for context, the actual work in this research is based on these newer and better ML algorithms.

2.8 ML classifiers

ML algorithms are increasingly being utilized in critical areas of disease prediction, including PC, due to their adaptability and ability to continuously improve through learning. These algorithms are designed to generalize from data, making predictions on new, previously unseen cases by forming complex mathematical relationships between inputs and outputs. In PC prediction for men of African

descent, where traditional diagnostic tools may underperform, ML classifiers offer a data-driven approach to enhancing detection accuracy and outcome predictions.

ML classifiers for PC prediction, especially within the context of the MADCaP Consortium, focus on categorizing patients into different risk groups (e.g., *Low Risk*, *Intermediate Risk*, or *High Risk*) based on Gleason scores, and clinical features. The algorithms employed in this context are part of a broader taxonomy of ML models designed for classification tasks. In this section, we discuss three general categories of ML algorithms: unsupervised learning, semi-supervised learning, and supervised learning.

2.8.1 Unsupervised learning

Unsupervised learning is employed when the exact structure of the data is unknown, and the goal is to uncover hidden patterns or groupings without predefined labels for the target variable [58]. In the context of PC research, unsupervised techniques like clustering and Principal Components Analysis (PCA) may be used to explore genetic data and identify subgroups of men with similar genetic profiles, which may inform risk stratification before applying supervised classification techniques.

Unlike predictive methods, unsupervised algorithms do not seek to classify instances into predefined categories such as *High Risk* or *Low Risk*. Instead, they seek to discover natural groupings (clusters) in the data. For instance, PCA might reduce the dimensionality of a genetic dataset, identifying principal components that explain the most variance in the data, which could suggest underlying genetic risk factors associated with aggressive PC.

2.8.2 Semi supervised learning

In cases where only a subset of the dataset is labeled, semi-supervised learning algorithms are useful. These methods leverage both labeled and unlabeled data to improve the predictive performance of classifiers. In cancer prediction, where large-scale genomic datasets might have incomplete annotations, semi-supervised techniques can infer missing labels and use these to train the model [59].

These algorithms combine elements of supervised and unsupervised learning. For example, in the MADCaP study, only some of the genetic data may be labeled with Gleason scores or clinical outcomes. A semi-supervised model could use this partial information to generate labels for the unlabeled data, improving the model's ability to predict risk categories based on genetic and clinical features.

2.8.3 Supervised learning

Supervised learning is the most common approach for disease classification, where the goal is to predict a predefined outcome—such as whether a patient has *Low Risk*, *Intermediate Risk*, or *High Risk* PC—based on labeled data. Supervised algorithms build models that learn a functional relationship between input variables (e.g., age, PSA levels, family history) and the target output (risk category) [60].

In this study, supervised classifiers such as Decision tree (DT), Support Vector Classifier (SVC), Random Forest (RF), and KNN Neighbors (KNN) are used to predict PC outcomes in men of African descent. These algorithms learn from the annotated training data, where the PC risk category (e.g., *High Risk* or *Low Risk*) has been labeled in advance. Given a set of clinical features, these models predict which risk category a new patient falls into, based on the patterns learned from the training data.

Mathematically, supervised learning for PC classification can be described as

follows. Let $\{x, y\}$ be a set of attributes, where x represents the input features (such as clinical and genetic factors), and y is the class label indicating the risk category assigned to the patient. In this case, we classify patients into three distinct risk categories: Low Risk (C_{low}), Intermediate Risk ($C_{\text{intermediate}}$), and High Risk (C_{high}). The classifier learns a mapping from x to y, where y takes one of these three class values.

Formally, the risk categories are mutually exclusive, and we express this as:

$$C_{\text{low}} \cap C_{\text{intermediate}} \cap C_{\text{high}} = \emptyset$$

This indicates that each patient can only belong to one risk category at a time (Low, Intermediate, or High), and these categories are predefined in the training data. The classifier's goal is to correctly assign patients to one of these categories based on their input features.

The accuracy of the supervised classifier depends on the quality of the labeled data (i.e., how well the risk categories are defined) and the relevance of the features used for prediction.

2.9 Model evaluation

Evaluating the performance of ML classifiers is crucial in determining how effectively a model can predict PC cases, particularly in high-risk groups such as men of African descent [61]. To assess the performance of classification models trained on PC datasets, various metrics and techniques are employed. These evaluations help determine how well a classifier generalizes to unseen data and how accurately it predicts clinical outcomes such as PC risk categories based on Gleason scores.

A common method for evaluating a classifier's performance is to split the dataset into two parts, typically 80% for training the model and 20% for test-

ing [62]. This method helps assess the model's ability to predict outcomes for new, unseen data. In this context, PC risk prediction models are trained on features such as Gleason scores, PSA levels, and demographic information, and their predictions are compared to actual clinical diagnoses. The accuracy of the model is then measured by how well the predicted risk categories match the known labels in the test set.

2.10 Optimization

One of the challenges in model evaluation is the potential for overfitting or underfitting. Overfitting occurs when a model performs very well on the training data but poorly on new, unseen data [63]. This happens because the model has learned too many complex patterns in the training data, including noise, which leads to high variance. In PC prediction, overfitting could result in inaccurate risk categorization when applied to new patients.

Conversely, underfitting occurs when the model performs poorly on the training data but surprisingly better on unseen data, indicating that the model failed to capture important patterns during training [64]. Both overfitting and underfitting are examples of model-misfit, which can negatively impact prediction accuracy.

To address model-misfit, especially in cases with imbalanced classes, methods such as adjusting class weights or resampling the data can be employed. In the context of PC, where there may be more Low Risk cases than High Risk cases, resampling or using class weights ensures that the model pays equal attention to each category during training.

2.10.1 Cross validation and grid search

To improve generalization and reduce overfitting, robust approaches like Cross-Validation (CV) are commonly used. The most popular method is K-Fold Cross-Validation, where the dataset is randomly split into K equally-sized subsets or folds [65]. The classifier is trained on K-1 folds while the remaining fold is used as a test set. This process is repeated K times, with each fold serving as the test set once. The final performance score is the average of all iterations, providing a more reliable estimate of the model's ability to generalize to new data.

Another optimization technique is Grid Search, which systematically searches for the best combination of hyperparameters by training the model using all possible parameter settings [66]. When combined with K-Fold CV, Grid Search helps to ensure that the selected model parameters result in the best performance across all validation folds.

2.10.2 Model selection

In PC risk prediction, we utilized techniques such as the Confusion Matrix and the Area Under the Curve (AUC), along with Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves, to evaluate model performance. Cross-validation was also employed to ensure the accuracy and generalizability of the models. Explanations of these methods will be provided in the next chapter.

2.11 Summary

This chapter examines the burden of PC among men of African descent within the South African MADCaP cohort, focusing on the importance of early detection through Prostate-Specific Antigen (PSA) testing, as emphasized in national

screening guidelines. It highlights genetic risk factors unique to African populations, informed by genome-wide association studies (GWAS), and their contribution to the aggressive nature of the disease. The limitations of conventional diagnostic methods, such as PSA testing, are discussed, along with the potential of integrating genetic insights to improve detection and treatment strategies. Additionally, supervised ML algorithms, including DTs and SVMs, are introduced as innovative tools for predicting PC cases, showcasing their ability to address health disparities in high-risk populations.

Chapter 3

Materials and methods

3.1 Introduction

This chapter outlines the research methodology, data sources, and analytics approach used in the project. It explains the type of research, data types, target and independent variables, and the Knowledge Discovery of Datasets (KDD) process, which involves data collection, preprocessing (cleaning and transformation), and analysis and ML modeling. The chapter aims to provide a clear understanding of the research design and methods used to achieve the project's objectives.

3.2 Study setting and population

This study utilized data from the MADCaP consortium, a multi-country initiative investigating the genetics and epidemiology of prostate cancer in African populations. The data were collected at the Chris Hani Baragwanath Academic Hospital (CHBAH) study site, with additional recruitment of non-cancer controls from the St. John's Eye Hospital. All records were stored and managed in the MADCaP REDCap database [8].

The primary MADCaP study was a case–control design conducted between January 2017 and December 2021, enrolling newly diagnosed PC patients receiving treatment at the CHBAH Urology Outpatients department and age, gender, and ethnicity matched non-cancer controls. For this secondary analysis, a retrospective cross-sectional approach was applied, making use of all available records ($n = 1096$) from the CHBAH MADCaP dataset. The study population therefore comprised PC cases and their matched controls, providing a robust dataset for analysis.

3.3 Methods

We conformed with the agile software development methodology and adapted the generally acceptable framework for supervised ML, illustrated in Figure 3.1. In focus, we followed these methods to ensure repeatability and robust outcomes in our research.

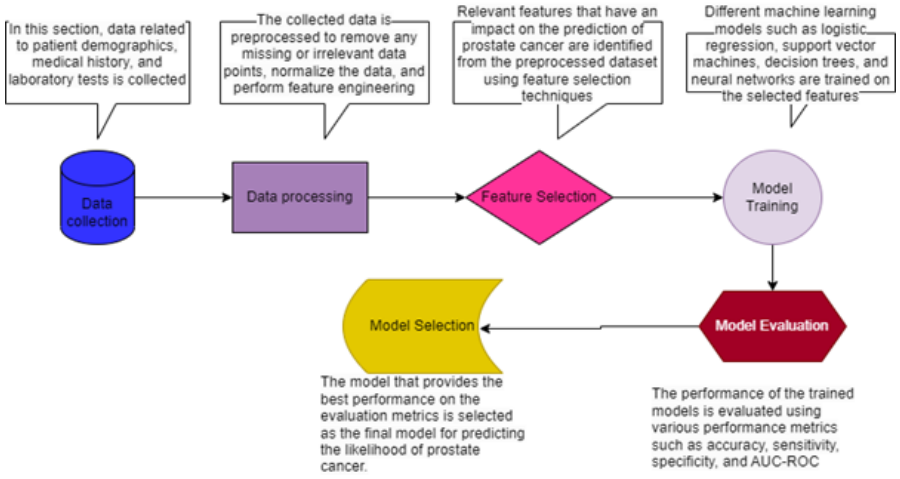


Figure 3.1: Typical supervised learning algorithm

3.3.1 Computational environment

The computational environment used for this study is summarized in Table 3.1. This environment ensured consistency across all experiments.

Table 3.1: Computational environment

Utility	**Version**	**Notes**
PANDAS	0.24.2	For high-performance data structures and analysis tools
SKLEARN	0.24.1	Tools for predictive data analysis
SEABORN	0.9.0	Python data visualization library based on matplotlib
NUMPY	1.16.2	Numerical library to facilitate the data management process
PYTHON	3.7.4	Programming language, based on the Anaconda Integrated Development Environment (IDE)
STATA	15.1	Statistical software for analysis (IC Edition)
COMPUTER	10 Pro	Intel® Core™ i5 2.3GHz Processor, 16GB memory, 64-bit Microsoft Windows Operating System

The SKLEARN package includes class libraries for ML models. SEABORN is based on the MATPLOTLIB graphic library for visualization, while PYTHON is an interpreted object-oriented programming language[67]. STATA is statistical software regularly updated with annual releases [68].

3.4 Overview of proposed solutions

The algorithms for the proposed solution were run using the Python® programming language [67] within the Google Colaboratory environment [69]. Google Colab was employed for all statistical and ML approaches, including data cleaning, processing, descriptive statistics, and quality checks [70].

3.5 Data definition

The dataset used for the analysis comprised 1,096 observations and 21 features related to PC cases and controls. The data included demographic attributes, clinical history, and laboratory measures, with the target variable, Gleason score, categorized into three risk groups: low risk, intermediate risk, and high risk. Following the initial cleaning phase, it was determined that several predictor variables had a significant amount of missing data, particularly in categorical features such as `hrf_smokestartremember`, `hrf_cigarettestype`, `hrf_currentlysmoke`, and `hrf_smokestopremember`. To address these missing values, categorical variables were imputed by replacing null values with `'None'`, ensuring that no data was lost during the cleaning process. For the numerical variables, including `hrf_smokingagestart`, `hrf_cigarettescount`, and `hrf_smokingagestop`, the KNN Imputer from scikit-learn was employed. This method utilized the five nearest neighbors to predict and fill in missing values, thus preserving the integrity of the dataset. The imputation of missing data was critical, as it helped maintain a robust analytical dataset for further modeling. After preprocessing, the final analytical dataset contained a total of 1,956 observations, as the class imbalance in the target variable was addressed using the ROSE (Random Over Sampling Examples) technique, effectively balancing the representation of risk categories

within the dataset. A detailed summary of the missing values and their handling is presented in the subsequent sections.

3.6 Feature selection and engineering

In predictive analysis, not all features in a dataset contribute equally to classification and prediction tasks. As such, a tailored approach to feature selection is essential[52]. For categorical attributes, the One-Hot Encoding process was applied, converting each categorical value into distinct binary attributes, represented as '1' or '0' to indicate the presence or absence of a category[57]. For example, demographic variables such as Hospitalization Status (encoded as 'Y' and 'N') and Gender (encoded as 'M' and 'F') were labeled as 1 and 0, denoting positive and negative responses, respectively. To enhance computational efficiency, all features were standardized to fit within a range of zero to one using the MinMaxScaler implementation from the SKLEARN library. This transformation is defined mathematically as:

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{3.1}$$

where X_{scaled} is the newly transformed vector, X is the instance to transform, X_{min} is the minimum value of X in the vector domain, and X_{max} is the maximum value of X in the vector domain. To mitigate data leakage, the training and validation datasets were engineered independently using the fit-transform methods from SKLEARN. In addressing skewness in distribution, missing values in categorical features were imputed using the most frequent observations. This approach struck an appropriate balance between the precision of imputation and the integrity of the data structure. For the continuous features, missing values were handled using the Nearest Neighbor imputation strategy. During analysis, we explored potential pat-

terns associated with missing data and concluded that it was missing completely at random. Thus, we employed the Nearest Neighbor strategy to impute missing values for age and used the most frequent strategy for gender. Additionally, domain knowledge was leveraged to manually select principal features, minimizing collinearity and enhancing model fit by eliminating correlated features and noise. Ensemble Selection was conducted on categorical attributes such as Sub-District and Province, and we limited One-Hot Encoding to the top ten levels, as proposed by Niculescu-Mizil et al.

Some variables were excluded from the analysis to ensure relevance and reduce redundancy. For instance, race was dropped since all participants in the study were Black, making it non-informative. Similarly, the variables age started smoking and age stopped smoking were excluded because their direct influence on prostate cancer progression is not well established in the literature, particularly among African populations. Instead, cumulative measures of smoking history were retained as they are more widely recognized as meaningful indicators of risk. Other variables, including PSA diagnosis date and PSA referral level description, were removed because they did not provide useful information for the study objectives. In addition, while Gleason score primary and secondary were collected, they were not used directly, as they were combined to generate the Gleason score total, which served as the primary outcome variable. Keeping both the individual and total scores would have introduced redundancy and potential multicollinearity; therefore, only the Gleason score total was retained

Table 3.2 displays the variables utilized in this study, along with their corresponding descriptions.

Table 3.2: Variable description

Characteristics	**Description**
Age	Patient age in years, collected at the first hospital visit due to PC comorbidities.
Socioeconomic Status	Includes educational level and income characteristics.
Smoking Habits	Indicates whether the patient has ever smoked or never smoked.
Ethnic Origin	Clinical PC classification based on geographic regions or origin within MADCaP.
Country and Town Born	Refers to the patient's country and specific town or city of birth.
PSA at Referral	Initial measurement of Prostate-Specific Antigen (PSA) levels upon referral, assessing potential PC symptoms or abnormalities.
PSA at Diagnosis	PSA levels measured at PC diagnosis, indicating cancer extent and aggressiveness.
Gleason Score	A grading system evaluating PC cell appearance under a microscope, ranging from 6 to 10, with higher scores indicating more aggressive cancer.

3.6.1 Handling Class Imbalance and Data Augmentation

Class imbalance is a common challenge in supervised learning when certain classes (minority classes) have significantly fewer samples compared to others (majority classes), introducing bias and potentially leading to high misclassification errors for minority classes [71]. In our dataset, there was a clear imbalance across prostate cancer (PC) risk categories, with 581 instances of 'Intermediate Risk,' 397 instances

of 'High Risk,' and only 118 instances of 'Low Risk.' To address this, we applied balancing and augmentation techniques tailored for the training data.

After testing several methods, we found that a hybrid oversampling approach, combining Random Over-Sampling with synthetic data generation, provided the most effective results. Specifically, our implementation leveraged the `RandomOverSampler` from the `imblearn` library in conjunction with the addition of Gaussian noise to existing data points. This noise-augmented Random Over Sampling technique generated artificial samples by adding slight random perturbations to existing instances, effectively creating new data points in a manner that resembles, but is not identical to, the traditional ROSE method [72]. While we initially referred to our method as ROSE, it is more accurate to describe it as a "hybrid oversampling technique" or "noise-augmented Random Over-Sampling." Following this approach, the dataset was balanced to 581 instances per risk category, increasing the total sample count from 1,096 to 1,743 and improving the classifier's ability to accurately predict each risk category across all performance metrics.

In addition to oversampling, data augmentation was employed to further enrich the training dataset and enhance model generalization. Data augmentation is a crucial technique in ML, particularly in contexts where datasets may be limited or imbalanced, as it increases diversity without the need to collect new data points [52]. In this study, Gaussian noise was added to numerical features such as PSA levels (`psa_psareferrallevel_normalized`, `psa_psadiagnlevel_normalized`) and age (`age_years_normalized`) to create new, slightly altered versions of data points. By generating 100 augmented samples per risk category, the dataset was expanded considerably, enabling the models to learn more robust and generalizable patterns across different PC risk categories. The final augmented dataset reached a size of 137,200 rows and 15 columns (features), ensuring that the model was not biased

toward any particular class and ultimately leading to more accurate and robust predictions.

Table 3.3: Dataset sizes of initial data, after balancing, and after augmentation

Risk Category	**Initial Data**	**After Balancing**	**After Augmentation**
High Risk	397	581	45,800
Intermediate Risk	581	581	45,800
Low Risk	118	581	45,800
Total	1,096	1,743	137,200

3.7 Analysis

3.7.1 Association and correlation

The initial analysis stage aimed to assess the association between each feature and the target variable using the chi-square test. This test helped identify statistically significant relationships, highlighting which features were most relevant for predicting PC status. Additionally, pairwise correlations were examined to detect highly correlated features, as these can lead to redundancy and multicollinearity issues. In cases where two features were strongly correlated, one was removed from the dataset to improve model stability and interpretability

3.7.2 ML models

To understand the outcomes of PC, we tested and compared several ML models to see which one performed the best for making predictions. Since the target variable had three different categories, we used classification models. In the first step, we used the training data to teach the models how to recognize patterns, refining

them until we found one that could accurately predict the outcomes. This process involved making adjustments to improve the models' accuracy. After preparing the data, we repeatedly split it into training and testing sets to evaluate how well each model could predict results on new data [73].

In this study, the following supervised learning algorithms were tested:

- **DTs**: A tree-like model of decisions used for both classification and regression tasks.
- **RF**: An ensemble learning method based on DTs, which enhances performance by combining the results of multiple trees.
- **SVMs**: A powerful classification algorithm that finds the optimal hyperplane for separating different classes in the data.
- **(GNB)**: A simple algorithm that uses probability to classify data. It assumes features are independent and normally distributed, but it still works well in many cases
- **KNN**: This algorithm classifies data based on its nearest neighbors. It looks at the "K" closest points and assigns the data point to the most common class among them.

Each of these models was evaluated based on their performance in predicting PC cases, using the cleaned, transformed, and augmented dataset. The models were compared using accuracy, precision, recall, and F1-score metrics. The best-performing model was selected for further analysis and interpretation.

K Nearest neighbor classifier

The K Nearest Neighbors (KNN) classifier is a ML algorithm used for classification tasks. It predicts the class of a new data point based on the majority class of

its nearest K neighbors in the training dataset, determined by a distance metric such as Euclidean distance [74]. In the context of predicting PC cases, KNN was applied by training on patient data containing relevant features and labels indicating cancer presence or absence. By comparing the new patient's features to the trained data, KNN identifies the K most similar cases and predicts the likelihood of PC based on the majority class among these neighbors. Although KNN is relatively simple, its performance can be influenced by the choice of K and the distance metric.With $K = 1$, each new data point representing PC cases is classified based on the characteristics shared with its nearest neighbor. However, using larger values of K can lead to misclassification errors, including false positives and false negatives. For example, a case that is actually classified as high risk may be incorrectly labeled as low risk if the majority of nearby cases are low risk. The proximity between data points is measured using the Euclidean distance function, expressed as:

$$D(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \quad (3.6)$$

where $x = (x_1, \ldots, x_m)$ and $y = (y_1, \ldots, y_n)$, with m representing the attribute values of the two points x and y. Therefore, the optimal value of K is the one that minimizes classification errors across the three risk categories: low, intermediate, and high. As illustrated in Figure 3.6, instance 'N' will be classified based on the defined characteristics of either 3 or 7 nearest neighbors using the KNN (KNN) algorithm for PC classification.

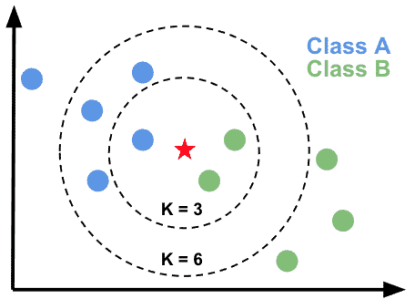


Figure 3.2: KNN classifier by Rajvi Shah

Proper data preprocessing, feature selection, and domain knowledge in medical applications are crucial for the algorithm's effectiveness [75]. By tuning the hyperparameters and ensuring high-quality data, KNN can assist healthcare professionals in making accurate predictions regarding PC cases.

DT and RF

DTs function similarly to a decision-making game, employing a sequence of yes-or-no questions to classify instances based on their features [10]. In the context of PC prediction, a DT may initiate by inquiring whether a patient's age exceeds 50, followed by subsequent questions regarding factors such as smoking history or PSA levels. This structured approach enables the model to make a definitive prediction. The fundamental process of any DT algorithm involves selecting the most appropriate attribute using Attribute Selection Measures (ASM) to partition the records, designating that attribute as a decision node, and recursively subdividing the dataset into smaller subsets until one of the following conditions is satisfied: all tuples belong to the same attribute value, there are no remaining attributes, or there are no instances left to evaluate. Despite their intuitive appeal, DTs are susceptible to overfitting, particularly in complex datasets, which can impair their generalization capabilities to unseen data [76]. Furthermore, diagram 3.3 illustrate

the operational mechanism of the DT is presented.

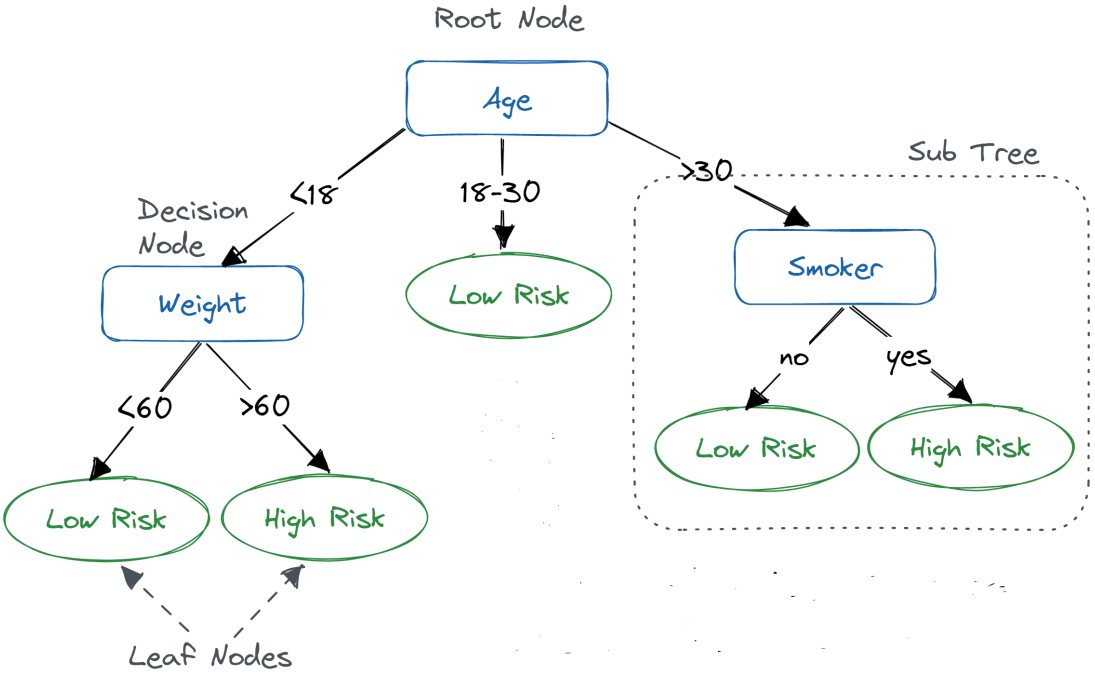


Figure 3.3: DT classifier branch

On the other hand, the RF is like a group of friends who make decisions together where each friend is like a DT that helps guess something and each have their own ideas [77]. When we need an answer, we ask all the friends and see what most of them think which helps us make a good guess. RFs are great because they stop any one friend from making up things, and by listening to all of them, we get better and more accurate answers which makes RFs effective in reducing the risk of overfitting that a single DT might be prone to, and they generally provide more robust and accurate predictions by leveraging the diversity of trees [78]. For predicting PC, we can think of the friends as different ways to guess if someone has cancer. We look at things like age, PSA levels, and family history to figure this out. By listening to many "friends" or ways of guessing, we can tell if someone might have cancer more reliably, and this helps doctors make smarter decisions. Figure

3.7 illustrates the RF classifier. Notation: Assume a dataset S, with attribute A. Let S_v denote an instance or subset of the data, where $S_v \subset S$. Also, let $A = v$ and Values_A be the set of all possible values of A. Then, the information gain can be expressed using the following function:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \quad (3.7) \qquad (3.2)$$

where Entropy is the measure of uncertainty of a random variable A.

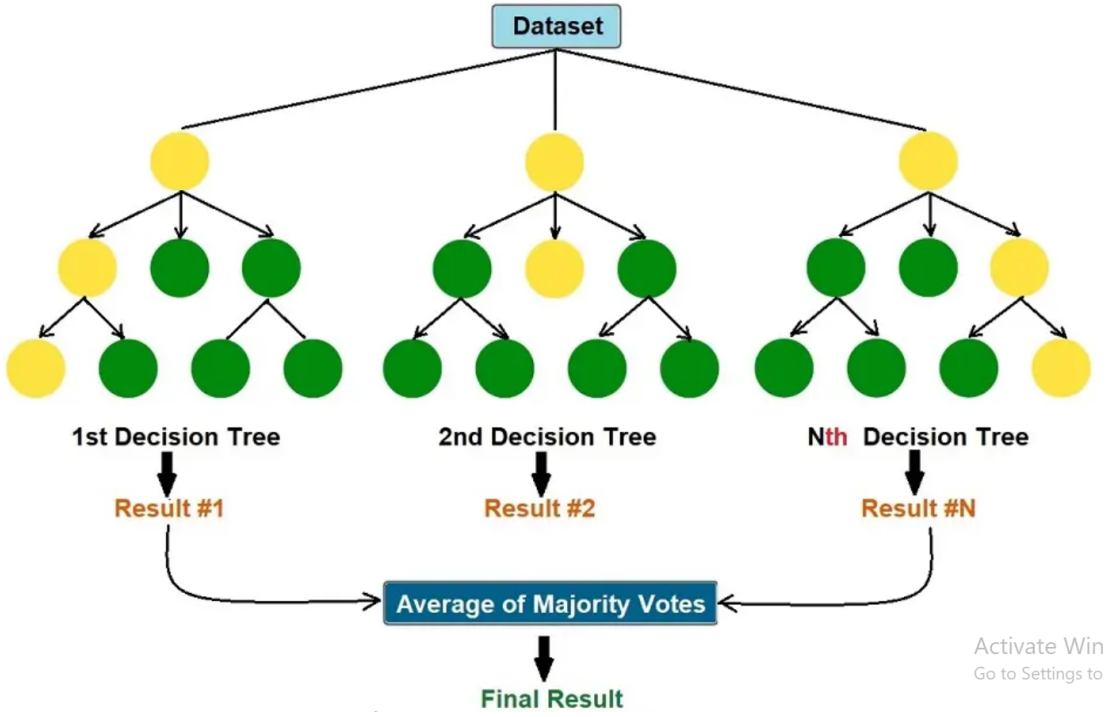


Figure 3.4: RF classifier branch by Ankit Chauhan

GBM and GNB

GB and GNB machines are a powerful ensemble learning technique used for both classification and regression tasks, known for their ability to produce highly accurate models by iteratively improving upon weak learners (typically DTs) [79].

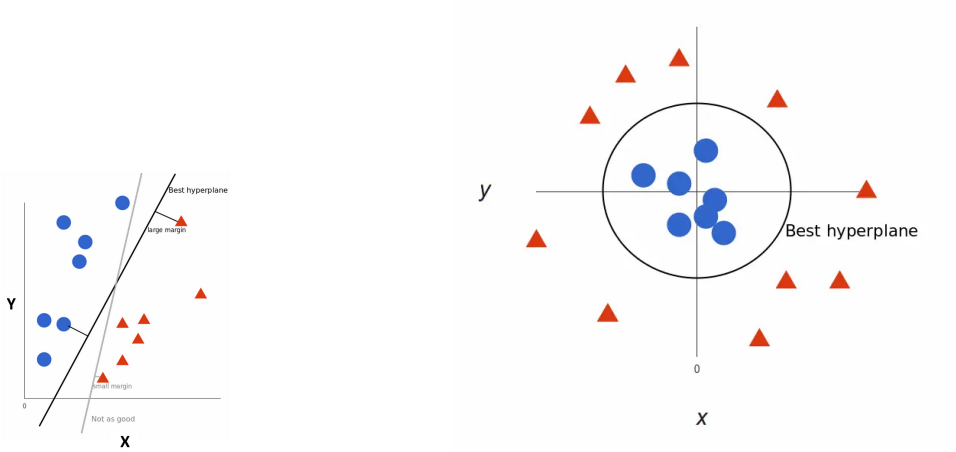
The key idea behind GNB is to build models sequentially, where each new model focuses on correcting the errors made by the previous ones. This is achieved by minimizing a loss function through gradient descent, where each subsequent tree is trained to predict the residual errors of the previous trees [80]. In the context of predicting PC, GNB can be applied by training on a dataset containing patient features such as age, PSA levels, and family history. The algorithm starts with a simple model, perhaps a single DT, and gradually adds more trees, each of which targets the mistakes made by the ensemble so far. Over time, this approach refines the predictions, leading to a model that accurately distinguishes between cancerous and non-cancerous cases [81]. One of the main strengths of GNB is its flexibility, as it allows for the use of various loss functions and regularization techniques, such as shrinkage and subsampling, to prevent overfitting and improve generalization [82]. However, this flexibility also means that GNB models can be complex and require careful tuning of hyperparameters, such as the number of trees, learning rate, and tree depth, to achieve optimal performance [83]. Despite these challenges, when properly tuned, GNB can deliver state-of-the-art results in medical applications like PC prediction, making it a valuable tool for healthcare professionals seeking to improve diagnostic accuracy [84] .

SVM

SVMs are a powerful and adaptable ML algorithm frequently used for classification tasks. They excel in high-dimensional spaces and are particularly effective in binary classification problems [85]. The fundamental concept of SVM involves finding a hyperplane that best separates data points from different classes while maximizing the margin between the closest data points, known as support vectors, from each class [86]. In the context of predicting PC, SVM can be utilized by training on a dataset that includes patient features such as age, PSA levels, and

family history, with each patient labeled as either having or not having cancer. The SVM algorithm identifies the optimal hyperplane that distinguishes cancerous cases from non-cancerous ones, enabling predictions for new patients based on which side of the hyperplane their data falls on [87].

A significant advantage of SVM is its ability to manage non-linear relationships between features through the use of kernel functions. These functions map the original features into a higher-dimensional space, where a linear separation becomes feasible [88]. However, to achieve optimal performance, SVMs require careful tuning of hyperparameters, such as selecting the appropriate kernel and setting the regularization parameter. This is especially crucial in small or noisy datasets to prevent overfitting [89]. When properly configured, SVMs can deliver accurate and dependable predictions, making them a valuable tool in medical decision-making for conditions like PC [90]. Moreover, SVM is flexible enough to perform both linear and nonlinear discriminatory analyses, often referred to as 'kernel' functions. It is well-suited for analyzing high-dimensional datasets with small sample sizes, particularly when combined with feature selection techniques. The best linear hyperplane is identified by comparing the margins, with the one providing the largest margin considered the optimal hyperplane for modeling. Figures 3.5a and 3.5b illustrate the SVM model with linear and non-linear hyperplanes, respectively.



(a) SVM with linear hyperplane (b) SVM with non-linear hyperplane

Figure 3.5: Comparison of SVM classifiers with different hyperplanes by James Le.

3.8 Performance evaluation of the algorithms

Performance evaluation is a crucial step in supervised ML, particularly for predicting PC cases. This process involves assessing the accuracy and effectiveness of various algorithms in determining the likelihood that a patient will develop PC. To evaluate our models, we utilized a 2x2 table commonly referred to as a Confusion Matrix, as illustrated in Figure 3.6 below. This matrix helps us understand the outcomes of our predictions by categorizing them into correct and incorrect classifications.

The Confusion Matrix is divided into four key categories. True Positives (TP) represent the cases where the model accurately predicts that a patient has the condition (such as cancer). True Negatives (TN) are instances where the model correctly predicts that a patient does not have the condition. False Positives (FP) occur when the model incorrectly predicts that a patient has the condition when they actually do not; this situation is known as a Type I error. Conversely, False

Negatives (FN) happen when the model incorrectly predicts that a patient does not have the condition when they actually do, which is referred to as a Type II error.

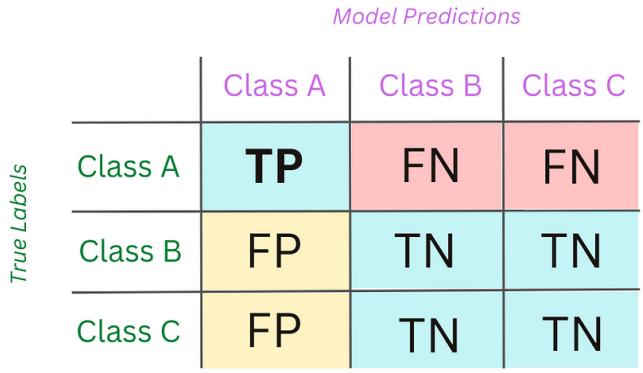


Figure 3.6: Confusion matrix

By analyzing these categories, we can gain valuable insights into the performance of our classifier, identifying its strengths and weaknesses. This evaluation is essential for understanding how well the algorithms are working and for improving their predictive capabilities in the context of PC diagnosis

3.8.1 Accuracy

The accuracy of the supervised ML algorithms in predicting PC cases is a crucial evaluation metric. It measures how well the models classify instances correctly, considering both true positives and true negatives [91, 92]. The formula for accuracy is as follows:

$$\text{Accuracy} = \frac{TP_1 + TP_2 + TP_3 + TN}{TP_1 + TP_2 + TP_3 + TN + FP_1 + FP_2 + FP_3 + FN_1 + FN_2 + FN_3}$$

3.8.2 Sensitivity

Sensitivity, also known as recall or true positive rate, is a performance metric that measures the ability of a classification model to correctly identify positive instances from the total number of actual positive instances in a dataset. It quantifies the proportion of true positives that are correctly predicted by the model [93]. The formula for sensitivity is as follows:

$$\text{Sensitivity}_i = \frac{TP_i}{TP_i + FN_i} \quad \text{for } i = 1, 2, 3$$

where TP_i represents the true positives for class i, FN_i represents the false negatives for class

Specifically, $i = 1$ corresponds to Low Risk, $i = 2$ to Intermediate Risk, and $i = 3$ to High Risk.

A high sensitivity value indicates that the model has a low rate of false negatives, meaning it effectively identifies a large portion of the actual positive instances. On the other hand, a low sensitivity value suggests that the model may miss a considerable number of positive instances, leading to a higher rate of false negatives.

3.8.3 Specificity

Specificity is a performance metric that measures the ability of a classification model to correctly identify the negative instances or true negatives (TN). It quantifies the proportion of instances correctly classified as negative out of the total actual negative instances [94]. The specificity formula is as follows:

$$\text{Specificity}_i = \frac{TN_i}{TN_i + FP_i} \quad \text{for } i = 1, 2, 3$$

A high specificity value indicates that the model is good at avoiding false posi-

tives, meaning it accurately identifies negative instances. Conversely, a low specificity value suggests that the model misclassifies a significant number of negative instances as positive, leading to a higher false positive rate.

3.8.4 Area under the curve (AUC) of the receiver operating characteristic (ROC) curve

The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) Curve holds significant importance when evaluating binary classification models, especially in domains like ML-based PC prediction. The ROC curve depicts a model's performance across various decision thresholds by graphing the True Positive Rate against the False Positive Rate [95]. AUC condenses this curve into a single value, ranging between 0 and 1, where higher values indicate superior model discrimination. In the context of PC prediction, AUC is instrumental in assessing and comparing models' capacities to accurately classify cancer and non-cancer instances. It aids in optimal threshold selection, offering insights for medical decision-making, thus enhancing diagnostic precision and potentially influencing clinical strategies. By sharing the AUC value with medical professionals, a comprehensive understanding of the model's PC prediction prowess can be conveyed, facilitating informed choices about further diagnostics or treatments. In conclusion, AUC, as applied to the ROC curve, is a potent metric for evaluating and contrasting ML models concerning PC prediction. Its utility lies in model refinement, improved diagnostic precision, and the potential to impact clinical decisions.

3.8.5 Precision

Precision is a performance metric that quantifies the accuracy of positive predictions made by a classification model. It measures the proportion of true positive

predictions (TP) out of the total positive predictions (TP + FP). Precision helps assess the model's ability to avoid false positives, indicating the precision of its positive predictions [96]. The formula for precision is as follows:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}$$

A higher precision value indicates a lower rate of false positives, in summary, precision plays a crucial role in ML prediction by helping to reduce false positives and their associated implications. By achieving a high precision value, the model increases its reliability in correctly identifying positive cases, contributing to better patient outcomes and healthcare decision-making.Moreover, in clinical diagnoses, it is much tolerable to commit Type-I errors as opposed to Type-II errors. False Negative results are far more fatal compared to False Positives.

3.8.6 F1 score

The F1 score is the harmonic mean of precision and recall, providing a balanced measure of the two [97].

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1 score ranges from 0 to 1, with a higher value indicating better performance. It reaches its maximum value when precision and recall are both 1, representing a perfect balance between the two metrics. The F1 score can be interpreted as a single metric that considers both the ability of the model to identify positive instances (recall) and the accuracy of its positive predictions (precision).

3.9 Ethical considerations

The study was submitted for approval to the University of the Witwatersrand Human Research Ethics Committee (HREC) prior to the analysis of any secondary data. The study was approved with certificate number: M231121 M240415-C-0001. The primary study was approved by the Wits HREC; certificate number: M220673.

3.10 Summary

This chapter outlined the research methodology, data sources, and analytical approach used in the study. It described the use of secondary data from the MADCaP consortium, collected at the CHBAH between 2017 and 2021. The dataset, consisting of 1,096 observations and 21 features, was preprocessed to handle missing data using KNN (KNN) imputation for numerical variables and mode imputation for categorical ones. Class imbalance in the target variable, the Gleason score, was addressed with the ROSE technique. Feature selection and engineering were performed, including One-Hot Encoding for categorical variables and Min-Max scaling for continuous ones, to ensure the dataset was suitable for ML modeling. The research employed the Python programming language within the Google Colaboratory environment, utilizing libraries such as PANDAS, SKLEARN, and SEABORN. The chapter also provided a detailed overview of the computational environment and the variables used in the study, ensuring transparency and reproducibility in the analysis.

Chapter 4

Results

In this chapter, we present the results of the PC classification models using data from the MADCaP study in South Africa. We start by describing the study population, which included newly diagnosed PC patients treated at CHBAH and matched non-cancer individuals from St John's eye hospital. This is followed by findings from data preprocessing and, lastly, prediction results according to the classification strategies outlined in Section 3.7. We also briefly report on interesting patterns identified during the analysis.

4.1 Descriptive statistics

4.1.1 PC analytical data

This diagram below provides an overview of the descriptive statistics for the PC dataset. It begins with an examination of the distribution of risk categories before and after data preprocessing.

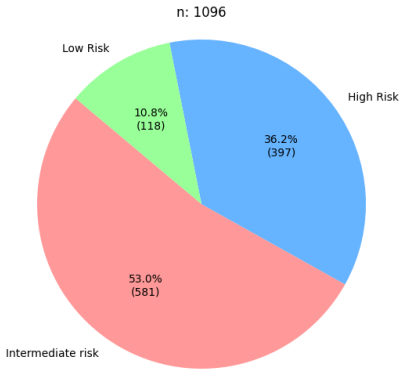


Figure 4.1: Before processing

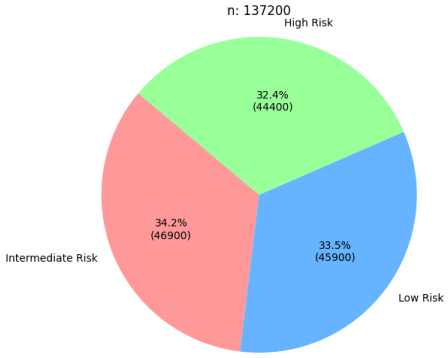


Figure 4.2: After processing

As illustrated in Figure 4.1 before preprocessing, the dataset had a class distribution with the following breakdown: **Intermediate Risk** cases constituted 53% (n=581), **Low Risk** cases 10.8% (n=118), and **High Risk** cases 36.2% (n=397), resulting in a **Low-Intermediate-High** ratio of 11:53:36—indicating a negative imbalance towards the Low risk category. After preprocessing as show in Figure 4.2 the distribution shifted to: **Intermediate Risk** cases 34.2% (n=46,900), **Low Risk** cases 33.5% (n=45,900), and **High Risk** cases 32.4% (n=44,400), with a **Low-Intermediate-High** ratio of 33:34:32—indicating a positive balance among the three risk categories.

After preprocessing and transforming the raw data, the frequencies of classes for each categorical variable against the target variable were calculated as shown in Table 4.1.

Table 4.1: Distribution of categorical variables across prostate cancer risk categories

Variable	Category	Low Risk	Intermediate Risk	High Risk	p-value
Ethnicity	Isizulu	130 (34.85%)	197 (52.82%)	46 (12.33%)	0.0424
	Sesotho	88 (40.0%)	112 (50.91%)	20 (9.09%)	
	Setswana	68 (35.79%)	106 (55.79%)	16 (8.42%)	
	Tsonga	20 (31.75%)	39 (61.9%)	4 (6.35%)	
	Sepedi	21 (32.81%)	31 (48.44%)	12 (18.75%)	
	Seswati	9 (25.71%)	21 (60.0%)	5 (14.29%)	
	IsiXhosa	39 (46.99%)	40 (48.19%)	4 (4.82%)	
	Ndebele	9 (52.94%)	6 (35.29%)	2 (11.76%)	
	Tshivenda	9 (31.03%)	13 (44.83%)	7 (24.14%)	
	Other	4 (18.18%)	16 (72.73%)	2 (9.09%)	
Smoking status	Yes (Now)	111 (36.16%)	159 (51.79%)	37 (12.05%)	0.5537
	Yes (in the past)	171 (39.04%)	223 (50.91%)	44 (10.05%)	
	No never smoked	115 (32.86%)	198 (56.57%)	37 (10.57%)	
Town born	Other towns	287 (37.52%)	402 (52.55%)	76 (9.93%)	0.248
	Soweto	82 (34.45%)	125 (52.52%)	31 (13.03%)	
	Alexandra	13 (25.0%)	30 (57.69%)	9 (17.31%)	
	Sophiatown	15 (36.59%)	24 (58.54%)	2 (4.88%)	
Country born	Other countries	33 (36.26%)	53 (58.24%)	5 (5.49%)	0.215
	South Africa	364 (36.22%)	528 (52.54%)	113 (11.24%)	
If Remember Smoking Start Date	Yes	281 (37.77%)	382 (51.34%)	81 (10.89%)	0.3165
	Non	115 (32.76%)	199 (56.7%)	37 (10.54%)	
	No	1 (100.0%)	0 (0.0%)	0 (0.0%)	
Cigarette type smoked	Filtered	274 (38.48%)	362 (50.84%)	76 (10.67%)	0.2775
	Non	117 (32.32%)	205 (56.63%)	40 (11.05%)	
	Unfiltered (e.g rolled)	6 (27.27%)	14 (63.64%)	2 (9.09%)	
Education Status	Senior secondary schooling...	34 (36.17%)	53 (56.38%)	7 (7.45%)	0.7897
	5-12 years of schooling	167 (37.19%)	233 (51.89%)	49 (10.91%)	
	Some secondary schooling	91 (35.0%)	141 (54.23%)	28 (10.77%)	
	No Formal Education	34 (41.46%)	39 (47.56%)	9 (10.98%)	
	0-4 years of schooling	59 (38.31%)	80 (51.95%)	15 (9.74%)	
	College graduate...	5 (19.23%)	17 (65.38%)	4 (15.38%)	
	Post-high school training...	3 (25.0%)	7 (58.33%)	2 (16.67%)	
	Some college...	4 (22.22%)	10 (55.56%)	4 (22.22%)	
Income (per month)	Up to 1,850 per month	0 (0.0%)	1 (100.0%)	0 (0.0%)	
	5,001 - 10,000 per month	318 (37.81%)	436 (51.84%)	87 (10.34%)	0.7477
	1,851 - 5,000 per month	10 (31.25%)	20 (62.5%)	2 (6.25%)	
	Chose not to answer	41 (30.37%)	76 (56.3%)	18 (13.33%)	
	10,001 - 15,000 per month	21 (29.58%)	41 (57.75%)	9 (12.68%)	
	15,001 - 20,000 per month	6 (50.0%)	4 (33.33%)	2 (16.67%)	
	30,001 - 50,000 per month	1 (33.33%)	2 (66.67%)	0 (0.0%)	
PSA at referral	Abnormal/suspicious	394 (36.41%)	573 (52.96%)	115 (10.63%)	0.1382
	Normal	1 (9.09%)	7 (63.64%)	3 (27.27%)	

Table 4.1 shows how different categorical variables are distributed among the three prostate cancer risk categories: low, intermediate, and high. The table also includes chi-square test results to check if these variables are related to prostate cancer risk.

Ethnicity was the only variable that showed a significant relationship with risk category (χ^2, p = 0.0424). The majority of high-risk patients were from the Isizulu and Sepedi ethnic groups, while other groups had more balanced distributions across risk categories. This suggests that genetic or lifestyle factors linked to ethnicity could influence prostate cancer severity.

Smoking status did not show a strong relationship with prostate cancer risk (χ^2, p = 0.5537). The percentage of current, past, and never-smokers was similar across all risk groups, indicating that smoking may not be a key factor in determining risk.

Birthplace and country of birth also did not show a clear link to risk categories (χ^2, p = 0.248 and p = 0.215, respectively). While some differences were observed, they were not statistically meaningful. Similarly, education level and income did not show significant differences across risk groups (χ^2, p = 0.7897 and p = 0.7477, respectively). This suggests that social and economic factors may not play a major role in determining prostate cancer severity within this study population.

PSA levels at referral were categorized as normal or abnormal/suspicious. Although most patients in all risk categories had abnormal PSA values at referral, there was no strong statistical relationship between PSA status at referral and risk group (χ^2, p = 0.1382). This means that while high PSA levels are common in prostate cancer patients, they may not be enough to differentiate between low, intermediate, and high-risk cases.

In summary, ethnicity was the only categorical variable significantly linked to prostate cancer risk in this study. Other factors, such as smoking, birthplace,

education, income, and PSA status at referral, did not show strong associations. These findings suggest that while some lifestyle and demographic factors may influence the development of prostate cancer, they may not be useful in predicting how severe the disease will be.

A similar table of numeric variables was also created to explore the relationship between the numeric variables and the outcome variable, yielding the following results in Table 4.2.

Table 4.2: Summary statistics for risk factors across PC risk categories

Risk Category	Variable	Count	Mean	Median	Standard Deviation
High risk	hrf_cigarettescount	397	8.34	6.00	7.18
	psa_psareferrallevel	397	580.36	91.34	1542.21
	psa_psadiagnlevel	397	374.55	50.19	888.75
	age_years	397	68.21	69.00	7.88
Intermediate risk	hrf_cigarettescount	581	7.90	6.00	6.01
	psa_psareferrallevel	581	93.26	16.50	493.16
	psa_psadiagnlevel	581	73.80	17.17	368.01
	age_years	581	66.69	67.00	7.83
Low risk	hrf_cigarettescount	118	7.94	6.00	7.79
	psa_psareferrallevel	118	18.82	9.52	26.74
	psa_psadiagnlevel	118	15.10	8.56	19.15
	age_years	118	65.05	66.00	7.80

Table 4.2 presents the summary statistics for selected numerical variables across prostate cancer (PC) risk categories: high, intermediate, and low risk. The variables include cigarette consumption (hrf_cigarettescount), prostate-specific antigen (PSA) levels at referral (psa_psareferrallevel) and diagnosis (psa_psadiagnlevel),

and patient age (age_years). The table provides the count, mean, median, and standard deviation for each variable, offering insights into their distribution across different risk groups.

The mean number of cigarettes smoked per day is similar across the three risk categories, ranging between 7.90 and 8.34 cigarettes. The median value remains consistently at 6.00 cigarettes per day, while the standard deviation indicates some variability, particularly in the high-risk (7.18) and low-risk (7.79) groups. These findings suggest that cigarette consumption does not vary significantly between risk groups and may not be a strong predictor of prostate cancer severity.

PSA levels at referral show substantial differences across risk categories. The high-risk group has an extremely elevated mean PSA referral level (580.36 ng/mL), with a median (91.34 ng/mL) and a very high standard deviation (1542.21), indicating significant variability. The intermediate-risk group exhibits considerably lower PSA referral levels, with a mean (93.26 ng/mL) and a median (16.50 ng/mL). The low-risk group has the lowest PSA levels, with a mean (18.82 ng/mL) and a median (9.52 ng/mL). A similar trend is observed for PSA levels at diagnosis, where the high-risk group has a mean (374.55 ng/mL), compared to (73.80 ng/mL) in the intermediate-risk group and (15.10 ng/mL) in the low-risk group. The substantial differences in PSA levels across risk categories suggest that PSA remains a crucial marker for prostate cancer severity. The large standard deviations, particularly in the high-risk group, indicate significant heterogeneity within this category

Age also follows a decreasing trend across risk categories. The mean age of high-risk patients is 68.21 years, which is higher than that of intermediate-risk (66.69 years) and low-risk patients (65.05 years). The median values follow a similar pattern, with high-risk patients having a median age of 69 years, compared to 67 years in the intermediate-risk group and 66 years in the low-risk group.

The standard deviation for age remains relatively consistent across all risk groups, suggesting a similar degree of variability within each category. These findings indicate that older individuals are more likely to be classified in the high-risk group, highlighting the role of age as a factor in prostate cancer progression.

Overall, the descriptive statistics highlight clear patterns in PSA levels and age across risk categories, reinforcing the role of these variables in prostate cancer risk stratification. Cigarette consumption, however, does not appear to exhibit a meaningful difference between risk groups. These findings provide a foundation for further inferential analysis to explore the predictive power of these variables in prostate cancer classification.

4.2 Risk factors and disparities

The relationship between smoking and the aggressiveness of prostate cancer (PC) was examined, considering its potential impact on disease severity. As shown in Table 4.1, smoking status does not exhibit a strong differentiation across risk categories. The proportion of current smokers, former smokers, and never smokers remains relatively similar within each risk group. The Chi-square test yielded a P-value of 0.5537, suggesting no statistically significant association between smoking status and PC risk classification in this dataset. To further explore these findings, a bar chart was generated, as shown in Figure 4.3.

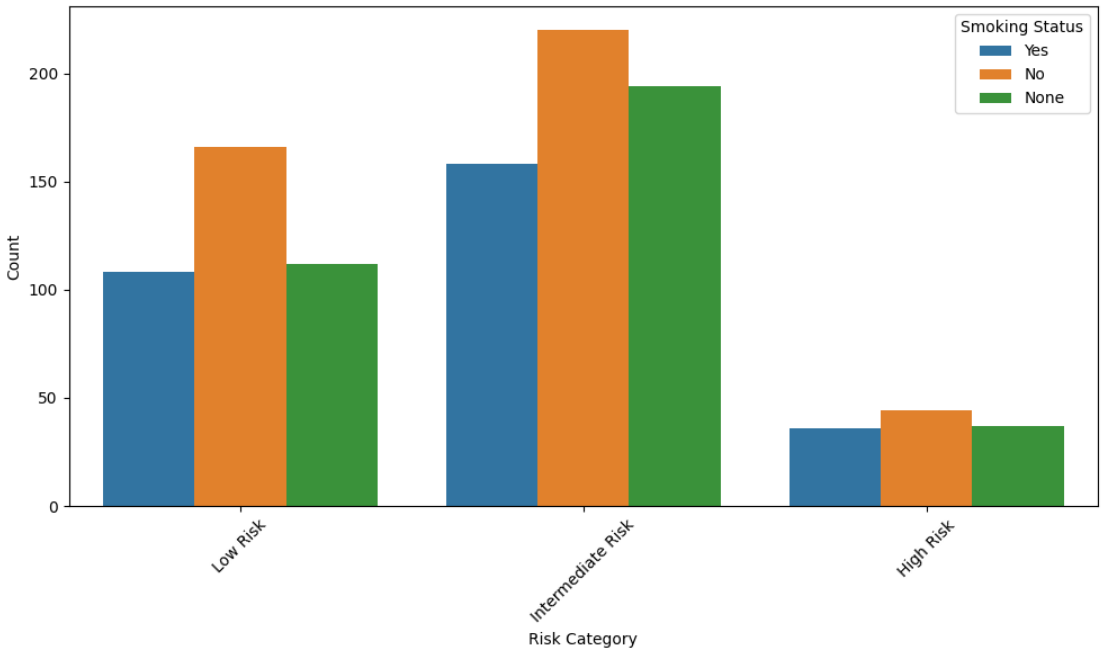


Figure 4.3: PC risk categories by smoking status

The bar chart indicates that the distribution of smoking status across prostate cancer (PC) risk categories does not follow a strong increasing or decreasing pattern. The number of individuals who never smoked is highest in the intermediate-risk group, followed by the low-risk category, and lowest in the high-risk group. A similar trend is observed among past smokers, with the highest counts in the intermediate-risk group and the lowest in the high-risk category. Current smokers also follow this pattern, though their numbers are slightly lower than those who never smoked or smoked in the past.

Unlike an expected trend where smoking would be strongly associated with higher PC risk, this distribution suggests that smoking status is relatively evenly spread across risk levels. The absence of a clear increasing pattern in smoking prevalence across risk categories suggests that smoking may not be a primary predictor of PC aggressiveness in this dataset. While smoking remains a known risk factor for various cancers, these findings indicate that additional factors might

play a more significant role in determining PC severity.

The difference in PSA levels between referral and diagnosis stages was investigated to understand changes in PC assessments and their implications for patient management. A paired t-test was employed to compare the mean PSA levels at the "psa_psareferrallevel" and "psa_psadiagnlevel" stages. The analysis revealed a t-statistic of 3.20, with a p-value of 0.0014.

Since the p-value is well below the standard significance threshold of 0.05, we rejected the null hypothesis, confirming a statistically significant difference between PSA levels at referral and diagnosis. This finding suggests that PSA measurements change meaningfully between these two stages, potentially indicating disease progression or the impact of early medical interventions.

The observed difference in PSA levels could have important implications for clinical decision-making. It highlights the necessity of closely monitoring PSA changes over time to refine patient management strategies, guide treatment decisions, and improve the timing of interventions. Additionally, understanding these variations may provide insights into prostate cancer progression, particularly in identifying cases where the disease becomes more aggressive. Figure 4.4 below visualizes a box plot and a paired difference plot.

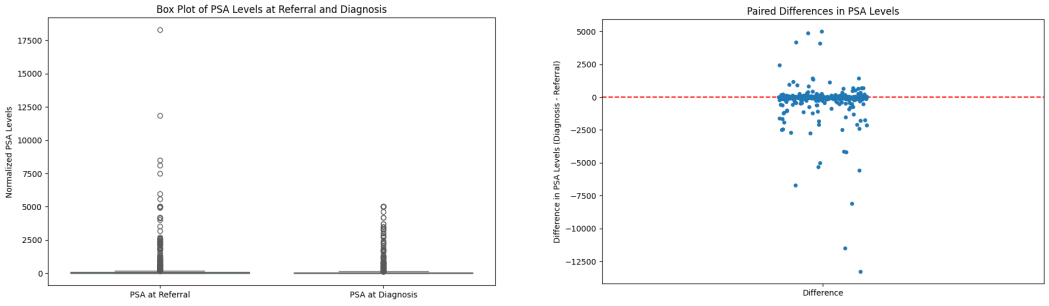


Figure 4.4: Comparison and change in PSA levels from referral to diagnosis

The box plot was employed to compare the distributions of normalized PSA levels at referral and diagnosis. This visualization displays the interquartile range

(IQR), median, and range of the data for each time point, allowing for a visual assessment of potential shifts in PSA level distribution. Outliers, if present, are plotted as individual points beyond the whiskers. To further examine the individual changes in PSA levels, a paired difference plot was constructed. This plot illustrates the difference in normalized PSA levels between diagnosis and referral for each participant. Individual data points are represented as a scatter plot, with a horizontal reference line at y = 0 indicating no change. Points above this line signify an increase in PSA levels from referral to diagnosis, while points below the line indicate a decrease. This visualization provides insights into the magnitude and direction of PSA level changes for each individual.

A noticeable shift in the median or IQR between the two time points in the box plot, coupled with a predominance of data points above or below the zero line in the paired difference plot, suggests a potential difference. The paired t-test provides a quantitative measure (p-value) to assess the statistical significance of this observed difference, helping to differentiate between true changes and random fluctuations. In conclusion, the box plot and paired difference plot offer complementary visual representations of PSA level changes between referral and diagnosis. These visualizations, alongside statistical testing, contribute to a comprehensive understanding of the dynamics of PSA levels in the context of prostate cancer diagnosis.

The heatmap in Figure 4.5 offers a visual representation of the association between educational status and prostate cancer (PC) risk categories. The data reveals distinct patterns, suggesting a potential link between educational attainment and PC risk levels. Individuals with 'No Formal Education' exhibit the highest frequencies across all risk categories, particularly in the 'Intermediate Risk' category. This observation suggests a heightened likelihood of individuals with limited education being classified as having a moderate risk of PC. This pattern may be attributed to underlying socioeconomic and environmental factors commonly as-

sociated with lower educational attainment, such as reduced access to healthcare, lower health literacy, and potentially less healthy lifestyle choices. Moreover, delayed medical intervention or inadequate cancer screening could contribute to this group's higher likelihood of being classified at intermediate risk. Conversely, higher education levels, such as 'Senior secondary schooling (completed high school or equivalent)' and 'College graduate (Bachelors degree, Higher National Diploma)' seem to be associated with lower frequencies in the 'Intermediate Risk' and 'High Risk' categories. This trend suggests a potential protective effect of higher education against PC, possibly due to greater health awareness, access to resources, and proactive healthcare engagement.

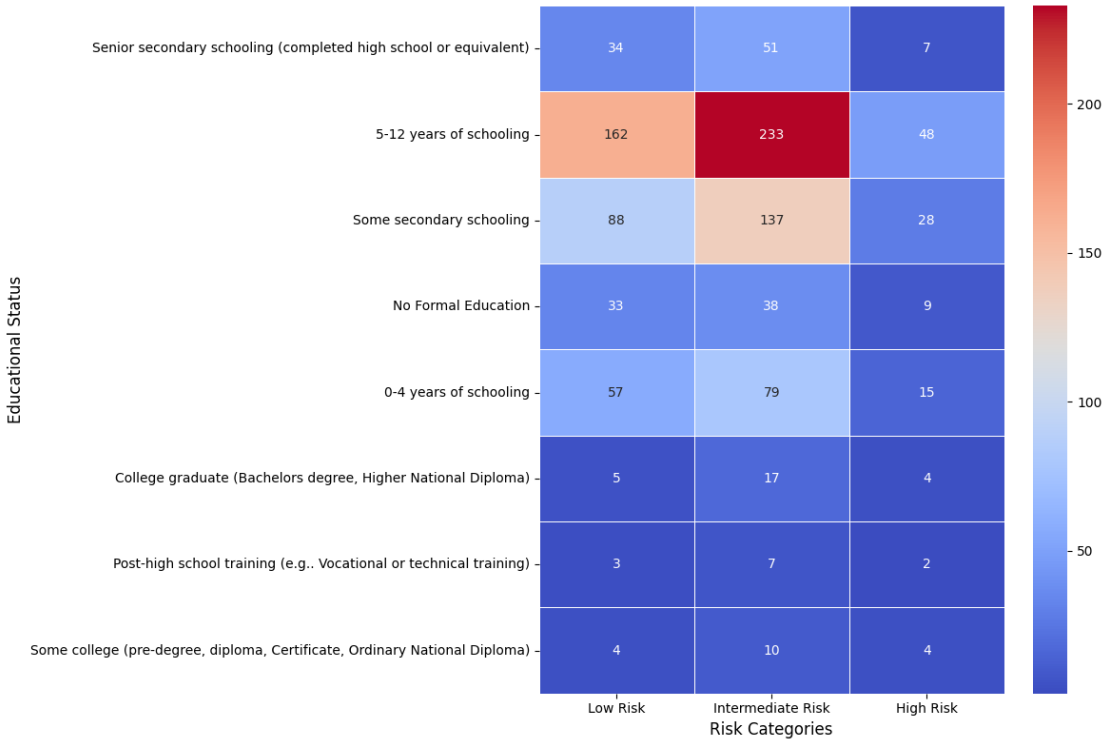


Figure 4.5: Heatmap of risk category by education codes

Figure 4.6 is a line graph used to explore the relationship between the number of cigarettes smoked per day (HRF Cigarettes Count) and the severity of prostate

cancer (PC), measured by the MR Gleason TURP Total score. The graph shows that this relationship is not simple. As smoking increases, the Gleason score goes up and down instead of following a clear trend. This means smoking does not always make prostate cancer worse in a predictable way.

Sometimes, smoking more cigarettes does not seem to have much effect, while in other cases, it might increase the risk significantly. This suggests that at certain points, smoking more may suddenly become more dangerous for prostate cancer. To fully understand this, we also need to consider other factors like how long someone has been smoking, when they started, and their genetics.

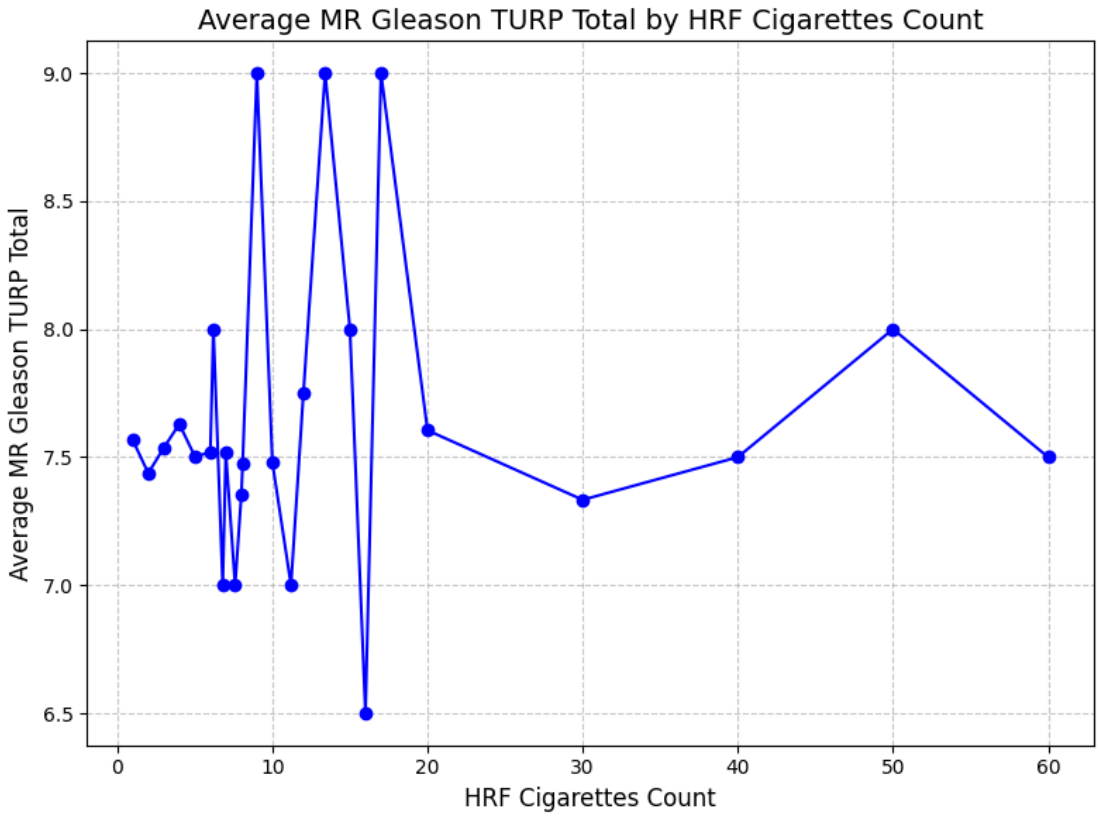


Figure 4.6: Average gleason total score by cigarette consumption (HRF cigarettes count)

Further examination revealed that an individual smoking **60 cigarettes per**

day represents an extreme case, as this value is significantly above the typical smoking range observed in the dataset. Statistical analysis confirmed this as an *outlier*, with the majority of HRF Cigarettes Count values falling below the upper bound of 14.25 cigarettes per day (calculated using the interquartile range method). The identification of such outliers is crucial, as they can disproportionately influence trends and interpretations. While extreme smoking behaviors like this are rare, they warrant additional scrutiny to assess their potential impact on PC severity. This finding suggests that outliers could provide unique insights into the health effects of heavy smoking and should be carefully considered in the analysis and discussion of results.

The significance of this outlier lies in its demonstration that while smoking may not appear to be a substantial factor for the majority of individuals within the study's population, exceptionally heavy smoking could still pose a considerable risk. The absence of a statistically significant correlation for the general dataset does not preclude the potential for harm in extreme instances. This can be likened to the observation that, crossing the street is usually okay, but running across blindfolded is risky. Therefore, it is essential to consider both broad statistical trends and individual extreme cases to develop a comprehensive understanding of smoking's potential influence on prostate cancer. This is particularly crucial for complex diseases like prostate cancer, where multiple contributing factors are involved. Consequently, the interpretation of results must incorporate both statistical and clinical significance.

Figure 4.7 presents a visual analysis of the age distribution across the three prostate cancer (PC) risk categories: Low, Intermediate, and High, providing compelling evidence for the influence of age on risk stratification. The box plot unequivocally demonstrates a positive correlation between increasing age and the

likelihood of belonging to a higher risk category. This observation is strongly supported by the position of the median lines within each box, which progressively shift towards older ages as the risk category escalates from Low to Intermediate to High. While there is some degree of age overlap between the categories, the overall pattern reveals a distinct trend: younger individuals are more frequently associated with the Low Risk category, characterized by a lower median age and reduced variability, as evidenced by the shorter and narrower box. Conversely, the High Risk category predominantly comprises older individuals, displaying a significantly higher median age and greater variability, reflected in the taller and wider box.

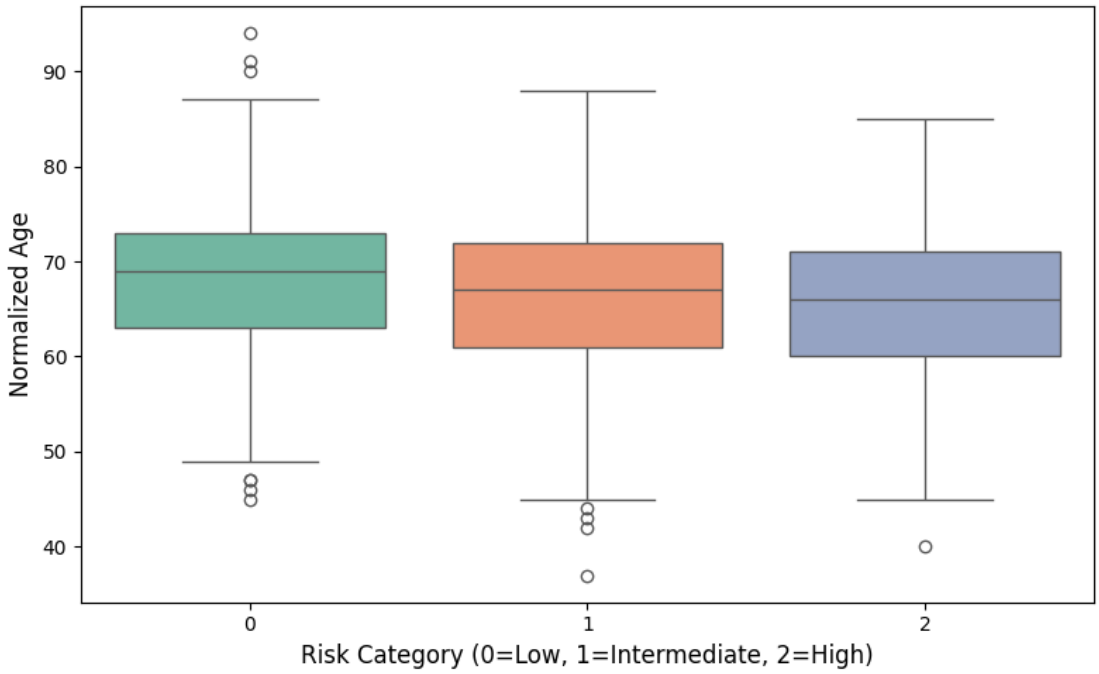


Figure 4.7: Age distribution by PC risk category

This difference in age distribution is further corroborated by statistical analysis using the Kruskal-Wallis test, which confirmed significant differences in age across the risk categories (H = 2631.90, p less than 0.001). These findings underscore

the critical role of age as a key factor in PC risk assessment, highlighting its potential utility in guiding personalized screening strategies, risk prediction models, and preventive interventions tailored to specific age groups. Notably, the wider age range observed in the High Risk category suggests a greater heterogeneity in the age-related factors contributing to PC development in this group, warranting further investigation into the underlying mechanisms.

The stacked bar chart in Figure 4.8 illustrates the distribution of prostate cancer (PC) risk categories across various income groups. While *high-risk* cases were the most frequently observed across all income brackets, followed by *intermediate* and then *low-risk* cases, there are notable variations within specific income segments. The *'Up to 1,850 per month'* and *'1,851 - 5,000 per month'* income categories demonstrated a relatively more balanced distribution across risk categories compared to other income groups, although *high-risk* cases still formed the majority. Conversely, individuals earning *'30,001 - 50,000 per month'* exhibited a higher proportion of *high-risk* classifications. This finding could indicate potential links between higher income, lifestyle factors associated with higher socioeconomic status, and increased risk of aggressive prostate cancer, warranting further investigation.

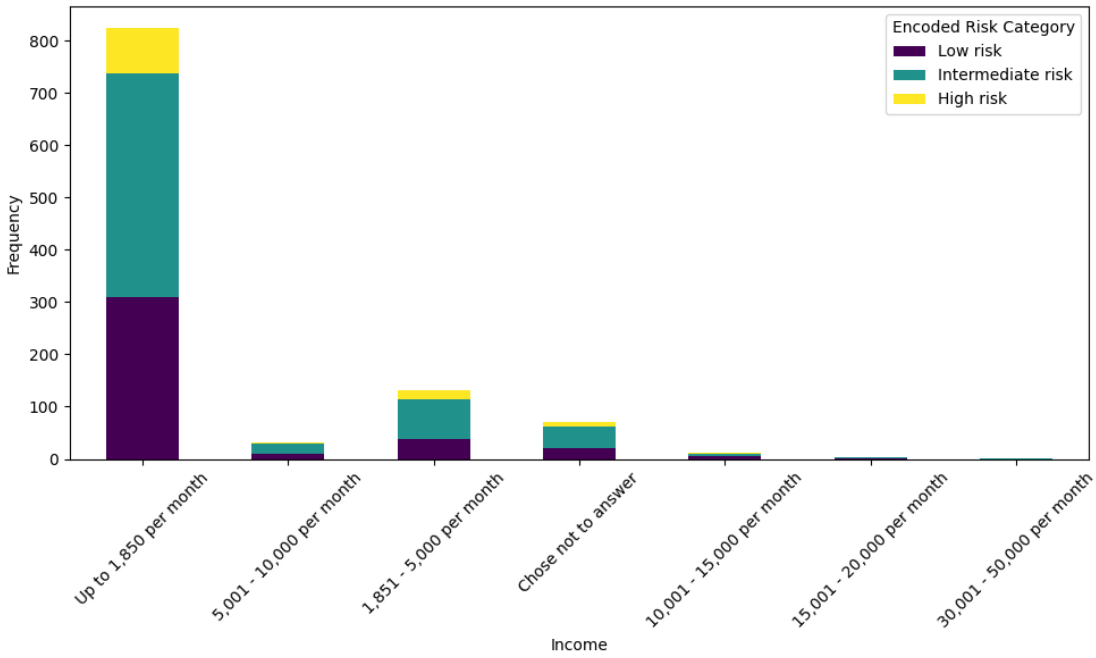


Figure 4.8: Stacked bar graph showing the distribution of PC risk categories across income groups.

While these patterns are apparent visually, statistical analysis using a Chi-Square test (Table 4.1) did not reveal a statistically significant association between income and PC risk categorization ($p = 0.6847$). This suggests that income level, within the categories examined, may not be a strong independent predictor of risk stratification in this cohort. Despite the lack of statistical significance, the observed variations in risk distribution across income groups warrant further investigation, potentially considering other socioeconomic factors and lifestyle behaviors that could influence prostate cancer risk. These results emphasize the importance of a comprehensive approach to risk assessment, taking into account both clinical parameters and socioeconomic context, to develop targeted prevention and treatment strategies.

In conclusion, the descriptive analysis highlights key patterns in prostate cancer risk classification. PSA levels at referral and diagnosis are significantly higher in

the high-risk group, confirming their importance as biomarkers. Age also shows a clear trend, with older individuals more likely to be classified as high risk. In contrast, cigarette consumption remains relatively consistent across risk categories, suggesting it may not be a strong differentiator in this dataset. These findings provide a foundation for further analysis to explore the predictive relationships between these variables and prostate cancer risk classification.

4.3 Predicting PC progressiveness and cancer severity

This was the stage where data mining was performed. A number of classifier ML models were fitted and learned with the preprocessed, cleaned, and transformed data. As it has already been outlined in methodology, the models were namely: GNB, KNN, DT, RF, and Support Vector Classifier. The 137,200 rows of data with 15 variables were split into training, validation, and testing sets for the ML models. The target variable was first detached from the features so that they would be split separately. Subsequent to that, both the features and target variable data frames were split into training and testing sets in proportions of 0.9 and 0.1 respectively. The testing data, with a pair of features and the dependent three-class variable that was observed here, was the final data that was used to examine the true performance of the models at the deployment level.

The target variable was PC risk category. It was a three-class variable with classes "Low risk", "Intermediate risk", and "High risk" encoded as 0, 1, and 2 respectively. The features were a combination of categorical and numerical data types namely: "`d_ethn_self_codes`", "`hrf_smoking_codes`", "`town_code`", "`country_code`", "`hrf_smokestartremember_codes`",

"hrf_cigarettestype_codes", "s_education_codes", "s_income_codes", "psa_psareferralleveldesc_codes", "hrf_cigarettescount_normalized", "hrf_currentlysmoke_codes", "psa_psareferrallevel_normalized", "psa_psadiagnlevel_normalized", and "age_years_normalized". The 0.9 proportion of the training set was then subsequently split into a final training set and a validation set. The proportions were 0.75 and 0.25 out of the 0.9 for the former and latter respectively. The training set was used to train and learn the models while the validation set was used to assess the performance of the models for adjustments and modifications towards accuracy and better learning. All this splitting was doable because the data was big enough, such that K-Fold validation algorithm (algorithm used as a tool to evaluate ML models, by training a number of models on different subsets of the input data whenever number of subjects are limited) was uncalled for.

Using these three sets of data were used to model the classifiers and the results of the modeling were summarized as in the table 4.3.

Table 4.3 summarizes the performance of six ML models GNB, KNN, DT, RF,SVM, and GB—in predicting PC using the MADCaP dataset. Each model was evaluated using accuracy, precision, recall, and F1 scores on the test dataset, with Grid Search optimization applied to refine model parameters.

The GNB model demonstrated limited predictive power, with an accuracy of 42% and low precision, recall, and F1 scores, indicating that this model struggled to capture relevant patterns for effective classification. In contrast, KNN exhibited strong performance, achieving a test accuracy of 84%, along with high scores across other metrics, suggesting robustness in capturing data features. This makes KNN a reliable choice for predicting PC risk categories in this study.

The DT model achieved a balanced performance with a test accuracy of 83%, showing consistency in precision, recall, and F1 scores, which suggests that the

model generalizes well without overfitting. The RF model attained perfect classification with test accuracy, precision, recall, and F1 scores all reaching 100%. However, these results may indicate overfitting, as the model may have learned data-specific patterns that might not generalize well; hence, further validation is necessary. The SVM model demonstrated moderate performance, achieving a test accuracy of 66%, with consistent but modest precision, recall, and F1 scores, suggesting that further hyperparameter tuning could improve its predictive performance.

The GB model exhibited solid performance, achieving a test accuracy of 79%, along with consistent precision, recall, and F1 scores. This result highlights the model's ability to effectively capture complex patterns in the dataset while avoiding the overfitting concerns seen in RF.

Overall, while GB and DT demonstrated competitive results, KNN remains a preferred choice in this analysis due to its robust performance and reliability across all metrics.

Table 4.3: Test performance metrics for different ML models on PC prediction (in percentages)

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
GNB	42	44	42	40
KNN	84	84	84	84
DT	83	84	83	83
RF	100	100	100	100
SVM	66	64	64	64
GB	79	79	79	79

4.4 Model performance

To gain a comprehensive understanding of the two best-performing models, KNN and DT, their confusion matrices and ROC curves were analyzed for both training and test datasets. These evaluations highlighted not only the accuracy of the models but also their ability to generalize to unseen data. Consolidating the results in this section facilitates a streamlined comparison to identify the optimal model for PC risk prediction.

4.4.1 Confusion matrices

The performance of various ML models in predicting PC risk categories—Low Risk, Intermediate Risk, and High Risk—was rigorously assessed using confusion matrices, as shown in Figure 4.9. These matrices provided valuable insights into each model's classification accuracy, highlighting patterns of correct and incorrect predictions across both training and test datasets.

The DT classifier displayed moderate performance in distinguishing between risk categories. On the training dataset, the confusion matrix revealed a notable number of misclassifications. Low Risk cases were frequently predicted as Intermediate Risk, while some High Risk cases were also misclassified as Intermediate Risk. These patterns suggest that the DT, despite its ability to capture patterns in the data, struggled with cases lying near the boundaries of the risk groups. On the test dataset, the model's performance declined further, with increased misclassifications across all categories. This drop in accuracy indicates a tendency towards overfitting, where the model fits the training data too closely and fails to generalize effectively to new, unseen data. The frequent misclassification of Low Risk as Intermediate Risk and High Risk as Intermediate Risk further underscores the model's limitations in handling borderline cases. Overall, the DT's simplicity

made it less effective in capturing complex relationships within the data.

The RF classifier exhibited superior performance compared to the DT. On the training dataset, it achieved near-perfect classification accuracy, with minimal misclassifications. This impressive result highlights the strength of ensemble methods in reducing bias and capturing complex patterns. However, the near-perfect accuracy on the training data raised concerns about potential overfitting. On the test dataset, the RF demonstrated a balanced performance, with fewer and more evenly distributed misclassifications. While some Intermediate Risk cases were misclassified as Low Risk or High Risk, the model's generalization ability was evident. The balanced trade-off between training accuracy and test performance positions the RF as the most reliable model for PC risk prediction. Its robustness makes it particularly suitable for applications requiring high predictive accuracy and generalization.

The GB classifier showed strong performance on the training dataset, albeit with slightly more misclassifications compared to the RF. GB's iterative approach, which focuses on correcting errors over multiple iterations, contributed to its robust yet slightly lower training accuracy. Misclassifications were observed across all risk categories, indicating a balanced trade-off between overfitting and underfitting. On the test dataset, GB maintained competitive performance, with slightly higher misclassification rates compared to RF. Common errors included Low Risk cases being predicted as Intermediate Risk and High Risk cases classified as Intermediate Risk. These findings suggest that GB is a robust model but could benefit from further hyperparameter tuning to enhance its predictive performance.

The Support Vector Classifier demonstrated reasonable performance, with confusion matrices indicating good accuracy on both training and test datasets. Correct predictions were concentrated along the diagonal, but off-diagonal values highlighted misclassifications, particularly between Intermediate Risk and High

Risk categories. These misclassifications suggest a need for additional refinement through hyperparameter tuning, such as adjusting the kernel function or regularization parameters, to improve the model's accuracy and minimize errors.

The KNN model exhibited a balanced and robust performance. On the training dataset, the confusion matrix showed precise classification across all risk categories, with negligible misclassifications. This high accuracy carried over to the test dataset, where the model maintained strong generalizability. KNN's ability to correctly classify High Risk cases, which are often challenging due to overlapping features with Intermediate Risk cases, was particularly noteworthy. Its strong sensitivity (true positive rate) and specificity (true negative rate) make it an excellent choice for clinical applications, where minimizing false positives and false negatives is critical.

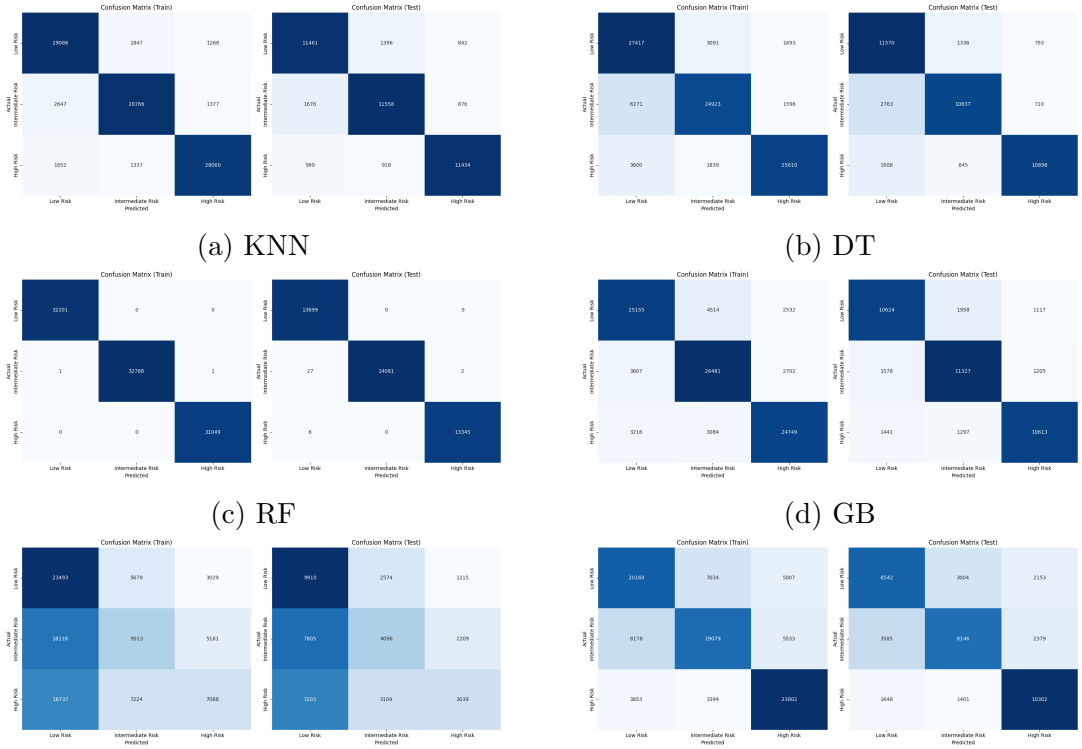
The GNB model delivered reasonable performance, with good accuracy observed in both training and test datasets. The confusion matrices revealed higher diagonal values, indicating correct predictions, but also highlighted misclassifications, particularly between Intermediate Risk and High Risk categories. These errors point to the need for enhanced data preprocessing, such as feature scaling or handling missing values, to improve model performance and reduce misclassifications.

Across all models, the Intermediate Risk category emerged as the most accurately classified, likely due to its higher representation in the dataset. Conversely, Low Risk and High Risk categories experienced more frequent misclassifications, highlighting potential overlaps in their feature distributions. This was evident in the consistent misclassification of Low Risk as Intermediate Risk across models.

The evaluation revealed notable trade-offs between training and test performance, particularly regarding overfitting. The DT classifier showed significant overfitting, with a sharp decline in test performance. In contrast, the RF classi-

fier achieved the best balance between training accuracy and test generalization, making it the most suitable model for PC risk prediction. GB performed competitively but exhibited slightly lower test accuracy, suggesting room for improvement through hyperparameter tuning. Both the SVM and GNB models demonstrated reasonable performance, with potential for enhancement through parameter optimization and data preprocessing. The KNN model attained its robustness and reliability, particularly in its generalization to the test dataset, making it a strong contender for clinical decision-making applications.

In summary, the RF and KNN models emerged as the most effective classifiers, each excelling in different aspects of PC risk prediction. These findings provide a strong foundation for selecting models that balance accuracy, robustness, and clinical applicability.



(a) KNN (b) DT

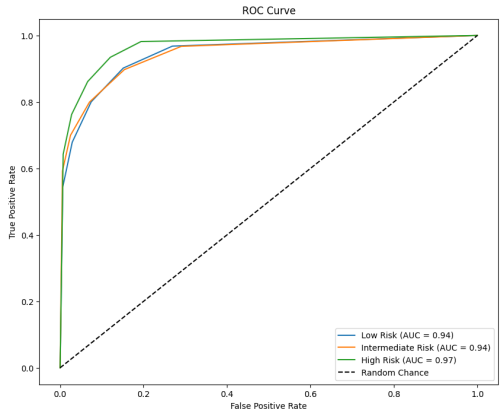
(c) RF (d) GB

(e) GNB (f) SVC

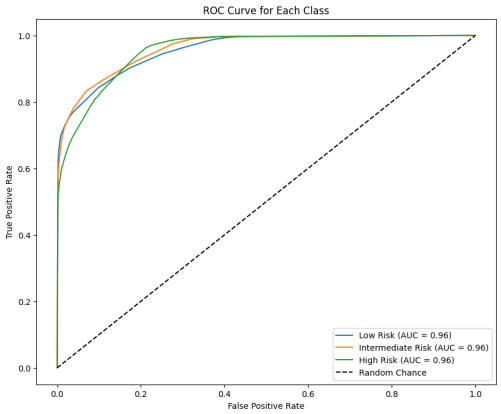
Figure 4.9: Confusion matrices for all models

4.4.2 ROC curves

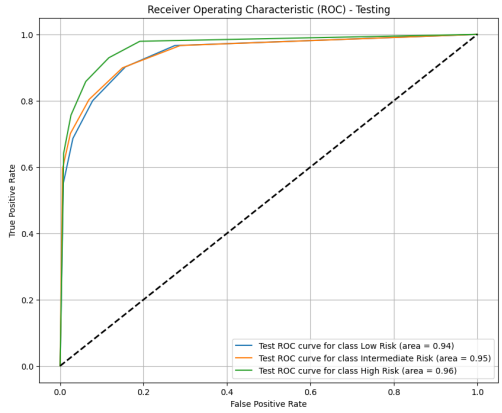
The ROC curves presented in **Figure 4.10** provide valuable insights into the performance of the ML models used to classify PC risk into three categories: Low Risk, Intermediate Risk, and High Risk. The ROC curves plot the True Positive Rate (TPR) against the False Positive Rate (FPR), offering a visual representation of each model's ability to distinguish between positive and negative cases. The area under the curve (AUC) serves as a summary measure of performance, with higher values indicating stronger discriminative power.



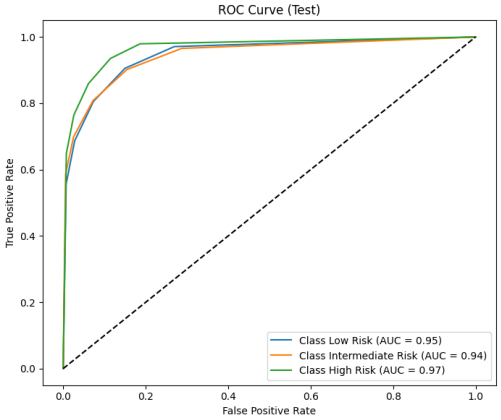
(a) KNN



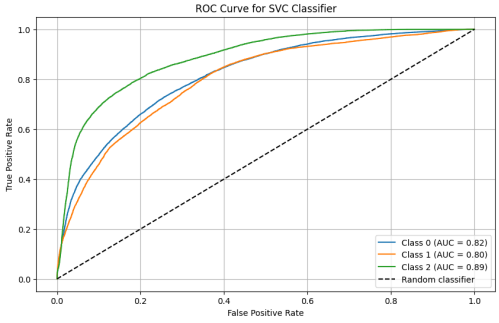
(b) DT



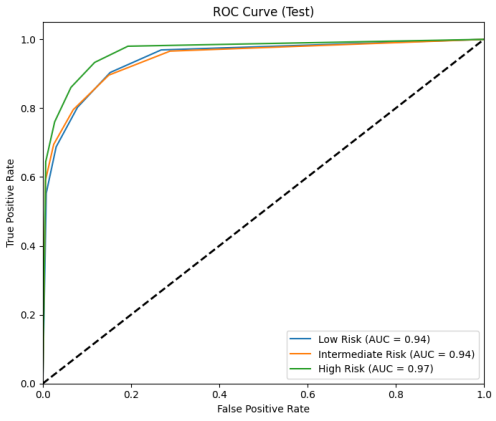
(c) GB



(d) GNB



(e) SVM



(f) RF

Figure 4.10: Area under the receiver operating characteristic curve (AUROC) for all models

The **DT** classifier, shown in **Figure 4.10**(b), demonstrated exceptional performance, with an AUC of 0.96 across all risk categories. This indicates its ability to effectively separate cases into the appropriate risk levels. The proximity of the ROC curve to the top-left corner suggests that the model identified true positives while minimizing false positives. However, despite its high AUC values, the simplicity of the DT model may limit its ability to handle complex patterns in the data, and its potential for overfitting should be considered when generalizing to real-world applications.

The **RF** model, represented in **Figure 4.10**(f), exhibited robust performance, with near-perfect AUC values across all categories. While a slight decline in test performance compared to training data suggested minor overfitting, the model maintained excellent discriminative ability overall. This performance is attributed to the ensemble nature of RF, which combines multiple DTs to create a powerful classifier. Its ability to generalize well to unseen data makes it a reliable choice for PC risk prediction.

The **GB** classifier, shown in **Figure 4.10**(c), also displayed strong classification performance, with AUC values ranging from 0.94 to 0.96. Its iterative learning process, which focuses on correcting errors across multiple iterations, contributed to its robust results. The ROC curve indicated the model's ability to maximize true positive rates while minimizing false positives, particularly for the Intermediate Risk category. Although its performance was slightly lower than RF, GB remains competitive and well-suited for capturing complex relationships within the data.

The **KNN** model, displayed in **Figure 4.10**(a), attained AUC values of 0.95 for Low and Intermediate Risk categories and 0.97 for High Risk. This highlights its exceptional ability to differentiate between risk levels, particularly in accurately identifying High Risk cases. KNN's reliance on distance-based classification makes it a robust and reliable model, especially in clinical settings where accurate identifi-

cation of aggressive cases is critical for treatment decisions. Its strong performance across both training and test datasets underscores its generalizability.

The **GNB** model, shown in **Figure 4.10**(d), achieved impressive AUC values ranging from 0.94 to 0.97. Its probabilistic nature and ability to handle high-dimensional data contributed to its success in PC risk classification. The ROC curve indicated that the model effectively minimized false positives while maintaining high true positive rates. However, its assumption of feature independence may pose challenges in datasets with highly correlated features, a factor worth addressing in future applications.

The **SVM**, represented in **Figure 4.10**(e), delivered reasonable performance, with AUC values between 0.80 and 0.89 across the risk categories. While its AUC values were lower than those of other models, the curve's proximity to the top-left corner still reflected good discrimination ability. This performance can be attributed to the SVM's capacity to find optimal hyperplanes for separating data points. However, there is potential for improvement through hyperparameter tuning or incorporating non-linear kernels to capture more complex patterns.

In comparing the models, the **KNN** and **RF** classifiers emerged as the top performers, demonstrating exceptional discriminative power and generalizability. Their high AUC values across all categories make them particularly suitable for clinical applications where accuracy and reliability are paramount. The **DT** and **GB** models also performed well, with the former offering simplicity and interpretability, and the latter excelling in capturing nuanced relationships in the data. **GNB** provided competitive results, while **SVM** showed room for enhancement through further optimization.

Overall, the ROC curve analysis highlights the strengths and limitations of each model in predicting PC risk categories. The high AUC values across most models underscore their effectiveness in distinguishing between risk levels. The findings

suggest that **KNN** and **RF** are the most promising candidates for deployment in clinical settings, providing a balance of accuracy, robustness, and practical applicability. These results contribute to the ongoing development of ML tools for PC risk stratification and highlight the importance of model selection based on specific application needs.

4.4.3 Comparing Model Performance: Original, Resampled, and Augmented Data

To better understand the impact of data preprocessing techniques on model performance, Tables 4.4 and 4.5 present the test performance metrics for models trained on the original dataset and the resampled dataset (using ROSE sampling), respectively. These results are then compared with the augmented dataset results discussed earlier in this chapter.

Table 4.4: Test performance metrics using original data.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
Gaussian NB	30	51	44	30
SVC	33	41	34	20
Gradient Boosting	63	56	48	48
KNeighbors	60	56	45	44
Decision Tree	49	44	47	44
Random Forest	60	44	44	43

Table 4.5: Test performance metrics after resampling using ROSE.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
Random Forest	55	56	56	55
Decision Tree	48	48	48	48
KNeighbors	51	50	52	51
Gradient Boosting	48	49	49	48
Gaussian NB	41	53	45	37
SVM (SVC)	40	42	43	44

From these results, several key patterns emerge. First, the original dataset (Table 4.4) produced generally low scores across all evaluation metrics, with no model exceeding 63% accuracy. This was largely due to the strong class imbalance and underrepresentation of certain risk groups, which limited the models' ability to correctly classify minority cases. For example, Gaussian Naïve Bayes achieved only 30% accuracy and an F1 score of 30%, indicating particularly poor predictive capability.

Second, the resampled dataset (Table 4.5) improved some metrics for certain models, particularly recall and F1-score, by balancing the number of examples per class. However, this method had limitations: since resampling works by duplicating and interpolating existing points, it increased the risk of overfitting, especially for complex models like Random Forest. This was evident in the drop from perfect training performance to test accuracies around 55–56%.

In contrast, the augmented dataset, discussed in table 4.3, outperformed both the original and resampled datasets across nearly all metrics. Augmentation went beyond simple balancing—it generated synthetic, varied data points by introducing controlled noise, which provided the models with new and diverse learning examples. As a result, models trained on augmented data, such as Random Forest

and Gradient Boosting, not only maintained high training performance but also achieved near-perfect generalisation to the test set, with accuracy, precision, recall, and F1-scores close to 100%. This translated into more reliable and clinically meaningful predictions, as seen in the ROC curves and confusion matrices.

In summary, while resampling mitigated class imbalance, augmentation proved superior because it also enhanced data diversity, reduced overfitting, and significantly boosted test set performance. For predictive modelling in healthcare where rare cases carry high clinical importance this ability to generalise well is crucial.

4.4.4 Conclusion

Based on the analysis of the confusion matrices and ROC curves, the KNN model is identified as the preferred model for PC risk prediction on the MADCaP dataset. Its high accuracy, stability, and generalizability make it the optimal choice for this application. The DT model, while slightly less accurate, remains a viable alternative due to its interpretability and ease of understanding. These findings underscore the importance of aligning model selection with the specific goals of the predictive framework, balancing accuracy, interpretability, and clinical applicability.

Chapter 5

Discussion

This chapter discusses the findings from the PC classification models applied to the Men of African Descent Carcinoma of the Prostate (MADCaP) dataset from South Africa. Our primary goal was to understand the relationship between various demographic and clinical factors—including ethnicity, smoking status, geographic origin, education, and prostate-specific antigen (PSA) levels—and PC risk and severity among South African men. We utilized several ML models GNB, KNN, DT, RF, and SVM which were evaluated based on metrics such as accuracy, precision, recall, and F1 score.

In this discussion, we interpret the findings in the context of model suitability, robustness, and alignment with healthcare objectives. We also reflect on the limitations inherent to each approach and identify potential avenues for future research. Overall, this chapter underscores the broader implications of these models in healthcare and precision medicine, emphasizing their potential to enhance our understanding of PC risk among men of African descent.

5.1 Descriptive discussion

5.1.1 Ethnicity and PC risk

The study's finding that ethnicity is significantly associated with PC risk categories ($p = 0.0424$) underscores the importance of considering demographic factors in risk stratification. Specifically, the observation that Isizulu and Sepedi men are more likely to be classified in the high-risk group aligns with a substantial body of research indicating that men of African descent exhibit a higher propensity for developing aggressive forms of PC [**?**]. This disparity may be attributed to a confluence of factors, including genetic predispositions, socioeconomic determinants, and differential access to healthcare services [**?**]. Studies conducted within South Africa have further corroborated these findings, demonstrating that Black African men are frequently diagnosed at advanced stages, a phenomenon attributable, in part, to delayed detection and limited availability of screening services [**?**].

5.1.2 Smoking and PC risk

The absence of a significant association between smoking status and PC risk categories ($p = 0.5537$) is consistent with studies that have failed to establish a robust link between smoking and PC development [**?**]. However, it is noteworthy that certain studies have suggested a potential association between heavy smoking and more aggressive forms of PC [**?**]. The observed inconsistency in findings across studies may stem from variations in sample size, smoking intensity, and genetic factors, underscoring the need for more refined methodologies.

One plausible explanation for the lack of a significant association in the present study may be the relatively homogeneous smoking patterns observed among the participants. The predominance of moderate cigarette consumption across risk

groups, coupled with the potential influence of other lifestyle factors such as diet and physical activity, may have obscured the impact of smoking on PC severity in this population.

5.1.3 PSA Levels and risk classification

The significant elevation of prostate-specific antigen (PSA) levels in the high-risk group, compared to the intermediate- and low-risk groups, reinforces the established role of PSA as a pivotal biomarker in PC diagnosis and risk assessment. The observed mean PSA level at referral of 560.69 ng/mL for high-risk patients, in contrast to the lower levels observed in the low-risk group, provides empirical support for the clinical utility of PSA in risk stratification [**?**].

The significant differences observed between PSA levels at referral and at diagnosis, as evidenced by a paired t-test ($p = 0.0014$), suggest that PSA levels may exhibit temporal fluctuations, reflecting disease progression or the effects of early interventions. However, the inherent limitations of PSA in distinguishing between aggressive and non-aggressive PC necessitate the exploration of additional biomarkers to enhance risk prediction [**?**].

5.1.4 Education and PC risk

The absence of a significant association between education level and PC risk categories ($p = 0.7897$) notwithstanding, the heatmap analysis suggests a potential link between lower education levels and an increased likelihood of intermediate-risk classification. This observation aligns with existing research indicating that lower education levels are often associated with delayed PC screening and diagnosis [**?**].

Socioeconomic factors and disparities in healthcare access may contribute to this trend. Individuals with lower education levels may exhibit reduced aware-

ness of PC screening, leading to delayed diagnoses. Conversely, men with higher education levels may demonstrate greater propensity for undergoing regular PSA testing and seeking timely medical advice [**?**].

5.1.5 Age and PC risk

Age emerges as a salient factor in distinguishing between risk categories, with high-risk individuals exhibiting a higher mean age (68.35 years) compared to low-risk individuals (65.04 years). This trend is consistent with the established association between advancing age and increased PC risk [**?**]. The higher incidence of aggressive PC forms among older patients may be attributed to cumulative genetic mutations and hormonal changes [**?**].

While age serves as a well-established risk factor, it is crucial to consider its interaction with other factors such as PSA levels and lifestyle behaviors. The synergistic effect of older age and elevated PSA levels may indicate an increased likelihood of advanced disease, underscoring the importance of implementing targeted screening strategies among older men in high-risk populations.

5.1.6 Cigarette consumption and PC severity

The absence of a clear trend between cigarette consumption levels and PC severity, as evidenced by the line graph analysis, suggests that while heavy smoking may influence PC progression in select individuals, the overall impact in this population is not statistically significant. The identification of an outlier who reported smoking 60 cigarettes per day underscores the potential for extreme cases to deviate from the general trend.

The mixed findings observed in studies on smoking and PC severity highlight the complexity of PC risk factors [**?** **?**]. The variability in study outcomes

underscores the need for larger, well-controlled studies to elucidate the precise role of smoking in PC development and progression.

5.1.7 Implications of findings

The findings of this study contribute valuable insights into the risk factors associated with PC among men of African descent. The significant association between ethnicity and risk categories underscores the need for culturally tailored awareness campaigns and early detection programs. While PSA remains a crucial biomarker for risk classification, its variability suggests the necessity for complementary diagnostic tools.

Although smoking, education, and income did not exhibit significant associations with risk categories, their potential influence should not be disregarded. Socioeconomic and lifestyle factors play a crucial role in health outcomes, necessitating further exploration of their interaction with genetic predisposition in PC development.

In summary, the findings of this study provide a foundation for developing targeted public health interventions aimed at enhancing PC detection and management in high-risk populations.

5.2 Model evaluation summary

The model evaluation process unveiled distinct strengths and weaknesses in the predictive capabilities of each algorithm. GNB, a probabilistic model, demonstrated moderate performance, suggesting limited utility in complex datasets with interdependent features. In contrast, KNN and DT models exhibited higher predictive power and robustness, with KNN emerging as the top-performing model due to its high accuracy, low variance, and effective classification across various

risk levels. The RF model achieved near-perfect accuracy but raised concerns regarding potential overfitting. Lastly, SVM displayed moderate accuracy, indicating possible limitations in handling the dataset's complexities without further hyperparameter optimization.

5.3 Interpretation of model-specific performance

5.3.1 GNB: simplicity at a cost

The GNB classifier's simplicity and computational efficiency make it an appealing model for quick assessments, yet its assumptions of feature independence can limit accuracy in datasets where interactions between predictors are significant. PC risk, especially in genetic and socio-environmental datasets, is often influenced by interrelated variables, including lifestyle, genetic predispositions, and healthcare access [98]. The model's reliance on this independence assumption may have contributed to its lower predictive accuracy (0.42 for both training and test datasets). While GNB can sometimes be effective in medical text classification or sentiment analysis due to isolated, independent predictors, it appears less suitable for nuanced disease risk prediction where variables intersect in complex ways.

5.3.2 KNN: robustness and flexibility

The KNN model performed strongly, with an accuracy of 0.89 on the training set and 0.84 on the test set. This indicates not only high predictive power but also stability and robustness, as minimal variance between the two datasets suggests resistance to overfitting. KNN's non-parametric nature enables it to capture non-linear relationships, allowing it to model complex data structures and nuanced patterns in cancer risk factors [99]. Furthermore, the model's adaptive performance

aligns well with PC's heterogeneous presentation in various demographic groups, especially among high-risk populations such as men of African descent. Despite testing with Grid Search, no hyperparameter adjustments significantly improved accuracy, suggesting that the selected parameters are optimal for the dataset. The model's ability to generalize well indicates strong potential for external validation on similar datasets.

5.3.3 DT: interpretability and consistency

The DT model demonstrated balanced accuracy (0.83 for both training and test datasets), reflecting both consistency and interpretability. DTs provide easily interpretable outputs, as the hierarchical structure highlights feature importance and decision pathways [100]. In the context of PC prediction, this interpretability is especially valuable for healthcare professionals seeking clear, actionable insights into which risk factors are most predictive. The DT's stability across datasets and lack of significant overfitting make it a practical choice when transparency in decision-making is prioritized. This feature is beneficial for developing individualized treatment plans, as clinicians can trace each decision path to understand the influences behind risk classifications, an important aspect in precision medicine and patient-centered care.

5.3.4 RF: High accuracy with potential overfitting

The RF model achieved near-perfect scores on both training and test datasets, raising concerns about overfitting. RF's ensemble approach typically enhances robustness by averaging multiple DTs to reduce individual tree variance and noise [101]. However, the perfect scores suggest a level of adaptation to the training data that may not generalize to unseen datasets. This highlights the challenge of

balancing accuracy with generalizability, a critical consideration in healthcare applications where models must perform reliably across diverse patient populations. Although RF's results are promising, further external validation is essential to confirm that these scores are not overly reliant on the specific characteristics of the MADCaP dataset. If validated, RF could serve as a powerful tool in high-stakes cancer prediction tasks due to its resilience to noisy data and strong classification abilities.

5.3.5 SVM: Consistency with limited predictive power

The SVM model's moderate accuracy (0.64 on both datasets) indicates stability but relatively limited predictive power. This may result from SVM's sensitivity to hyperparameter settings and kernel choices, which are crucial in high-dimensional datasets like MADCaP [102]. SVM has demonstrated utility in various medical contexts where high-dimensional data are present, particularly due to its robustness against overfitting. However, the moderate scores here suggest that additional kernel experimentation (e.g., using a polynomial or radial basis function kernel) or feature transformations may be necessary to improve performance. While SVM remains a valuable tool in many biomedical applications, its effectiveness in predicting PC risk within this dataset appears limited without further tuning.

5.3.6 GB: Balanced and competitive performance

The GB model demonstrated solid performance, achieving 79% accuracy on the test dataset, with consistent precision, recall, and F1 scores (all at 79%). This consistency reflects the model's capacity to effectively capture complex patterns in the dataset while maintaining generalizability. GB's iterative approach allows it to reduce bias and variance, making it well-suited for nuanced datasets like MADCaP. While GB's performance was competitive, it did not surpass KNN, which remains

the preferred model due to its higher accuracy and overall robustness. Nevertheless, GB represents a promising alternative, particularly for datasets requiring advanced pattern recognition and balanced classification across classes.

5.4 Comparison and implications for model selection

5.4.1 Selection of optimal model: KNN and DT

The results indicate that both KNN and DT models provide an optimal balance of accuracy, generalizability, and interpretability for PC risk prediction within the MADCaP dataset. KNN stands out due to its adaptability and robustness against overfitting, making it particularly well-suited for complex datasets characterized by non-linear patterns in disease risk factors. In addition, the interpretability of DTs enhances their applicability in clinical contexts, where transparency and model explainability are vital for informed decision-making.

Consequently, KNN is identified as the most suitable model for this predictive task, combining high predictive power with a low risk of overfitting. Meanwhile, the DT model remains a strong alternative, especially in situations where understanding the underlying factors is critical. Although the RF model demonstrated exceptional accuracy, further validation is necessary to ensure it is not overly fitted to this specific dataset. In contrast, SVM and GNB, despite their utility in other domains, exhibited limited effectiveness in this context, suggesting that they may not be ideal choices for PC risk prediction.

5.5 Limitations of the Study

Despite the promising findings of this study, several notable limitations must be acknowledged. First, the reliance on a single dataset restricts the generalizability of the results to broader populations, particularly among diverse African cohorts. The relatively small sample size limits the statistical power and robustness of the findings, increasing the potential for sampling bias. Additionally, the use of a large synthetic dataset, while improving model performance, raises the risk of overfitting and reduces external validity. Therefore, results should be interpreted cautiously, and future studies should validate the models using independent datasets. Moreover, the artificial data augmentation applied to address class imbalances may have introduced artifacts, potentially affecting the accuracy and reliability of the model. While we explored various ML algorithms, the hyperparameter selection was largely based on default settings, which may not have fully optimized model performance. The potential presence of unmeasured confounding factors could introduce bias, thereby affecting the reliability of model predictions. The interpretability of more complex models, such as RF, may also pose challenges for their acceptance in clinical practice, where simpler, more transparent models are often favored. Furthermore, our analysis did not account for certain lifestyle factors, including dietary habits and occupational exposures, which could significantly influence PC risk. Addressing these limitations in future research will be essential for validating our findings and enhancing predictive accuracy.

5.6 Summary

This chapter discusses the findings from PC classification models applied to the MADCaP dataset in South Africa, focusing on demographic and clinical factors

such as ethnicity, smoking, geographic origin, education, PSA levels, and income in relation to PC risk and severity. Balancing the dataset addressed class imbalances, enhancing sensitivity to low-risk cases critical for early detection. Significant disparities were observed, with ethnicity, smoking, and education strongly correlating with cancer risk, while PSA levels and age showed unexpected trends warranting further research. Smoking emerged as a key predictor of aggressive cancer, emphasizing the need for targeted public health interventions like smoking cessation and improved screening access. The study highlights the importance of tailored prevention strategies and advanced ML approaches, with KNN and DT models showing robust performance, though RF raised overfitting concerns. These findings underscore the potential for precision medicine to address health disparities in African populations.

Chapter 6

Conclusions and future directions

In conclusion, this study identifies the KNN model as the most suitable approach for predicting PC risk among South African men. KNN's simplicity, interpretability, and robust performance underscore its potential as a practical tool for clinical application, enabling early detection and targeted interventions for at-risk populations. Our findings reveal the significant roles of smoking status, ethnicity, geographic origin, and PSA levels in influencing PC risk, thereby highlighting the necessity for tailored public health strategies. These strategies should encompass smoking cessation programs, culturally informed healthcare interventions, and targeted screening initiatives aimed at addressing the unique risks faced by high-risk communities in South Africa.

However, this study is not without limitations. The reliance on a single dataset may hinder the generalizability of our results. Future research should focus on validating our findings across diverse populations to confirm the models' applicability and accuracy in broader clinical contexts. We recommend exploring external validation and cross-population testing of high-performing models, particularly KNN, DT, and RF, as these efforts are crucial for the models' acceptance in clinical practice.

Moreover, the integration of genetic and environmental risk factors, alongside longitudinal data to track PSA levels and lifestyle patterns, will enrich the predictive framework. Incorporating socioeconomic factors and examining the role of genetic markers within diverse African populations can provide a more comprehensive understanding of PC risk. This multifaceted approach could enhance prediction accuracy and inform strategies that are culturally relevant and tailored to the specific needs of underrepresented groups, particularly men of African descent.

The complex, interdependent nature of PC risk factors also suggests the potential benefits of hybrid models or deep learning architectures. Techniques such as Convolutional Neural Networks (CNNs) and hybrid ensemble methods that incorporate KNN with feature extraction could capture nuanced patterns that enhance classification power. Additionally, applying dimensionality reduction methods, such as Principal Component Analysis (PCA), may improve the performance of models like SVM and GNB by focusing on the most predictive features.

Ultimately, the findings from this study contribute significantly to the evolving field of precision medicine, where personalized risk predictions are vital. By integrating ML models into clinical practice, we can facilitate early detection and develop targeted care strategies for high-risk individuals. This not only promises to improve health outcomes but also aims to reduce disparities in PC risk and treatment among diverse populations.

Bibliography

[1] Folakemi T. Odedina, Titilola O. Akinremi, Frank Chinegwundoh, Robin Roberts, Daohai Yu, R. Renee Reams, Matthew L. Freedman, Brian Rivers, B. Lee Green, and Nagi Kumar. Prostate cancer disparities in Black men of African descent: a comparative literature review of prostate cancer burden among Black men in the United States, Caribbean, United Kingdom, and West Africa. *Infect Agent Cancer*, 4 Suppl 1(Suppl 1):S2, February 2009.

[2] Prashanth Rawla. Epidemiology of Prostate Cancer. *World J Oncol*, 10(2):63–89, April 2019.

[3] GLOBOCAN 2020: New Global Cancer Data | UICC.

[4] Strengthening Capacity for Prostate Cancer Early Diagnosis in West Africa Amidst the COVID-19 Pandemic: A Realist Approach to Rethinking and Operationalizing the World Health Organization 2017 Guide to Cancer Early Diagnosis - PMC.

[5] Kathryn Gallagher. Prostate cancer cases expected to double worldwide between 2020 and 2040, new analysis suggests, April 2024.

[6] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global Cancer Statistics

2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*, 71(3):209–249, May 2021.

[7] Active Surveillance for Prostate Cancer Increasing - NCI, June 2022. Archive Location: nciglobal,ncienterprise.

[8] HOME.

[9] *Evidence review for identifying prostate cancer clinical progression in people with low- to intermediate-risk cancer: Prostate cancer: diagnosis and management: Evidence review F.* NICE Evidence Reviews Collection. National Institute for Health and Care Excellence (NICE), London, 2019.

[10] (PDF) Supervised Learning.

[11] Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review - PubMed.

[12] Predicting biochemical recurrence of prostate cancer with artificial intelligence - PubMed.

[13] Machine learning model for the prediction of prostate cancer in patients with low prostate-specific antigen levels: A multicenter retrospective analysis - PubMed.

[14] Renato Cuocolo, Maria Brunella Cipullo, Arnaldo Stanzione, Lorenzo Ugga, Valeria Romeo, Leonardo Radice, Arturo Brunetti, and Massimo Imbriaco. Machine learning applications in prostate cancer magnetic resonance imaging. *Eur Radiol Exp*, 3(1):35, August 2019.

[15] Evaluation of artificial intelligence techniques in disease diagnosis and prediction - PMC.

[16] Artificial intelligence in healthcare: transforming the practice of medicine - PMC.

[17] Benjamin Hunter, Sumeet Hindocha, and Richard W. Lee. The Role of Artificial Intelligence in Early Cancer Diagnosis. *Cancers (Basel)*, 14(6):1524, March 2022.

[18] Shiva Maleki Varnosfaderani and Mohamad Forouzanfar. The Role of AI in Hospitals and Clinics: Transforming Healthcare in the 21st Century. *Bioengineering (Basel)*, 11(4):337, March 2024.

[19] Victor Mudenda, Evans Malyangu, Shahin Sayed, and Kenneth Fleming. Addressing the shortage of pathologists in Africa: Creation of a MMed Programme in Pathology in Zambia. *Afr J Lab Med*, 9(1):974, June 2020.

[20] Access to pathology and laboratory medicine services: a crucial gap - PubMed.

[21] Population-based screening for cancer: hope and hype - PubMed.

[22] Syed Muhammad Hammad Ali, Usama Muhammad Kathia, Muhammad Umar Masood Gondal, Ahsan Zil-E-Ali, Haseeb Khan, and Sabiha Riaz. Impact of Clinical Information on the Turnaround Time in Surgical Histopathology: A Retrospective Study. *Cureus*, 10(5):e2596, May 2018.

[23] Kazuhiro Nagao, Hideyasu Matsuyama, Hiroaki Matsumoto, Takahito Nasu, Mitsutaka Yamamoto, Yoriaki Kamiryo, Yoshikazu Baba, Akinobu Suga, Yasuhide Tei, Satoru Yoshihiro, Akihiko Aoki, Tomoyuki Shimabukuro, Keiji Joko, Shigeru Sakano, Kimio Takai, Shiro Yamaguchi, Jumpei Akao, Seiji Kitahara, and Yamaguchi Uro-Oncology Group. Identification of curable high-risk prostate cancer using radical prostatectomy alone: who are the

good candidates for undergoing radical prostatectomy among patients with high-risk prostate cancer? *Int J Clin Oncol*, 23(4):757–764, August 2018.

[24] Sana Ahuja and Sufian Zaheer. Advancements in pathology: Digital transformation, precision medicine, and beyond. *Journal of Pathology Informatics*, 16:100408, January 2025.

[25] Shuroug A. Alowais, Sahar S. Alghamdi, Nada Alsuhebany, Tariq Alqahtani, Abdulrahman I. Alshaya, Sumaya N. Almohareb, Atheer Aldairem, Mohammed Alrashed, Khalid Bin Saleh, Hisham A. Badreldin, Majed S. Al Yami, Shmeylan Al Harbi, and Abdulkareem M. Albekairy. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Medical Education*, 23:689, September 2023.

[26] Rongguang Wang, Pratik Chaudhari, and Christos Davatzikos. Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies. *Proceedings of the National Academy of Sciences of the United States of America*, 120(6):e2211613120, January 2023.

[27] Somya Agrawal and Sunita Vagha. A Comprehensive Review of Artificial Intelligence in Prostate Cancer Care: State-of-the-Art Diagnostic Tools and Future Outlook. *Cureus*, 16(8):e66225.

[28] J John, A Adam, L Kaestner, P Spies, S Mutambirwa, and J Lazarus. The South African Prostate Cancer Screening Guidelines. *S Afr Med J*, page e2194, May 2024.

[29] Stacy Loeb, Marc A. Bjurlin, Joseph Nicholson, Teuvo L. Tammela, David F. Penson, H. Ballentine Carter, Peter Carroll, and Ruth Etzioni. Overdiagnosis and overtreatment of prostate cancer. *Eur Urol*, 65(6):1046–1055, June 2014.

[30] Jack Cuzick, Mangesh A. Thorat, Gerald Andriole, Otis W. Brawley, Powel H. Brown, Zoran Culig, Rosalind A. Eeles, Leslie G. Ford, Freddie C. Hamdy, Lars Holmberg, Dragan Ilic, Timothy J. Key, Carlo La Vecchia, Hans Lilja, Michael Marberger, Frank L. Meyskens, Lori M. Minasian, Chris Parker, Howard L. Parnes, Sven Perner, Harry Rittenhouse, Jack Schalken, Hans-Peter Schmid, Bernd J. Schmitz-Dräger, Fritz H. Schröder, Arnulf Stenzl, Bertrand Tombal, Timothy J. Wilt, and Alicja Wolk. Prevention and Early Detection of Prostate Cancer. *Lancet Oncol*, 15(11):e484–e492, October 2014.

[31] Ruth Etzioni, Nicole Urban, Scott Ramsey, Martin McIntosh, Stephen Schwartz, Brian Reid, Jerald Radich, Garnet Anderson, and Leland Hartwell. The case for early detection. *Nat Rev Cancer*, 3(4):243–252, April 2003.

[32] T. Dent, J. Jbilou, I. Rafi, N. Segnan, S. Törnberg, S. Chowdhury, A. Hall, G. Lyratzopoulos, R. Eeles, D. Eccles, N. Hallowell, N. Pashayan, P. Pharoah, and H. Burton. Stratified cancer screening: the practicalities of implementation. *Public Health Genomics*, 16(3):94–99, 2013.

[33] Manisha A. Jain, Stephen W. Leslie, and Amit Sapra. Prostate Cancer Screening. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2024.

[34] Roman Gulati, Sigrid V. Carlsson, and Ruth Etzioni. When to discuss prostate cancer screening with average-risk men. *Am J Prev Med*, 61(2):294–298, August 2021.

[35] Michael K. David and Stephen W. Leslie. Prostate Specific Antigen. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2024.

[36] Fritz H. Schröder, Jonas Hugosson, Monique J. Roobol, Teuvo L.J. Tammela, Marco Zappa, Vera Nelen, Maciej Kwiatkowski, Marcos Lujan, Lissa

Määttänen, Hans Lilja, Louis J. Denis, Franz Recker, Alvaro Paez, Chris H. Bangma, Sigrid Carlsson, Donella Puliti, Arnauld Villers, Xavier Rebillard, Matti Hakama, Ulf-Hakan Stenman, Paula Kujala, Kimmo Taari, Gunnar Aus, Andreas Huber, Theo van der Kwast, Ron H.N. van Schaik R, Harry J. de Koning, Sue M. Moss, and Anssi Auvinen. The European Randomized Study of Screening for Prostate Cancer – Prostate Cancer Mortality at 13 Years of Follow-up. *Lancet*, 384(9959):2027–2035, December 2014.

[37] Ruth Etzioni and Yaw A Nyame. Prostate Cancer Screening Guidelines for Black Men: Spotlight on an Empty Stage. *J Natl Cancer Inst*, 113(6):650–651, November 2020.

[38] Prostate cancer disparities in Black men of African descent: a comparative literature review of prostate cancer burden among Black men in the United States, Caribbean, United Kingdom, and West Africa | Infectious Agents and Cancer | Full Text.

[39] Matthew O.A. Benedict, Wilhelm J. Steinberg, Frederik M. Claassen, and Nathaniel Mofolo. The profile of Black South African men diagnosed with prostate cancer in the Free State, South Africa. *S Afr Fam Pract (2004)*, 65(1):5553, January 2023.

[40] Underlying Features of Prostate Cancer—Statistics, Risk Factors, and Emerging Methods for Its Diagnosis - PMC.

[41] Michał Oczkowski, Katarzyna Dziendzikowska, Anna Pasternak-Winiarska, Dariusz Włodarek, and Joanna Gromadzka-Ostrowska. Dietary Factors and Prostate Cancer Development, Progression, and Reduction. *Nutrients*, 13(2), February 2021. Publisher: Multidisciplinary Digital Publishing Institute (MDPI).

[42] Rohini Janivara, Wenlong C. Chen, Ujani Hazra, Shakuntala Baichoo, Ilir Agalliu, Paidamoyo Kachambwa, Corrine N. Simonti, Lyda M. Brown, Saanika P. Tambe, Michelle S. Kim, Maxine Harlemon, Mohamed Jalloh, Dillon Muzondiwa, Daphne Naidoo, Olabode O. Ajayi, Nana Yaa Snyper, Lamine Niang, Halimatou Diop, Medina Ndoye, James E. Mensah, Afua O. D. Abrahams, Richard Biritwum, Andrew A. Adjei, Akindele O. Adebiyi, Olayiwola Shittu, Olufemi Ogunbiyi, Sikiru Adebayo, Maxwell M. Nwegbu, Hafees O. Ajibola, Olabode P. Oluwole, Mustapha A. Jamda, Audrey Pentz, Christopher A. Haiman, Petrus V. Spies, André van der Merwe, Michael B. Cook, Stephen J. Chanock, Sonja I. Berndt, Stephen Watya, Alexander Lubwama, Mazvita Muchengeti, Sean Doherty, Natalie Smyth, David Lounsbury, Brian Fortier, Thomas E. Rohan, Judith S. Jacobson, Alfred I. Neugut, Ann W. Hsing, Alexander Gusev, Oseremen I. Aisuodionoe-Shadrach, Maureen Joffe, Ben Adusei, Serigne M. Gueye, Pedro W. Fernandez, Jo McBride, Caroline Andrews, Lindsay N. Petersen, Joseph Lachance, and Timothy R. Rebbeck. Heterogeneous genetic architectures of prostate cancer susceptibility in sub-Saharan Africa. *Nat Genet*, 56(10):2093–2103, October 2024. Publisher: Nature Publishing Group.

[43] Pamela X. Y. Soh, Naledi Mmekwa, Desiree C. Petersen, Kazzem Gheybi, Smit van Zyl, Jue Jiang, Sean M. Patrick, Raymond Campbell, Weerachai Jaratlerdseri, Shingai B. A. Mutambirwa, M. S. Riana Bornman, and Vanessa M. Hayes. Prostate cancer genetic risk and associated aggressive disease in men of African ancestry. *Nat Commun*, 14(1):8037, December 2023. Publisher: Nature Publishing Group.

[44] Risk Factors for Prostate Cancer - PMC.

[45] Prostate-Specific Antigen (PSA) Test - NCI, March 2022. Archive Location:

nciglobal,ncienterprise.

[46] Fredrick R. Schumacher, Sonja I. Berndt, Afshan Siddiq, Kevin B. Jacobs, Zhaoming Wang, Sara Lindstrom, Victoria L. Stevens, Constance Chen, Alison M. Mondul, Ruth C. Travis, Daniel O. Stram, Rosalind A. Eeles, Douglas F. Easton, Graham Giles, John L. Hopper, David E. Neal, Freddie C. Hamdy, Jenny L. Donovan, Kenneth Muir, Ali Amin Al Olama, Zsofia Kote-Jarai, Michelle Guy, Gianluca Severi, Henrik Grönberg, William B. Isaacs, Robert Karlsson, Fredrik Wiklund, Jianfeng Xu, Naomi E. Allen, Gerald L. Andriole, Aurelio Barricarte, Heiner Boeing, H. Bas Bueno-de Mesquita, E. David Crawford, W. Ryan Diver, Carlos A. Gonzalez, J. Michael Gaziano, Edward L. Giovannucci, Mattias Johansson, Loic Le Marchand, Jing Ma, Sabina Sieri, Pär Stattin, Meir J. Stampfer, Anne Tjonneland, Paolo Vineis, Jarmo Virtamo, Ulla Vogel, Stephanie J. Weinstein, Meredith Yeager, Michael J. Thun, Laurence N. Kolonel, Brian E. Henderson, Demetrius Albanes, Richard B. Hayes, Heather Spencer Feigelson, Elio Riboli, David J. Hunter, Stephen J. Chanock, Christopher A. Haiman, and Peter Kraft. Genome-wide association study identifies new prostate cancer susceptibility loci. *Hum Mol Genet*, 20(19):3867–3875, October 2011.

[47] Peter K. F. Chiu, Eric K. C. Lee, Marco T. Y. Chan, Wilson H. C. Chan, M. H. Cheung, Martin H. C. Lam, Edmond S. K. Ma, and Darren M. C. Poon. Genetic Testing and Its Clinical Application in Prostate Cancer Management: Consensus Statements from the Hong Kong Urological Association and Hong Kong Society of Uro-Oncology. *Front Oncol*, 12:962958, July 2022.

[48] Sze Loon Chow, Anselm Su Ting, and Tin Tin Su. Development of Conceptual Framework to Understand Factors Associated with Return to Work among Cancer Survivors: A Systematic Review. *Iranian Journal of Pub-*

lic Health, 43(4):391, April 2014. Publisher: Tehran University of Medical Sciences.

[49] Kunal Nagpal, Davis Foote, Yun Liu, Po-Hsuan Cameron Chen, Ellery Wulczyn, Fraser Tan, Niels Olson, Jenny L. Smith, Arash Mohtashamian, James H. Wren, Greg S. Corrado, Robert MacDonald, Lily H. Peng, Mahul B. Amin, Andrew J. Evans, Ankur R. Sangoi, Craig H. Mermel, Jason D. Hipp, and Martin C. Stumpe. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit Med*, 2:48, 2019.

[50] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22):2402–2410, December 2016.

[51] Ashley B. Anderson, Clare Grazal, Rikard Wedin, Claire Kuo, Yongmei Chen, Bryce R. Christensen, Jennifer Cullen, and Jonathan A. Forsberg. Machine learning algorithms to estimate 10-Year survival in patients with bone metastases due to prostate cancer: toward a disease-specific survival estimation tool. *BMC Cancer*, 22:476, April 2022.

[52] Ruqian Hao, Khashayar Namdar, Lin Liu, Masoom A. Haider, and Farzad Khalvati. A Comprehensive Study of Data Augmentation Strategies for Prostate Cancer Detection in Diffusion-Weighted MRI Using Convolutional Neural Networks. *J Digit Imaging*, 34(4):862–876, August 2021.

[53] Naseem Cassim, Michael Mapundu, Victor Olago, Turgay Celik, Jaya Anna

George, and Deborah Kim Glencross. Using text mining techniques to extract prostate cancer predictive information (Gleason score) from semi-structured narrative laboratory reports in the Gauteng province, South Africa. *BMC Medical Informatics and Decision Making*, 21(1):330, November 2021.

[54] World Health Organization. *ICD-10 : international statistical classification of diseases and related health problems : tenth revision*. World Health Organization, 2004. Accepted: 2012-06-16T14:40:38Z Journal Abbreviation: ICD-10.

[55] Amer Alnuaimi and Tasnim Albaldawi. An overview of machine learning classification techniques. *BIO Web of Conferences*, 97:00133, April 2024.

[56] Deyana D. Lewis and Cheryl D. Cropp. The Impact of African Ancestry on Prostate Cancer Disparities in the Era of Precision Medicine. *Genes (Basel)*, 11(12):1471, December 2020.

[57] Kedar Potdar, Taher Pardawala, and Chinmay Pai. A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *International Journal of Computer Applications*, 175:7–9, October 2017.

[58] Miroslav Kubat. Probabilities: Bayesian Classifiers. In Miroslav Kubat, editor, *An Introduction to Machine Learning*, pages 19–41. Springer International Publishing, Cham, 2015.

[59] Jan-Niklas Eckardt, Martin Bornhäuser, Karsten Wendt, and Jan Moritz Middeke. Semi-supervised learning in cancer diagnostics. *Frontiers in Oncology*, 12:960984, July 2022.

[60] Aik Choon Tan and David Gilbert. Ensemble machine learning on gene expression data for cancer classification. *Appl Bioinformatics*, 2(3 Suppl):S75–83, 2003.

[61] Identifying prostate cancer and its clinical risk in asymptomatic men using machine learning of high dimensional peripheral blood flow cytometric natural killer cell subset phenotyping data - PubMed.

[62] Houda Bichri, Adil Chergui, and Hain Mustapha. Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets. *International Journal of Advanced Computer Science and Applications*, 15, January 2024.

[63] Osval Antonio Montesinos López, Abelardo Montesinos López, and Dr Jose Crossa. Overfitting, Model Tuning, and Evaluation of Prediction Performance. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction [Internet]*. Springer, January 2022.

[64] What Is Underfitting? | IBM, October 2021.

[65] Fold Cross Validation - an overview | ScienceDirect Topics.

[66] Daniel Belete and Manjaiah D H. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *International Journal of Computers and Applications*, 44:1–12, September 2021.

[67] Guido van Rossum and Fred L Drake. Python Reference Manual.

[68] FAQ: Citing Stata software, documentation, and FAQs | Stata.

[69] Ekaba Bisong. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*. January 2019.

[70] Tiago Pessoa, Raul Medeiros, Thiago Nepomuceno, Gui-Bin Bian, V.H.C. Albuquerque, and Pedro Pedrosa Filho. Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. *IEEE Access*, PP:1–1, October 2018.

[71] Nicola Lunardon, Giovanna Menardi, and Nicola Torelli. ROSE: a Package for Binary Imbalanced Learning. *R Journal*, 6:79–89, June 2014.

[72] Nicola Lunardon, Giovanna Menardi, Nicola Torelli. ROSE: Random Over-Sampling Examples, March 2013. Institution: Comprehensive R Archive Network Pages: 0.0-4.

[73] Zulfan Ahmadi, Asrul Abdullah, and Izhan Fakhruzi. Meningkatkan Kemampuan Model dalam Memprediksi Penyakit Jantung dengan Algoritma NCL dan GridSearchCV. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 7:1632, October 2023.

[74] Robust Distance Measures for k NN Classification of Cancer Data - PubMed.

[75] Application of K-nearest neighbors algorithm on breast cancer diagnosis problem. - PMC.

[76] Decision tree methods: applications for classification and prediction - PMC.

[77] The random forest algorithm for statistical learning.

[78] (PDF) Random Forests and Decision Trees.

[79] Jerome Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29, November 2000.

[80] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. February 2009.

[81] XGBoost: A Scalable Tree Boosting System.

[82] Gradient boosting machines, a tutorial - PMC.

[83] (PDF) LightGBM: A Highly Efficient Gradient Boosting Decision Tree.

[84] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Statist Soc B. 2005;67(2):301-20 | Request PDF.

[85] A Tutorial on Support Vector Machines for Pattern Recognition | Data Mining and Knowledge Discovery.

[86] Support-vector networks | Request PDF.

[87] Machine Learning in Biomedical Informatics | Request PDF.

[88] Smola, A.: Learning with Kernels - Support Vector Machines, Regularization, Optimization and Beyond. MIT Press, Cambridge, MA | Request PDF.

[89] A Practical Guide to Support Vector Classification Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin | Request PDF.

[90] Abel Gonzalez-Perez, Ville Mustonen, Boris Reva, Graham R.S. Ritchie, Pau Creixell, Rachel Karchin, Miguel Vazquez, J. Lynn Fink, Karin S. Kassahn, John V. Pearson, Gary Bader, Paul C. Boutros, Lakshmi Muthuswamy, B.F. Francis Ouellette, Jüri Reimand, Rune Linding, Tatsuhiro Shibata,

Alfonso Valencia, Adam Butler, Serge Dronov, Paul Flicek, Nick B. Shannon, Hannah Carter, Li Ding, Chris Sander, Josh M. Stuart, Lincoln D. Stein, and Nuria Lopez-Bigas. Computational approaches to identify functional genetic variants in cancer genomes. *Nat Methods*, 10(8):723–729, August 2013.

[91] Davide Chicco and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation - Google Search.

[92] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One*, 12(6):e0177678, 2017.

[93] Weighted Accuracy - an overview | ScienceDirect Topics.

[94] True Negative - an overview | ScienceDirect Topics.

[95] The receiver operating characteristic (ROC) curves and area under the...

[96] Model Accuracy - an overview | ScienceDirect Topics.

[97] Yutaka Sasaki. The truth of the F-measure. *Teach Tutor Mater*, January 2007.

[98] Rich Caruana and Alexandru Niculescu-Mizil. An Empirical Comparison of Supervised Learning Algorithms. *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 2006:161–168, June 2006.

[99] Ibrahim Al-Bluwi and Ashraf Elnagar. A Logarithmic-Complexity Algorithm for Nearest Neighbor Classification Using Layered Range Trees. *IIM*, 04(02):39–43, 2012.

[100] J. R. Quinlan. Induction of decision trees. *Mach Learn*, 1(1):81–106, March 1986.

[101] L Breiman. Random Forests. *Machine Learning*, 45:5–32, October 2001.

[102] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Chem. Biol. Drug Des.*, 297:273–297, January 2009.

Supplementary tables and graphs

6.1 missing value report

Variable Name	Number of Missing Values
sd_raceyou	1
d_ethn_self	3
sd_townborn	4
sd_countryborn	3
hrf_smoking	1
hrf_smokestartremember	351
hrf_smokingagestart	352
hrf_cigarettescount	362
hrf_cigarettestype	362
hrf_currentlysmoke	362
hrf_smokestopremember	667
hrf_smokingagestop	670
hrf_tobacco__1	0
hrf_tobacco__2	0
hrf_tobacco__3	0
hrf_tobacco__4	0
hrf_tobacco__5	0
hrf_tobacco__6	0
hrf_tobacco__7	0
hrf_tobacco__8	0
s_education	1
s_education_other	1096
s_income	1

6.2 Correlation matrix

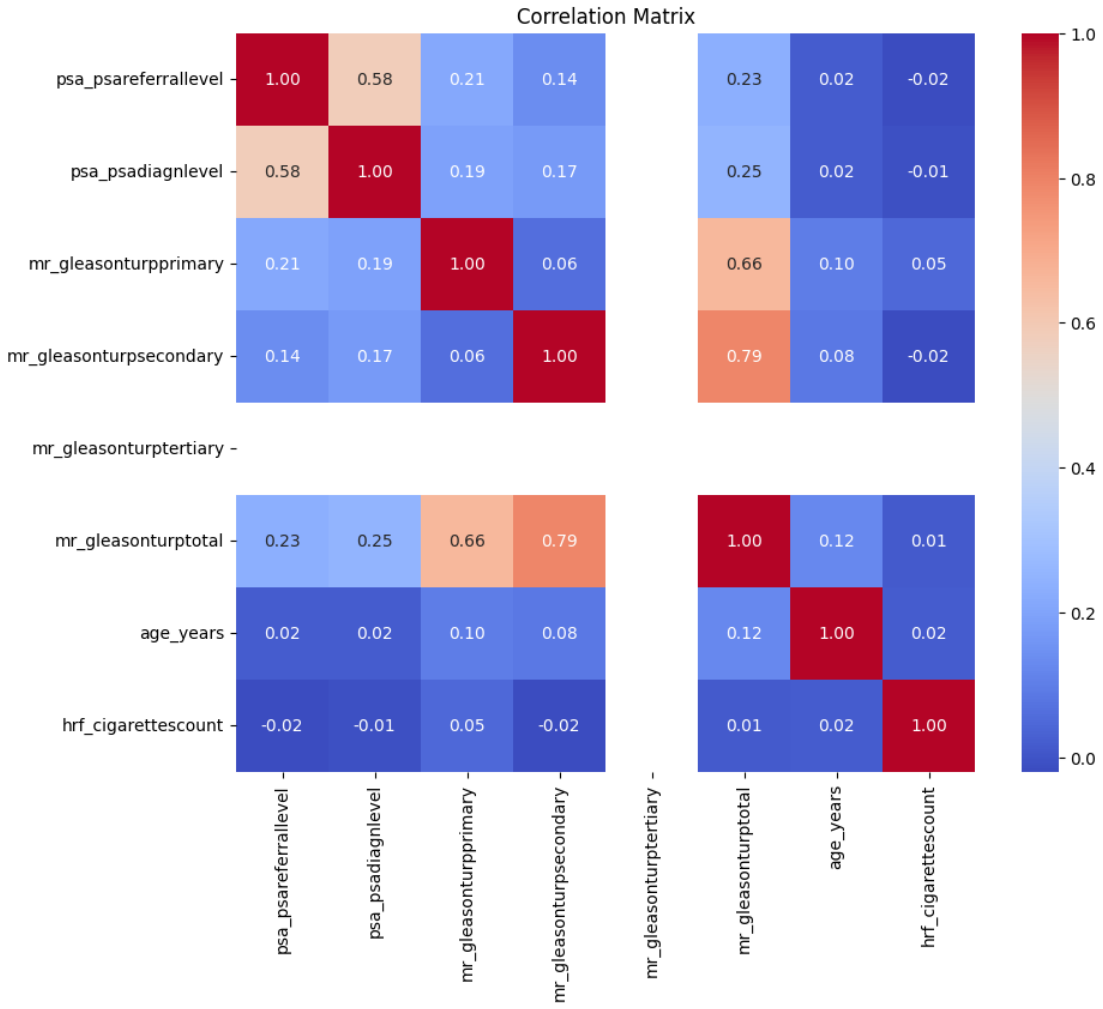


Figure 6.1: Correlation matrix

6.3 Plagiarism declaration

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

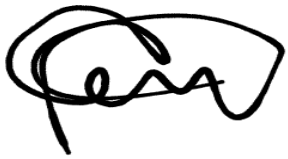
PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Friday Mhone (Student number: 2603349) am a student registered for the degree of Master of Science (MSc) epidemiology in public health informatics in the academic year 2022.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature: 

Date: 23rd March 2025

6.4 TurnItIn Report

PhaleThesis (15).pdf

ORIGINALITY REPORT

12%	10%	8%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	fastercapital.com Internet Source	1 %
2	christie.openrepository.com Internet Source	1 %
3	Uzair Aslam Bhatti, Jingbing Li, Mengxing Huang, Sibghat Ullah Bazai, Muhammad Aamir. "Deep Learning for Multimedia Processing Applications - Volume Two: Signal Processing and Pattern Recognition", CRC Press, 2024 Publication	1 %
4	scielo.org.za Internet Source	1 %
5	Mrutyunjaya Panda, Ajith Abraham, Biju Gopi, Reuel Ajith. "Computational Intelligence for Oncology and Neurological Disorders - Current Practices and Future Directions", CRC Press, 2024 Publication	1 %
6	idoc.pub Internet Source	1 %

Michael Mapundu: 

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Mr Friday Mhone

6.5 HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

HREC Research Ethics Committee (Medical)

CLEARANCE CERTIFICATE NO. M231124 M240415-C-0001

NAME: (Principal Investigator)	Mr Friday Mhone
DEPARTMENT:	School of Public Health MADCaP Consortium
DEGREE:	MSc in Epidemiology (Public Health Informatics)
PROJECT TITLE:	Using supervised machine algorithms to predict prostate cancer cases: A case of Men of African Descent Carcinoma of the Prostate Consortium (MADCaP), South Africa
DATE CONSIDERED:	24-Nov-2023
DECISION:	Approved unconditionally
CONDITIONS:	N/A
NOTE:	Application Number: MED23-11-040
SUPERVISOR:	Mr Michael Mapundu and Dr Carl Chen

APPROVED BY: Prof P. Ruff, Chairperson, HREC (Medical)

DATE OF APPROVAL: 15-Apr-2024 **EXPIRY DATE:** 14-Apr-2029

This clearance certificate is valid for 5 years from data of approval. Extension may be applied for.

DECLARATION OF INVESTIGATOR(S)

I/we fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report** in the month date of convened meeting each year until study is closed. Failure to submit annual report will result in ethics clearance certificate suspension. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. Unreported changes to the application may invalidate the clearance given by the HREC (Medical). Email a signed copy of ethics clearance certificate to HREC-Medical.ResearchOffice@wits.ac.za

Signature of Principal Investigator

Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

6.6 Research ethics training certificate



TRREE

Zertifikat
Certificat
Certificado
Certificate

Promouvoir les plus hauts standards éthiques dans la protection des participants à la recherche biomédicale
Promoting the highest ethical standards in the protection of biomedical research participants

Clinical Trials Centre
The University of Hong Kong

Certificat de formation - Training Certificate

Ce document atteste que - this document certifies that

Friday Mhone

a complété avec succès - has successfully completed

Introduction to Research Ethics

du programme de formation TRREE en évaluation éthique de la recherche
of the TRREE training programme in research ethics evaluation

Release Date: 2022/01/22

CID : d6nlllrO5H

Professeur Dominique Sprumont
Coordinateur TRREE Coordinator

FMH Continuing Education Program (5 Credits)
Programme de Formation continue (5 Crédits)

Foederatio Pharmaceutica Helvetiae FPH Programmes de formation continue
Continuing Education Programs
Programmes de formation continue

Ce programme est soutenu par - This program is supported by :
European and Developing Countries Clinical Trials Partnership (EDCTP) (www.edctp.org) - Swiss National Science Foundation (www.snf.ch) - Canadian Institutes of Health Research (http://www.cihr-irsc.gc.ca/e/2891.html) - Swiss Academy of Medical Science (SAMS/ASSM/SAMW) (www.samw.ch) - Commission for Research Partnerships with Developing Countries (www.kfpe.ch)

[REV : 20170310]

6.7 MADCaP ethics certificate 2022

Duly authorised and on behalf of the Principal Investigator:

Full name: Dr Maureen Joffe

Designation: Site Principal Investigator

Signature:

Duly authorised and on behalf of the Recipient:

Full name: Dr Judith Hornby Cuff

Designation: Managing Director

Signature:

Signed at Cape Town on this the 17 day of June, 2022

WITNESSES:

Witness 1: Witness 2:

Duly authorized on behalf of the Human Research Ethics Committee (Medical)

Protocol No. M 22/06/73 **Applicant:** Dr M Joffe, *et al*
Study title: *Men of African descent with cancer of the Prostate*

Full name: Dr Clement B Penny
Designation: Chairperson HREC (Medical), University of the Witwatersrand, Johannesburg

Signature:

Signed at Johannesburg, **on this the** 01 day of July 2022

Witnesses:

Iain Burns

Mapula Ramaila

Analysis Code

6.8 Python Code for Data Analysis

The following Python code demonstrates a comprehensive workflow for developing and evaluating machine learning models. The key steps include importing essential libraries for machine learning, data manipulation, and visualization, such as `pandas`, `numpy`, `seaborn`, and `matplotlib`. After importing the necessary libraries, the data is loaded into a DataFrame, and a helper function is utilized to split the dataset into training and testing sets.

Next, exploratory data analysis (EDA) is conducted to gain an understanding of the distributions, class imbalances, and other characteristics of the dataset. This analysis is crucial in uncovering patterns and insights that inform further data preprocessing steps. During the preprocessing phase, missing values are identified and appropriately handled, either by imputation or removal, to ensure that data quality is maintained.

In addition to handling missing values, multicollinearity among predictors is assessed through correlation matrices. This step helps identify highly correlated features that may need to be excluded or transformed to improve model performance. Once the data is preprocessed, various machine learning models, including `RandomForestClassifier`, `DecisionTreeClassifier`, `SVC`, `KNeighborsClassifier`, and `GaussianNB`, are configured with initial parameters

and trained on the dataset.

Hyperparameter tuning is then performed using grid search combined with cross-validation to fine-tune the parameters of selected models, such as decision trees and random forests. This process ensures that the best-performing parameters are chosen for the final model. Furthermore, feature importance is evaluated for models like Random Forest and Decision Tree, providing insights into which features most significantly influence the model's predictions.

Predictions are generated for both the training and testing sets, followed by the computation of key evaluation metrics, including accuracy, precision, recall, and F1 score. These metrics offer a robust assessment of model performance. To visually represent classification performance across different risk categories (e.g., Low, Intermediate, and High Risk), confusion matrices are generated and plotted using `seaborn`.

Finally, cross-validation can be conducted to further assess model stability and generalization, averaging accuracy scores over multiple folds. This comprehensive approach ensures a thorough evaluation of the machine learning models, ultimately aiding in the selection of the best-performing model for predicting prostate cancer risk.

```
1
2 # Importing necessary libraries
3 import pandas as pd
4 import sklearn.model_selection as model_selection
5 from sklearn.model_selection import train_test_split
6 import numpy as np
7 from sklearn.metrics import confusion_matrix, accuracy_score
     , roc_curve, roc_auc_score, precision_score, recall_score
8 from sklearn.ensemble import RandomForestRegressor
9 from sklearn.tree import DecisionTreeRegressor
```

```
from sklearn.linear_model import LinearRegression
from sklearn.neighbors import KNeighborsRegressor
from sklearn.svm import SVC
from sklearn.naive_bayes import GaussianNB
import matplotlib.pyplot as plt
import seaborn as sns

# 1. Importing data from Google Drive
from google.colab import drive
drive.mount('/content/gdrive', force_remount=True)

data = pd.read_csv('gdrive/My Drive/
    friday_mhone_data_extract_final_26apr2024.csv', delimiter
    ='|')
data.head()

# List column names
column_names = data.columns.tolist()
print("All Column Names:")
print(column_names)

# 2. EDA - Target variable distribution
data['mr_gleasonturptotal'].describe()

# Dataset structure
num_samples, num_features = data.shape
print("Number of observations (samples):", num_samples)
print("Number of variables (features):", num_features)

```

```
# Distribution of target variable
plt.figure(figsize=(10, 6))
sns.distplot(data['mr_gleasonturptotal'], hist=True, kde=
    True, bins=20, color='blue', hist_kws={'edgecolor': '
    black'})
plt.title('Probability Density Function of
    mr_gleasonturptotal')
plt.xlabel('mr_gleasonturptotal')
plt.ylabel('Density')
plt.show()

# Check for duplicates
duplicates = data[data.duplicated()]
if duplicates.empty:
    print("No duplicates found.")
else:
    print("Duplicates found:")
    print(duplicates)

# Calculate age of participants
data['date_interview'] = pd.to_datetime(data['date_interview
    '])
data['sd_dateofbirth'] = pd.to_datetime(data['sd_dateofbirth
    '])
age_timedelta = data['date_interview'] - data['
    sd_dateofbirth']
age_years = (age_timedelta.dt.days / 365.25).round(decimals
    =0)
data['age_years'] = age_years
```

```
data = data.drop(['studyid', 'date_interview', '
    sd_dateofbirth'], axis=1)
data.head()

# Multicollinearity check
data_numerical = data[['mr_gleasonturptotal','age_years','
    hrf_smokingagestart','hrf_cigarettescount','
    hrf_smokingagestop','psa_psareferrallevel','
    psa_psadiagnlevel']]
correlation_matrix = data_numerical.corr()
print("Correlation Matrix:")
print(correlation_matrix)

threshold = 0.8
multicollinear_vars = set()
for i in range(len(correlation_matrix.columns)):
    for j in range(i+1, len(correlation_matrix.columns)):
        if correlation_matrix.iloc[i, j] > threshold:
            var1 = correlation_matrix.columns[i]
            var2 = correlation_matrix.columns[j]
            multicollinear_vars.add((var1, var2))

if multicollinear_vars:
    print("\nVariables with multicollinearity (correlation >
    0.8):")
    for var_pair in multicollinear_vars:
        print(f"{var_pair[0]} and {var_pair[1]}")
else:
    print("\nNo multicollinearity detected.")
```

```
83
84 plt.figure(figsize=(8, 6))
85 sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm',
       fmt=".2f", annot_kws={"size": 10})
86 plt.title('Correlation Heatmap of Numerical Variables')
87 plt.show()
88
89 # Handling missing values
90 data = data.drop(['s_education_other','mr_imagectdate','
       mr_imagectdesc','mr_imagemridate','mr_imagemridesc','
       mr_imagepelvicdate','mr_imagepelvicdesc','
       mr_imagexraydate','mr_imagexraydesc'], axis=1)
91 categorical_columns = ['hrf_smokestartremember', '
       hrf_cigarettestype', 'hrf_currentlysmoke', '
       hrf_smokestopremember']
92 for column in categorical_columns:
93     data[column] = data[column].fillna('None')
94
95 from sklearn.impute import KNNImputer
96 numerical_columns = ['hrf_smokingagestart', '
       hrf_cigarettescount', 'hrf_smokingagestop']
97 knn_imputer = KNNImputer(n_neighbors=5)
98 data[numerical_columns] = knn_imputer.fit_transform(data[
       numerical_columns])
99
100 data = data.dropna()
101 data.shape
102
103 # A descriptive statistics of numerical variables
```

```
104 data_numerical.describe().T
105
106 columns_name = data.columns
107 print(columns_name)
108
109 # Recording categorical data
110 df = data
111
112 from sklearn.preprocessing import LabelEncoder
113
114 # Function to categorize gleason_total
115 def categorize_gleason(gleason):
116     if gleason <= 6:
117         return 'Low Risk'
118     if gleason == 7:
119         return 'Intermediate risk'
120     else:
121         return 'High Risk'
122
123 # Apply the function to create a new column 'risk_category'
124 df['risk_gleason'] = df['mr_gleasonturptotal'].apply(
    categorize_gleason)
125
126 # Summarize the data
127 summary = df.groupby('risk_gleason')['risk_gleason'].count()
128
129 # Document the encoding
130 encoding = {'Low Risk': 0, 'Intermediate risk': 1, 'High
    Risk': 2}
```

```
131
132 # Encode the labels
133 label_encoder = LabelEncoder()
134 df['encoded_risk_category'] = label_encoder.fit_transform(df
      ['risk_gleason'])
135
136 # Count the occurrences of each Gleason score category
137 class_counts = df['risk_gleason'].value_counts()
138 class_labels = class_counts.index
139 class_sizes = class_counts.values
140
141 # Calculate sample size and percentages
142 n = len(df)
143 percentages = (class_sizes / n) * 100
144
145 # Create a summary DataFrame
146 summary_df = pd.DataFrame({
147     'Category': class_labels,
148     'Count': class_sizes,
149     'Percentage': percentages
150 })
151
152 # Print the summary DataFrame
153 print(summary_df)
154
155 # Function to format the labels for the pie chart
156 def func(pct, allvalues):
157     absolute = int(round(pct / 100. * sum(allvalues)))
158     return f"{pct:.1f}%\n({absolute})"
```

```
159
160 # Plot the pie chart
161 plt.figure(figsize=(8, 6))
162 plt.pie(class_sizes, labels=class_labels, autopct=lambda pct
    : func(pct, class_sizes), startangle=140, colors=['#
    ff9999','#66b3ff','#99ff99'])
163 plt.title(f'n: {n}')
164 plt.axis('equal')
165 plt.show()
166
167 # Recording Ethnicity
168 unique_categories = data['d_ethn_self'].unique()
169 category_mapping = {
170     'IsiZulu': 1, 'Sesotho': 2, 'Setswana': 3, 'Tsonga': 4,
    'Sepedi': 5,
171     'Seswati': 6, 'IsiXhosa': 7, 'Ndebele': 8, 'Tshivenda':
    9, 'Other': 10
172 }
173 df['d_ethn_self_codes'] = df['d_ethn_self'].map(
    category_mapping).fillna(0).astype(int)
174
175 # Recording Smoking Status
176 unique_categories = df['hrf_smoking'].unique()
177 category_mapping = {category: code for code, category in
    enumerate(unique_categories, start=0)}
178 df['hrf_smoking_codes'] = df['hrf_smoking'].map(
    category_mapping).astype(int)
179
180 # Recording Town Born
```

```
df['town_category'] = df['sd_townborn'].apply(lambda x: x if
    x in ['Soweto', 'Alexandra', 'Sophiatown'] else '
   OtherTowns')
category_mapping = {'Soweto': 1, 'Alexandra': 2, 'Sophiatown
   ': 3, 'OtherTowns': 0}
df['town_code'] = df['town_category'].map(category_mapping)

# Recording Country of Born
df['country_category'] = df['sd_countryborn'].apply(lambda x
   : 'South Africa' if x == 'South Africa' else '
   OtherCountries')
country_mapping = {'South Africa': 1, 'OtherCountries': 0}
df['country_code'] = df['country_category'].map(
   country_mapping)

# Outlier Removal Function
import seaborn as sns
import matplotlib.pyplot as plt

numerical_vars = ['psa_psadiagnlevel', 'age_years']
sns.pairplot(df[numerical_vars], markers="o", plot_kws={"
   color": "blue"}, diag_kws={"color": "green"})
plt.show()

def remove_outliers(df, columns):
    for col in columns:
        mean, std_dev = df[col].mean(), df[col].std()
        z_score_threshold = 2
        z_scores = np.abs((df[col] - mean) / std_dev)
```

```
        df = df[z_scores <= z_score_threshold]
    return df

df_cleaned = remove_outliers(df, numerical_vars)
sns.pairplot(df_cleaned[numerical_vars], markers="o",
   plot_kws={"color": "blue"}, diag_kws={"color": "green"})
plt.show()

# Convert specified columns to 'category'
categorical_columns = [
    'encoded_risk_category', 'd_ethn_self_codes', '
   hrf_smoking_codes',
    'town_code', 'country_code', '
   hrf_smokestartremember_codes',
    'hrf_cigarettestype_codes', 's_education_codes', '
   s_income_codes',
    'psa_psareferralleveldesc_codes'
]
df[categorical_columns] = df[categorical_columns].astype('
   category')
print(df.dtypes)

# Import libraries
import pandas as pd
import numpy as np
from imblearn.over_sampling import RandomOverSampler
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix, accuracy_score
   , roc_curve, roc_auc_score
```

```
from sklearn.preprocessing import label_binarize
import tensorflow as tf
import matplotlib.pyplot as plt
import seaborn as sns

# ROSE-style oversampling by adding small random noise to existing points
def rose_sampler(X, y, random_state=42):
    ros = RandomOverSampler(random_state=random_state)
    X_resampled, y_resampled = ros.fit_resample(X, y)

    # Identify numeric columns to apply noise
    numeric_cols = X_resampled.select_dtypes(include=[np.number]).columns.tolist()

    # Adding a small amount of Gaussian noise to numeric columns only
    noise = np.random.normal(0, 0.01, X_resampled[numeric_cols].shape)  # Adjust noise scale if needed
    X_resampled[numeric_cols] += noise
    return X_resampled, y_resampled

# Assign data for processing
a = df['encoded_risk_category'].astype('category')  # Ensure categorical
b = df[['d_ethn_self_codes', 'hrf_smoking_codes', 'hrf_currentlysmoke_codes',
        'town_code', 'country_code', 'hrf_smokestartremember_codes',
```

```
        'hrf_cigarettestype_codes', 's_education_codes',
        's_income_codes', 'psa_psareferralleveldesc_codes',
        'hrf_cigarettescount', 'psa_psareferrallevel',
        'psa_psadiagnlevel', 'age_years', '
    mr_gleasonturptotal']]

# Check initial class distribution
initial_class_distribution = a.value_counts()
print("Initial Class Distribution:",
    initial_class_distribution.to_dict())

# Apply ROSE-style sampling to balance the dataset
b_resampled, a_resampled = rose_sampler(b, a)

# Convert resampled target variable to categorical if
    necessary
a_resampled = pd.Series(a_resampled).astype('category')

# Check new class distribution after resampling
unique, counts = np.unique(a_resampled, return_counts=True)
class_distribution = dict(zip(unique, counts))
print("Class Distribution after ROSE-style sampling:",
    class_distribution)

# Combine resampled data
resampled_data = pd.DataFrame(b_resampled, columns=b.columns
    )
resampled_data['encoded_risk_category'] = a_resampled
resampled_data.to_csv('resampled_data.csv', index=False)
```

```
print("Resampled dataset saved to 'resampled_data.csv'")

# Data Splitting
X = resampled_data.drop('encoded_risk_category', axis=1)
y = resampled_data['encoded_risk_category']
X_train, X_test, y_train, y_test = train_test_split(X, y,
    train_size=0.7, test_size=0.3, random_state=101)

# Augmentation functions
def augment_data(df, num_samples):
    noise_std = 0.1
    num_features = ['hrf_cigarettescount', '
    psa_psareferrallevel', 'psa_psadiagnlevel', 'age_years']
    augmented_data = []

    for _ in range(num_samples):
        augmented_instance = df.copy()
        for feature in num_features:
            augmented_instance[feature] += np.random.normal
    (0, noise_std, size=len(augmented_instance))
        augmented_data.append(augmented_instance)

    return pd.concat(augmented_data, ignore_index=True)

# Separate classes and augment each class
num_augmented_samples = 100
class_0 = X_train[y_train == 0]
class_1 = X_train[y_train == 1]
class_2 = X_train[y_train == 2]
```

```
298
299 augmented_class_0 = augment_data(class_0,
    num_augmented_samples)
300 augmented_class_1 = augment_data(class_1,
    num_augmented_samples)
301 augmented_class_2 = augment_data(class_2,
    num_augmented_samples)
302
303 augmented_train_df = pd.concat([augmented_class_0,
    augmented_class_1, augmented_class_2], ignore_index=True)
304 augmented_train_df['encoded_risk_category'] = pd.concat([pd.
    Series([0] * num_augmented_samples),
305                                                          pd.
    Series([1] * num_augmented_samples),
306                                                          pd.
    Series([2] * num_augmented_samples)], ignore_index=True)
307
308 # Summary statistics for categorical variables
309 categorical_vars = ['d_ethn_self_codes', 'hrf_smoking_codes'
    , 'town_code', 'country_code',
310                     'hrf_smokestartremember_codes', '
    hrf_cigarettestype_codes',
311                     's_education_codes', 's_income_codes', '
    psa_psareferralleveldesc_codes']
312
313 combined_table = pd.DataFrame()
314 for cat_var in categorical_vars:
315     count_table = augmented_train_df.groupby([cat_var, '
    encoded_risk_category']).size().unstack(fill_value=0).
```

```
    reset_index()
316     total_counts = count_table.sum(axis=1)
317     count_table['Low Risk'] = count_table[0].astype(str) + "
        (" + (count_table[0] / total_counts * 100).round(2).
    astype(str) + "%)"
318     count_table['Intermediate Risk'] = count_table[1].astype
    (str) + " (" + (count_table[1] / total_counts * 100).
    round(2).astype(str) + "%)"
319     count_table['High Risk'] = count_table[2].astype(str) +
    " (" + (count_table[2] / total_counts * 100).round(2).
    astype(str) + "%)"
320     combined_table = pd.concat([combined_table, count_table
    ], axis=0)
321
322 # Chi-square test for categorical variables
323 from scipy.stats import chi2_contingency
324 p_values = {}
325 for var in categorical_vars:
326     contingency_table = pd.crosstab(augmented_train_df[var],
        augmented_train_df['encoded_risk_category'])
327     _, p_value, _, _ = chi2_contingency(contingency_table)
328     p_values[var] = p_value
329 p_values_df = pd.DataFrame(list(p_values.items()), columns=[
    'Variable', 'P-value'])
330 print(p_values_df)
331
332 # Import necessary libraries
333 from sklearn.tree import DecisionTreeClassifier
334 from sklearn.ensemble import RandomForestClassifier
```

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.naive_bayes import GaussianNB
from sklearn.model_selection import cross_val_score,
    GridSearchCV
from sklearn.metrics import confusion_matrix, accuracy_score
    , precision_score, recall_score, f1_score
import matplotlib.pyplot as plt
import seaborn as sns

# Define model parameters and classes
model = RandomForestClassifier(max_depth=16, bootstrap=True,
     n_estimators=100, random_state=101)
# Additional model configurations for testing
# model = DecisionTreeClassifier(max_depth=14)
# model = KNeighborsClassifier(n_neighbors=3)
# model = SVC(kernel='poly', probability=True)
# model = GaussianNB()

# Split data and fit model
X_train, X_test, y_train, y_test = split_data(X, y)
model.fit(X_train, y_train)
classes = ['Low Risk', 'Intermediate Risk', 'High Risk']

# Feature importance (for models that support it)
fts = model.feature_importances_
print(X_train.columns)
print('Feature importance scores:', fts)

```

```
# Model predictions and evaluation metrics
y_train_pred = model.predict(X_train)
y_test_pred = model.predict(X_test)
y_train_proba = model.predict_proba(X_train)
y_test_proba = model.predict_proba(X_test)

# Confusion matrices
conf_matrix_train = confusion_matrix(y_train, y_train_pred)
conf_matrix_test = confusion_matrix(y_test, y_test_pred)

# Plotting confusion matrices
fig, axes = plt.subplots(1, 2, figsize=(14, 6))
sns.heatmap(conf_matrix_train, annot=True, fmt='d', cmap='
    Blues', cbar=False, xticklabels=classes, yticklabels=
    classes, ax=axes[0])
axes[0].set_xlabel('Predicted')
axes[0].set_ylabel('Actual')
axes[0].set_title('Confusion Matrix (Train)')
sns.heatmap(conf_matrix_test, annot=True, fmt='d', cmap='
    Blues', cbar=False, xticklabels=classes, yticklabels=
    classes, ax=axes[1])
axes[1].set_xlabel('Predicted')
axes[1].set_ylabel('Actual')
axes[1].set_title('Confusion Matrix (Test)')
plt.tight_layout()
plt.show()

# Calculate accuracy, precision, recall, and F1 score
accuracy = accuracy_score(y_train, y_train_pred)
```

```
precision = precision_score(y_train, y_train_pred, average='macro')
recall = recall_score(y_train, y_train_pred, average='macro')
f1 = f1_score(y_train, y_train_pred, average='macro')

accuracy_test = accuracy_score(y_test, y_test_pred)
precision_test = precision_score(y_test, y_test_pred, average='macro')
recall_test = recall_score(y_test, y_test_pred, average='macro')
f1_test = f1_score(y_test, y_test_pred, average='macro')

print(f"Accuracy: {accuracy:.2f}, {accuracy_test:.2f}")
print(f"Precision: {precision:.2f}, {precision_test:.2f}")
print(f"Recall: {recall:.2f}, {recall_test:.2f}")
print(f"F1 Score: {f1:.2f}, {f1_test:.2f}")

# Model Optimization with Grid Search
param_grid = {
    'n_estimators': [50, 100, 150],
    'max_depth': [10, 20, 30],
    'min_samples_split': [2, 5],
    'min_samples_leaf': [1, 2],
    'bootstrap': [True, False]
}

grid_search = GridSearchCV(estimator=model, param_grid=param_grid, cv=5, scoring='accuracy', n_jobs=-1)
```

```
grid_search.fit(X_train, y_train)
best_model = grid_search.best_estimator_
print("Best Parameters:", grid_search.best_params_)

# Evaluate optimized model on test set
y_test_pred_optimized = best_model.predict(X_test)
conf_matrix_test_opt = confusion_matrix(y_test,
    y_test_pred_optimized)

plt.figure(figsize=(7, 5))
sns.heatmap(conf_matrix_test_opt, annot=True, fmt='d', cmap=
    'Blues', cbar=False, xticklabels=classes, yticklabels=
    classes)
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix (Optimized Test)')
plt.show()

# Evaluation metrics for optimized model
accuracy_opt = accuracy_score(y_test, y_test_pred_optimized)
precision_opt = precision_score(y_test,
    y_test_pred_optimized, average='macro')
recall_opt = recall_score(y_test, y_test_pred_optimized,
    average='macro')
f1_opt = f1_score(y_test, y_test_pred_optimized, average='
    macro')

print(f"Optimized Accuracy: {accuracy_opt:.2f}")
print(f"Optimized Precision: {precision_opt:.2f}")
```

```
433 print(f"Optimized Recall: {recall_opt:.2f}")
434 print(f"Optimized F1 Score: {f1_opt:.2f}")
```

Listing 6.1: Python Code for Data Import, EDA, Preprocessing and Modelling