UNIVERSITY OF THE WITWATERSRAND

FACULTY OF HEALTH SCIENCES

SCHOOL OF PUBLIC HEALTH

DIVISION OF EPIDEMIOLOGY AND BIOSTATISTICS

MSc Report

Unsupervised Machine Learning Clustering of Prostate Cancer Cases in Soweto Using Clinical, Demographic, and Socioeconomic Variables (PSA, GS, TNM).

By

Tanaka Rwafa

Student Number- 2606612

Supervisors: Dr. Michael T. Mapundu
Co-Supervisor: Dr. Wenlong Carl Chen

This MSc Report was submitted to the Faculty of Health Sciences, University of the Witwatersrand, in partial fulfilment of the requirements for the Master of Science degree majoring in Public Health Informatics at the School of Public Health.

November 6, 2025

Declaration of Authorship

I, Tanaka Rwafa, declare that this research report titled, *"Unsupervised Machine Learning Clustering of Prostate Cancer Cases in Soweto Using Clinical, Demographic, and Socioeconomic Variables (PSA, GS, TNM)"*, 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 fulfilment 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: November 6, 2025

Abstract

Introduction: Prostate cancer (PCa) is one of the leading cause of morbidity among men worldwide, and men of African descent experience significantly higher incidence, prevalence, and mortality. Traditional diagnostic methods, for PCa typically depend on specific biomarkers, such as, prostate specific antigen (PSA). These are gathered through procedures, such as the digital rectal exam (DRE). DRE is non-invasive, dependent on clinician's expertise, thus highly subjective, often leading to variability in detection rates. Furthermore, the excessive cost and time associated with the procedures often create barriers to timely diagnosis, particularly in low income countries (LIC). This highlights the need for innovative, cost-effective, and scalable approaches to PCa diagnosis. Traditional diagnosis methods often overlook the wealth of information in high-dimensional data collected from cases over time, such as PSA, Gleason scores (GS), and socio-demographic data. The primary objective of this study was to investigate factors associated with PCa morbidity among men of African descent using unsupervised machine learning (UML) techniques. UML offers a promising approach by identifying correlations, patterns, and risk profiles in complex datasets, which facilitates faster, precise and cost effective diagnostics.

Methods: This study utilized UML techniques to address these problems, using a robust dataset from the Men of African Descent and Carcinoma of the Prostate (MADCaP) study, conducted in Soweto, South Africa; during the survey period 2016 to 2020. Stratification of cases into clinically meaningful groups was performed using a combination of key prognostic variables: GS, $PSA\ level$, and TNM stage. The TNM criteria included primary Tumour extent (T), regional lymph node involvement (N), and the presence of distant metastasis (M). This multi-factorial approach was designed to elucidate distinct risk profiles and generate novel insights into PCa heterogeneity. Key methodologies included data pre-processing, clustering using advanced UML algorithms, and evaluation using the silhouette score to validate the robustness of the clusters. Applying clustering algorithms including hierarchical, K-means, and spectral clustering, this research sought to uncover hidden correlations and stratify patients into meaningful risk groups. This approach not only complements traditional diagnostic methods, but also provides a pathway for more efficient, accessible, and personalized PCa management, particularly in populations with limited access to advanced healthcare resources.

Results: The dataset comprised clinical, behavioural, and demographic information for $(n = 1,096)$ PCa cases. Case ages ranged from $40 - 94$ years, with a median age of 67 $(IQR : 62 - 94)$. Based on GS, 36.6% of cases were classified as high-risk, 52.9% as intermediate-risk, and 10.8% as low-risk. Reported cigarette consumption varied widely, reaching up to 60 cigarettes per day, with a median of 8 $(IQR : 5 - 8)$. Clustering analysis revealed clinically significant subgroups, underscoring the combined influence of GS, PSA levels, and TNM staging on PCa severity.

These findings underscore the utility of UML in case stratification and highlight its potential to inform risk-adaptive clinical guidelines. To validate the clustering models applied to the PCa dataset, the Silhouette score was employed. This metric evaluates both intra-cluster cohesion and inter-cluster separation, with values ranging from –1 to $+1$. Scores approaching $+1$ indicate well-defined and clearly separated clusters, values near 0 suggest overlapping clusters, and values approaching –1 reflect poor or incorrect clustering assignments.

Conclusion: The results of this study demonstrate the value of applying UML to large-scale clinical datasets, providing a pathway to uncover underlying subgroups that may enhance clinical decision-making, support personalized PCa management strategies, and inform policymaking. Future research should aim to validate these PCa clusters using supervised machine learning approaches and incorporate additional dataset features to further explore their implications for treatment outcomes.

Acknowledgments

The completion of this master's thesis has been one of the most challenging yet rewarding academic undertakings of my life. It would not have been possible without the unwavering support, patience, and guidance of many individuals to whom I owe my deepest gratitude.

First, I extend my heartfelt appreciation to my academic advisors, Dr. Michael T. Mapundu and Dr. Wenlong Carl Chen, for their invaluable insights, encouragement, and unwavering support throughout this project.

I am deeply grateful to Tatenda Winston Wushoma (a dear friend, who has since passed away), on August 8^{th}, 2016, said, "I see you doing health based IT" which has always stayed with me. I would also like to thank my sisters, Tatenda Rwafa and Teurai Rwafa-Ponela, and my brother-in-law, Ronald Ponela, for their prayers and constant encouragement. Their belief in me to give me this opportunity to embark on this journey and their reminder that "all things are possible" kept me motivated.

A special thank you goes to my parents, Mr. Justen Rwafa and Mrs. Marjorie Rwafa, for being my pillars of strength and a source of comfort during challenging times. Their love and support mean the world to me.

Finally, my utmost gratitude goes to the Lord God Almighty for His endless provision, guidance, and comforting words during this journey. His presence has been my anchor through the toughest moments.

To anyone whose name I may have inadvertently omitted, please know that your contribution has not gone unnoticed. I greatly appreciate your role in helping me reach this milestone.

Contents

List of Figures

List of Tables

List of Abbreviations

AC	**A**gglomerative **C**lustering
AI	**A**rtificial **I**ntelligence
BD	**B**ig **D**ata
BMI	**B**ody **M**ass **I**ndex
CT	**C**omputerized **T**omography
CHBAH	**C**hris **H**ani Baragwaneth **A**cademic **H**ospital
DHRC	**D**'Amico's **H**igh **R**isk **C**riteria
DBSCAN	**D**ensity **B**ased **S**patial **C**lustering **A**pplication **N**oise
EDA	**E**xploratory **D**ata **A**nalysis
EHR	**E**lectronic **H**ealth **R**ecords
EAU	**E**uropean **A**ssociation of **U**ology
GS	**G**leason **S**core
HC	**H**ierarchical **C**lustering
IDE	**I**ntegrated **D**evelopment **E**nvironment
LIC	**L**ow **I**ncome **C**ountries
LDA	**L**inear **D**iscrimination **A**nalysis
ML	**M**achine **L**earning
MRI	**M**agnetic **R**esonance **I**maging
MADCaP	**M**en **A**frican **D**escent **C**arcinoma of the **P**rostate
NCD	**N**on **C**ommunicable **D**iseases
PCa	**P**rostate **C**ancer
PCA	**P**rincipal **C**omponent **A**nalysis
PSA	**P**rostate **S**pecific
REDCaP	**R**esearch **E**lectronic **D**ata **C**apture
RL	**R**einforcement **L**earning
SEER	**S**urveillance **E**pidemiology **E**nd **R**esult
SSL	**S**emi **S**upervised **L**earning
SML	**S**upervised **M**achine **L**earning
SC	**S**pectral **C**lustering
SL	**S**upervised **L**earning
SSA	**S**ub **S**aharan **A**frica
SAUA	**S**outh **A**frican **U**rological **A**ssociation
SAPCS	**S**outh **A**frican **P**rostate **C**ancer **S**tudy
ISUP	**I**nternational **S**ociety of **U**rological **P**athology
t-SNE	**t**-Distributed **S**tochastic **N**eighbor **E**mbedding
TURP	**T**rans-**U**rethra **R**esection of the **P**rostate
UML	**U**nsupervised **M**achine **L**earning

Chapter 1

INTRODUCTION

Chapter 1 provides an introduction and overview of this research project, focusing on the application of big data (BD), Artificial intelligence (AI), machine learning (ML) and unsupervised machine learning (UML) in prostate cancer (PCa) research. This study is part of a larger research within men of African descent carcinoma of the prostate (MADCaP). It highlights the urgent need to innovate screening, diagnostic, and prognostic approaches to address the morbidity and mortality posed by the disease.

This chapter outlines the purpose of the research, which is to promote PCa research integrating BD and UML techniques to analyse PCa data from the MADCaP database of men in sub-Saharan Africa (SSA), find hidden patterns, and stratify cases into clinically defined risk groups for severe PCa. The aim is to contribute to improving public health outcomes in underprivileged societies within the greater part of Africa. Specific objectives are described that indicate how the integration of innovative technology on PCa may assist precision medicine and patient care to improve public health outcomes. This research represents a step toward addressing a gap in understanding PCa to improve the outcomes of PCa in men.

1.1 Background

Cancer remains one of the leading causes of morbidity and mortality worldwide, with nearly 20 million new cases reported in 2022 [1]. Recent estimates from the Global Cancer Observatory (GLOBOCAN) indicate that the most frequently diagnosed cancers were lung cancer $40.9\ per\ 100,000$, breast cancer $209.0\ per\ 100,000$, colorectal cancer $73.2\ per\ 100,000$, and prostate cancer PCa; $126.7\ per\ 100,000$ [2]. Cancer is broadly defined as a group of diseases characterised by uncontrolled division of abnormal cells, which can affect various areas of the body [3]. The spread of cancer within the body, known as metastasis, is a critical factor in disease progression. Cancer cells invade tissues beyond their site of origin, metastasis accounts for the majority of cancer-related mortality [4].

The ability of cancer to spread depends on multiple factors, including the type and stage of the disease. Once metastasis occurs, the invasion of new tissues disrupts normal organ function, contributing substantially to the morbidity and mortality associated with cancer [4]. Understanding the molecular and genetic mechanisms underlying cancer spread, as well as the processes involved in the establishment of Tumour cells at secondary sites, is essential for the development of targeted therapies and for improving morbidity and mortality outcomes [5].

In 2022, Sub-Saharan Africa (SSA), with an estimated population of 1.23 billion, reported a cancer incidence rate of $127.8\ per\ 100,000$ and a mortality rate of $88.2\ per\ 100,000$, highlighting cancer as a significant public health concern in the region [6]. The most prevalent cancers in the region were breast cancer $54.5\ per\ 100,000$, cervical cancer $46.0\ per\ 100,000$, PCa $27.4\ per\ 100,000$, and liver cancer $5.7\ per\ 100,000$ [7]. The increasing cancer burden in Africa can be attributed to several risk factors, including rising life expectancy, urbanization, lifestyle changes, and limited awareness campaigns on prevention and early detection [8]. For example, disparities in cancer awareness programs often favour urban areas, leaving rural districts underserved [9].

PCa demonstrated the highest incidence rate among men across 40 SSA countries, with an estimated $27.4\ per\ 100.000$ reported in 2022 [10]. This substantial burden positions PCa as a critical public health concern, reflecting not only improved awareness and advances in diagnostic methods, but also by demographic shifts, particularly ageing population, a well established risk factor for the disease [11]. Within the MADCaP consortium, Nigeria recorded the highest PCa incidence at $41.2\ per\ 100,000$, a figure that may be partially attributable to its large and diverse population base [3].

In 2022, South Africa reported an estimated PCa incidence of $62.0\ per\ 100,000$, indicating a substantial disease burden that may partially be attributable to more comprehensive cancer registries and advanced diagnostic infrastructure [7]. In contrast, Ghana reported an estimated incidence of $30.4\ per\ 100,000$, a figure that may reflect the impact of centralized screening initiatives [3]. Senegal's reported incidence of $27.1\ per\ 100,000$ was notably lower than that of other nations in the region. However, this apparent discrepancy is more likely to reflects constraints in healthcare access and diagnostic capacity rather than a genuinely reduced underlying risk of PCa [12].

According to the South African National Cancer Registry 2022, PCa was the most commonly diagnosed malignancy among men, with an incidence rate of $51.9\ per\ 100,000$, followed by basal cell carcinoma at $35.8\ per\ 100,000$ [2]. PCa continues to be the leading male malignancy across SSA. However, a substantial proportion of cases are diagnosed at advanced stages, frequently following metastatic progression, which significantly contributes to elevated morbidity rates [13].

Advancement in digital data collection and storage technologies have greatly enhanced the capacity for large scale aggregation and management of medical data [14]. These developments have enabled the identification of trends and patterns within medical datasets through the application of AI [15]. AI refers to a computation system capable of autonomously identifying complex data correlations and generating predictive insights through techniques such as deep learning, and robotic process automation [15]. Within the healthcare domain, AI has demonstrated substantial utility across a range of applications, including diagnostic support, personalized treatment recommendations, and optimization of administrative workflows [16].

In recent decades, rapid technological advancements, have led the widespread integration of digital systems across industries, with particularly transformative effects in the healthcare sector. One notable shift has been the transition from paper based record-keeping to Electronic Health Records (EHR), enabling the creation of large scale repositories of clinical data [17]. These repositories encompass diverse datasets, including genomic, imaging, and demographic information, enabling health organizations to leverage advanced data analytics and ML techniques in pursuit of improved population health outcomes [18,19].

The digitization of data collection and storage has led to the emergence of data science, as a pivotal discipline in data-driven decisions making, particularly through the use of descriptive and predictive analytics [20]. The primary objective of analytics is to enhance operational efficiency and improve performance by uncovering meaningful patterns within complex datasets [21]. In healthcare, the broad applicability of data science has enabled institutions to interrogate diverse clinical datasets; covering laboratory results, imaging, and patient demographics, to identify latent trends and generate predictive insights, such as estimating the probability of disease recurrence following treatment [18,21,22].

Data analysis has four key aspects, the first being descriptive analytics; this is the foundation of data insights through summarizing the frequencies of the data that allows the identification of key performance indicators [23]. Diagnostics analytics inspects the results of descriptive analysis and seeks to infer patterns of behaviour within the data [21]. Predictive analysis uses historical data within repositories to make informed predictions, through statistical modelling, ML algorithms, and data mining methods to anticipate future trends [17].

The implementation of digital health technologies, particularly longitudinal databases, has enabled the accumulation of substantial healthcare data, resulting in the emergence of big data (BD) [22]. The MADCaP study, utilizes a standardized Research Electronic Data Capture (REDCaP) system for data collection across participating institutions capturing clinical , demographic, and lifestyle variables. Key variables included were age, GS, PSA $(diagnosis\ and\ referral)$, $income$, $smoking\ history$ and $Tumour\ staging$, alongside family history of PC [24,25]. REDCaP is a secure, web-based platform systematically collects and stores demographic, clinical and treatment modalities, and outcome data

[26]. REDCaP enables real time data validation and harmonization across study sites, facilitating robust cross national data sharing while adhering to the POPIA Act and data security [13].

BD are large complex datasets whose scale and complexity exceed conventional analytical capabilities, necessitating ML [27]. ML employs adaptive algorithms that autonomously improve their performance through exposure to data, utilizing statistical models to generate predictive outputs with minimal explicit programming [22]. ML is further subdivided into subcategories: supervised learning (SL), semi-supervised learning (SSL), unsupervised machine learning (UML) and reinforcement learning (RL) [27]. Among these, UML specialises in detecting inherent patterns from a dataset in which similar or dissimilar outcomes are grouped to form a clusters [28]. This approach facilitates the extraction of knowledge from raw, unstructured data through automated pattern recognition techniques [28].

UML is a subdivision of ML that focuses on analysing and presenting data without predefined labels or outputs, through the application of clustering [17]. Clustering is a process of grouping an array of objects such that objects with similar features are placed in similar clusters [27]. The aim is to show similarities where they exist and differences where they exist [27]. Before applying ML algorithms, the data must be cleaned for impurities that can hinder the accuracy of the analysis, through data pre-processes, such as imputation and label encoding, allowing for well-rounded analysis with high predictive quality [18].

Clinical data repositories are often multi-dimensional in nature, comprising several features or attributes, for example, body mass index (BMI), age, date of birth, demographic information, allergies and vitals [21]. To reduce multi-dimensionality, preparation methods were be applied, such as, dimensionality reduction, To reduce multi-dimensionality, preparation methods were applied to reduce the number of variables within the dataset while retaining vital information [17]. This is essential for simplifying complex datasets for visualizing high dimensional data in UML methods.

This research aims to apply UML clustering techniques to MADCaP datasets, including Hierarchical Clustering (HC), Agglomerative Clustering (AC), Density-based spatial clustering of application noise (DBSCAN), K-means and Spectral Clustering (SC) techniques to uncover latent patterns and associations within PCa data [28]. These analytical methods are instrumental in identifying hidden substructures and subgroups in clinical datasets, particularly given the heterogeneous nature of PCa presentations across individual cases [16]. These observed variations underscore the necessity of isolating case specific features with clinical relevance within the clustering framework. By identifying variables that meaningfully contribute to subgroup differentiation, this approach holds promise for enhancing the precision of personalized treatment and management strategies for PCa [8].

This research employed iterative UML clustering on PCa case data without predefined input parameters, utilizing internal validity metrics to algorithmically identify similarities and dissimilarities within clusters, external validation metrics were subsequently applied to validate the accuracy and robustness of the clustering outcomes [29]. Once associations are drawn between groups for clinically relevant outcomes, MADCaP data could be used to classify cases into distinct well-defined risk groups to assist in clinical decision making [18]. UML offers a powerful, unbiased framework; by posing targeted computational queries, these algorithms can uncover subgroups that may enhance traditional hypothesis driven approaches, revealing novel insights into PCa heterogeneity [30].

1.2 Problem Statement

In SSA, the prevalence of PCa is a growing public health concern due to barriers in access to healthcare, prolonged diagnostic times, and the inaccessibility of prompt and efficient treatment, thus contributing to higher morbidity and mortality among men. Diagnosis relies on the histological assessment of tissue samples, which, while effective, maybe further improved by integrating structured and unstructured clinical (TNM staging) and laboratory data (PSA and GS) alongside demographic and socioeconomic factors for accurate PCa diagnosis.

Although existing detection methods have shown considerable efficacy. These conventional approaches fail to fully exploit the potential predictive value inherent in the multivariate relationships between clinical parameters, demographic factors and socioeconomic determinants of health. In resource constrained settings like Soweto, South Africa, challenges lie in providing timely access to these services. Furthermore, disparities in diagnostic infrastructure persist increasing PCa morbidity and mortality. Thus, there is a need to investigate the feasibility of using UML to supplement current detection methods to identify high-risk PCa cases in an African setting to improve prognosis.

UML methods could provide low cost, scalable approach for identifying high-risk PCa case clusters and improve triage efficacy. Unstructured and semi-structured PCa clinical and laboratory data are often overlooked, and the systematic examination of such data through ML may provide a greater perspective on PCa diagnosis and management and reduce PCa morbidity and mortality.

1.3 Justification

To efficiently uncover hidden patterns inside PCa datasets, five clustering algorithms were selected K-means, HC, DBSCAN, and SC each offering distinct advantages associated with the complexity of clinical and laboratory data [31]. K-Means was chosen for its computational efficiency and suitability for large scale, continuous datasets such as PSA levels, though its assumption of spherical clusters necessitates careful preprocessing [32]. HC provides a dendrogram based view of patient similarity, enabling exploration of nested relationships without requiring a predefined number of clusters [33]. DBSCAN was included for its robustness in detecting noise and outliers, which is critical when identifying rare or atypical PCa presentations [34]. SC, leverages graph based techniques, excels in capturing nonconvex structures and complex relationships in high dimensional data, such as histological data [29].

All algorithms underwent rigorous parameter tuning to optimize performance and ensure clinical relevance. For K-Means, the optimal number of clusters was determined using the elbow method, Silhouette Score [17]. HC was tuned by comparing linkage criteria ward, complete, and average to identify the most stable structures [35]. DBSCAN's and parameters were refined through k-distance plots and local density estimation to balance cluster granularity and noise detection [35]. SC affinity matrix and Laplacian eigenmaps were analysed to determine the optimal cluster count through eigenvalue gap analysis [34].

To evaluate clustering quality and clinical utility, a combination of statistical and domain specific metrics were employed. These included Silhouette Score to assess how well data points fit into assigned clusters and how distinct it is from other clusters [19]. Visualization technique PCA was applied to interpret cluster structures and validate separability [36]. Together, these methods ensure that the clustering outputs are not only statistically robust but also clinically actionable, supporting early diagnosis, risk stratification, and personalized treatment planning in PCa management [37].

1.4 Study Aim and Objectives

1.4.1 Research question

Can unsupervised ML on routine clinical/demographic data (PSA, GS, TNM, age, income, smoking, family history) identify data-driven clusters aligned with or extending guideline risk strata?

1.4.2 Aim

The aim of this study was to identify factors associated with prostate cancer severity among men of African descent by apply unsupervised machine learning techniques on unlabelled prostate cancer clinical data to identify prostate cancer morbidity.

1.4.3 Objectives

The specific objectives of this study were to:

1. Apply unsupervised machine-learning to identify high-risk prostate cancer clusters based on demographic (age), socioeconomic (income, smoking history), prostate specific antigen (at referral and diagnosis), digital rectal examination (tumour and metastasis staging), family history and prostate biopsy Gleason score.

2. Explore how unsupervised machine learning can detect clinically meaningful clusters that support patient risk stratification and tailored management.

3. Evaluate algorithmic performance using quantitative cluster validity measures, such as silhouette analysis.

1.5 Summary

This section outlines the global health burden of PCa, highlighting epidemiological trends and significant healthcare disparities. It emphasizes that men of African descent face disproportionately high incidence rates, more aggressive disease, and worse outcomes. These disparities underscore the urgent need for targeted research and interventions. The analysis also notes that technological and big data advances are enabling more comprehensive PCa research, particularly in Sub-Saharan Africa.

Chapter 2

LITERATURE REVIEW

This literature review provides a thorough examination of existing research and knowledge associated with PCa, with a specific focus on its epidemiology, risk factors, and progression. We explore recent progress in data science approaches, including the application of UML techniques, to address challenges in clinical data analysis. This section highlights the intersection of AI, BD, data analysis and PCa research, emphasizing the possibility of UML to uncover hidden relationships, stratify patients, and inform clinical decision-making. By combining past studies, this review ascertains the basis and justification for leveraging UML in investigating PCa data, addressing gaps in current methods, and evolving analytical and diagnostic capabilities.

2.1 Epidemiology of Prostate Cancer

PCa is the second most prevalent malignancy among men globally, with an incidence of $91.5\ per\ 100,000$ in men over 39 years and a mortality rate of $22.6\ per 100,000$ [2]. In SSA, the incidence and mortality rates were substantially higher in 2022, at $113.7\ per\ 100,000$ and $68.4\ per\ 100,000$, respectively, with PCa representing the highest incidence among men across 40 SSA countries [10]. These figures reflect a significant increase in cancer incidence, reaching levels in 2022 that were initially projected to occur only by 2040 [10].

PCa mortality is strongly associated with age, particularly among men aged between $60 - 70$ years [8]. Additionally, African American men exhibit higher incidence rates and significantly greater mortality compared to men of other ethnic backgrounds [6]. Recent epidemiological research indicates a change in PCa diagnosis patterns, with the average age at detection reported at 45 years [4]. These findings, supported by data from the GLOBOCAN project; a comprehensive repository of global cancer statistics covering 185 countries and 36 cancer types; suggest a changing demographic landscape of the disease [6,7]. Disparities in PCa incidence and outcomes, including such epidemiological shifts, are largely attributed to complex interactions among social, dietary, environmental, and genetic factors [38]

2.2 Histopathologic Yields and Concordance of Prostate Cancer Diagnosis

PSA is a serine protease produced by the epithelial cells of the prostate gland and is widely used as a biomarker for PCa screening, diagnosis, and monitoring [25]. In clinical practice, serum PSA levels are interpreted alongside other diagnostic modalities, such as DRE and prostate biopsy, rather than in isolation [39]. According to the European Association of Urology (EAU) and the South African Urological Association (SAUA), PSA levels are stratified as follows: low-risk, < 10ng/mL; intermediate-risk, $10 - 20$ ng/mL; and high-risk, > 20 ng/mL, to guide the diagnosis of adverse disease features [40].

Interpretation of PSA levels must account for biological and analytical variability, age-specific reference ranges, and the influence of non-malignant conditions such as benign prostatic hyperplasia and prostatitis, which can also elevate PSA levels [40]. In research settings, PSA values are often analysed as continuous variables; however, categorizing PSA into clinically relevant risk groups facilitates direct comparison with established clinical guidelines [39]. The integration of PSA classifications with established clinical pathological parameters, notably clinical stage and International Society of Urological Pathology (ISUP) grade, significantly enhances their prognostic utility [4]. This approach facilitates a more precise and individualized assessment of biochemical recurrence risk, moving beyond the limitations of PSA level alone [41].

Within specific diagnostic zones, the free to total PSA ratio provides critical discriminatory value. Particularly when total PSA levels fall within the borderline range of 4–10 ng/mL, a lower percentage of free PSA is strongly correlated with an increased probability of malignancy, thereby serving as an invaluable diagnostic tool [42]. It is imperative to contextualize these diagnostic thresholds within specific populations. The SAUA acknowledges the established benchmarks while emphasizing the necessity for additional considerations [43, 44]. This is due to the unique epidemiological profile of the African demographic, which includes a higher incidence and potentially more aggressive disease biology among Black South African men, factors that may influence risk stratification and clinical decision-making [44].

2.3 Clinical Staging Systems in Prostate Cancer: Gleason Score, ISUP Grade Groups, and TNM Classification

The GS system remains a cornerstone of histological grading for prostate adenocarcinoma. It is determined by summing the primary (most prevalent) and secondary (second-most prevalent) architectural patterns, each scored from 1 to 5 based on glandular differentiation. [45]. ISUP Grade Groups (GG), provide enhanced prognostic stratification for PCa [25]. These GG align with low, intermediate, and high-risk disease categories and hold substantial prognostic value for predicting biochemical recurrence, metastasis, and survival. Incorporating GS or ISUP grade groups into data-driven clustering ensures alignment with clinically validated risk stratification system [39]. Table 2.1 shows ISUP GG mapped from GS.

Table 2.1: ISUP Grade Groups mapped from Gleason scores

ISUP Grade Group	Gleason Score
1	≤ 6 (3+3)
2	7 (3+4)
3	7 (4+3)
4	8 (4+4, 3+5, 5+3)
5	9–10 (4+5, 5+4, 5+5)

Table 2.2: TNM Classification for Prostate Cancer (based on DRE assessment)

T – Primary Tumour (stage based on digital rectal examination only)	
TX	Primary tumour cannot be assessed
T0	No evidence of primary tumour
T1	Clinically inapparent tumour that is not palpable
T1a	Tumour incidental histological finding in 5% or less of tissue resected
T1b	Tumour incidental histological finding in more than 5% of tissue resected
T1c	Tumour identified by needle biopsy (due to elevated PSA)
T2	Tumour that is palpable and confined within the prostate
T2a	Tumour involves one half of one lobe or less
T2b	Tumour involves more than half of one lobe, but not both lobes
T2c	Tumour involves both lobes
T3	Tumour extends palpably through the prostatic capsule
T3a	Extracapsular extension (unilateral or bilateral)
T3b	Tumour invades seminal vesicle(s)
T4	Tumour is fixed or invades adjacent structures other than seminal vesicles: external sphincter, rectum, levator muscles, and/or pelvic wall
N – Regional (pelvic) Lymph Nodes[1]	
NX	Regional lymph nodes cannot be assessed
N0	No regional lymph node metastasis
N1	Regional lymph node metastasis
M – Distant Metastasis[2]	
M0	No distant metastasis
M1	Distant metastasis
M1a	Non-regional lymph node(s)
M1b	Bone(s)
M1c	Other site(s)

[1] Nodal metastasis no larger than 0.2 cm can be designated pNmi.
[2] When more than one site of metastasis is present, the most advanced category is used. (p)M1c is the most advanced category.

The TNM system provides anatomical staging based on three components. T (Tumour): extent of the primary tumour in the prostate and surrounding structures (T1c – Tumour identified by needle biopsy due to elevated PSA; T2 – Tumour confined within the prostate; T3 – extracapsular extension; T4 – invasion of adjacent structures) [4]. N (Node): regional lymph node involvement (N0 – no regional node metastasis; N1 – metastasis in regional lymph nodes). M (Metastasis): presence of distant metastasis (M0 – none; M1 – distant spread, including bone) [43].

The EAU guidelines recommend combining PSA, ISUP grade group, and TNM stage to classify patients into low, intermediate, or high-risk disease classification for treatment planning [40]. In exploratory ML, these variables serve as robust, clinically interpretable features that can be cross validated against data-driven clusters for their potential to uncover prognostic insight [23]. TNM staging is universally applied, including in South African where guidelines also emphasize the need for context based specific management, particularly in low resource setting where advanced stage diagnosis is frequent [46]. This integrated approach plays a critical role guiding decisions and prog-

nostic assessment. Table 2.2 shows the TNM classification applied in the study.

2.4 Disparities in Prostate Cancer

The largest global health disparity exists in PCa, with black men having a two-fold increased risk of developing PCa compared to other ethnic races [47]. These disparities stem from a complex interplay of genetic, environmental, and socioeconomic factors. Additionally, PCa tends to be diagnosed at younger ages among Black men, further compounding the burden of disease in this population [48]. In the United States of America black men are disproportionately affected by PCa, with an incidence rate of $185,5\ per\ 100,000$, a higher burden of PCa compared to Caucasian men with $108.0\ per\ 100,000$, and a mortality rate of 2.2 times higher [48].

A meta-analysis of a randomized clinical trial showed that black men have a median age of 68 years compared to 71 in Caucasian men, underscoring the susceptibility of black men to PCa [48]. Geographical disparities further exacerbate these inequalities. The southeastern states of America are characterized by black populations and a high poverty rate, reported elevated PCa-related morbidity and mortality, and similar patterns are observed in SSA in black men more than other ethnic groups [49].

Historical exclusion of black men and systemic racism have led to distrust of medical systems within black communities, resulting in lower screening and diagnosis in men of African descent [49]. Addressing systematic issues, such as access to healthcare and socioeconomic inequalities, through policy changes and community-based interventions will promote equity and improve disease outcomes [48].

In view of what has been mentioned so far, one may suppose that there is an urgent need for solutions to bridge racial gaps in PCa care. By addressing the various causes of these disparities, such as biological, access to healthcare, and systematic inequalities, healthcare providers and policymakers may be able to work toward equitable outcomes for black men.

2.5 Prostate cancer Disparities and Management South Africa

Disparities in PCa outcomes arise from a complex interaction of medical practices, cultural norms, and systemic healthcare constraints [48]. In the African context, clinicians face challenges in implementing screening and treatment strategies that adequately reflect region specific risk profiles. At the same time, cultural values; particularly norms surrounding masculinity, often discourage men from seeking timely medical care [8]. The Southern African Prostate Cancer Study (SAPCS) provided valuable insights into these dynamics by investigating both epidemiological and social determinants of health among $1,387$ men aged 40 to 107 years as part of a multi-ethnic study of PCa [45].

The SAPCS cohort consisted of 741 PCa cases (53.4%) and 505 controls (36.4%). The ancestral distribution among cases was highly reflective of the South African population, with 78.1% identified as Black South African, 8.2% as admixed/Asian, and 9.2% as European ancestry. Among controls, 75.8% were Black South African, 8.3% admixed/Asian, and 10.5% European ancestry [50]. Black South African men were significantly more likely to present with advanced disease, with (ISUP$\geq$,4:

OR = 2.25, 95% CI 1.49–3.40 and ISUP≥3: OR = 2.02, 95% CI: 1.41–2.90) [50]. Furthermore, residence in provinces with high poverty rates was associated with increased likelihood of advanced disease (ISUP≥4: OR = 1.51, 95% CI: 1.07–2.13). In contrast, men of admixed/Asian ancestry were less likely to present with ISUP ≥3 compared to men of European ancestry [45].

2.6 Contemporary Evaluation of the D'Amico Risk Classification of Prostate Cancer

One of the most significant events of the was the formation of D'Amico's model for stratifying PCa patients into those with low, intermediary and high-risk of biomedical recurrence after surgery [51]. A high incidence of PCa, paired with high mortality emphasized the need for accurate PCa risk stratification due to the increased number of treatment options that were being made available [51]. The prediction model was based on clinical tumour stage (t-stage), PSA and GS [52]. Patients with clinical T1c-T2a, a PSA of 10 ng/mL , and GS of less than or equal to 6 were regarded as low risk, those with clinical stage T2b disease or PSA of $10.1$to 20 ng/mL or biopsy GS 7 were intermediate risk, and those with clinical stage T2c or PSA ≥ 20 ng/mL or biopsy GS 8 to 10 were high-risk [52]. D'Amico's classification remains key to stratifying PCa risk. Part of the aim of this project is to train UML models on PCa stratification of newly diagnosed patients based on the heterogeneity that may exist due to the risk criteria they may harbor [53]. However, D'Amico's classification did not consider differences in PCa phenotypes according to ethnicity; a large growing body of literature has investigated the susceptibility of diverse ethnic groups [53].

2.7 Exploring Ethnic Disparities in the Distribution and effect and Distribution of High-Risk Prostate Cancer undergoing Radiotherapy

Investigating disparities in the PCa phenotype based on ethnicity is a continuing concern within public health. This article identified $31,002$ PCa patients treated using D'Amico's high-risk criteria (DHRC); $20,894$ were Caucasian, $5,256$ African Americans, $2,868$ Hispanic-Latino and $1,984$ Asian [53]. PCa data was stored within the surveillance, epidemiology, end result (SEER) database consisting of approximately 26% from the United States of America in terms of demographic composition aged ≥ 18 years with histologically diagnosed PCa [51]. The researchers aimed to evaluate the variations in the distribution of DHRC and their effect on cancer specific mortality among PCa patients treated with external beam radiography and stratified by ethnicity [51]. Although DHRC has been widely applied in clinical decision making, there is a partial comprehension of how ethnicity influences the distribution and impact of DHRCs [53]. A retrospective analysis was utilizing SEER, and key findings of the analysis revealed significant differences in the distribution of DHRC among the ethnic strata [54]. The findings of the study provide valuable insights into the complex contribution of ethnicity, DHRC distribution, and cancer-specific mortality in high-risk PCa patients experiencing radiotherapy [45]. This knowledge can assist in targeted treatment strategies to deliver evidence-based healthcare solutions tailored to an individual, therefore contributing to efforts aimed

at addressing disparities in PCa care [30].

2.8 Men of African Descent and Carcinoma of the Prostate

The MADCaP consortium was formed as an international, multi-centre, and hospital-based PCa research, collaboration to study the genetic aetiology of PCa in men of African descent [13]. The MADCaP study focused on researching PCa in men of black ethnic origin residing in the same catchment [55]. The MADCaP network was established to encourage scientific research, knowledge and data exchange among the seven participating centres [13].

The MADCaP study was a case-control study. The cases were men newly diagnosed with PCa at one of the seven participating urological clinics and controls where PCa-free men of matching ancestry, age (within 5-year age groups), and geographical location [56]. The focus was to compare genetic and epidemiological factors between the cases and controls to identify PCa associated risk factors [13]. The seven participating study sites included Urological centres in Nigeria, Senegal, Ghana and South Africa [56].

Suitable participants, both in cases and controls, provided information on demographics, anthropometric, physical activity, lifestyle, medical and family history [56]. Surveys and medical records were captured using REDCaP. A total of $1,000$ PCa cases and $1,000$ ethnic and age-matched population controls were enrolled from Soweto for this study [13]. A map of all participating MADCaP centres is shown in Figure 2.1.

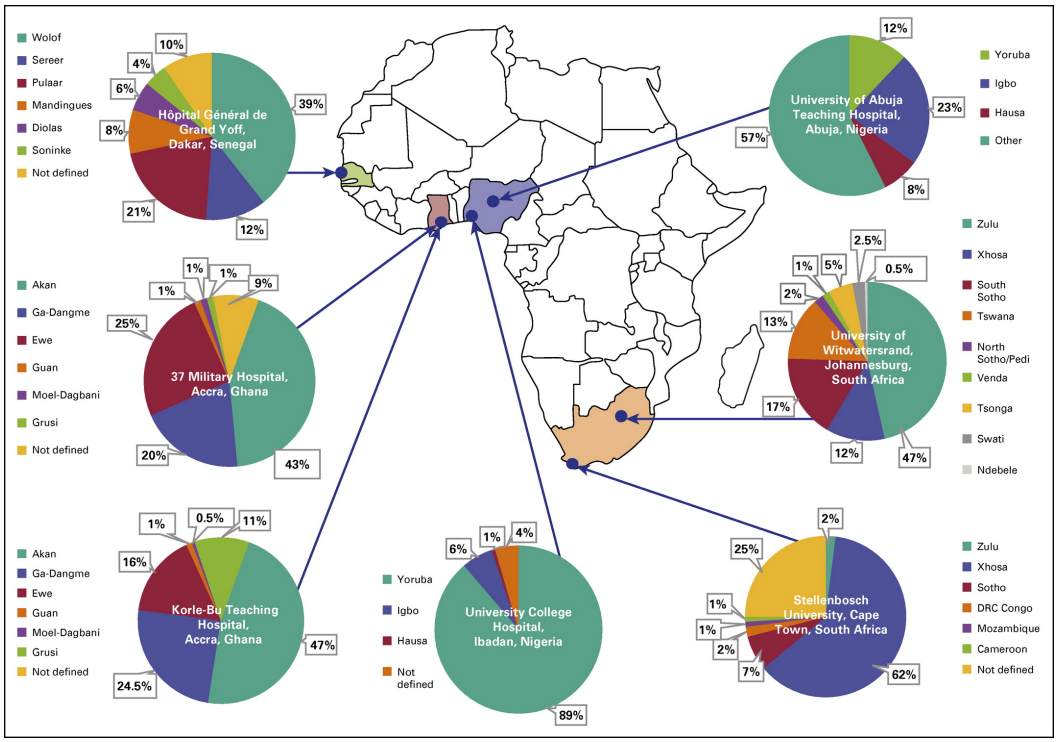

Figure 2.1: *Population allocation of the Black African men registered at multiple African and implementation centres, highlighting ethnic composition of both cases and controls across various MADCaP centres*

2.9 Big Data and Artificial Intelligence in Public Health

BD refers to large, complex data characterised by volume, velocity, variety and veracity [57]. In Healthcare, these datasets are generated from diverse sources such as, the internet and health surveillance networks. The complex, often unstructured nature of this data, includes systematic semantic, social and economic interrelationships, renders conventional analytical techniques inadequate [58, 59]. Consequently, managing and extracting value from BD requires advanced computational techniques to handle, most notably those in the field of AI. [57].

Managing BD is crucial in healthcare as a growing field that continues to evolve and influence health measures [18]. The quality BD analysis can be significantly hindered by challenges such as data entry errors, duplicated or incomplete records, and noise pollution [60]. Therefore, robust data management and preprocessing are essential steps before any meaningful analysis can be run [61, 62]. The global landscape is increasingly emphasizing on BD within the internet domain, parallel the steady rise of AI in healthcare diagnostics and decision-making [35].

The relevance of BD is supported by current findings that provide a comprehensive analysis of the fundamental concepts and application of AI in the field of Public Health, with a detailed overview of the theoretical underpinnings of AI, ML algorithms and data modelling techniques [63,64]. Additionally, it highlights the potential brought by BD in healthcare research, such as improved disease monitoring and prevention, precision medicine, and epidemiological surveillance [65].

2.10 Data Science in Action

Data Science in recent years has emerged as a new and important field of study. Data science is a combination of disciplines such as statistics, data mining, databases, and distributed systems [21]. Data science methods consist of combining data from a variety of sources to encourage a large, complex database for individuals, corporations, and society [37]. This contains linking principles, procedures, and techniques for understanding a phenomenon by way of automating methods of exploration of data to improve decision making in organisation and health institutions [18].

As explained in Chapter 1, civilisation has shifted from primarily paper based or analogue storage to digital systems, such as databases [66]. Data science can become a cornerstone of public healthcare, disease surveillance and policy delivery [51]. Researchers have demonstrated the use of data science in analysing large data, emphasizing the importance of predictive analysis for gaining a competitive advantage in industries such as finance, retail, and manufacturing [36]. These studies have shown the transition from conventional statistical approaches to a dynamic ML model that adapts to the ever-changing data and technology environments [18].

2.11 Data Analysis in Health

Data analysis has become an integral component across industries, and healthcare is the primary poised domain to benefit from its application [18]. The use of EHRs, statistical techniques, and global health surveillance systems has resulted in BD has enabled researchers to be able to extract meaningful insights from the health data [28].

This development has led to significant advancements in the utilisation of data analysis for predicting, managing and implementing policy [51]. However, despite these developments there remain areas that are still unaddressed, principally in the situation of addressing health challenges in low-income nations and tackling health [17]. Much of the present research application focuses on high-income countries leaving low-income nations undeserved.

Data analysis is characterized into descriptive, prescriptive, and predictive analytics [23]. In descriptive analytics the focus is on summarizing data to provide insights or outlines on trends. Prescriptive provides recommendations by examining potential outcomes that may arise with interventions; this may play a pivotal role in personalized care and resource allocation [23]. Predictive data analytics utilize historical data to make future predictions such as treatment outcomes, disease outbreaks, and remissions [17].

While data analysis has made notable headway in healthcare, its application remains focused on infectious diseases, while non-communicable disease's (NCD) have received less attention contributing to the high burden of disease [8]. Moreover, current data analysis models are developed for structured data, leaving unstructured and semi-structured data such as clinical notes and social determinants of health underutilised [11]. These data have the potential to understand disease progression, treatment response, and patient outcomes.

The application of data analysis within healthcare has transformative potential and limitations, while there have been groundbreaking developments in understanding and managing disease [58]. There are inequities in the application of data analysis in low-income nations and addressing these disparities will require investment into innovation and integration of analytical methods to prioritize the development of appropriate solutions for improved disease mitigation in underdeveloped countries, thus advancing global health outcomes [27].

2.12 Leveraging Artificial Intelligence to Improve Cancer Detection

AI represents a computational model designed to emulate human cognitive functions, including complex decision making and problem-solving capabilities, through advanced algorithmic processes [16]. AI field encompasses several specialized domains, such as deep learning (DL), neural networks (NN), and machine learning (ML) [34]. Recently, AI has been integrated into many aspects of oncology demonstrating transformative potential, with significant implementation in areas such as cancer screening, gene mutation identification, MRI-biopsy imaging, and, most recently, the acceleration of drug detection [15].

AI frameworks are employed to systematically analyse structured structured PCa records, facilitating data-driven predictions for optimal patient treatment strategies [34]. In radiology, for instance, AI-assisted images processing enhancing efficiency, allowing clinicians to focus on complex aspects of diagnostic tasks requiring expert judgement [63]. However, unstructured and semi-structured datasets are often underutilised, resulting in an abundant wealth of information left unexplored that has the potential to give deeper meaning to cancer diagnoses [29].

By analysing large volume of PCa patient data, AI-based solutions may provide a comprehensive and automated diagnosis, minimizing the risk of human error and improving the quality of patient care [15]. In addition, AI can help public health professionals identify measures aimed at early diagnosis and prompt treatment to prevent PCa progression and complications, thus reducing PCa prevalence [51]. The integration of AI into PCa detection represents a promising direction for the future of PCa diagnostic methods and has the potential to revolutionize the field of Public Health [67].

2.13 Harnessing Big Data in Prostate Cancer Translational Research

BD refers to exceptionally large and complex datasets that are beyond the understanding of the human mind; necessitating computational tools for processing and interpretation [31]. A critical challenge lies in the standardization and integration of data collected from multiple institutions, for instance, in the case of MADCaP, where diverse data types ranging from images and texts, on PCa diagnosis and treatment are stored [13]. To unlock BD's potential, researchers seek to discover disparities in PCa burden and outcomes across different stages of progression [60].

BD applications in PCa research remain at an early developmental stage, with significant unexplored potential in this domain [16, 21]. Thus, BD initiatives have strategically prioritised crucial analysis on PCa leading to conceptual advances in our understanding of cancer biology impacting PCa diagnosis and treatment [60]. As the application of BD in PCa is relatively new, there remains a plethora of opportunities for exploration. Overall, there seems to be some evidence to indicate that BD may further advance critical analysis fundamental concepts associated with PCa biology [59].

2.14 Machine Learning

ML is a rapidly growing subfield of AI where the computer-based scientific application of algorithms and statistical modelling techniques are utilized to make predictions from BD [28]. Specifically, it has been adopted in the healthcare industry due to its ability to process and analyse BD produced by surveillance systems and EHR [68]. Collected from EHR, has led to supervised machine learning algorithms being trained and applied to labelled data to predict the future using historical data [18]. The potential of ML in healthcare stems from the diverse structure of healthcare data, which also represents a substantial challenge in modelling precise and valuable healthcare ML models [69].

UML utilises clustering algorithms inputted with unlabelled data to find similar clusters of data points. Clustering assists when nothing about the data is fully known, and exploring the data helps in uncovering patterns that stand out about the dataset [28]. ML in summation, maybe an invaluable apparatus for up to the minute data analysis, providing public health researchers with effective predictive and exploratory abilities [70].

2.15 Unsupervised Machine Learning on Prostate Cancer

UML is uniquely suited for identifying new patterns and relationships in complex, unlabelled datasets, such as those commonly found in EHRs [28]. This is particularly relevant for conditions such as PCa were identification of novel patterns and associations in irregularly stored data can reveal essential insights for treatment and prognosis [23, 42]. By utilizing UML techniques to uncover hidden substructures within data and classification of potential biomarkers can improve accuracy of prognostic models and treatment plans [17].

While UML models do not provide complete accuracy due to the lack of labelled outcomes and are often used as a pre-processing step for supervised learning, they remain an effective and essential instrument for the exploration analysis of new patterns and associations in PCa research [69]. In the context of PCa, clustering techniques enable grouping cases according to risk factors, and disease characteristics [19]. Such grouping provides clinically meaningful stratification that may not be apparent through traditional risk models [28]. To support UML, dimensionality reduction methods are employed to simplify high dimensional data, thereby improving computational efficiency and enhancing the interpretability of results through clear visualization [36].

2.16 Summary

The literature review emphasizes two central themes: the significant global burden of PCa pronounced disparities in incidence and morbidity for men of African descent and the transformative role of technology (AI, BD, UML) in addressing these challenges. Specifically, UML provides powerful tools for analysing complex datasets without predefined labels. Techniques such as clustering and dimensionality reduction are being used to identify novel patient subtypes, analyse biopsy interactions, and improve risk stratification based on integrated clinical and genetic information.

Chapter 3

METHODS

3.1 Research Approach

In this section, we discuss the methods that were applied to the MADCaP dataset in conjunction with UML analysis. We detail the criterion of participant inclusion and exclusion criterion, data preparation process, encompassing pre-processing techniques, feature selection, and dimensionality reduction techniques, with specific algorithms utilized to discover hidden correlation within the MADCaP data. The methodological steps were taken to ensure reproducibility and offer clarity on how the analysis was systematized to report the objectives of the research dissertation. Additionally, the motivation for selecting UML methods is deliberated, in combination with the criteria used to evaluate the models.

3.2 Study Design

This study presents a secondary analysis of retrospective cases collected enrolled into the MADCaP database over a period of four years (2016 to 2020). The data was sourced from the Urology Out-Patient Department, Chris Hani Baragwanath Academic Hospital (CHBAH), which serves as a collaborator centre of the MADCaP consortium.

This analysis explored various parameters and variables contained within the data set and seek to derive meaningful insights and outcomes that can contribute to the larger body of scientific knowledge related to PCa. This study aimed to advance of medical science within the field of urology through robust analysis of PCa dataset.

CHBAH is a tertiary academic hospital located in Soweto, South Africa, is the largest hospital in Africa spanning 173 acres, and operates with approximately $3,200$ beds while employing $6,760$ staff members [71]. The hospital serves a densely populated socio-economically diverse population from Soweto, surrounding peri-urban and rural communities with Gauteng Province [72]. This includes a large proportion of historically underserved populations, many of whom experience limited access to private healthcare services. CHBAH provides a critical point of access for specialized medical care, including oncology services, for a significant portion of low income households [73]. The hospitals diverse patient base offers a unique opportunity to examine disease patterns with representative

segment of the population most affected by health disparities in the region [13].

3.3 Study Site

The study was be based on the MADCaP cohort enrolled at the Urology Out-Patient Department, Chris Hani Baragwanath Academic Hospital (CHBAH) study site. Cases with newly diagnosed primary PCa, irrespective of stage, grade, or pathological classification, were deemed eligible and subsequently enrolled in the study [55]. The research collected data for men aged 40 years and older to garner valuable insights into the characteristics, risk factors, and potential management strategies associated with primary PCa [13]. Overall, this study contributes to the development of a nuanced understanding of PCa , offering insights that may inform improved approaches to management and treatment [13].

3.4 Study Population

The dataset comprised of men of African descent diagnosed with primary PCa, who were followed from the point of initial diagnosis, through to conclusion of the follow up period, as part of their enrolment in the MADCaP study [13]. The majority of cases underwent either radiation therapy or hormone deprivation therapy, Though treatment modalities varied across MADCaP study centres [74].

The research study conducted an comprehensive evaluation of secondary PCa cases among black African men in Soweto, South Africa utilizing data collected by the MADCaP consortium at CHBAH. The study assessed their current state of health and quality of life; case data was mined from a STATA file format [13]. The study made significant strides to mitigate the influence of extraneous factors, ensuring integrity and reliability of the results. To achieve this, the data anonymisation, and all personal identifiers were meticulously removed. This method safeguarded the privacy and minimized the risk of bias and potential confounding factors.

The MADCaP study enrolled participants from Soweto diagnosed with PCa in terms of age, ethnicity and clinical characteristics [13]. Study data were systematically collected using standardized protocols and stored in a secure REDCaP database, ensuring POPIA compliant data management. A STATA file containing de-identified eligible cases was exported for analysis. See appendix F for data dictionary.

3.4.1 Variables

Table 3.1 below presents the set of variables that selected for clustering analysis, based on their relevance and contribution to models discriminatory power. These selected variables were chosen through feature selection techniques, guided by their clinical relevance, predictive strength, and statistical correlation with case outcomes. The data consisted of several types, Integers, Float and Boolean variables. To ensure consistency in the data for modelling, mapping and conversion facilitated for seamless integration, for instance, categorical variables currently smoke "Yes" or "No" were converted to binary "1" or "0" denoting whether patient are still smoking [75]. Ordinal variables such

as smoking with "Yes", "Yes (in the past)" and "No" were label encoded to 0, 1, and 2 respectively denoting smoking behaviours [34].

Table 3.1: Variables Chosen for Analysis

Characteristic	Description
Age	Age at diagnosis
Cigarette count	Average number of cigarettes smoked per day
Smoking	Current smoking status (Yes/No)
Income	Monthly earnings
Final Staging–T	Primary tumour stage (based on digital rectal examination [DRE])
Final Staging–M	Presence of distant metastasis
PSA at Diagnosis	Prostate-specific antigen (PSA) level at diagnosis
PSA at Referral	Prostate-specific antigen (PSA) level at referral
Gleason Total	Calculated as Gleason primary + secondary + tertiary (if present)
Father's Cancer	Whether the father was diagnosed with prostate cancer
Brother's Cancer	Whether any biological brother was diagnosed with prostate cancer
Son's Cancer	Whether any son was diagnosed with prostate cancer

3.4.2 Inclusion and Exclusion Criteria

The MADCaP study collected data related to Men of African descent who were previously diagnosed with PCa and had received prior treatment. Specifically, the study analyses secondary data from men with treatment naive, histologically diagnosed PCa, enrolled from the Department of Urology, at CHBAH. Only cases were included in this UML analysis. The study leveraged existing clinical and demographic variables from the MADCaP REDCaP database, focusing exclusively on cases with confirmed PCa to identify potential subtypes or patterns within the population. The strictness of the inclusion criteria made the findings of this research robust and reduced bias.

3.4.3 Sampling

This study constitutes a secondary data analysis of PCa records from the MADCaP study, covering the period from January 2016 to December 2020. The sample size consisted of $n = (1,096)$ cases that met the inclusion criteria ensuring maximum data utilization for exploratory ML. Since this was unsupervised analysis of secondary, the determination of a sufficient sample size did not rely on traditional statistical power calculations. Rather, it was guided by considerations of cluster stability and analytical precision [76,77]. Techniques such as the elbow method and silhouette scoring were employed to evaluate the robustness and quality of the derived clusters. Using this dataset, the study aimed to explore risk factors and variables associated with PCa diagnosis and prognosis over a five-year period.

3.5 Experimental Environment

The experimental environment for this research utilized a suite of open-source software tools deployed on a Microsoft Windows Operating System. Python 3 served as a core programming language for the algorithm development and training. The Jupyter Notebook is the preferred Integrated Design Environment (IDE) selected for its flexibility and and support for iterative analysis. To support ML workflows, libraries from the Scikit-learn ecosystem. Widely recognized as gold standard in python-based ML. Table 3.2 below summarizes the specific hardware and software tools employed in this study.

Table 3.2: Hardware and Software Tools List

Apparatus	Version	Description
Laptop	Windows 11	Intel Core i9, 3.60GHz base frequency, 16GB RAM
NumPy	1.24.4	Provides support for large, multidimensional arrays and matrices, along with mathematical functions to operate on them
Pandas	1.3.10	Offers data structures and tools for data manipulation and analysis, particularly tabular data (DataFrames)
SciPy	1.8.5	Provides algorithms for optimization, integration, interpolation, eigenvalue problems, and statistics, widely used in ML
Matplotlib	3.8.4	Comprehensive library for creating static, animated, and interactive visualizations in Python
Seaborn	0.13.1	High-level interface for statistical data visualization, built on Matplotlib, with emphasis on heatmaps and correlation plots

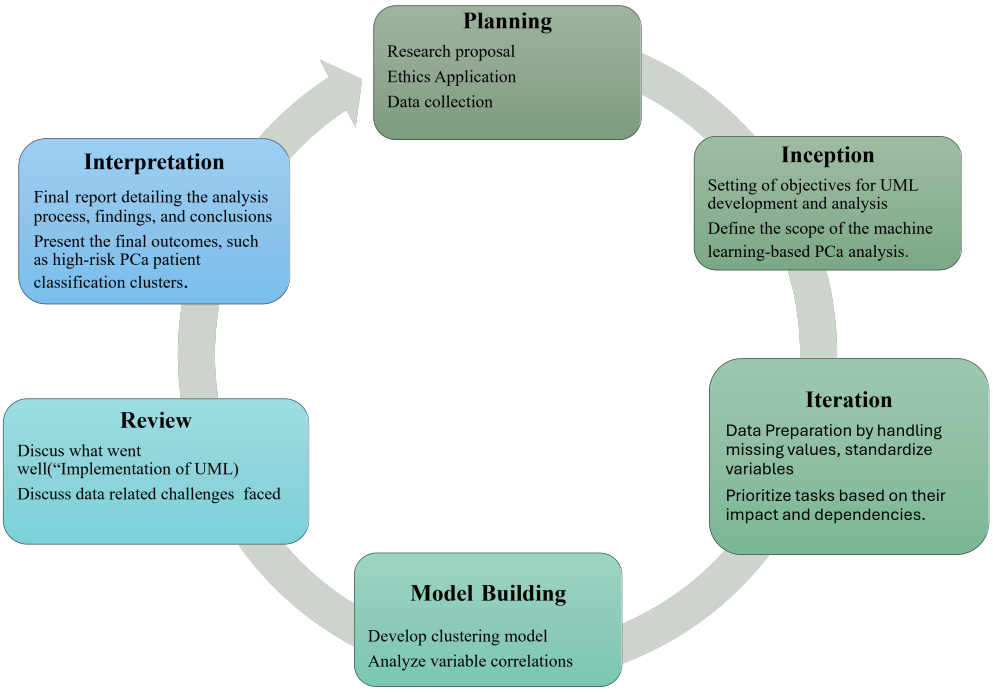

Figure 3.1: *Shows agile iterative cycles and steps followed for UML modelling process, planning and design to develop, review, and interpretation.*

3.5.1 Software Engineering Practices

The Agile development methodology was adopted in this research to accommodate the need for flexibility and iterative refinement during data preparation and analysis allowing for continuous testing of assumptions and adjustments to methods as new insights emerged [78]. Unlike traditional research methods, agile analysis approach emphasizes collaboration, adaptability and responsiveness to change. This process followed key phases, such as planning, execution and reflection enabling findings to be integrated and refined throughout the study. The aim was to map UML practices adopted and the challenges encountered thereby supporting navigation of evolving requirements for the study [79].

This method facilitated analysis through iterative cycles, allowing for adaptability to changing requirements and supporting continuous model improvement [78]. As outlined in Chapter 1, the research objectives were centred on unsupervised modelling, which aims were defined at the initial stage of the study. Regular goal assessment provided a more detailed understanding of project components in iteration ensuring alignment with the broader study. The analytics portion of data analysis typically followed a sequential waterfall process, The agile methodology facilitated the breakdown of larger tasks into smaller ones for the overall process to be manageable, flexible and responsive [79]. The overarching goal was to reach a point of optimality in the UML models through iterative experiments and refinement of models [78]. The iterative process for PCa analysis using agile methodology is illustrated above in Figure 3.1.

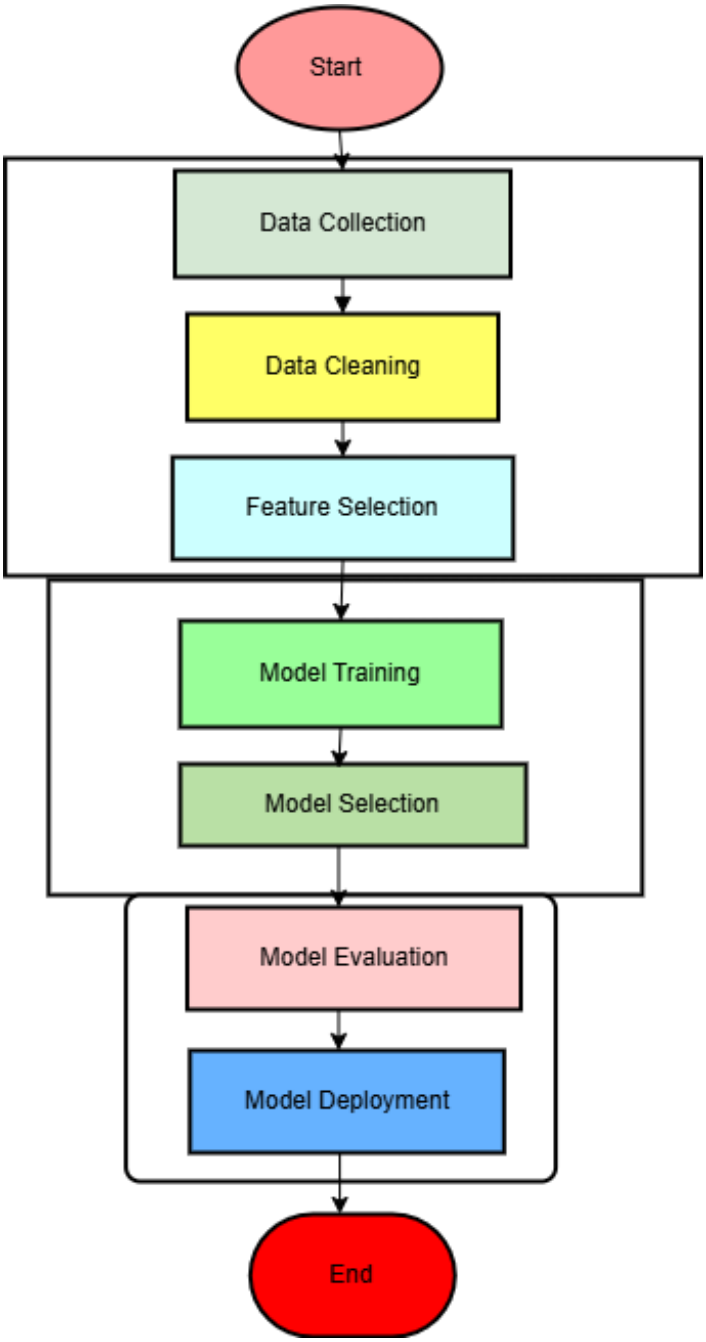

Figure 3.2: *Flow diagram illustrating the stages of UML applied to PCa data, from preprocessing and feature selection to clustering and clinical interpretation of case subgroups*

3.5.2 Data Generation

The data for this study were obtained from a single MADCaP site, selected from among seven locations of MADCaP consortium, see Figure 2.1. Specifically, the data were sourced from the CHBAH, and subsequently extracted from the REDCaP database, and where they are archived and stored.

The dataset employed in this study consists of cases diagnosed with PCa who were admitted into study site at CHBAH. These datasets are stored in a database, ensuring systematic structure in retrieval and management. Key variables were extracted ,included demographic, clinical and lifestyle factors outlined in Table 4.4. To ensure consistency, categorical variables were processed according to predefined coding schemes. For instance, monthly earnings were grouped into standardized income bands (<ZAR $5,000$: ZAR $5,000 - 10,000$: >$10,000$), while smoking status was classified as

No (never smoked), Yes (now), Yes (in the past) based on self reported histories. Family history of PCa was recorded as (Don't know, NO and Yes) indicating whether first degree relative had been diagnosed with the disease.

3.6 Data Preprocessing

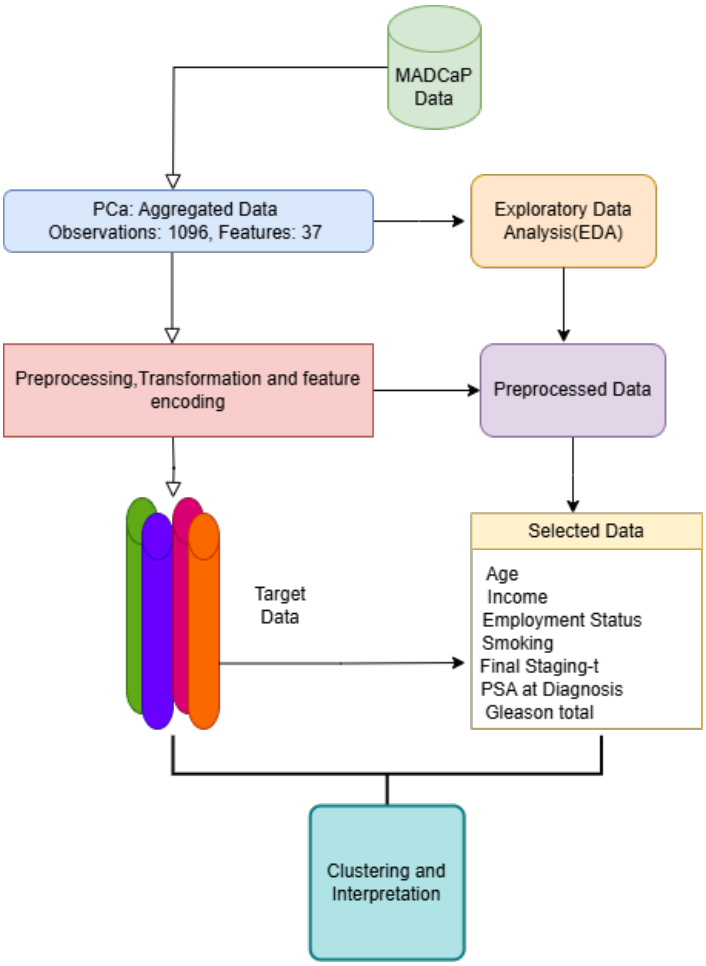

Figure 3.3: *Pre-processing workflow for MADCaP data, including data cleaning, normalization, feature selection, and handling of missing values prior to unsupervised machine learning analysis*

Jupyter Notebook (version 6.5.4) was utilized to access statistical and data visualization packages and libraries required for data pre-processing. The scikit-learn package included several transformer modules for converting datasets with discrepancies into standardized data suited for unsupervised learning algorithms [80]. The dataset necessitated various pre-processing techniques, for instance handling missing values and Gaussian rounding. Within the dataset, compatibility issues arose with estimators assuming each value in a variable contained meaningful information, essential for ML algorithms to run successfully. In such circumstances, imputation techniques for handling missing values were employed [17].

The proposed pre-processing steps were essential prior to applying UML techniques, as they enhanced the likelihood of uncovering meaningful underlying patterns in the dataset [31]. However, a significant challenge with such clinical data is the inherent presence of biases and formatting inconsistencies. These issues can skew clustering results by disproportionately emphasizing certain data

features, thereby constraining the discovery of novel insights [36].

To identify anomalies and potential sources of noise, this study employed exploratory data analysis (EDA) on the MADCaP dataset. EDA is a descriptive statistical approach that facilitates hypothesis generation and refinement through visual techniques such as outlier detection and distributional assessment [29]. These approaches were applied to extract maximal insight from the dataset, enabling an unbiased examination of data structure and variability. This approach is particularly valuable for uncovering underlying phenomena within the MADCaP dataset without relying on preconceived assumptions [34].

Quite a few variables in the dataset exhibited a wide range of values, including extreme and potentially implausible observations. To address this, outlier detection was performed to identify and exclude values that deviated from established biological plausibility or clinical norms, for example, PSA levels presented with markedly aberrant values [81]. This step was essential to mitigate the influence of data entry errors, measurement anomalies, and false observations, thereby preserving the integrity and reliability of subsequent analytical results [17].

3.6.1 Data Imputation

The dataset contained missing values due to unrecorded information and instances of data corruption, which were represented as NaN. To address this, data imputation was applied as a strategy to substitute missing values with alternate values from the dataset, thereby preserving most of the data for analysis. [18]. To ensure that the dataset was generally undamaged and usable for modelling purposes subsequently reducing bias and impairing the model's analysis [63]. Additionally, the dataset was not very large, and avoiding further loss is very important because deleting a substantial amount of the dataset could significantly affect the final model [36]. A common technique for data imputation is to compute a statistical value for each column attaining the arithmetic mean of each column, and replacing missing values in the columns [27].

To handle the missing values, the mean imputation method was applied to numerical columns, such as PSA levels, GS, and cigarette count [75]. Categorical variables such as smoking history, employment status and income were also imputed using mode imputation, and median imputation was applied to variables with skewed distributions such as age and father's age [34]. The combination of these imputation techniques facilitated the dataset to maintain its essential statistical properties while addressing missing values effectively, thus providing a robust foundation for subsequent UML analysis in the context of PCa research. This pre-processing approach is critical in medical data analysis, were accurate handling of missing information can significantly influence model performance and the interpretation of results.

3.6.2 Feature Extraction

The PCa dataset was characterized by substantial amount of features which could result in subpar classification performance therefore feature extraction was applied [36]. Feature extraction is the process of examining the relevant importance of each feature in the dataset and removing redundant features [18]. Through the feature-importance scores technique, decision trees were used to describe the most relevant features, which assisted in identifying the variables that contributed significantly to the model [27]. When the dimensionality and computational cost of UML techniques were reduced, only the most relevant features were extracted and retained, while unnecessary characteristics were discarded [18]. In the PCa dataset, certain variables were excluded to ensure that the analysis focused on factors most relevant to disease outcomes, with the selection guided by existing literature and expert input, thereby avoiding the inclusion of irrelevant variables. [82].

Feature extraction was conducted through a systematic process that combined domain expertise, clinical relevance, and statistical analysis to ensure that the most meaningful variables were retained for subsequent clustering [18]. The dataset initially included 47 variables encompassing demographic, behavioural, clinical, and familial domains. However, not all variables were relevant for modelling PCa outcomes. Therefore, the feature extraction strategy was guided by literature, expert driven exclusion of biological, clinically irrelevance variables, and data driven dimensionality reduction informed by statistical significance and redundancy.

The variable $participanttype$, maternal and paternal ethnicity d_ethn_mother, d_ethn_father, and place of birth $sd\_townborn$, $sd\_countryborn$, as these are considered to have no direct or indirect biological relevance to PCa progression or outcomes. Variables such as $record\_id$ and $study\_id$ serve as unique identifiers for records or participants. These identifiers were removed, as they do not hold any predictive or biological relevance to PCa, serving solely to label individual cases. Similarly, $sd\_languageinterview$, which specifies the language used in interviews, is unlikely to impact PCa risk, progression, or treatment outcomes, and thus provides little utility in our research. Date fields $psa\_psareferraldate$, $psa\_psadiagndate$ are also removed as they were contextually redundant and not relevant to the disease prediction in this study.

On the other hand, feature engineering enabled transformation of existing features within the dataset to a new variable. The GS was computed by summing the primary, secondary, and, when applicable, tertiary architectural patterns of the highest grade. For analytical purposes, a novel categorical variable, PCa_risk, was generated to capture clinically meaningful risk groups. Low-risk was defined as GG $1 \leq 6$ intermediate-risk as GG 2 (GS $3 + 4 = 7$) and GG 3 (GS $4 + 3 = 7$); and high-risk as GG 4 (GS 8) and GG 5 (GS 9 –10). Similarly, to facilitate risk stratification, the variable psa_psadiagnlevel was transformed into an ordinal variable, psa_category, based on clinically validated thresholds: low-risk (PSA < 10 ng/mL), intermediate-risk (PSA 10-20ng/mL), and high-risk (PSA >20ng/mL). By adhering to both locally and internationally accepted guidelines for GS and PSA classification, the study ensured consistency with established clinical protocols and enhanced comparability with existing case data [18,25,36,83]

When performing UML, it was essential avoid redundant variables, as they can introduce noise and inflate dimensionality without contributing to meaningful insight [27]. Retaining $mr\_finalstagingt$,

which encapsulates the definitive clinical stage, ensured the dataset remained concise and free of unnecessary duplication. In contrast, removing $mr\_dreclinicalstagingt$, $mr\_dreclinicalstagingn$, and $mr\_dreclinicalstagingm$ improved computational efficiency, by reducing the number of features the model must process, thereby minimizing dimensionality and potential noise [18]. Table 3.3 shows feature extraction summary.

Table 3.3: Feature Extraction Summary

Variable(s)	Decision	Rationale
study_id, record_id, Participant type	Excluded	Administrative identifiers with no analytical value
Language interview, ountry born, Town born	Excluded	Non-predictive demographic variables
ethnicity mother, father, Race father, Race mother	Excluded	Lacked biological plausibility for PCa prediction
Employ status_other, Employ type_other	Excluded	Qualitative free-text inputs with inconsistent values
PSA referraldate, PSA diagndate	Excluded	Redundant date variables with limited predictive value
Age, Currently smoke,Income, Cigarette count	Retained	Demographic and modifiable behavioural risk factors
PSA diagnosis level, PSA referral level	Retained	PSA levels as key clinical biomarkers
Gleason turp total	Retained	Standard indicator of tumour aggressiveness
Final staging (TM)	Retained	Final tumour and metastasis staging
Father, brothers cancer, Sons cancer	Retained	Family history variables with established PCa associations

The succeeding steps encompassed statistical evaluation aimed at dimensionality reduction. Univariate and Multivariate analyses were conducted to assess the strength and significance of each variable's association with the outcome of interest [61]. This approach facilitated the identification and preservation of clinically significant and potentially modifiable variables, guided by both literature evidence and expert validation [4, 84]. For instance, PSA levels at both diagnosis and referral, GS and staging TM were retained as essential clinical indicators due to their established prognostic value. Smoking behaviour and income were retained as potentially modifiable risk factors, given their relevance to public health interventions. Additionally, family history of PCa among fathers, brothers and sons were also preserved due to their known predictive characteristic value in assessing individual risk. By extracting only relevant features the MADCaP dataset became more streamlined, focusing on clinically and potentially modifiable factors.

3.6.3 Label Encoding

The PCa dataset was comprised of variables of various data types, such as categorical, ordinal and numerical data. Because of this discovery, label encoding was selected as an appropriate data preprocessing technique to enable the effective processing and analysis of the dataset [34]. By applying label encoding, categorical variables in this dataset were assigned unique integer values, where each category was allotted a specific numerical symbol [29]. This transformation ensured that categorical data could be efficiently processed by UML algorithms that typically require numerical data [18]. Through encoding categorical features, the dataset became suitable for computations, enabling the models to recognize patterns and associations in the data [23]. For instance, variables like $employstatus$, and $income$ represent categorical data that contain a limited number of possible values, such as different employment statuses, and income brackets. Label encoding allowed these values to be transformed into a numerical form while preserving their categorical distinctions, making it feasible to use them in clustering algorithms or principal component analysis (PCA) for dimensionality reduction [34].

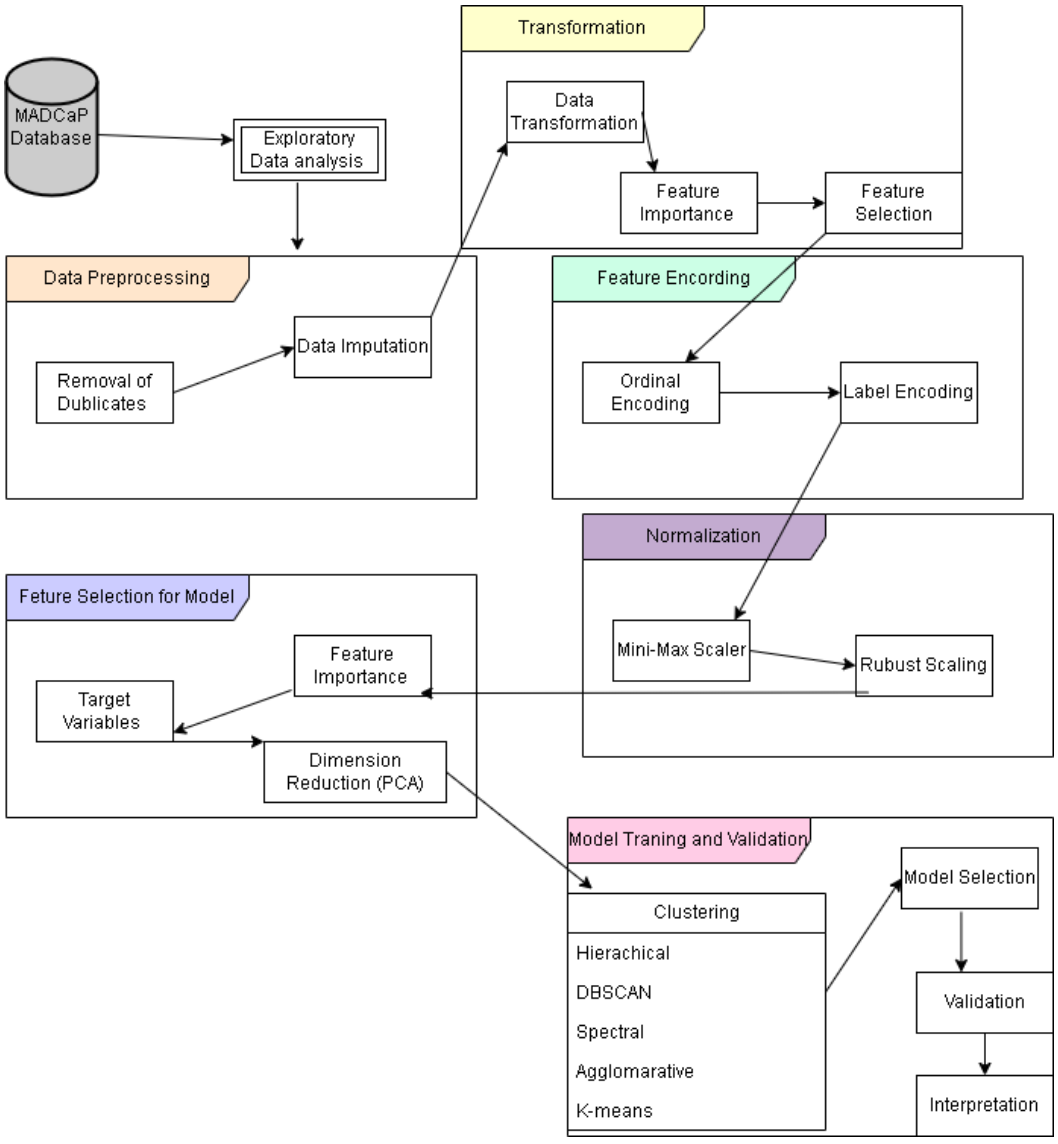

Figure 3.4: *Framework for UML modelling process, outlining stages from data collection, exploration, and preparation to iterative refinement for robust and interpretable clustering.*

Contrasting one-hot encoding, which generates split-up binary columns for each category, label encoding keeps the dimensionality lower, which is beneficial for UML techniques, as it reduces computational complexity and avoids issues with the "curse of dimensionality" [28]. For variables with no natural ordering, such as $income$ or $psa - referralleveldesc$, label encoding provided an efficient transformation, as it simply mapped each category to an integer label without adding extra columns [18]. This simplification is valuable in UML modelling as it made it better to measure similarity or distance between data points. Moreover, UML algorithms like k-means clustering and hierarchical clustering can operate more effectively with lower-dimensional, integer-coded data, as this representation allows for more straightforward computations of similarity or dissimilarity between data points [70].

In summing up, label encoding was chosen for these variables to efficiently transform categorical data into a form harmonious with UML algorithms, maintaining distinctions between categories without adding unnecessary dimensionality, which is crucial for improving model performance and interpretable in clustering data.

3.6.4 Data Standardization and Normalization

ML algorithms perform effectively when variables are scaled to standard and normalization in a popular numerical data scaling technique applied before to modelling. Standardization is a data pre-processing method that is used to transform the characteristics of a data set so that they have a mean of 0 and a standard deviation of 1 [69]. This process is attained by subtracting the mean from each feature and dividing by the standard deviation [28]. Normalization is principally important in ML analysis as it places features on a similar scale, ensuring that features with larger ranges or values do not disproportionately influence models [34]. It is commonly applied in algorithms sensitive to feature magnitude, such as Principal Component Analysis (PCA) and clustering techniques. Standardization is expressed by the following formulae.

The formulae for Z-score normalization of feature X is:

$$Z = \frac{X - \mu}{\sigma} \tag{3.1}$$

where:

- Z is the standardized value,
- X is the original value of the feature,
- μ is the mean of the feature values, and
- σ is the standard deviation of the feature values

This phase safeguards that each feature is capable of contributing equally to the model and avoids features with superior scales from dominating the ML process leading to bias [69]. In the context of the PCa dataset, normalization ensures that all data variables are similar in scale, for the algorithms to efficiently learn the patterns and associations within the data without being subject to bias in the scale of the variables [85]. This procedure is a fundamental practice in ML, and guarantees

the robustness and dependability of the models in training and evaluation [34]. This measure is particularly important as various ML models, including PCA and clustering algorithms, perform better when features are on a comparable scale [18].

This harmonizing of data allows distance-based models, such as k-Means or DBSCAN, to assess correlations among data points more accurately, optimizing model performance and reliability of the models [28]. Through the fit transform a consistent scale across all numeric features, standardization strengthens the foundations for UML and meaningful insights in PCa by reducing bias in the data [57].

3.7 Dimensionality Reduction

In ML, problems may arise due to the dataset having a large number of variables, statistical and ML methods have difficulty in handling high-dimensional data [34]. Dimensionality reduction is the method that minimizes the overall number of features in a dataset while still preserving the majority of the essential data as feasible [29]. There are two techniques of conducting dimension reduction, feature selection and feature extraction.

Feature selection was a needed component for ML modelling, encompassing identifying and retaining a set of variables that are most relevant from the original data while removing less noteworthy features [18]. Variables without statistical significance for the outcome of the disease were removed; for example, the mother's ethnicity, dates, and language of interview, were excluded as they held no biological predictive relevance [8]. It was essential to avoid redundant variables as they increase noise within the dataset. Literature strategies proposed optimal features, refer to Table **??** on selected variables as the best predictors and correlation analysis to pick out relevant clinical features associated with the outcomes of the patients [74].

Varieties of algorithms are used in dimensionality reduction, such as PCA and t-Distributed Stochastic Neighbors Embedding (t-SNE) for visualization of unlabelled and linear discrimination analysis (LDA) for labelled data [18]. Each has its advantages and drawbacks. In this research, the adopted method was PCA, which is a statistical technique that utilizes an orthogonal conversion on associated variables to unassociated variables [18].

3.7.1 Principal Component Analysis

PCA is a statistical technique that utilizes an orthogonal conversion on associated variables to associated variables [86]. PCA is a UML linear dimensional reduction-algorithm utilising singular value decomposition of data to project in lower dimensional space [35]. This is achieved by transforming variables into a new set of variables, which are uncorrelated and ordered by the variance to explain the data [17]. PCA generates new orthogonal coordinates within the data, such that the utmost variance that develops is set as the initial coordinate referred to as principal component one (PC1), whereas the subsequent notable variance coordinate is placed in second (PC2) [86]. PC1 embodies the principal source in variation the eigenvector containing the largest eigenvalue, and PC2 inhibited the least important contribution to variation in PCa data [86]. Part of the study aims to develop an algorithm and assess the correlation of unlabelled MADCaP dataset through application of UML

dimension reduction technique PCA [18].

The explained variance ratio is the measure of the fraction of the sum variance in each eigenvector through assigned eigenvalues from the original dataset, explained for each principal component (PC) [86]. PCA explains that the variance ratio is given to the i-th PC as it is taken from scikit-learn to explain the total variance of the MADCaP dataset using the following mathematical expression:

$$\text{EVR}_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} \tag{3.2}$$

Whereλ_i is the eigenvalue conforming to the i-th value of PC, and p is the total number of PC [87]. The explained variance ratio assisted in verifying the importance of every single PC, and portion of variability in the data, thereby identifying the PC that captured the most significant patterns [87]. This approach assisted in revealing the core structure of MADCaP data and enhancing the execution of consequent UML models [86].

3.8 Clustering Algorithms

This section describes the clustering techniques applied comprising data pre-processing to facilitate optimal ML algorithms extensively employed in analysing large volumes of MADCaP dataset. Part of the aim of this project is to develop clustering algorithms to effectively analyse and interpret data. These techniques include but are not limited to PCA, agglomerative clustering (AC), hierarchical clustering (HC), K-means, DBSCAN and Spectral Clustering (SC) [28]. The application of these techniques will enable the identification of patterns and relationships within the data, leading to a deeper understanding of PCa risk factors, associations and aid in decision-making processes [51]. The versatility of these techniques allows for the exploration of different perspectives and approaches to data analysis, which may ultimately yield more comprehensive results [23].

3.8.1 K-means Clustering

This technique of cluster analysis seeks to partition the observations in K clusters in which each data point fits to the nearing mean [17]. The data points are given into the algorithm without any predefined outcome to create clustering results [18]. The final clusters depend on selecting initial centroids that create separate clusters [34]. The algorithm has two phases: firstly, to determine initial centroids and secondly designating data points to the proximate cluster and then recalculating the cluster mean assigning individual points to the nearest centroid [31]. To attain the number of centroids to select the elbow approach is applied to determine the K-value of centroids for optimal performance by iterating for $K = 1$ to $K = n$ [18]. To determine the optimum number of clusters the elbow approach was applied. The method looked to find the best number of clusters to the k-means model for superior PCa data segmentation and visual inspection of PCa aggressiveness amongst cases [88].

The core purpose of the algorithm on PCa is to discover related cases, consequently assigning them together to form clusters or segments according to similarity conditions, which are defined between data [32]. To attain this, the model employs distance measure to compute comparisons and variations in the PCa data points. The algorithm relies on centroids to determine the clusters. Three centroids, as recommended by the elbow method, followed by an iterative cycle that continues until the predefined conditions are met [36]. To determine the optimal number of clusters, the elbow method score was used to identify the K value of the centroids that correspond to the highest peak on the silhouette coefficient plot [88]. This assessment signifies the number of clusters that best fit the underlying pattern while maintaining variance between clusters. A high score denotes a testimony of well-formed clusters with marginal overlay or miss-assignment. By applying the silhouette method we can determine how well the model performed on the PCa data [88].

The application of UML to PCa datasets provides indispensable information on features that are unrecognizable by the human eye [89]. Through grouping, recognition, and segmentation methods, UML may assist in diagnosing PCa by indicating patients at high risk of PCa [17]. However, the models are not the only source of information on PCa diagnosis [27]. ML models explore enormous datasets to discover anomalies [28]. The trends identified by UML in the data set can provide insight into the risk factors for PCa, thus helping to diagnose PCa early and improving the prognosis [81]. Therefore, early diagnosis would reduce the risk of PCa morbidity by detecting early-stage cases, leading to increased chances of curative treatments [22]. The application of ML can contribute to the management and care of PCa, stemming from diagnosis, treatment recommendations and administrative activities [15]. Therefore, by revealing at risk patients through a constellation of symptoms, primarily in patients with underlying PCa, this study may bridge the gap in PCa morbidity [89]. Figure 3.2 shows how MADCaP datasets were managed and pre-processed in the analysis using UML techniques.

Figure 3.3 shows the flow chart of data sourced from the MADCaP centres on men afflicted by PCa. Inclusion parameters encompasses all men affected with PCa. Utilising the datasets, clusters were generated and classified into numerous groups. The data from MADCaP was loaded into Jupyter Notebook in a Python environment, using pandas and Numpy software libraries; the data will be refined using data munging into formats better suited for unsupervised machine learning [27]. Clustering variables with similar or different characteristics using unsupervised machine learning techniques on unlabelled data [27]. Using clustering techniques, the algorithm is trained to extract patterns from input data to derive meaning. Points are assigned to the closest cluster centroids that reflect certain parts of the data using the K-means algorithm, which detects centre points within the data [17].

3.8.2 Agglomerative Clustering

AC builds upon pre-defined principles; the algorithm will first declare data points in its structure and merge the two systems until a criterion is met and ceases when the criterion is satisfied [17]. Clusters are linked by a specification of similarity and how they are measured that is defined by an existing cluster [28]. In a bottom-up systematic practice where each data point is a cluster of its own, paired clusters appear as the data points move upward the hierarchy, while pairs of different clusters move far apart [18]. DBSCAN does not require several clusters to be set and operates by identifying in crowded points areas of a feature space [17]. Points within a densely populated region are the core sample clusters that grow until all points have been picked within a lying distance [49].

To comprehend the AC process, it is imperative to acknowledge the existence of various approaches, the agglomerative approach emphasizes proximity of data points end up in the same cluster, this bottom-up approach starts with grouping singleton points to create a new instance of the cluster from individual points closest to each other [18, 28]. It repeatedly finds pairs closest to each other, merging them to produce a new cluster and, at the same time, removing individual points previously represented [63].

This research sought to implement AC to uncover features within the MADCAaP data using the ward method. This study provided the opportunity to advance the understanding of PCa using AC [18]. In the research, two clusters were merged such that the variance within the clusters was not increased; primarily each point stood as an individual cluster, and the closest clusters merged according to the given metrics [34]. The ward linkage method is applied in agglomerative clustering, which uses at times a non-euclidean similarity; this measure was used to generate clusters in dissimilar PCa datasets [19]. Ward linkage joins by initially setting the distances calculated as follows.

$$D(a,u) = \sqrt{\frac{|v|+|s|}{T}D(a,s)^2 + \frac{|v|+|t|}{T}D(v,t)^2 - \frac{|v|}{T}D(s,t)^2} \tag{3.3}$$

Where $T = \|v| + |s| + |t|$

The ward method is commonly implemented using the nearest neighbor chain algorithm. This approach does not precisely express a measure of distance among the datum of clusters [18]. The technique is an ANOVA-based methodology. One-way uni-variate ANOVAs are implemented for individual variables with groups characterized by the clusters at that phase of the process. In each phase, two clusters merge with a small increase in the combined error sum of squares [27].

3.8.3 Hierarchical Clustering

HC results in a dendrogram as shown in Figure 4.16, which depicts the hierarchical structure of the cases within the MADCaP dataset [90]. This method aided in identifying potential disease subtypes based on the PCa clinical characteristics in Table 4.4. These subtypes may have different prognoses and treatment responses enabling patient stratification and predicting patient outcomes [82]. This analysis provides an exciting opportunity to advance our knowledge of PCa in identifying unique subtypes of the disease that may have varied prognoses and responses to treatment [82]. Due to the

scarcity of PCa research employing HC dendrogram analysis researchers may not fully understand HC and its full application. The initial step of the model was to determine the statistic to be used to calculate distance or similarity between cases [91]. The chosen distance measure is the Euclidean distance, calculated by the formulae.

Where:

$$\sum_{j=1}^{k}(a_j - b_j)^2 \tag{3.4}$$

In principle, a and b are two cluster cases being measured up on the variable j, where K is the total number of variables included in the analysis [65]. The algorithm allowed for distance amongst two cases to be calculated across multiple variables, whereby each phase, the sets of clusters with the least squared Euclidean distance were merged [65].

3.8.4 DBSCAN

DBSCAN exploits the densities of the data points in the dataset to facilitate the identification of meaningful trends and patterns by recognizing a random point in the data as a core point [27]. A user-defined radius is used to measure discrete points proximity to the core point, consequently, points that are close to the core point are added to extend the cluster [92]. On the other hand points not in proximity to the cluster are classified as non-core points, thus serving as definitive markers for the limits of the cluster [92], [36]. The model produces clusters that are arbitrary shapes in the dataset. DBSCAN centres its clusters not simply by the highest density values but also considers the distances between points. This dual consideration allows DBSCAN to perform efficiently on complex datasets [84].

The algorithm applies two parameters to determine how the clusters are defined: epsilon, which sets the radius of the area around a point, and the minimum number of points required to form a dense region (Min Pts) [84]. The algorithm begins at any unprocessed case points and if the points satisfy the density criterion of Min Pts neighbor, all the neighboring points are added to the same cluster [34]. The model continues expanding the cluster at each dense neighbor point. Once all the expansions have been achieved, a cluster is formed, and the model can take up again at any other unprocessed point to locate additional clusters [17]. Non- dense points that discovered are either noise or border points.

3.8.5 Spectral Clustering

SC involves two phases; it is a two-phase process. Data points are treated as nodes of a graph, whereas the succeeding step calculates the eigenvalue decomposition of the corresponding similarity matrix [66]. The number of clusters are specified in advance in SC component = 2 were 21 defined within the algorithm to produce 2 clusters for the results [66]. The nodes in the affinity graph correspond to data points and edges between the nodes represent the similarity among the respective data point measured using a Gaussian kernel or distance metric [34]. The consequential affinity matrix beneath describes these comparisons and aids as a basis for the next phase of algorithmic computation [66]. The eigenvalue decomposition algorithm calculates the Laplacian matrix

derived from the affinity matrix [93]. When the Laplacian matrix is constructed, its eigenvalues and eigenvectors are calculated, and the eigenvectors equivalent to the smallest eigenvalues develop a new depiction of the data in lower dimensional space [93].

The Laplacian matrix is computed as follows:

$$L = D - W \tag{3.5}$$

Where L is the Laplacian matrix, D is the degree matrix where one diagonal component represents the sum of the similarities, and the affinity matrix of the pairwise similarities are presented by W.

3.9 Performance Testing

A popular exercise is to adjust the features so that the data interpretation is more suitable for these algorithms [28]. The PCA is an approach that revolves around the dataset in a way such that the revolved features are statistically uncorrelated [34]. The conception behind feature extraction is that it is probable to find a representation of the data better suited to analysis than the raw representation given [17]. Several critical stages must be taken in the process of establishing and analysing a UML models for PCa in the research. We identified an appropriate evaluation criterion the elbow, Silhouette Score and epsilon score [27]. This method provides an objective method for evaluating clustering methods.

3.10 Data Management

All the data was stored in a secure external hard disc drive, which was closely monitored, password encrypted and accessed by authorized personnel, ensuring the highest level of data confidentiality. These measures were implemented to safeguard patient information and prevent any unauthorized access, maintaining compliance with ethical and legal standards of data protection. Furthermore, no identifiable data was shared or accessed during the study, thereby preserving anonymity.

To ensure adherence to ethical research practices, an application for an ethical clearance certificate was made to the Human Research Ethics Committee (Medical) at the University of Witwatersrand. This application was supported by the MADCaP study, which oversees the use of the dataset. Table 3.2 provides a detailed overview of variables included in our study, outlining the relevance of the data used for analysis. This framework guaranteed that all research activities were conducted ethically and responsibly while protecting the integrity of the patient data.

3.11 Ethical Considerations

An application for ethical clearance will be submitted to the Human Resource Ethics Committee (Medical) at the University of the Witwatersrand. Additionally, permission to use the available dataset was sought from the National Cancer Registry. Approval from the National Health Research Database (NHRD) was not required, as this project was conducted as a sub-study under the MADCaP study, which had already received the necessary approvals. The ethics certificate for the MADCaP study is provided in the appendix protocol number $M220673$. These steps were undertaken to ensure a compliance with ethical research practices, safeguard the integrity of the research, and uphold the confidentiality and privacy of patient data. Please see the research ethics certificate in Appendix A.

This study adhered to ethical and legal standards governing the use of personal health information, with particular attention to participant anonymity, confidentiality, and responsible data management. The dataset used was retrieved through authorized institutions in accordance with the Protection of Personal information Act (POPIA) No. 4 of 2013 of South Africa. For this research, only anonymized, de-identified secondary data were used. As such, no additional consent was required for secondary data analysis, in line with the guidance for research using de-identified data under the POPIA Act.

To guard participant confidentiality, all direct identifiers, including names, ID numbers, and contact details were removed before data access and analysis. The dataset was fully anonymized, and no attempts were made to re-identify any individual. All data were securely stored on encrypted password following the data protection principles outlined in the POPIA through protected systems, with limited access to the researchers. All results are presented in aggregated form only, with no individual level data disclosed in tables, figures, or narrative descriptions.

3.12 Limitations of the Study

However, approaches of this kind carry with them several limitations unlabelled data presented a significant challenge, as there are small statistical tests applicable to measure the accuracy of the clustering techniques; the PCA technique rotated the dataset, and therefore, statistical features became uncorrelated [34]. The reduction of the dataset into two dimensions led to the loss of information; this practice may have discarded imperative features that are necessary for the underlying structure [31]. The initial limitation was the size of the data as it has been well recommended that UML models require a vast majority of data to for them to be efficient [28]. Another problem with this approach is that it fails to take missing variables into account, unrecorded variables in the dataset meant the dataset could not be fully utilized and the removal of variables during the data cleaning and feature extraction, contributed to the model not performing at optimum [18].

UML analysis is dependent on the quality of input data; problems with data might cause bias limiting the ability of ML approaches to generate insightful results [31]. Additionally, ambiguous models create a challenge of interpretability. Considering the models are trained from unlabelled variables it could be problematic in the interpretation of these outcomes [27]. Through substantive manual summarization of the cluster features in PCa data, we can to garner interrelations from the patterns generated. The project was limited to a sample of Soweto representatives of cases mainly at the MADCaP study site at CHBAH and would tend to miss participants diagnosed from other provinces and cities in South Africa.

3.13 Summary

The application of UML followed a systematic process to extract meaningful patterns from PCa data. This began with data collection, followed by pre-processing steps including cleaning, handling missing values, and normalization to ensure data consistency. Feature selection then identified the most relevant clinical and biological variables, such as GS, PSA levels, and Tumour stage. These tailored steps addressed the complexity and heterogeneity of PCa biopsies, ensuring robust and interpretable results. Leveraging UML techniques like dimensionality reduction and clustering, the analysis provided insights into risk stratification based on key factors GS, PSA, age, lifestyle, and staging. Cluster quality was evaluated using metrics like the Silhouette score to ensure robustness. Ultimately, this approach identified clinically significant patient groupings, revealed potential stratification strategies, and supported the development of data-driven public health initiatives.

Chapter 4

RESULTS

4.1 Introduction

This study was exploratory and interpretive in nature; results of applying a clustering assessment of the MADCaP dataset for PCa in South African men are represented here. The outcomes highlight patterns, clusters and associations identified within the data, which contribute significant understanding into potential risk stratification and underlying attributes of PCa. Results start with an overview of the cluster analysis of the UML procedures. Interpretation of key variables, including the newly generated risk variables categories. Visualizations, such as, dendrogram, scatter plots, and heatmaps are made use of to illustrate the structure and association revealed by the analysis.

4.2 Descriptive Statistics

This section presents a descriptive overview of the study dataset through univariate, bivariate, and multivariate analyses. Univariate distributions of continuous variables were examined using histograms, box plots, and violin plots, while categorical variables were summarized with pie charts and stacked bar charts presented in ranges and interquartile range(IQR). Relationships between variables were further explored through correlation analyses, scatter plots, and multivariate heatmaps. This EDA was undertaken to validate data quality, guide the choice of appropriate statistical models, and provide a baseline characterization of the cohort. Specifically, EDA highlighted the clinical and demographic features of men diagnosed with PCa within the MADCaP dataset.

4.2.1 Results of Imputed Variables

Table 4.1 summarizes the extent of missing data across key study variables prior to imputation. Overall, the dataset demonstrated excellent completeness, with very low levels of missing values for most clinical parameters. Demographic variables such as age and smoking-related measures each exhibited 0.09% missing values. Tumour staging variables $T - stage$ and $M - stage$) showed complete data capture with no missing observations 0%. For PSA biomarkers, PSA at diagnosis demonstrated slightly higher missing-ness 1.46% compared with PSA at referral 0.18%. $Gleasontotal$ also showed minimal missing values 0.18%, ensuring sufficient reliability for risk stratification. In contrast, family history variables exhibited the highest degree of missing values, particularly regarding second-generation relatives, with 15.42% for sons and 12.77% for brothers diagnosed with cancer. These findings highlight strong overall data quality, while emphasizing the need for cautious interpretation of family history variables in subsequent analyses.

Table 4.1: Imputation by Variable

Variable	Missing (n)	Imputed (%)
Age	1	0.09%
Cigarette count	1	0.09%
Smoking	1	0.09%
Income	1	0.09%
Final Staging-t	0	0%
Final Staging-m	0	0%
PSA at Diagnosis	16	1.46%
PSA at Referral	2	0.18%
Gleason Total	2	0.18%
Fathers	2	0.18%
Brothers Cancer	140	12.77%
Sons with Cancer	169	15.42%

4.2.2 Univariate Analysis

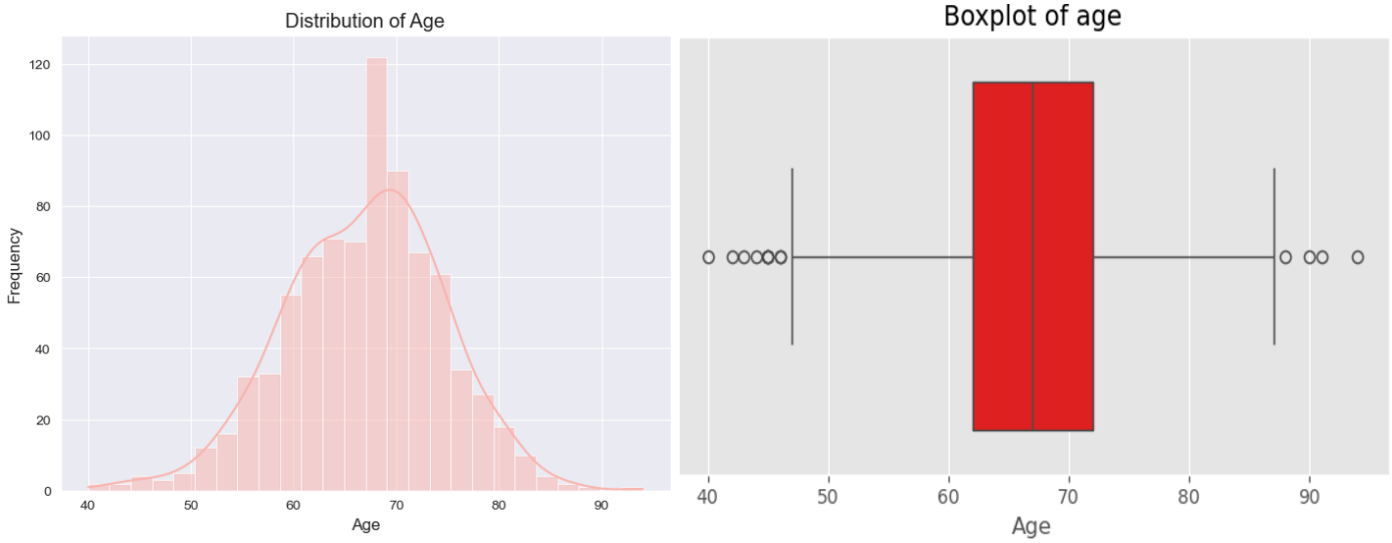

Figure 4.1: *The age distribution of prostate cancer cases, reflecting the predominance of PCa in older men*

Figure 4.1 presents the age distribution of PCa cases in the cohort, ranging from $40 - 94$ years. The histogram reveals a near-normal distribution, with the majority of cases concentrated between ages $55 - 75$. Peak incidence was observed between $68 - 70$ years, aligning with global epidemiological trends that characterized PCa as a malignancy predominantly affecting older males. Cases below the age 50 were rare, and incidence declined gradually beyond age 80, further reinforcing the well established association between advancing age and increased PCa risk.

The accompanying box plot indicates a median age of 68 $(IQR : 62 - 72)$, showing that half of the cases were diagnosed before and the other half after this age. The interquartile range captures the middle 50% of cases within a narrow age band, typical of an older population. The whiskers extend from approximately $45 - 95$ years, with outliers indicating diagnoses at unusually young or old ages. The distribution is slightly right-skewed, further emphasizing the concentration of cases in older age groups while acknowledging uncommon early and late diagnoses.

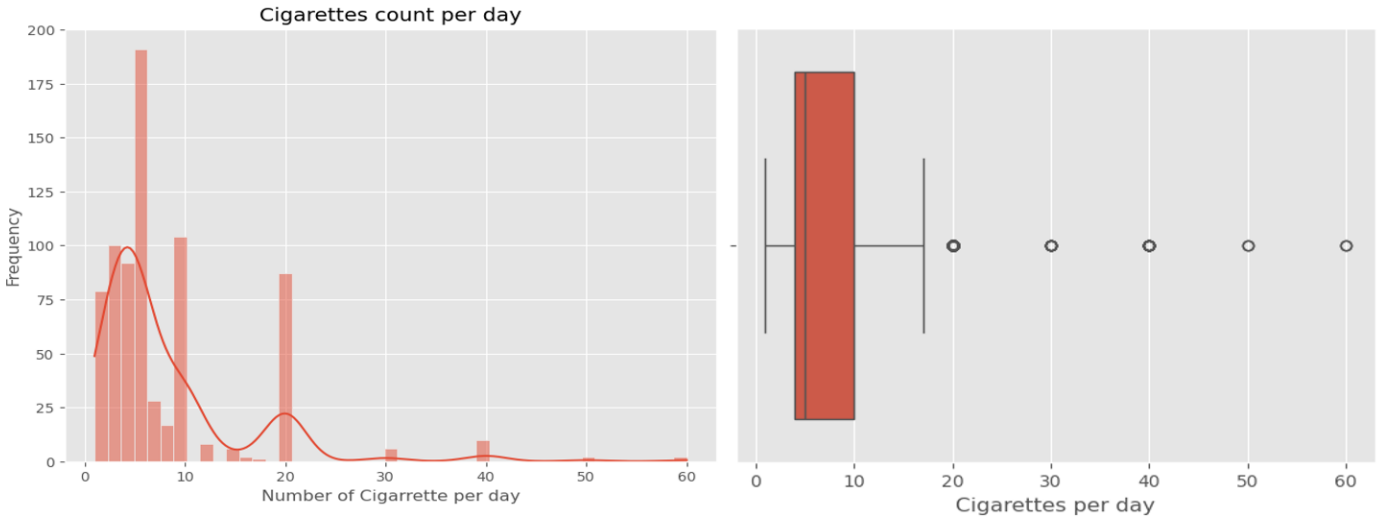

Figure 4.2: *Visualization illustrates distribution represents the count of cigarettes smoked by MADCaP participants.*

Figure 4.2 depicts the distribution of daily cigarette consumption among cases, ranging from 0 - 60 cigarettes per day. The histogram demonstrates a right-skewed distribution, with most individuals concentrated on fewer than 10 cigarettes per day. A clear peak is observed at approximately 5 cigarettes per day, with smaller secondary peaks at 10 and 20 cigarettes per day, reflecting distinct smoking subgroups within the cohort. Heavy smoking > 20 cigarettes per day was rare, though a few cases extended up to 60 cigarettes per day, highlighting the presence of extreme outliers.

The accompanying box plot illustrates a median daily cigarette consumption of 6 $(IQR : 3 - 12)$, indicating that half of the cohort smoked fewer than 6 cigarettes per day, while the other half smoked more. The IQR captures the middle 50% of smokers within a relatively low-consumption band, suggesting that light-to-moderate smoking predominates in this population. The whiskers extend to approximately 18 cigarettes per day, with outliers representing individuals who reported smoking between 20 and 60 cigarettes per day. This distribution highlights the predominance of light smokers, while also identifying a small subset of heavy smokers who may exert a disproportionate impact on clinical outcomes..

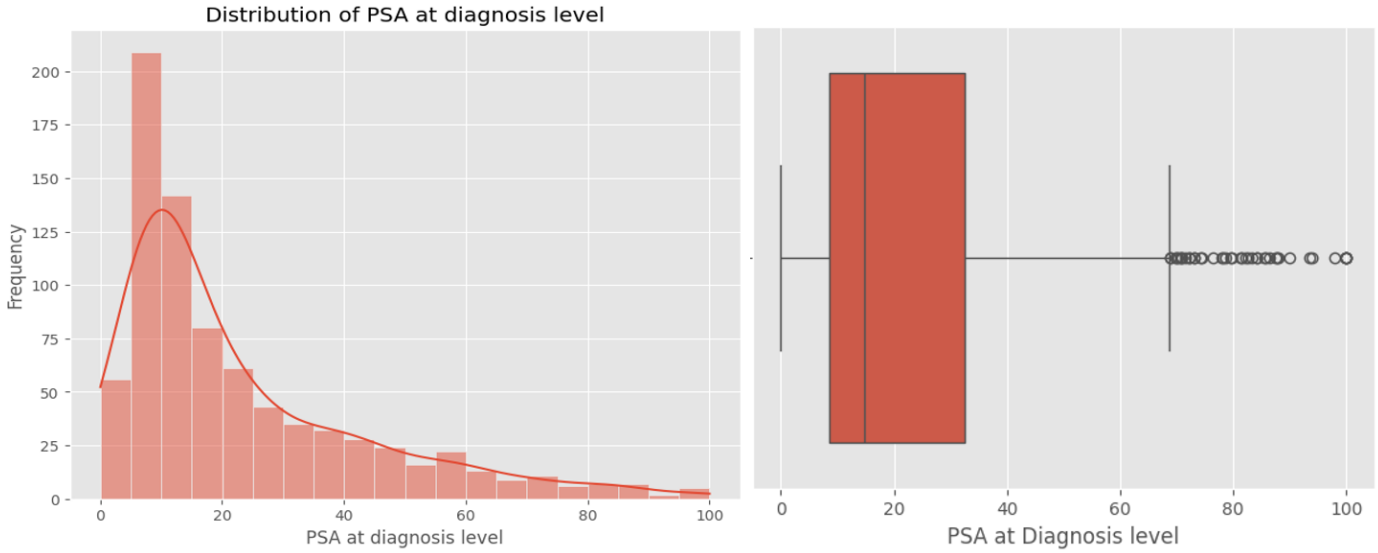

Figure 4.3: *Prostate specific Antigen distribution at diagnosis.*

Figure 4.3 demonstrates the distribution of PSA levels at diagnosis within the cohort. The distribution is right-skewed, with values ranging from $0.00 - 100.00$ ng/mL, indicating a concentration of cases at lower PSA levels and an extended tail toward higher values. This pattern reflects the heterogeneity of disease presentation at diagnosis. Half of the cohort exhibited PSA levels below the central threshold, while the middle 50% of observations fell within a relatively broad range—highlighted by a median of 14.00 $(IQR : 8.00 - 32.00)$ ng/mL

Approximately 50% of cases demonstrated PSA levels within the clinically borderline range of 4–10 ng/mL, whereas 25% of cases displayed markedly elevated levels exceeding 32.00 ng/mL. These elevated concentrations may be associated with advanced disease stages or high-risk pathological characteristics. The presence of extreme values; > 100.00 ng/mL further underscores the heterogeneity of PCa at diagnosis and highlights the limitations of PSA as a stand-alone biomarker for risk stratification. This distribution aligns with established epidemiological patterns and reinforces the necessity for integrated diagnostic approaches.

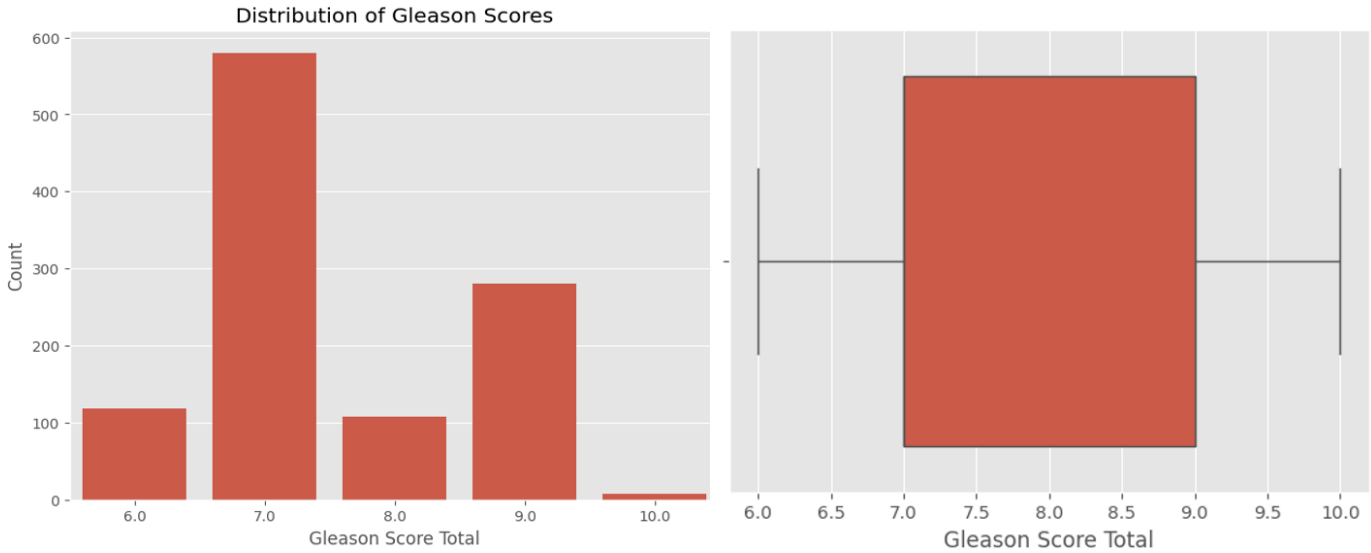

Figure 4.4: *Total Gleason score distribution Calculated as primary + secondary + tertiary*

Figure 4.4 presents a histogram revealing a bimodal, right-skewed distribution of GS, with a prominent peak at GS 7 $(3+4)$ or $(4+3)$, consistent with the clinical predominance of intermediate-risk PCa. A secondary peak is observed at GS 6 $(3+3)$, representing low-risk cases, while higher scores $(8-10)$ appear with decreasing frequency, reflecting the expected distribution of aggressive disease within PCa populations. The corresponding boxplot supports these findings, showing a concentration of scores within the intermediate-risk range, with whiskers extending from GS $6-10$ and potential outliers beyond these bounds. This distribution underscores the heterogeneity of pathological aggressiveness in PCa and strengthens the GS role as a key prognostic indicator, consistent with global histopathological trends in PCa diagnostics. The median GS was 7 $(IQR: 6-8)$.

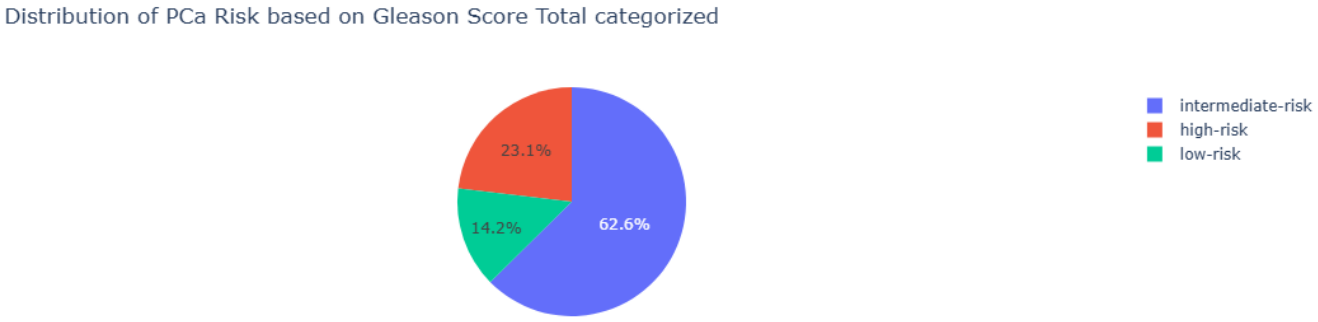

Figure 4.5: *Evaluation of Gleason total categorized into three groups to represent low-risk, intermediate-risk and high-risk*

Figure 4.5 depicts the distribution of PCa risk levels among cases, categorized by GS into low-risk (green), intermediate-risk (blue), and high-risk (red) groups. Most cases fell within the intermediate-risk category 62.6%, followed by high-risk cases 23.1%, resulting in a combined 85.8% classified as clinically significant PCa. This distribution suggests that most cases in the cohort present with disease requiring active intervention rather than conservative management. In contrast, only 14.2% of cases were classified as low-risk, indicative of a minority with indolent Tumours typically managed through active surveillance. This pattern may reflect delayed diagnosis due to limited routine screening, as well as potential genetic and environmental factors contributing to more advanced disease presentation.

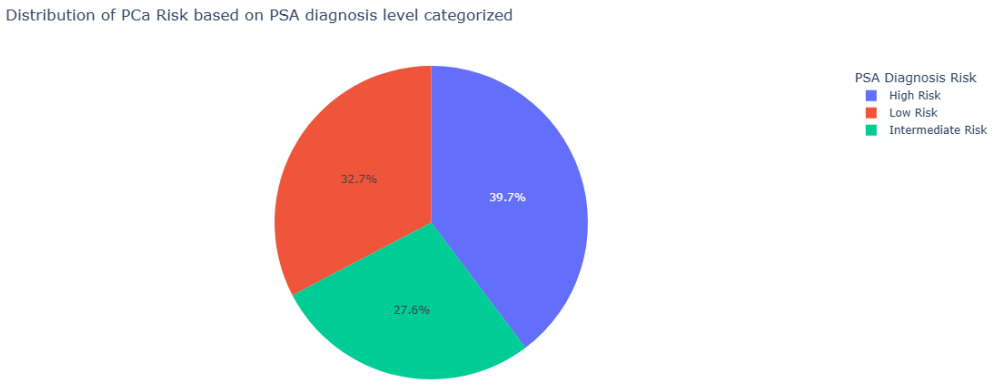

Figure 4.6: *PSA category classification categorized into three groups to represent low-risk, intermediate-risk and high-risk.*

The pie chart Figure 4.6 presents the distribution of PCa risk categories at diagnosis, stratified by PSA levels. The cohort is predominantly characterized by high-risk cases 39.7%, defined by PSA levels exceeding 100 ng/mL, a threshold indicative of aggressive disease. In contrast, intermediate- and low-risk categories constitute 32.7% and 27.7% of cases, respectively, reflecting markedly lower prevalence. The pronounced predominance of high-risk cases underscores a significant clinical and public health concern: the continued trend of late-stage disease presentation, which may reflect delays in detection, screening accessibility, or diagnostic follow-up. This distribution highlights the urgent need for enhanced early detection strategies and targeted interventions to mitigate advanced prostate cancer burden.

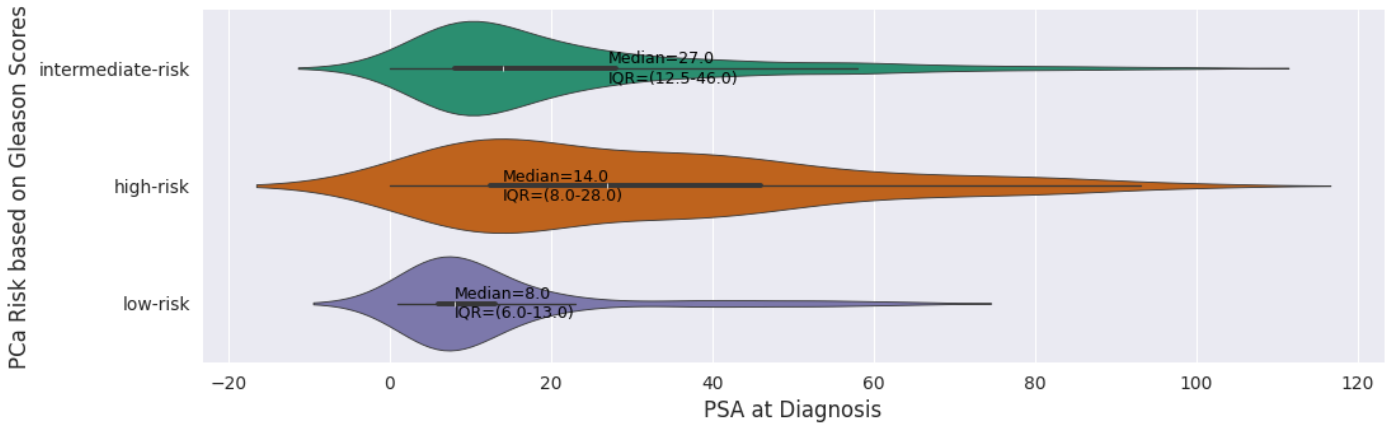

Figure 4.7: *Violin plot of PSA and GS to evaluate the visualizes the distribution of PSA at diagnosis across different risk categories for PCa represented by different colour groups: purple represents low-risk, orange intermediate-risk, and green represents high-risk.*

Table 4.2: Kruskal–Wallis and Pairwise Dunn Test Results (Bonferroni-adjusted p-values)

Comparison	High-risk	Intermediate-risk	Low-risk
High-risk	1.00	0.00	0.00
Intermediate-risk	0.00	1.00	0.00
Low-risk	0.00	0.00	1.00

Kruskal–Wallis Test: $H = 74.901, p < 0.0001$

Figure 4.7 presents a violin plot illustrating the distribution of PSA levels at diagnosis across three PCa risk categories. This visualization combines box plot elements with kernel density estimates, offering a depiction of both central tendency and distribution shapes within each group. The low-risk, (purple), shows a relatively narrow and symmetric distribution, suggesting a consistent PSA profile with minimal outliers. The intermediate-risk group (green) displays a broader distribution with an extended upper tail, indicating considerable heterogeneity and the presence of elevated PSA values despite a comparable central tendency to the low-risk group. The high-risk (orange) reveals a distinct right-skewed distribution, reflecting the clinical reality that high-risk cases frequently present with markedly elevated PSA levels. These patterns highlight the variability in PSA expression across risk groups and underscore the necessity of an integrated diagnostic approach; incorporating GS, imaging, and other clinical parameters, to improve risk stratification and guide treatment decisions. The median PSA levels and (IQR) were as follows: low-risk, 8.0 $(IQR: 6.0 - 13.0)$; intermediate-risk, 7.7 $(IQR: 5.2 - 46.0)$; and high-risk, 14.0 $(IQR: 8.0 - 82.0)$.

The Kruskal–Wallis test revealed a significant difference in PSA levels across risk groups $(H = 74.901, p < 0.0001)$. Pairwise Dunn tests confirmed that all group comparisons were statistically significant $(p < 0.001)$, indicating clear distributional separation between high, intermediate, and low-risk categories. These findings validate PSA as a strong stratification marker and support the clinical relevance of the risk groupings.

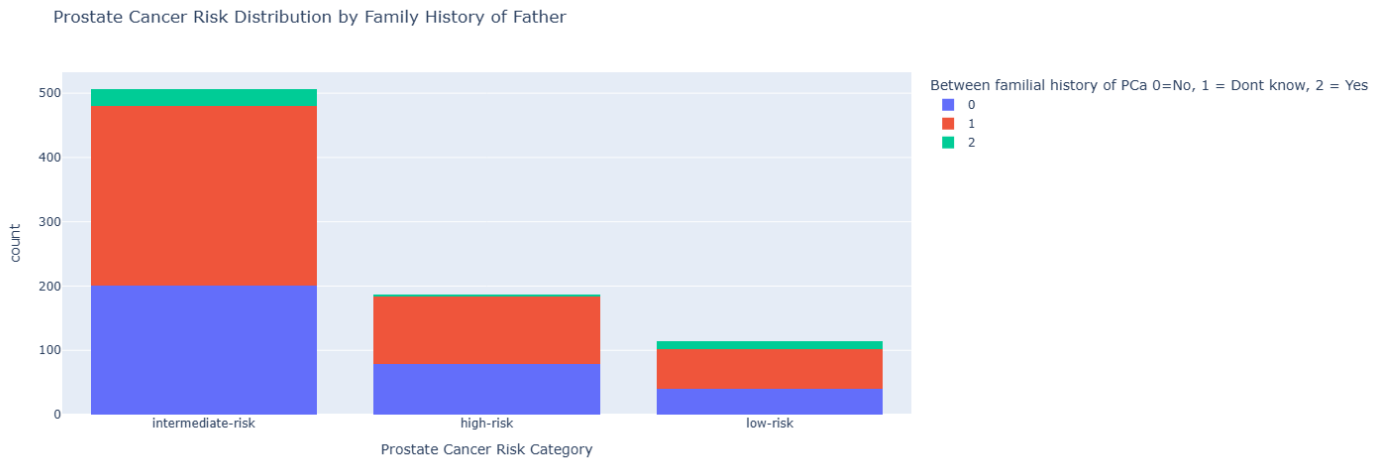

Figure 4.8: *A stacked bar chart visualizing the association between familial history PCa and severity, blue (0): No history of PCa in the father. red (1): Uncertain/Don't know. Green (2): Yes, the father had PCa*

Figure 4.8 illustrates participants' knowledge of their fathers' PCa history. The majority of respondents reported no known paternal history of PCa (blue), while a substantial proportion indicated uncertainty (red, "Don't know"). A confirmed paternal history of PCa (green, "Yes") was the least frequently reported. Notably, the prevalence of uncertainty was high across all risk categories 53.9% in the low-risk group, 55.3% in the intermediate-risk group, and 56.9% in the high-risk group; highlighting a significant gap in awareness of family medical history. This lack of knowledge may undermine the effectiveness of early detection and prevention strategies that depend on accurate familial risk assessment. Interestingly, a small subset of participants with a confirmed paternal history was observed across all disease severity levels: 34.7% in the low-risk group, 5.1% in the intermediate-risk group, and 41.9% in the high-risk group. This distribution suggests that while hereditary factors may influence PCa risk, they are not the sole or dominant determinant of disease severity within this cohort

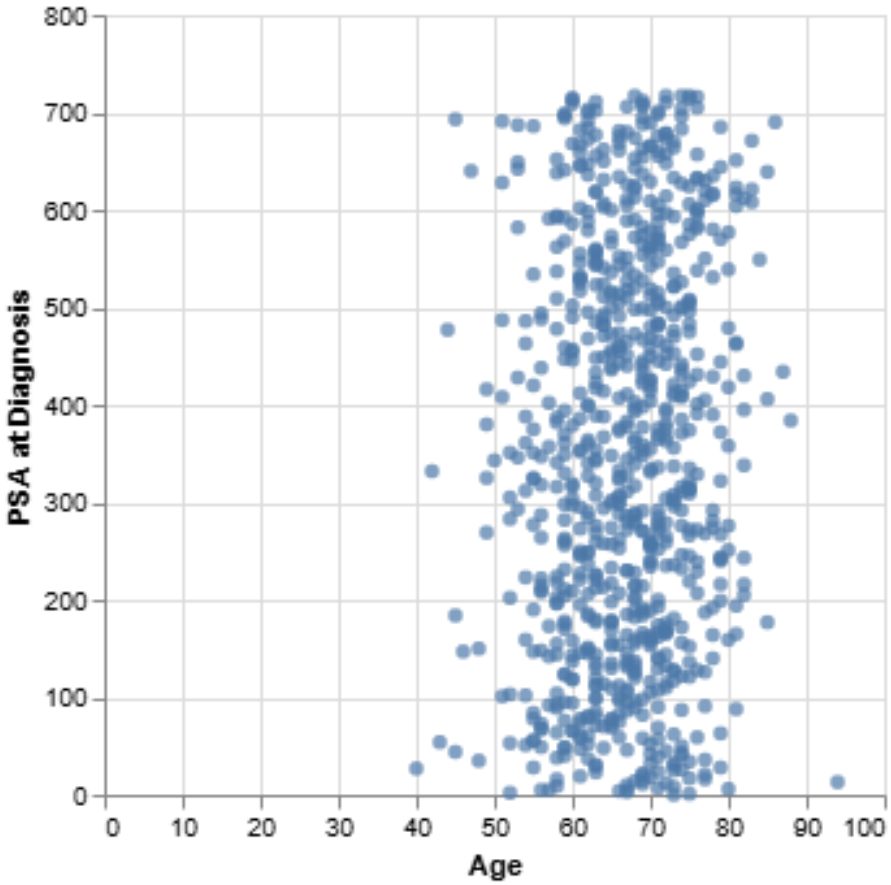

Figure 4.9: *Scatter plot showing the relationship between PSA levels at diagnosis and age among cases in the MADCaP database, illustrating variability in PSA distribution across age groups.*

The scatter plot in Figure 4.9 illustrates the distribution between age and PSA at diagnosis of men that were diagnosed PCa. The data reveal a concentrated distribution of diagnoses between the ages of approximately $50 - 90$, consistent with the epidemiological profile of PCa as a disease predominantly affecting older men. Within this age band, PSA levels exhibit substantial variability, with several cases exceeding 700 ng/mL: indicative of potentially advanced disease. The absence of a clear linear trend between age and PSA suggests that while age is a known risk factor for PCa, PSA elevation at diagnosis is not uniformly associated with increasing age. Instead, the wide dispersion of PSA values across the age scale underscores the heterogeneity of disease presentation and reinforces the need for individualized diagnostic assessment.

This pattern further supports the clinical imperative for age-appropriate screening protocols that account for both chronological age and PSA dynamics, rather than relying solely on age thresholds. The presence of markedly elevated PSA levels in cases across various age groups may reflect delayed diagnosis, biological aggressiveness, or disparities in healthcare access, particularly relevant in resource constrained settings.

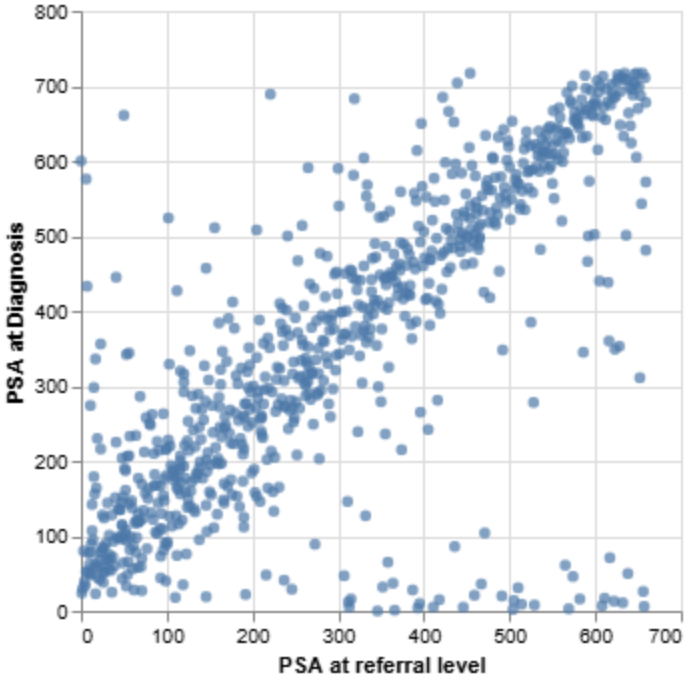

Figure 4.10: *Scatter plot for correlation between PSA referral level and PSA diagnostic level*

Figure 4.10 above displays a scatter plot that shows clear positive association between PSA referral and diagnosis. The plot reveals a dense clustering of points that generally follow a positive linear trend, suggesting a strong association between PSA values recorded at the point of referral and those measured at diagnosis. This observed pattern implies that elevated PSA levels at referral are predictive of similarly elevated levels at the time of formal diagnosis, reinforcing the clinical utility of referral PSA as an early indicator of disease burden. The consistency of this relationship across the cohort supports the reliability of PSA screening in identifying cases who may require expedited diagnostic workup. However, the presence of variability around the trend line also highlights potential delays or fluctuations in disease progression between referral and diagnosis, which may be influenced by healthcare system factors, biological variability, or timing of follow-up assessments. Clinically, the findings underscore the importance of timely referral and diagnostic confirmation

following elevated PSA readings, particularly in settings where delays may contribute to disease advancement.

Table 4.3: Correlation Tests for Prostate Cancer Dataset

Variable 1	**Variable 2**	**Test Type**	**Correlation Coefficient / χ^2**	**P-value**
Age (years)	PSA at Diagnosis (ng/mL)	Pearson	$r = 0.10$	0.0017
PSA at Referral Level (ng/mL)	PSA at Diagnosis (ng/mL)	Pearson	$r = 0.72$	5.91×10^{-172}
Final Staging – Tumour (T)	Final Staging – Metastasis (M)	Spearman	$r_s = 0.29$	9.72×10^{-23}
PCa Risk Category	Family History – Father with Cancer	Chi-Square	$\chi^2 = 16.66$	0.002
PCa Risk Category	Family History – Brother(s) with Cancer	Chi-Square	$\chi^2 = 11.08$	0.025
Cigarette Consumption (per day)	PSA at Diagnosis (ng/mL)	Spearman	$r_s = 0.35$	4.23×10^{-15}
PCa Risk Category	Family History – Son(s) with Cancer	Chi-Square	$\chi^2 = 1.80$	0.77

Table 4.3 shows correlation analysis of the PCa dataset, which uncovers associations between severity of the diseases and the range of clinical values and lifestyle factors. Age and PSA at diagnosis show a weak correlation $r = 0.10, p = 0.002$, indicating that, PSA levels increase with age. PSA at referral and diagnosis level is robust $r = 0.72, p = 5.91^{-172}$, signifying that PSA levels at referral are predictive of PSA at diagnosis and PCa severity.

Results from this study showed that there is moderate correlation between final staging (T) and metastasis (M) $r = 0.29, p = 9.72^{-23}$. This suggests that higher Tumour stages are associated with metastasis. Additionally, chi-square tests reveal further associations between PCa_risk and family history. For example, father's history of PCa wass related to severity of PCa $X^2 = 16.66, 0.03$. Brother's having PCa is also associated with increased $X^2 = 11.08, p = 0.03$. These results suggest a potential hereditary factors or shared environmental component influencing PCa severity. There is a moderate correlation $r = 0.35, p = 4.23^{-15}$, with cigarettes count per day, and PSA at diagnosis. This indicates that smoking might contribute PSA levels.

4.2.3 Multivariate Analysis

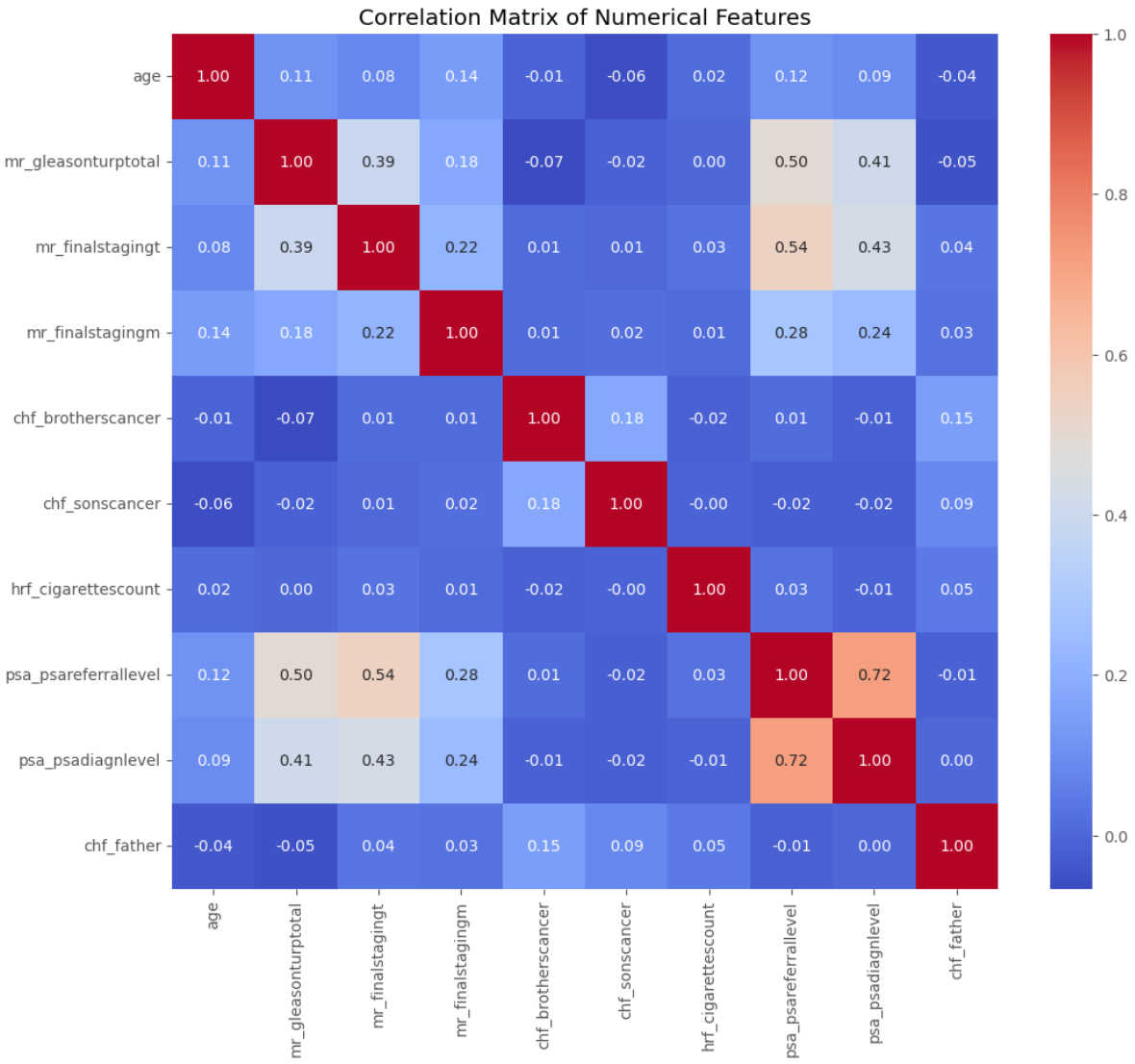

Figure 4.11: Correlation heat-map showing variable relationships

Figure 4.11 presents a correlation heatmap illustrating the interrelationships among key clinical and pathological variables within the MADCaP dataset. Variables with a correlation coefficient ≥ 0.8 were classified as highly correlated. Notably, $GS\ total$ and $final\ staging(T)$ exhibited a moderate correlation $r = 0.39$, suggesting a partial association between histological grading and Tumour extent. $PSA\ at\ referral$ and $PSA\ at\ diagnosis$ demonstrated a moderate correlation $r = 0.54$, while PSA at diagnosis and final staging (T) were similarly associated $r = 0.43$. These findings reinforce the role of PSA as a biomarker for PCa progression, with elevated PSA levels tending to coincide with more advanced disease stages.

A particularly strong correlation was observed between $PSA\ at\ referral$ and $PSA\ at\ diagnosis$ $r = 0.72$, which aligns with clinical expectations given the temporal proximity and biological continuity of these measurements. In contrast, age showed weak correlations with most variables: GS $r = 0.11$, $final\ staging(T)$ $r = 0.08$, $final\ staging(M)$ $r = 0.14$, $PSA\ at\ diagnosis$ $r = 0.09$, and $PSA\ at\ referral$ $r = 0.12$. These low coefficients suggest that age alone may not be a robust indicator of disease severity within this cohort. Nevertheless, given prior literature identifying age as a risk factor for prostate cancer incidence, the variable was retained in the analysis to preserve

epidemiological relevance.

4.3 Results of Unsupervised Machine Learning on PCa

Table 4.4: List of variables included in the PCA analysis.

No.	Variable
1	Cigarettescount
2	PSA at referral level
3	PSA at diagnosis level
4	Age
5	Income
6	Smoking
7	Final staging (T)
8	Final staging (M)
9	Father
10	Brother's cancer
11	Son's cancer
12	PCa risk

4.3.1 Principal Component Analysis

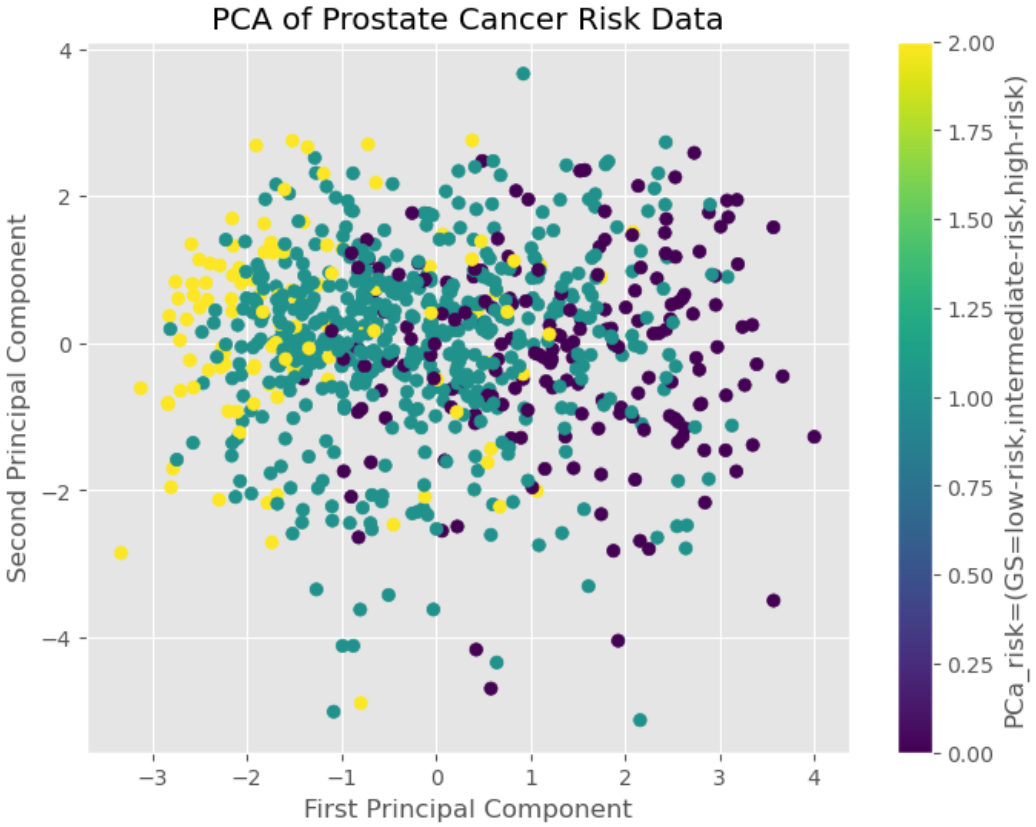

Figure 4.12: *Principal Component Analysis of selected prostate cancer risk profiles based on key clinical variables, including PSA levels, tumor staging, and Gleason Scores.*

PCA reduced dimensionality in data while preserving the variance, allowing these clinical features to cluster at different risk levels. From the scatter plot, data shows overlapping yet distinguishable clusters amongst the three risk groups. The first principal component (PC1) largely captures PCa variations are influenced by PSA, and Tumour staging risk factors, while the second principal component (PC2) may be influenced by age, smoking history, and familial PCa history. High-risk cases appear to be more concentrated in specific regions of the plot, possibly due to a combination of advanced Tumour staging, elevated PSA levels, and family history of PCa. These cases are likely to present with more severe PCa characteristics.

While intermediate-risk cases form a transitional region between high-risk and low-risk clusters 4.12. This suggests some cases share PCa characteristics with high-risk participants, while others may align closely with the low-risk individuals, depending on Tumour staging, PSA levels and lifestyle risk factors. Low-risk cases are more wide spread and blend into the the broader population, likely due to lower PSA levels, early stage Tumour classification and weaker associations with family history or smoking exposure 4.3. PCA provided a valuable dimensionality reduction method by reducing dimensional data into smaller set of summary indices to analyse complex PCa datasets. Improving efficiency by fully differentiating PCa risk categories.

4.4 K-means Clustering

4.4.1 K-means Clustering of dimensionally reduced PCa data

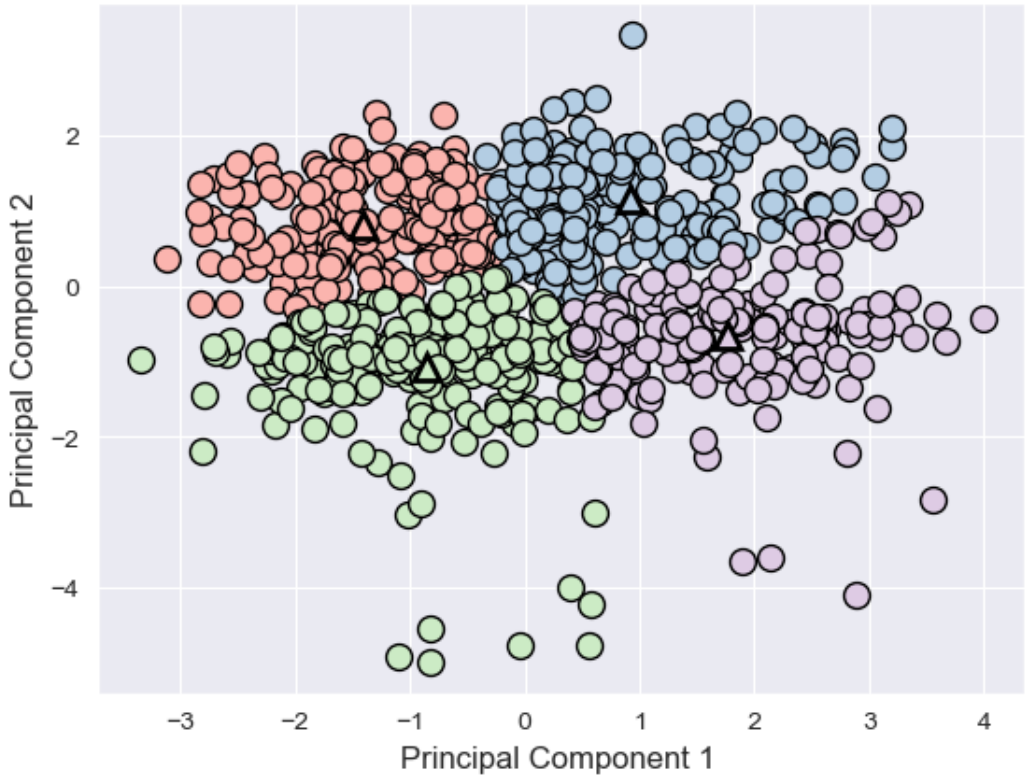

Figure 4.13: *K-means clustering applied to dimensionally reduced prostate cancer variables, illustrating the unsupervised grouping of cases based on underlying clinical features.*

K-means clustering Figure 4.13 provided insights on the distribution of PCa risk based on key clinical and demographic variables decomposed using PCA to reduce dimensionality while retaining the most informative aspects of the dataset. The plot has two axes labelled "Principal Component 1" and "Principal Component 2" representing two principal components of the data. The clusters were colour-coded into four colours: pink, blue, green, and purple. Each cluster has a centroid marked by a black triangle. K-means identified four significant groupings among PCa cases. The centroids of each cluster represent the average characteristics of each group, which may correspond to different PCa risk levels. The separation between clusters is influenced by the PCA decomposed variables, indicating that the chosen risk factors contribute significantly to risk stratification.

4.4.2 K-means clustering on PSA levels

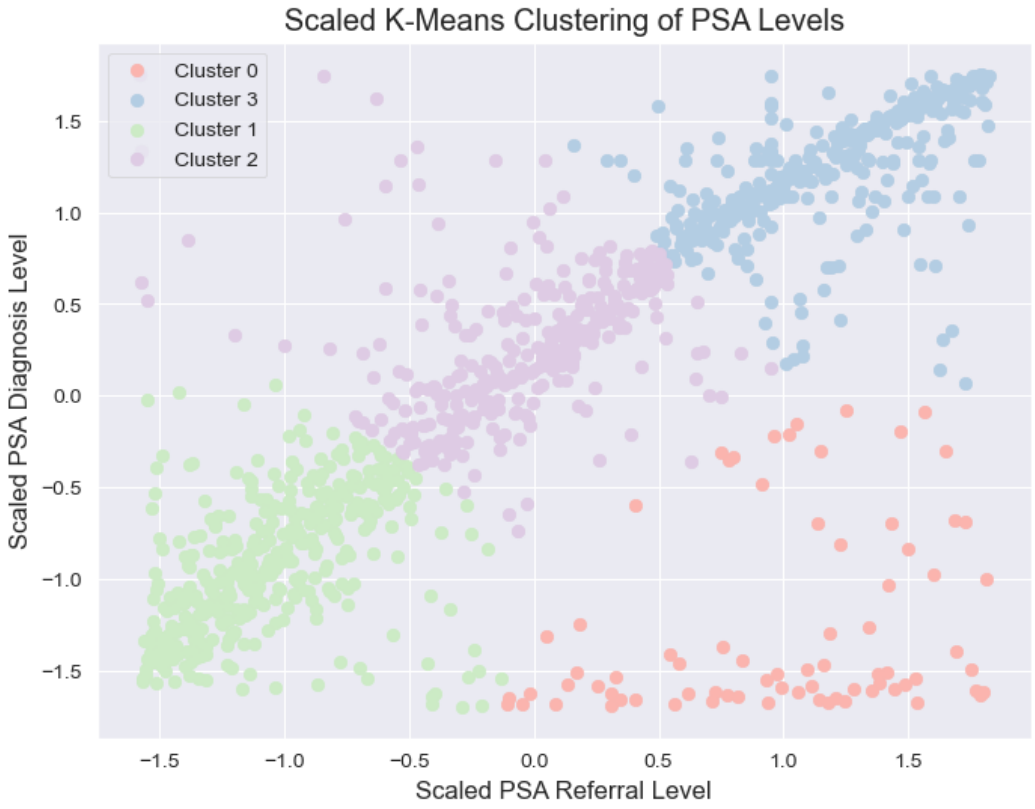

Figure 4.14: *Using K-means clustering on PSA at referral and diagnosis levels to extract PCa severity*

The PSA levels demonstrated a strong correlation of 0.72 Figure 4.11, necessitating further analysis using K-means clustering shown in Figure 4.14, to gain deeper insights into PCa severity based on PSA levels at referral and diagnosis. The clustering analysis identified four distinct groups, each representing different levels of PCa severity. Cluster 1 (green) likely comprises cases with low PSA levels at both referral and diagnosis. Cluster 2 (purple) may represent cases with intermediate PSA levels at both stages, indicating moderate disease severity. Cluster 3 (blue) includes cases with high PSA levels at both referral and diagnosis, potentially corresponding to those with more aggressive disease. Finally, Cluster 0 (red) may include cases of misdiagnosis or cases with a high GS, signifying a high-risk category.

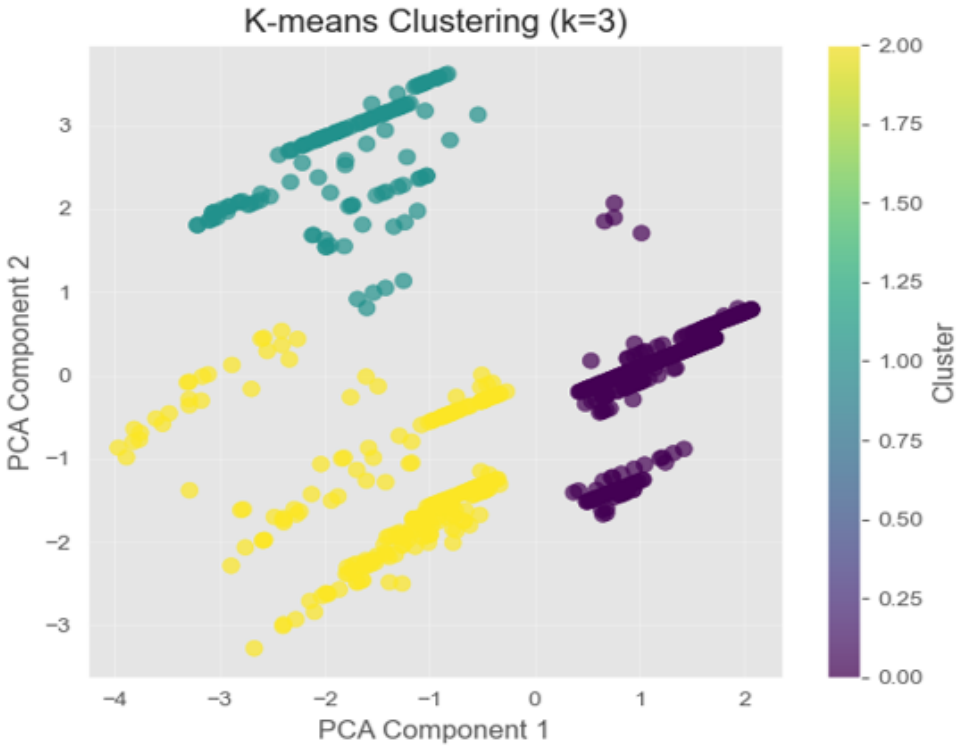

Figure 4.15: *K-means clustering of PCa cases (k=3) based on PSA, Tumour staging (TM), and GS. Following ISUP low-risk, intermediate-, and high-risk categories defined clinical guidelines.*

4.4.3 K-means Clustering of PSA, Tumour Staging, and Gleason Score

The clustering analysis Figure 4.15 revealed three distinct subgroups of PCa cases that closely align with recognized clinical risk stratifications. One subgroup was characterized by low PSA levels, organ-confined disease (stage T1–T2, M0), and ≤ 6, corresponding to a low-risk category. A second subgroup encompassed cases with intermediate PSA values, limited disease extension, and GS of 7 ($3+4$ and $4+3$), consistent with an intermediate-risk classification. The third subgroup comprised cases with markedly elevated PSA levels, advanced local or metastatic disease (stage T3–T4 or M1), and GS ≥ 8, indicative of high-risk disease.

These stratified clusters demonstrated strong coherence with established clinical guidelines, suggesting that unsupervised learning approaches can effectively capture and summarize conventional risk categories from underlying patient data. However, minor overlaps observed at the boundaries between clusters point to residual heterogeneity within standard classifications. This finding highlights the potential of data-driven methods to uncover more distinct PCa subtypes, which may inform improved risk stratification and personalized management strategies beyond current standardized frameworks.

4.4.4 Hierarchical Clustering Dendrogram

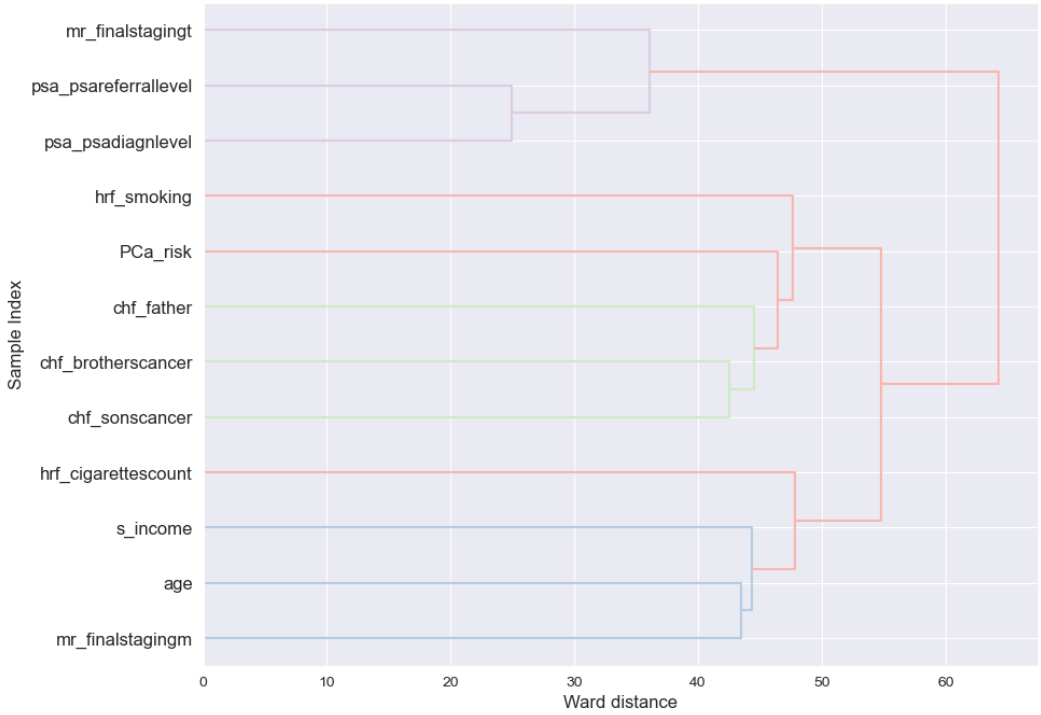

Figure 4.16: *Hierarchical clustering dendrogram of prostate cancer (PCa) risk factors, illustrating the stepwise grouping of variables based on clinical similarity.*

Figure 4.16 presents a HC of key clinical and demographic variables associated with PCa severity. The dendrogram reveals meaningful groupings based on feature similarity, offering insight into how different factors coincide in relation to disease progression. PSA-related variables, specifically $PSA\ at\ referral$ and PSA at diagnosis, cluster closely together, reflecting their shared role in assessing Tumour burden. $Tumour\ staging\ (TM)$ also form a distinct cluster, underscoring their combined relevance in defining disease extent. Interestingly, PCa risk is grouped alongside smoking-related and hereditary cancer factors, suggesting that both lifestyle and genetic predisposition may contribute to disease severity.

In terms of risk stratification, the clustering structure mirrors conventional PCa classifications. Low-risk cases tend to form isolated clusters, characterized by lower PSA levels and early-stage Tumour. Intermediate-risk cases exhibit overlaps with both low and high-risk groups, indicating a transitional severity profile. High-risk cases cluster with advanced staging, elevated PSA levels, and stronger familial predisposition. This multidimensional clustering highlights the complex interplay of clinical, behavioural, and genetic factors in PCa progression and reinforces the need for comprehensive screening approaches that integrate these domains.

4.5 Agglomerative Clustering

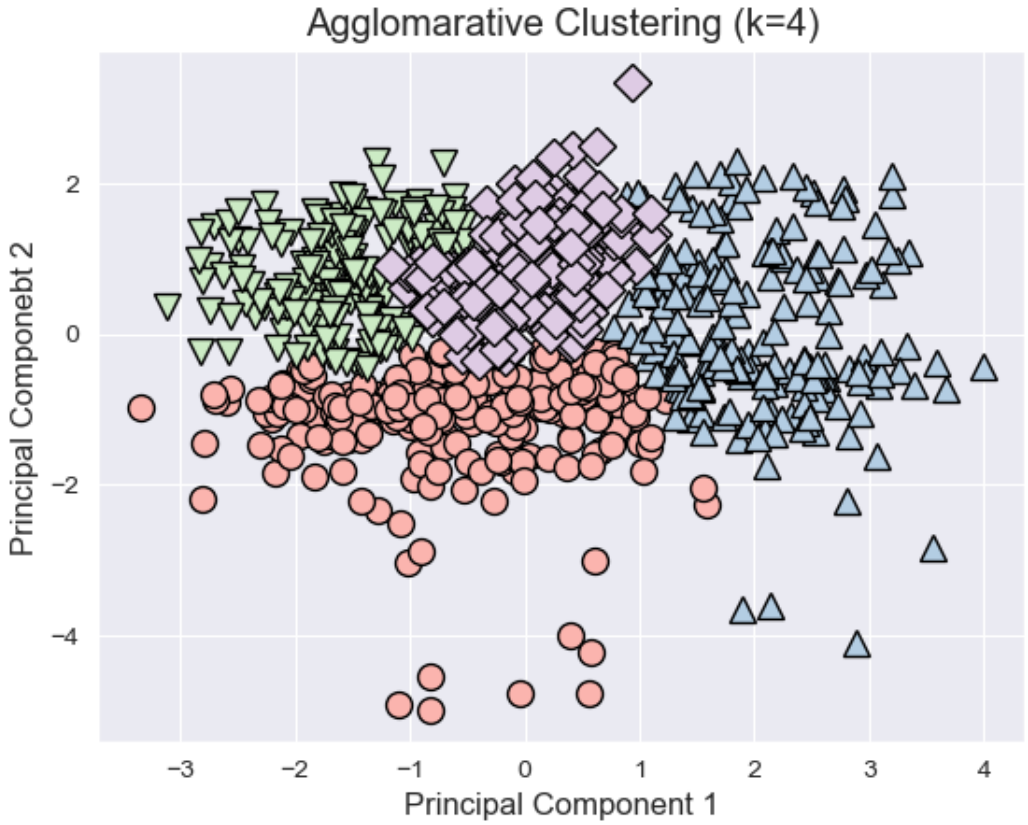

Figure 4.17: *Agglomerative Clustering analysis, applied in the context of PCa*

Results show that there were four distinct clusters which correspond to different PCa risk-profile shown in Figure 4.17 in colours blue, green, purple and red. AC was chosen because it builds nested clusters by successfully merging or splitting them based on similarity. The clusters represent the PCa risk groups Cluster 1 (pink) low-risk group, with lower PSA level, less aggressive Tumour staging. Cluster 2 (green) intermediate group with moderate levels of PSA, $Tumour\ staging$, mixed or positive family history. These necessitate closer monitoring and potentially more intensive treatments.

Cluster 3 (purple) are the high-risk cases requiring comprehensive and aggressive treatment plans due to the clinical presentation of high PSA, advanced Tumour staging, a family history of PCa, exposure to smoking and a low economic status. Cluster 4 (blue) may be transitional group containing PCa outliers or mixed cases, with inconsistent PSA level and a risk factors that may not align strongly with other clusters.

4.5.1 Agglomerative Clustering on PSA at diagnosis and Age

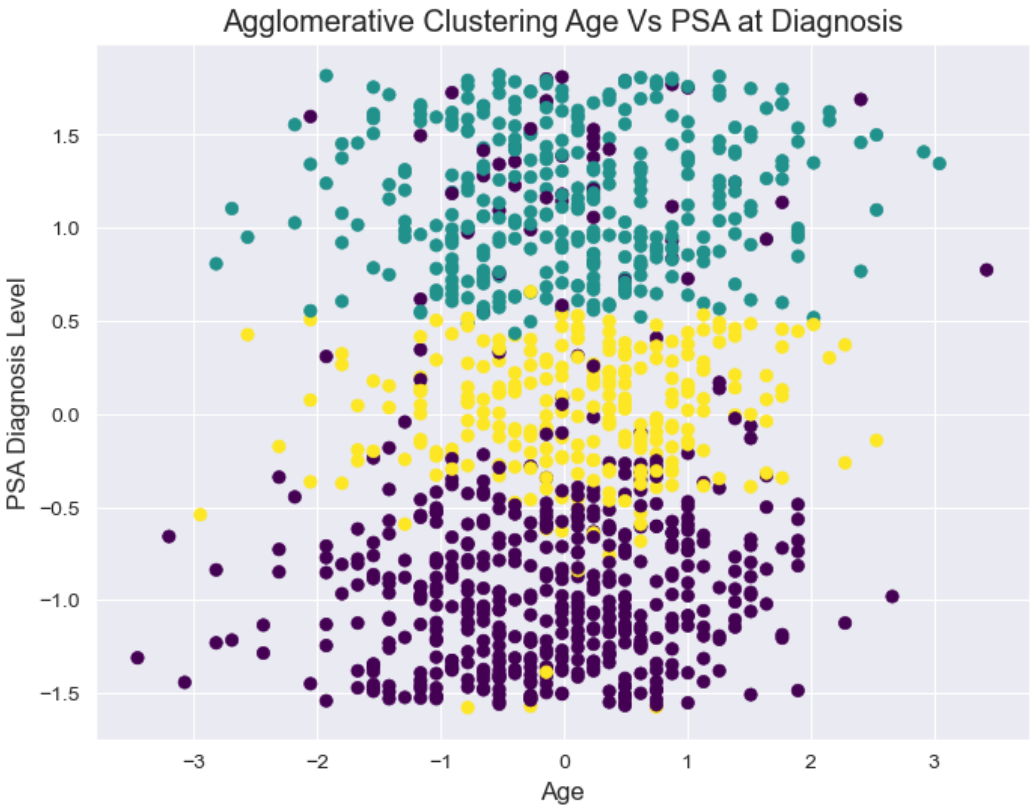

Figure 4.18: *A two dimensional scatter plot of Age and PSA at diagnosis, were three clusters of PCa cases identified by the agglomerative clustering algorithm, the values on the x-axis (Age) and y-axis PSA at diagnosis are expressed in terms of standard deviations from the mean rather than their original units.*

To further investigate the association between age and delayed PCa diagnosis, age and PSA levels were clustered for analysis. Figure 4.18 presents the outcomes of AC, offering valuable insights into the risk stratification of PCa cases based on age and PSA levels at diagnosis. The upper Cluster (green) comprises participants with PSA levels above 0.5, representing cases with severe PCa. Elevated PSA values in this group are indicative of advanced Tumour stages, with the cluster consisting of older men.

The middle Cluster (yellow), characterized by PSA levels between -0.5 and 0.5, represents an intermediate-risk category. These cases may correspond to localised PCa of moderate severity, with a broad age distribution. The overlap between this group and the low-risk category suggests that moderate PSA levels are observed across various age ranges. Lastly, the lower Cluster (purple) includes cases with PSA levels ranging from -0.5 to -1.5, indicative of lower disease severity, reflecting indolent PCa. This group consists of individuals across different age groups, suggesting the presence of early-diagnosed or less aggressive Tumour types.

4.5.2 Agglomerative Clustering on PSA, Tumour staging (T and M), and GS

Figure 4.19: *Agglomerative clustering of PCa cases (k=3) based on PSA, Tumour staging (T and M), and GS. Following ISUP low-risk, intermediate-, and high-risk categories defined clinical guidelines.*

The analysis yielded three distinct patient subgroups Figure 4.19 illustrates the emergence of three distinct case subgroups based on key clinical features associated with PCa severity. When compared to established risk stratification frameworks, such as those defined by ISUP and local clinical guidelines, a partial overlap was observed. One subgroup corresponded closely to the guideline-defined low-risk category, characterized by lower PSA levels, early-stage Tumours, and lower GS. In contrast, the remaining two subgroups comprised heterogeneous combinations of cases traditionally classified as intermediate and high-risk, suggesting that conventional criteria may obscure underlying variability within these categories.

This divergence indicates that the clustering approach captured latent heterogeneity not fully accounted for by standard classification systems. Specifically, the algorithm appears to have defined more granular disease patterns within the intermediate and high-risk divisions, revealing subpopulations that may differ in prognosis, treatment response, or underlying biology. These findings underscore the potential of UML methods to augment existing clinical frameworks by uncovering clinically relevant subgroups that transcend traditional boundaries. Such insights may inform more personalized risk stratification and guide tailored therapeutic strategies in diverse patient populations.

4.6 DBSCAN

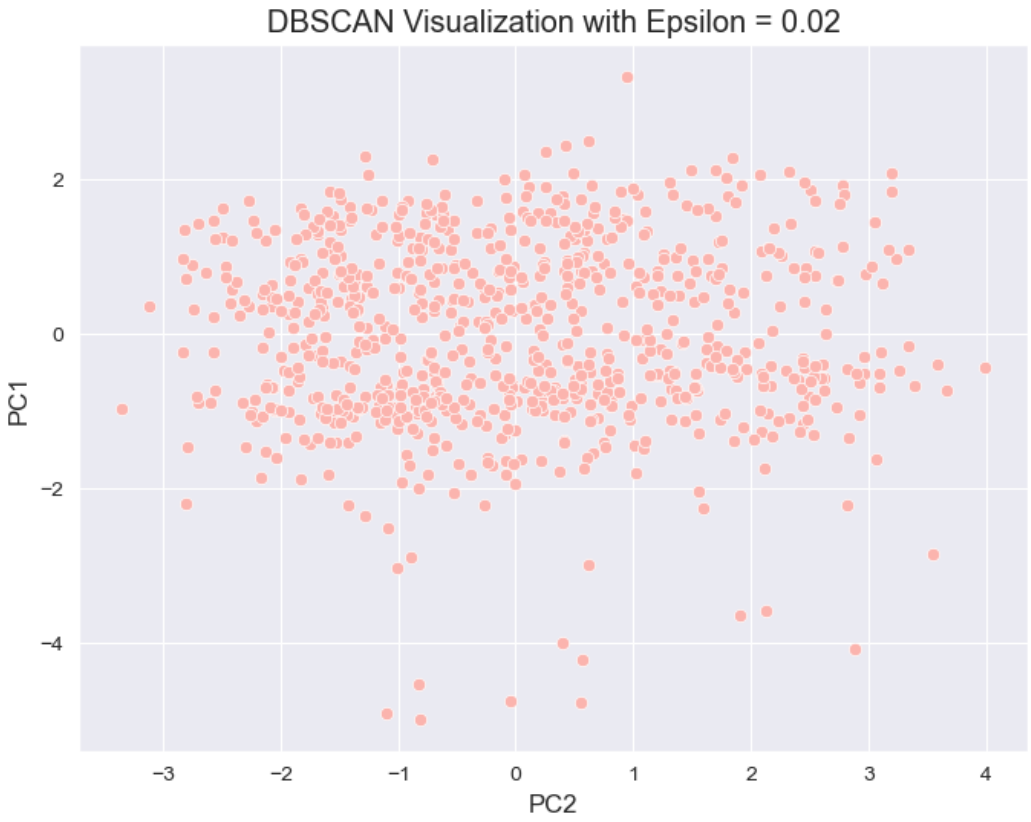

Figure 4.20: *DBSCAN on decomposed MADCaP dataset to identify outliers within the data*

DBSCAN clustering with an epsilon of 0.2, applied visualization showed a prominent horizontal cluster, indicating that many PCa cases have similar clinical features, represented along the first PC1. This suggests that PC1 plays a significant role in distinguishing risk groups related to clinical factors like $PSA\ levels$, Tumour staging (TM), and genetics. The scattered points, especially in the lower areas of the plot, highlight outliers who may have unique or extreme risk profiles, such as those with advanced Tumours, high PSA values, or a strong family history of PCa.

Results indicate outliers, identified as noise by DBSCAN, represent cases whose features significantly differ from the main clusters. Furthermore, the vertical spread along PC2, although less noticeable, may reflect secondary factors such as socioeconomic status, smoking history, or family history. The presence of unclassified cases points to clinical heterogeneity, where individuals do not fit neatly into specific groups, highlighting the complexity of classifying PCa severity.

DBSCAN clustering algorithm on Age and metastasis (M)

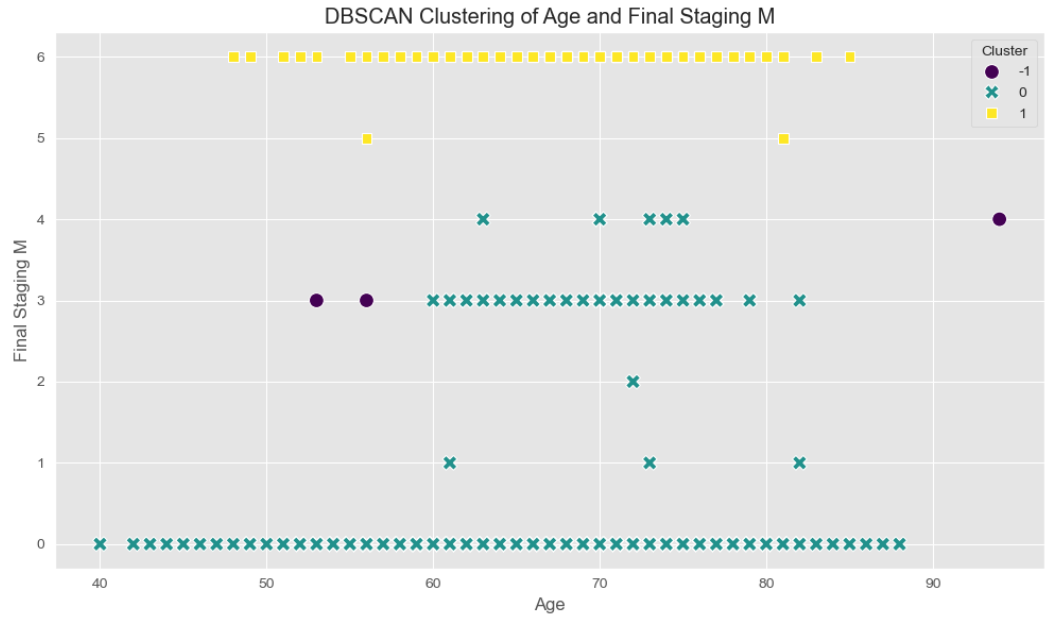

Figure 4.21: *DBSCAN algorithm on age and metastasis (M), This means that the values on the x-axis (Age) and y-axis (Final Staging M) are expressed in terms of standard deviations from the mean rather than their original units.*

The DBSCAN algorithm to standardized values of age (x-axis) and metastasis stage (M) y-axis with both variables scaled to express standard deviations from their mean generated two primary clusters alongside a limited set of outliers Cluster = – 1 in Figure 4.21. The first cluster consisted of cases exhibiting no evidence of metastasis M = 0 distributed across a broad range of ages, reflecting a consistent profile of localized disease independent of cases age. A second distinct cluster included cases with advanced metastatic disease, also spanning a wide age range, which suggests that metastatic burden rather than age served as the dominant factor in cluster formation. A small number of cases with intermediate metastatic stages were identified as noise, underscoring the presence of clinically heterogeneous cases that do not conform to densely grouped cluster patterns. This result emphasizes the utility of DBSCAN in distinguishing dominant case subgroups based on severe metastatic status, while also revealing less common transitional states that may represent unique pathological trails.

4.7 Spectral Clustering

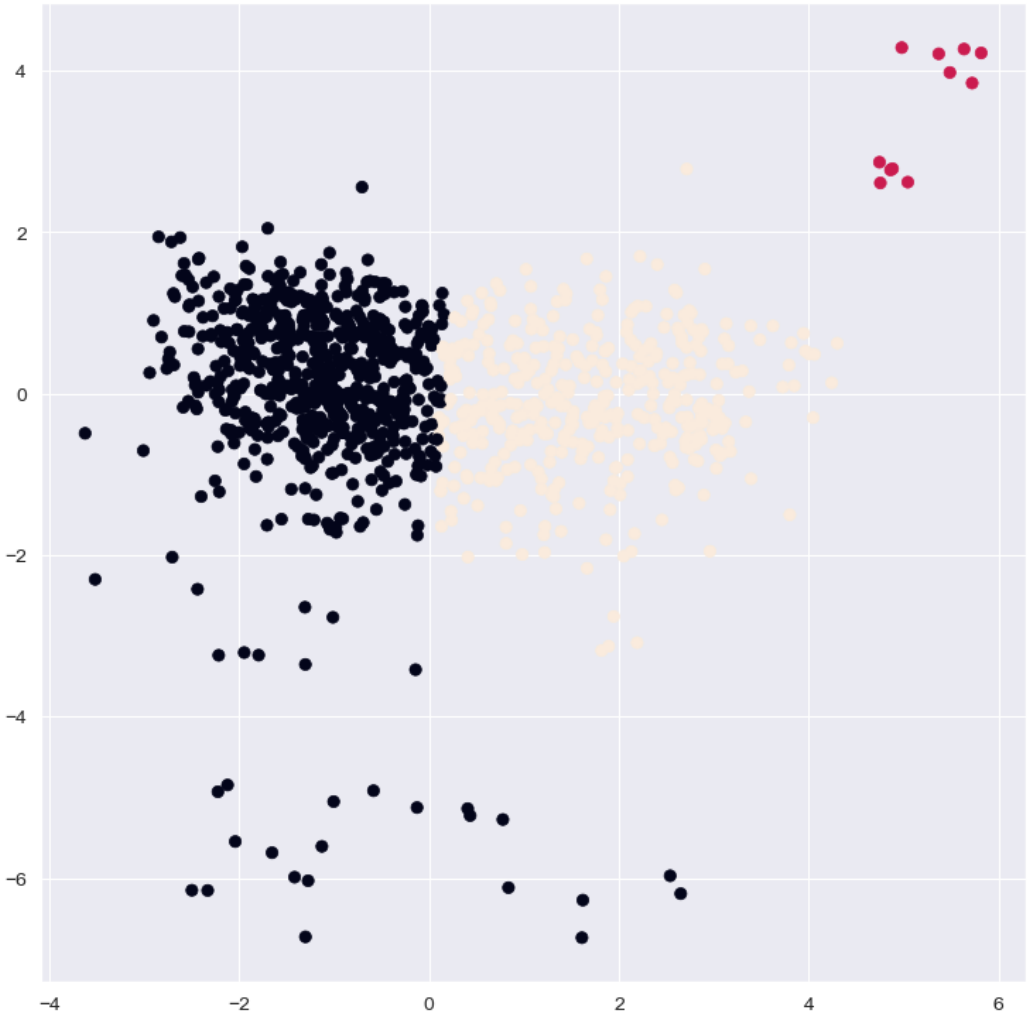

Figure 4.22: *Spectral clustering of decomposed data from PCA.*

SC on PCa dataset leveraging the eigenvalues similarity matrix to identify cluster, particularly those that are not necessarily spherical in shape. The plot reveals three dissimilar clusters, differentiated by colour: dark, teal, and yellow. The dark purple cluster is concentrated along the left side of the graph, showing a group of cases with less severe PCa, characterized by a younger demographic with lower risk profiles. The teal cluster appears vertically aligned, showing variability in PSA levels but remaining primary associated with younger ages, suggesting a subgroup with slightly elevated PSA levels, potentially identifying cases at increased risk for PCa due to elevated PSA reading.

The low-risk group consists of cases with early-stage Tumours, lower PSA levels, and a minimal risk of metastasis. These cases usually show a low disease severity and could be managed through active surveillance instead of immediate treatment. The intermediate-risk group includes those with a moderate Tumour burden, varying PSA levels, and certain risk factors that could lead to severe PCa. This group needs careful monitoring to assess whether they might move into a higher severity category.

In contrast, the high-risk group is clearly defined, encompassing cases with advanced PCa features, such as elevated PSA levels, aggressive Tumour staging, and a higher likelihood of metastasis. Individuals in this group face a significant risk of poor outcomes and require more aggressive treatment approaches, including surgery, radiation, or systemic therapies. SC proved to be particularly useful in identifying these severity-based groups because it can capture complex, non-linear relationships among variables. The algorithm's use of similarity graphs allows for a more accurate depiction of PCa severity gradients, it provides a refined, data-driven classification of PCa severity, which could improve personalized treatment strategies and risk-based patient management.

4.7.1 Spectral clustering on age and Tumour staging (T)

Figure 4.23: *Spectral clustering on age and Tumour staging (T)*

The SC visualization offers a clear breakdown of PCa severity, considering both age and the $final\ staging(T)$. How the cases are grouped suggests distinct categories that align with varying levels of PCa progression. The variable $Tumour\ staging\ (T)$ is an indicator of how far the Tumour has invaded, a key aspect of how severe the PCa has become.

Age also contributes significantly in PCa, affecting when PCa begins, and how it progresses. The clusters reveal that younger cases have lower-risk because of low Tumour, while older cases show a wider range of staging, indicative of different stages of progression. These clusters suggest potential risk-based groups, where Tumours at lower stages are separated from those that are advanced. Clusters, particularly those with higher staging (T), may represent cases with high-risk profiles who need close monitoring and more aggressive treatment strategies. See Figure 4.23

4.7.2 Performance Evaluation

Table 4.5 shows that the first PC1 captured 20.85% of the total variance, indicating that PC1 is the most influential in differentiating patterns in risk stratification. This indicates that a sizeable portion of the variance is driven by clinical variables, such as, PSA levels, Tumour staging and age. The second principal component (PC2) explains 10.62% of the variance, contributing to the second axis this could be hereditary factors.

Table 4.5: Explained Variance of Principal Components

Principal Component	Explained Variance (%)
PC1	20.85
PC2	10.62
PC3	10.36
PC4	9.72
PC5	8.93
PC6	7.28
PC7	6.99
PC8	6.21
PC9	6.06
PC10	5.87
PC11	4.75
PC12	1.97

Further than PC2, the next principal components are PC3, PC4 and PC5, capturing variance of 10.36%, 9.72%, and 8.93%, collectively 29.22% of the variance,which may represent risk factors such as socioeconomic status, smoking history and PSA levels. The first five components capture over $60.48\%$of the total variance, meaning that clustering methods applied within the reduced dimensions can effectively stratify PCa severity levels while reducing the noise from less informative variables.

4.7.3 Silhouette Scores

Clustering Method	Silhouette Score	Optimal K
K-Means	0.37	4
K-means on PSA levels	0.47	4
K-means on PSA,GS, Tumour staging (TM)	0.51	3
Agglomerative clustering	0.32	4
Agglomerative clustering Age and PSA at Diagnosis	0.28	3
Agglomerative on PSA,GS, Tumour staging (TM)	0.58	3
Hierarchical Clustering	0.29	4
DBSCAN Age and Tumour staging M	0.70	3
Spectral clustering	0.47	3
Spectral clustering Age and Tumour staging T	0.37	4

Table 4.6: Silhouette Scores for Different Clustering Methods on PCa Data

The silhouette score, which assesses the goodness of fit and cohesiveness was utilized to evaluate the clustering models on PCa dataset. The scores range from -1 to $+1$, where the best score is $+1$ indicates clusters are clearly separated and -1 signifies poor clustering in the dataset. DBSCAN achieved the highest silhouette score 0.70 when clustering based on age and Tumour staging (M), suggesting that this method effectively captures the underlying patterns within the data and forms well-separated clusters. Similarly, DBSCAN applied to the broader dataset yielded a silhouette score of 0.65, further demonstrating its effectiveness in detecting natural groupings without requiring a predefined number of clusters.

Among the partition-based approaches, K-means clustering on PSA levels produced a relatively high silhouette score 0.47 with four clusters, indicating that PSA levels play a significant role in stratifying PCa severity. Standard K-means clustering and AC both achieved moderate scores of 0.37 and 0.32, respectively, with four clusters, suggesting comparable clustering performance. However, when AC was applied to age and PSA at diagnosis, the silhouette score decreased to 0.28, likely due to increased heterogeneity in the data.

Spectral clustering yielded varying results depending on the variables used. When applied to Tumour staging (T) and age, it produced a silhouette score of 0.37 with four clusters, while its performance on the broader dataset resulted in a score of 0.47 with three clusters. These findings suggest that SC can provide reasonable separability but may be sensitive to the choice of input variables. HC produced a silhouette score of 0.29, comparable to AC on age and PSA, suggesting that the hierarchical approach may not provide the most optimal separation for this dataset.

4.8 Summary

The outcomes of the UML uncovered distinct case clusters constructed PCa risk, $PSA\ levels$, Age, $lifestyles$, and $Staging$ classifications. These clusters afforded insight into the heterogeneousness and disparities of PCa severity, feature relationships that align with clinical observations. Valuation metrics, such as, silhouette Score, authenticated the robustness of our clustering algorithms, revealing well-defined and interpretable classes. The outcomes stress the capability of UML methods to stratify cases into clinically significant subdivisions, offering a groundwork for enhancing risk assessment and personalized PCa treatment and management policies. This methodology draws attention to how effective UML models can uncover hidden interactions in PCa data, providing more effective and tailored PCa interventions.

Chapter 5

Discussion

5.1 Prostate Cancer Profile Evaluation

PCa represents a global public health challenge, characterized by heterogeneity in progression, severity, and outcomes in different populations. Traditional risk stratification system, while useful, fail to capture the complex interplay of clinical, demographic, and socioeconomic factors especially in under-represented populations such as men of African ancestry, who experience disproportionately high rates of aggressive and advanced stage PCa. A major difference in our understanding PCa progression rests in our ability to efficiently stratify PCa cases centred on PCa risk factors and disease attributes by means of large complex dataset from MADCaP. In the face of progression in clinical and epidemiological research, existing methods for classification very often overlook the potential of data driven decision making, specifically in addressing inequalities in populations, such as, men of African origins, who experience excessively advanced rates of aggressive PCa compared to other ethnic groups [48].

The true value of UML in this context lies not only in its analytical value, however, in its translation of clinical findings into insights. This translation can be achieved through sustained collaboration among urologist, data scientist, and software engineers [49, 54]. A promising pathway is the integration of UML into EHR platforms could enhance clinical workflows by providing automated risk stratification of cases. To translate these findings into clinical value, a practical deployment path would involve embedding the validated clustering model as a risk stratification widget within the existing EHR systems [13]. The widget would function as a passive clinical support tool, routinely analysing case data during or after consultation assigning cluster based risk profiles [36].

The required data features for inference are clinical, and socio-demographic routinely collected in urological procedures, facilitating integration [14]. The model would operate on a scheduled updates (quarterly or biannually), retrained on the growing institutional data to ensure it evolves with shifting demographics and clinical practices [18, 19]. Essentially, a model monitoring framework must be established to continuously track performance metrics cluster drift, silhouette score, and the distribution of input variables to ensure the tool remains accurate, reliable, and clinically relevant [89]. This method would allow for a low-friction implementation that supports the urologist's workflow without interrupting it, ultimately aiming to standardize and improve risk assessment at

the point of care [32, 49].

Through integrating UML techniques in the PCa research; our efforts were to provide a framework for leveraging BD, AI, and UML to widen clinical understanding of the disease heterogeneity and progression in African men. This study, analysed retrospective PCa clinical data of cases in Soweto, South Africa, focusing on their demographic, socioeconomic and clinical features to categorized into PCa risk groups (low, intermediate and high risk) using UML techniques Table **??** [6]. The current stratification is based on Table 2.1, ISUP guided features while the models align with these frameworks, they capture additional nuances in African men that are underrepresented within the local and global datasets [48]. The application of UML should be evaluated, such clustering could inform tailored risk models that better reflect local disease biology and presentation patterns [6]. These findings may reinforce the value of UML in advancing the standards of personalized medicine in oncology, where treatment decisions are increasingly guided by patient specific risk stratification and biological profiles [2, 69].

An open question remains is whether the patient subgroups identified through UML correspond to the established guideline risk groups. Our analysis showed broad alignment, for example, clusters characterized by high PSA, GS and advanced $staging$ were clearly mapped to the defined categories. However, UML added clinically meaningful sub-groups within these broad clusters. Within the guideline-defined risk groups, clustering identified sub-clusters differentiated by PSA, GS and $staging$ relationships. These findings suggests that the current triage system, while effective, may mask heterogeneity by grouping together cases with distinct disease phenotypes [36, 42, 94]. Rather than contradicting existing guidelines, UML offers a data-driven framework to enhance them, especially in populations with unique clinical presentations [16].

The impact of PCa in LICs is profound, as it not only affects the individuals diagnosed but also places a significant burden on families and the healthcare systems [25]. Older cases often face more aggressive forms of PCa, which leads to higher morbidity rates and a greater loss to productive years of life [95]. This is particularly concerning in LICs, where healthcare infrastructure is already strained, and resources for PCa treatment are limited. The most reasonable explanations for severe outcomes in PCa in LICs include, but are not limited to, the inaccessibility of life-prolonging treatments such as radiation therapy, hormone therapy, and targeted molecular treatments [66]. Furthermore, in many LICs, healthcare systems are often under-resourced, and access to specialized PCa care is limited [12, 95]. Hospitals in these regions, including tertiary institutions, often experience delays in patient diagnosis. As a result, cases often present with advanced stages of PCa which are more difficult to treat [4]. By uncovering this hidden heterogeneity, clustering can enable more nuanced triaging, for example, prioritizing select high-risk subgroups for immediate intervention while flagging certain intermediate-risk cases who may exhibit more aggressive disease and require closer surveillance [32, 50]. This method demonstrates how AI could complement established guidelines to move toward more personalized treatment and efficient resource allocation in LICs [59].

The younger age of PCa onset in observed in LICs, particularly in SSA, maybe partially attributed to a mixture of genetic, environmental, and lifestyle factors [14]. For instance, genetic susceptibility to more aggressive forms of PCa have been observed in certain populations, which could contribute to the earlier onset and faster progression of the PCa [6, 15]. Additionally, the absence

of widespread screening programs, such as PSA testing, often result in delayed diagnosis at advanced stages of PCa, further worsening outcomes [13]. For example, clustering outputs based on $PSA\ levels$, $Tumour\ staging$ and GS features trigger identification of high risk cases may assist clinicians towards guideline based interventions methods [14]. Such systems would need the expert collaboration between urologist, oncologist, data scientist and engineers to ensure clinical interpretability, usability, reliability and adaptable to diverse health systems contexts, particularly in LICs where PCa burden is high with limited healthcare resources [4,6,48,58].

In SSA, the burden of infectious diseases and other non-communicable diseases often reduces the attention on PCa, leading to a lack of awareness and delayed diagnosis. The dominance of other health priorities, together with insufficient resources for PCa care, means that PCa cases in these regions are less likely to receive timely and effective treatment [56]. This underscores the pressing need for improved healthcare infrastructure, increased access to diagnostic tools, and the implementation of targeted screening programs to address the growing burden of PCa in LICs. ML could play a significant role in bridging these gaps by enabling earlier detection, risk stratification, and personalized treatment strategies, even in resource-limited settings.

Many studies have established that ML algorithms are quite effective for grouping of PCa in cases [89,96,97]. However, fewer studies have explored the application of UML algorithms for PCa in men of African descent [64,81,98,99]. In a review of past studies on PCa classification using UML clusters were derived from clinical variables such as GS, PSA, and tumour staging (TNM). For instance, Sammounda R et al (2021), used K-means clustering with elbow method on pixels of historical and near-infrared of PCa images [100]. The findings of the study highlight the potential of UML techniques to uncover hidden meaningful segments within PCa cohort dataset, which can inform personalized treatment strategies and improved clinical outcomes. An optimized K-means based segmentation method may improve the accuracy of PCa segmentation, particularly distinguishing different regions effectively [17,64,81,100].

Clustering algorithms uncovered distinct subgroups of PCa based clinical and sociological variables. This data-driven approach offers, clinicians a guided risk profiling and tailoring strategies, thereby influencing positive PCa management. This study is unique providing a novel exploration of UML in the understanding PCa pattens, in SSA context. By employing PCA, K-means, AC, HC, SC, and DBSCAN, the study managed to identify robust clusters grounded in features with with established clinical relevance, including GS, PSA, $staging$, and age. The consistency of certain features across multiple algorithms underscores their prognostic importance. The resulting clusters move beyond mere academic exercise; they provide actionable, insights that could potentially be integrated into clinical decision-support systems to enable more precise risk profiling, tailor treatment strategies, and ultimately work towards mitigating outcome disparities in the region.

The application of multiple clustering methods ensured a widespread understanding of the data, capturing both linear and non-linear associations, as well as atypical cases critical to management of PCa. PCa reduced data dimensionality while preserving key features like age, PSA, GS and $Tumour\ staging$. This enabled effective clustering without losing much of the information. The results emphasized GS, PSA, and $Tumourstaging$ are contributors to PCa severity. Through detection of hidden nuances in clusters, these methods could improve clinical decision-making and

customize treatment strategies. Moreover, integrating UML techniques into clinical practices could assist mitigate disparates in PCa severity and treatment by providing a more detailed understanding of disease severity.

5.2 Limitations

This study has several important limitations that must be considered when interpreting its findings. First, in the initial phase of the clustering analysis, the GS, a cornerstone of PCa risk stratification, was omitted. Although the analysis was revised to incorporate GS, its initial exclusion may have influenced the early clustering outputs and interpretations. Future work must ensure that all key clinical markers are systematically integrated from the outset to build a biologically relevant clustering models.

Second, the generalizability of our findings is constrained by the single-source, hospital-based nature of the cohort. The data were drawn from a tertiary referral centre CHBAH, which manages a disproportionate number of cases presenting advanced or complex disease. This inherently introduces a selection bias toward more severe PCa presentations, which skews the identified clusters and risks overestimating the prevalence of high-risk disease. This limitation is particularly significant given the established context that Black men are more prone to present with advanced disease and experience worse outcomes. Therefore, our clusters may mask presentations more common in community or screening settings and may not be generalizable to the broader population. Therefore, the external validity and clinical utility of these clusters must be confirmed through future validation across multi-center cohorts, including primary and secondary care facilities, to ensure robustness and broader applicability.

Third, the rigorous data cleaning process, while necessary, directly impacted on the analytical cohort size and its characteristics. During preprocessing, we identified extreme, physiologically implausible values in key variables, most notably a PSA value of 5000,2 ng/mL. Such artifacts, resulting from data entry errors, vastly exceed the known biological range for PSA. Their removal was mandatory to ensure the internal validity and biological plausibility of the analysis; however, this process reduced the final analytic cohort from 1,096 to 808 cases. This reduction necessarily diminished the statistical power of the study and may have altered the distribution of variables within the dataset, potentially biasing the resulting clusters away from the original population's true characteristics. This underscores a fundamental trade-off between data quality and sample size in retrospective analyses and highlights the critical need for automated data validation at the point of entry in future prospective studies.

Finally, the unsupervised nature of the machine learning approaches used means that the clinical significance and prognostic value of the identified clusters remain inferential. Without prospective validation against long-term patient outcomes, such as metastasis, or cancer-specific mortality , the true utility of these clusters for guiding clinical decision-making cannot be fully ascertained.

UML methods do not rely on labelled data, making it challenging to validate the accuracy of the clusters. Unlike supervised learning, there is no ground truth to compare the results. While DBSCAN was effective at identifying outliers, other methods like K-means may have overlooked rare but

clinically important subgroups, such as cases with atypical disease presentations. AC can capture complex and non-linear relationships, the resulting clusters may be difficult to interpret, especially in high-dimensional datasets.

Small clusters or outlier correlation might not have adequately reflected the entire information, affecting the dependability of inferences drawn about the rare subgroups. Finally, the analysis centred on predefined set of clinical variables, which, while insightful, might have disregarded other effective aspects such as nutritional of ecological exposures. Impending research could address these confines by incorporating, more diverse dataset and integrating multi-feature data to provide a more comprehensible understanding of heterogeneity in PCa.

Chapter 6

Conclusion And Future Directions

This study, examined PCa data by means of UML algorithms to uncover clinically significant correlations and patterns. The results indicates the importance of data quality in finding clinically significant insights from PCa data, particularly in grouping and patient stratification. Demographic, clinical and lifestyle information contain substructures for understanding PCa severity.

This study demonstrated the potential of UML to identify distinct patient subgroups based on clinical features including age, tumour staging, family history, smoking history, PSA levels, and socioeconomic status contribute to differentiating PCa risk levels $GS \leq 6$ (low-risk), $GS = 7$ (intermediate-risk), and $GS \geq 8$ (high-risk). Through clustering algorithms on patient data based on these multi-dimensional clinical predictors. Adding on more clinical features, such as, imaging and molecular data, could provide more thorough understanding of of PCa risk profiles.

The findings underscore the importance of using ML techniques to extract actionable insight from the complex dataset, specifically in the context of improving patient outcomes and improving patient to patient treatment. Study highlighted the potential of ML methods to derive actionable insights from complex PCa datasets. The ability to classify patients into clinically applicable clusters indicates that clustering methods may facilitate better clinical decision-making and personalised healthcare treatment strategies.

Additionally, the study lays groundwork for incorporating ML models into clinical practice, setting the stage for robust classification models that can assist in severity prediction and treatment optimization. Future studies should aim to validate these results in larger, prospective cohorts and compare UML-derived clusters with long-term patient outcomes. Additionally, investigating hybrid ML methods that combine supervised learning with clustering could further improve risk prediction models.

In summary, this study underscores the importance of machine learning in furthering prostate cancer research and enhancing patient outcomes, highlighting the need for high-quality datasets to fully realize its potential impact.

References

[1] Klaus Pantel, Catherine Alix-Panabieres, and Sabine Riethdorf. Cancer micrometastases. *Nature reviews Clinical oncology*, 6(6):339–351, 2009.

[2] 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: a cancer journal for clinicians*, 71(3):209–249, 2021.

[3] World Health Organization. *Report of the 12th FAO/WHO joint meeting on pesticide management: 19–22 November 2019, Geneva, Switzerland*. World Health Organization, 2020.

[4] Stefanie Gerstberger, Qingwen Jiang, and Karuna Ganesh. Metastasis. *Cell*, 186(8):1564–1579, 2023.

[5] Elisa C Woodhouse, Rodrigo F Chuaqui, and Lance A Liotta. General mechanisms of metastasis. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 80(S8):1529–1537, 1997.

[6] Freddie Bray, D Maxwell Parkin, Freddy Gnangnon, Gontse Tshisimogo, Jean-Felix Peko, Innocent Adoubi, Mathewos Assefa, Lamin Bojang, Baffour Awuah, Moussa Koulibaly, et al. Cancer in sub-saharan africa in 2020: a review of current estimates of the national burden, data gaps, and future needs. *The Lancet Oncology*, 23(6):719–728, 2022.

[7] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[8] Jacques Ferlay, Murielle Colombet, Isabelle Soerjomataram, Donald M Parkin, Marion Piñeros, Ariana Znaor, and Freddie Bray. Cancer statistics for the year 2020: An overview. *International journal of cancer*, 149(4):778–789, 2021.

[9] Olabode Omotoso, John Oluwafemi Teibo, Festus Adebayo Atiba, Tolulope Oladimeji, Oluwatomiwa Kehinde Paimo, Farid S Ataya, Gaber El-Saber Batiha, and Athanasios Alexiou. Addressing cancer care inequities in sub-saharan africa: current challenges and proposed solutions. *International journal for equity in health*, 22(1):189, 2023.

[10] Wilfred Ngwa, Beatrice W Addai, Isaac Adewole, Victoria Ainsworth, James Alaro, Olusegun I Alatise, Zipporah Ali, Benjamin O Anderson, Rose Anorlu, Stephen Avery, et al. Cancer in sub-saharan africa: a lancet oncology commission. *The Lancet Oncology*, 23(6):e251–e312, 2022.

[11] Andrew Wooyoung Kim, Madeleine Lambert, Shane A Norris, and Emily Mendenhall. Perceptions and experiences of prostate cancer patients in a public tertiary hospital in urban south africa. *Ethnicity & Health*, 28(5):696–711, 2023.

[12] Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74(3):229–263, 2024.

[13] Emeka Odiaka, David W Lounsbury, Mohamed Jalloh, Ben Adusei, Thierno Amadou Diallo, Papa Moussa Sene Kane, Isabella Rockson, Vicky Okyne, Hayley Irusen, Audrey Pentz, et al. Effective project management of a pan-african cancer research network: men of african descent and carcinoma of the prostate (madcap). *Journal of global oncology*, 4:1–12, 2018.

[14] T Kue Young, Janet J Kelly, Jeppe Friborg, Leena Soininen, and Kai O Wong. Cancer among circumpolar populations: an emerging public health concern. *International Journal of Circumpolar Health*, 75(1):29787, 2016.

[15] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.

[16] Nicholas Berente, Bin Gu, Jan Recker, and Radhika Santhanam. Managing artificial intelligence. *MIS quarterly*, 45(3), 2021.

[17] Manan Agarwal, Khushboo K Rao, Kaushar Vaidya, and Souradeep Bhattacharya. Ml-moc: machine learning (knn and gmm) based membership determination for open clusters. *Monthly Notices of the Royal Astronomical Society*, 502(2):2582–2599, 2021.

[18] Ethem Alpaydin. *Machine learning*. MIT press, 2021.

[19] Marcel R Ackermann, Johannes Blömer, Daniel Kuntze, and Christian Sohler. Analysis of agglomerative clustering. *Algorithmica*, 69:184–215, 2014.

[20] Jamie E Anderson and David C Chang. Using electronic health records for surgical quality improvement in the era of big data. *JAMA surgery*, 150(1):24–29, 2015.

[21] Hadley Wickham and Hadley Wickham. *Data analysis*. Springer, 2016.

[22] Roland Gamache, Hadi Kharrazi, and Jonathan P Weiner. Public and population health informatics: the bridging of big data to benefit communities. *Yearbook of medical informatics*, 27(01):199–206, 2018.

[23] OF EDA. Exploratory data analysis. *Handbook of Psychology, Research Methods in Psychology*, 2:34, 2012.

[24] Emily F Patridge and Tania P Bardyn. Research electronic data capture (redcap). *Journal of the Medical Library Association: JMLA*, 106(1):142, 2018.

[25] Mohamed Jalloh, Lamine Niang, Medina Ndoye, Issa Labou, and Serigne M Gueye. Prostate cancer in sub saharan africa. *Journal of Nephrology and Urology Research*, 1(1):15–20, 2013.

[26] LA Harvey. Redcap: web-based software for all types of data storage and collection. *Spinal cord*, 56(7):625–625, 2018.

[27] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[28] Tom M Mitchell and Tom M Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.

[29] Bhavneet Bhinder, Coryandar Gilvary, Neel S Madhukar, and Olivier Elemento. Artificial intelligence in cancer research and precision medicine. *Cancer discovery*, 11(4):900–915, 2021.

[30] Malcolm Dewar, L Kaestner, Q Zikhali, K Jehle, S Sinha, and J Lazarus. Investigating racial differences in clinical and pathological features of prostate cancer in south african men. *South African Journal of Surgery*, 56(2):54–58, 2018.

[31] Kee Yuan Ngiam and Wei Khor. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, 20(5):e262–e273, 2019.

[32] Saptarshi Bej, Jit Sarkar, Saikat Biswas, Pabitra Mitra, Partha Chakrabarti, and Olaf Wolkenhauer. Identification and epidemiological characterization of type-2 diabetes sub-population using an unsupervised machine learning approach. *Nutrition & Diabetes*, 12(1):27, 2022.

[33] Alena Lukasová. Hierarchical agglomerative clustering procedure. *Pattern Recognition*, 11(5-6):365–381, 1979.

[34] Andreas C Müller and Sarah Guido. *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.", 2016.

[35] Sedat Ozer, Deanna L Langer, Xin Liu, Masoom A Haider, Theodorus H Van der Kwast, Andrew J Evans, Yongyi Yang, Miles N Wernick, and Imam S Yetik. Supervised and unsupervised methods for prostate cancer segmentation with multispectral mri. *Medical physics*, 37(4):1873–1883, 2010.

[36] John A Richards and John A Richards. Clustering and unsupervised classification. *Remote Sensing Digital Image Analysis*, pages 369–401, 2022.

[37] Tobias Schoenherr and Cheri Speier-Pero. Data science, predictive analytics, and big data in supply chain management: Current state and future potential. *Journal of Business Logistics*, 36(1):120–132, 2015.

[38] Prashanth Rawla. Epidemiology of prostate cancer. *World journal of oncology*, 10(2):63, 2019.

[39] Yuta Takeshima, Yuta Yamada, Taro Teshima, Tetsuya Fujimura, Shigenori Kakutani, Yuji Hakozaki, Naoki Kimura, Yoshiyuki Akiyama, Yusuke Sato, Taketo Kawai, et al. Clinical

significance and risk factors of international society of urological pathology (isup) grade upgrading in prostate cancer patients undergoing robot-assisted radical prostatectomy. *BMC cancer*, 21(1):501, 2021.

[40] P Gontero, E Comperat, JD Escrig, F Liedberg, P Mariappan, A Masson-Lecomte, AH Mostafid, BWG van Rhijn, M Rouprêt, T Seisen, et al. Eau guidelines. In *Edn. presented at the EAU Annual Congress Milan*, 2023.

[41] Kenneth A Iczkowski, Geert JLH Van Leenders, and Theodorus H Van Der Kwast. The 2019 international society of urological pathology (isup) consensus conference on grading of prostatic carcinoma. *The American Journal of Surgical Pathology*, 45(7):1007, 2021.

[42] Hendrik Van Poppel, Monique J Roobol, Christopher R Chapple, James WF Catto, James N'Dow, Jens Sønksen, Arnulf Stenzl, and Manfred Wirth. Prostate-specific antigen testing as part of a risk-adapted early detection strategy for prostate cancer: European association of urology position and recommendations for 2021. *European Urology*, 80(6):703–711, 2021.

[43] Dragan Ilic, Molly M Neuberger, Mia Djulbegovic, and Philipp Dahm. Screening for prostate cancer. *Cochrane database of systematic reviews*, (1), 2013.

[44] J John, Ahmed Adam, L Kaestner, P Spies, S Mutambirwa, J Lazarus, South African Urology Association, and the Prostate Cancer Foundation of South Africa. The south african prostate cancer screening guidelines. *South African Medical Journal*, 114(5):15–18, 2024.

[45] Elizabeth A Tindall, L Richard Monare, Desiree C Petersen, Smit Van Zyl, Rae-Anne Hardie, Alpheus M Segone, Philip A Venter, MS Riana Bornman, and Vanessa M Hayes. Clinical presentation of prostate cancer in black south africans. *The Prostate*, 74(8):880–891, 2014.

[46] Yasutaka Yamada and Himisha Beltran. The treatment landscape of metastatic prostate cancer. *Cancer letters*, 519:20–29, 2021.

[47] Giuliano Di Pietro, Ganna Chornokur, Nagi B Kumar, Chemar Davis, and Jong Y Park. Racial differences in the diagnosis and treatment of prostate cancer. *International neurourology journal*, 20(Suppl 2):S112, 2016.

[48] Arthur L Burnett, Yaw A Nyame, and Edith Mitchell. Disparities in prostate cancer. *Journal of the National Medical Association*, 115(2):S38–S45, 2023.

[49] Evangelia Christodoulou, Jie Ma, Gary S Collins, Ewout W Steyerberg, Jan Y Verbakel, and Ben Van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*, 110:12–22, 2019.

[50] Kazzem Gheybi, Naledi Mmekwa, Maphuti Tebogo Lebelo, Sean M Patrick, Raymond Campbell, Mukudeni Nenzhelele, Pamela XY Soh, Muvhulawa Obida, Massimo Loda, Joyce Shirindi, et al. Linking african ancestral substructure to prostate cancer health disparities. *Scientific Reports*, 13(1):20909, 2023.

[51] Anthony V D'Amico, April Desjardin, Ming-Hui Chen, Steve Paik, Delray Schultz, Andrew A Renshaw, Kevin R Loughlin, and Jerome P Richie. Analyzing outcome-based staging for clin-

ically localized adenocarcinoma of the prostate. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 83(10):2172–2180, 1998.

[52] David J Hernandez, Matthew E Nielsen, Misop Han, and Alan W Partin. Contemporary evaluation of the d'amico risk classification of prostate cancer. *Urology*, 70(5):931–935, 2007.

[53] Francesco Chierigo, Rocco Simone Flammia, Gabriele Sorce, Benedikt Hoeh, Lukas Hohenhorst, Andrea Panunzio, Zhe Tian, Fred Saad, Markus Graefen, Michele Gallucci, et al. Racial/ethnic disparities in the distribution and effect of type and number of high-risk criteria on mortality in prostate cancer patients treated with radiotherapy. *Arab Journal of Urology*, 21(3):135–141, 2023.

[54] Caroline Dickens, Ruth M Pfeiffer, William F Anderson, Raquel Duarte, Patricia Kellett, Joachim Schüz, Danuta Kielkowski, and Valerie A McCormack. Investigation of breast cancer sub-populations in black and white women in south africa. *Breast cancer research and treatment*, 160:531–537, 2016.

[55] Caroline Andrews, Brian Fortier, Amy Hayward, Ruth Lederman, Lindsay Petersen, Jo McBride, Desiree C Petersen, Olabode Ajayi, Paidamoyo Kachambwa, Moleboheng Seutloali, et al. Development, evaluation, and implementation of a pan-african cancer research network: men of african descent and carcinoma of the prostate. *Journal of global oncology*, 4:1–14, 2018.

[56] Ilir Agalliu, Caroline Andrews, Akindele Olupelumi Adebiyi, Mohamed Jalloh, Maureen Joffe, Timothy R Rebbeck, and Ann Hsing. Pilot study of prostate cancer survival among african patients in the men of african descent and cancer of the prostate (madcap) consortium. *Cancer Research*, 80(16_Supplement):3518–3518, 2020.

[57] Jasmine Zakir, Tom Seymour, and Kristi Berg. Big data analytics. *Issues in Information Systems*, 16(2), 2015.

[58] Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. Big data in healthcare: management, analysis and future prospects. *Journal of big data*, 6(1):1–25, 2019.

[59] Hnin Wint Khaing. Data mining based fragmentation and prediction of medical data. In *2011 3rd International Conference on Computer Research and Development*, volume 2, pages 480–485. IEEE, 2011.

[60] Leonard G Gomella. Big data equals big challenges for prostate cancer. *The Canadian Journal of Urology*, 23(4):8329, 2016.

[61] Ganesh Kumar, Shuib Basri, Abdullahi Abubakar Imam, and Abdullateef Oluwagbemiga Balogun. Data harmonization for heterogeneous datasets in big data-a conceptual model. In *Software Engineering Perspectives in Intelligent Systems: Proceedings of 4th Computational Methods in Systems and Software 2020, Vol. 1 4*, pages 723–734. Springer, 2020.

[62] Gerard George, Martine R Haas, and Alex Pentland. Big data and management, 2014.

[63] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8):500–510, 2018.

[64] Anja Thieme, Danielle Belgrave, and Gavin Doherty. Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(5):1–53, 2020.

[65] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview, ii. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(6):e1219, 2017.

[66] Krishna Kumar Sharma and Ayan Seal. Multi-view spectral clustering for uncertain objects. *Information Sciences*, 547:723–745, 2021.

[67] Alexandra Prados-Torres, Amaia Calderón-Larrañaga, Jorge Hancco-Saavedra, Beatriz Poblador-Plou, and Marjan van den Akker. Multimorbidity patterns: a systematic review. *Journal of clinical epidemiology*, 67(3):254–266, 2014.

[68] Lala Septem Riza, Rendi Adistya Rosdiyana, Asep Wahyudin, and Alejandro Rosales Pérez. The k-means algorithm for generating sets of items in educational assessment. *Indonesian Journal of Science and Technology*, 6(1):93–100, 2021.

[69] Rose S George, Arkar Htoo, Michael Cheng, Timothy M Masterson, Kun Huang, Nabil Adra, Hristos Z Kaimakliotis, Mahmut Akgul, and Liang Cheng. Artificial intelligence in prostate cancer: Definitions, current research, and future directions. In *Urologic Oncology: Seminars and Original Investigations*, volume 40, pages 262–270. Elsevier, 2022.

[70] Karen Kirk, Tracy L McClair, Sina Pascal Dakouo, Timothy Abuya, and Pooja Sripad. Introduction of digital reporting platform to integrate community-level data into health information systems is feasible and acceptable among various community health stakeholders: a mixed-methods pilot study in mopti, mali. *Journal of Global Health*, 11, 2021.

[71] Richard Nshuti, Deirdré Kruger, and Thifheli E Luvhengo. Clinical presentation of acute appendicitis in adults at the chris hani baragwanath academic hospital. *International journal of emergency medicine*, 7(1):12, 2014.

[72] Andrew Black, Freddy Sitas, Trust Chibrawara, Zoe Gill, Mmamapudi Kubanje, and Brian Williams. Hiv-attributable causes of death in the medical ward at the chris hani baragwanath hospital, south africa. *PloS one*, 14(5):e0215591, 2019.

[73] Uzma Khan, Colin N Menezes, and Nimmisha Govind. Patterns and outcomes of admissions to the medical acute care unit of a tertiary teaching hospital in south africa. *African journal of emergency medicine*, 11(1):26–30, 2021.

[74] Andrew May, Scott Hazelhurst, Yali Li, Shane A Norris, Nimmisha Govind, Mohammed Tikly, Claudia Hon, Keith J Johnson, Nicole Hartmann, Frank Staedtler, et al. Genetic diversity in black south africans from soweto. *BMC genomics*, 14:1–12, 2013.

[75] José M Jerez, Ignacio Molina, Pedro J García-Laencina, Emilio Alba, Nuria Ribelles, Miguel Martín, and Leonardo Franco. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*, 50(2):105–115, 2010.

[76] David Ruppert. The elements of statistical learning: data mining, inference, and prediction, 2004.

[77] James D Miller. *Statistics for data science: Leverage the power of statistics for data analysis, classification, regression, machine learning, and neural networks*. Packt Publishing Ltd, 2017.

[78] David Cohen, Mikael Lindvall, and Patricia Costa. An introduction to agile methods. *Adv. Comput.*, 62(03):1–66, 2004.

[79] Lars Rönnbäck, Olle Regardt, Maria Bergholtz, Paul Johannesson, and Petia Wohed. Anchor modeling—agile information modeling in evolving data environments. *Data & Knowledge Engineering*, 69(12):1229–1253, 2010.

[80] Nikita Silaparasetty. *Machine learning concepts with python and the jupyter notebook environment: Using tensorflow 2.0*. Springer, 2020.

[81] Arman Ghavidel and Pilar Pazos. Machine learning (ml) techniques to predict breast cancer in imbalanced datasets: a systematic review. *Journal of Cancer Survivorship*, 19(1):270–294, 2025.

[82] Mark S Litwin and Hung-Jui Tan. The diagnosis and treatment of prostate cancer: a review. *Jama*, 317(24):2532–2542, 2017.

[83] 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*, 21:1–11, 2021.

[84] Kinsuk Giri and Tuhin Kr Biswas. Determining optimal epsilon (eps) on dbscan using empty circles. In *International Conference on Artificial Intelligence and Sustainable Engineering: Select Proceedings of AISE 2020, Volume 1*, pages 265–275. Springer, 2022.

[85] Simon Nusinovici, Yih Chung Tham, Marco Yu Chak Yan, Daniel Shu Wei Ting, Jialiang Li, Charumathi Sabanayagam, Tien Yin Wong, and Ching-Yu Cheng. Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of clinical epidemiology*, 122:56–69, 2020.

[86] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[87] David M LaHuis, Michael J Hartman, Shotaro Hakoyama, and Patrick C Clark. Explained variance measures for multilevel models. *Organizational Research Methods*, 17(4):433–451, 2014.

[88] L Nitya Sai, M Sai Shreya, A Anjan Subudhi, B Jaya Lakshmi, and KB Madhuri. Optimal k-means clustering method using silhouette coefficient. 2017.

[89] Subrata Bhattacharjee, Yeong-Byn Hwang, Kouayep Sonia Carole, Hee-Cheol Kim, Damin Moon, Nam-Hoon Cho, and Heung-Kook Choi. Unsupervised, self-supervised, and supervised learning for histopathological pattern analysis in prostate cancer biopsy. In *Proceedings of the Future Technologies Conference*, pages 1–17. Springer, 2023.

[90] Subbulakshmi Pasupathi, Vimal Shanmuganathan, Kaliappan Madasamy, Harold Robinson Yesudhas, and Mucheol Kim. Trend analysis using agglomerative hierarchical clustering approach for time series big data. *The Journal of Supercomputing*, 77(7):6505–6524, 2021.

[91] John Clifford Gower. Properties of euclidean and non-euclidean distance matrices. *Linear algebra and its applications*, 67:81–97, 1985.

[92] Hui Chen, Man Liang, Wanquan Liu, Weina Wang, and Peter Xiaoping Liu. An approach to boundary detection for 3d point clouds based on dbscan clustering. *Pattern Recognition*, 124:108431, 2022.

[93] Eric W Weisstein. Laplacian matrix. *https://mathworld. wolfram. com/*, 1999.

[94] John T Wei, Daniel Barocas, Sigrid Carlsson, Fergus Coakley, Scott Eggener, Ruth Etzioni, Samson W Fine, Misop Han, Sennett K Kim, Erin Kirkby, et al. Early detection of prostate cancer: Aua/suo guideline part ii: considerations for a prostate biopsy. *The Journal of urology*, 210(1):54–63, 2023.

[95] Rajesh Sharma. The burden of prostate cancer is associated with human development index: evidence from 87 countries, 1990–2016. *EPMA Journal*, 10:137–152, 2019.

[96] Sergios Gatidis, Markus Scharpf, Petros Martirosian, Ilja Bezrukov, Thomas Küstner, Jörg Hennenlotter, Stephan Kruck, Sascha Kaufmann, Christina Schraml, Christian la Fougère, et al. Combined unsupervised–supervised classification of multiparametric pet/mri data: application to prostate cancer. *NMR in Biomedicine*, 28(7):914–922, 2015.

[97] Muktevi Srivenkatesh. Prediction of prostate cancer using machine learning algorithms. *Int. J. Recent Technol. Eng*, 8(5):5353–5362, 2020.

[98] Ankush D Jamthikar, Deep Gupta, Amer M Johri, Laura E Mantella, Luca Saba, Raghu Kolluri, Aditya M Sharma, Vijay Viswanathan, Andrew Nicolaides, and Jasjit S Suri. Low-cost office-based cardiovascular risk stratification using machine learning and focused carotid ultrasound in an asian-indian cohort. *Journal of medical systems*, 44:1–15, 2020.

[99] Amit Kumar, Harshika Bansal, Ayush Jaiswal, and Sovit Kumar Gupta. Early disease prediction using ml. *International Journal of Advanced Engineering and Nano Technology*, 11(10):123–134, 2023.

[100] Rachid Sammouda and Ali El-Zaart. An optimized approach for prostate image segmentation using k-means clustering algorithm with elbow method. *Computational Intelligence and Neuroscience*, 2021(1):4553832, 2021.

[101] H Gilbert Welch, David H Gorski, and Peter C Albertsen. Trends in metastatic breast and prostate cancer—lessons in cancer dynamics. *New England journal of medicine*, 373(18):1685–1687, 2015.

[102] Maphuti Tebogo Lebelo, Naledi Mmekwa, Weerachai Jaratlerdsiri, Shingai BA Mutambirwa, Massimo Loda, Vanessa M Hayes, and MS Riana Bornman. Associating serum testosterone levels with african ancestral prostate cancer health disparities. 2024.

[103] Amelia M Stanton, Ryan L Boyd, Conall O'Cleirigh, Stephen Olivier, Brett Dolotina, Resign Gunda, Olivier Koole, Dickman Gareta, Tshwaraganang H Modise, Zahra Reynolds, et al. Hiv, multimorbidity, and health-related quality of life in rural kwazulu-natal, south africa: A population-based study. *Plos one*, 19(2):e0293963, 2024.

[104] Witness Mapanga, Shane A Norris, Ashleigh Craig, Yoanna Pumpalova, Oluwatosin A Ayeni, Wenlong Carl Chen, Judith S Jacobson, Alfred I Neugut, Mazvita Muchengeti, Audrey Pentz, et al. Prevalence of multimorbidity in men of african descent with and without prostate cancer in soweto, south africa. *Plos one*, 17(10):e0276050, 2022.

[105] Kathryn L Hopkins, Khuthadzo E Hlongwane, Kennedy Otwombe, Janan Dietrich, Mireille Cheyip, Jacobus Olivier, Heidi van Rooyen, Tanya Doherty, and Glenda E Gray. The substantial burden of non-communicable diseases and hiv-comorbidity amongst adults: screening results from an integrated hiv testing services clinic for adults in soweto, south africa. *EClinicalMedicine*, 38, 2021.

[106] Myrna Van Pinxteren, Nonzuzo Mbokazi, Katherine Murphy, Frances S Mair, Carl May, and Naomi S Levitt. Using qualitative study designs to understand treatment burden and capacity for self-care among patients with hiv/ncd multimorbidity in south africa: A methods paper. *Journal of multimorbidity and comorbidity*, 13:26335565231168041, 2023.

[107] Marco Randazzo, Alexander Müller, Sigrid Carlsson, Daniel Eberli, Andreas Huber, Rainer Grobholz, Lukas Manka, Ashkan Mortezavi, Tullio Sulser, Franz Recker, et al. A positive family history as risk factor for prostate cancer in a population-based study with organized psa-screening: Results of the swiss erspc (aarau). *BJU international*, 117(4):576, 2015.

[108] Andrew W Roddam, Michael J Duffy, Freddie C Hamdy, Anthony Milford Ward, Julietta Patnick, Christopher P Price, Janet Rimmer, Cathie Sturgeon, Peter White, Naomi E Allen, et al. Use of prostate-specific antigen (psa) isoforms for the detection of prostate cancer in men with a psa level of 2–10 ng/ml: systematic review and meta-analysis. *European urology*, 48(3):386–399, 2005.

[109] Nathaniel Mofolo, Olwethu Betshu, Ogomoditse Kenna, Sarah Koroma, Tlalane Lebeko, Frederik M Claassen, and Gina Joubert. Knowledge of prostate cancer among males attending a urology clinic, a south african study. *Springerplus*, 4(1):67, 2015.

Appendix A

Human Research Ethics Certificates

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG	HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

Office of the Deputy Vice-Chancellor (Research and Innovation)

TO: Mr T Rwafa
School of Public Health
Division of Epidemiology and Biostatistics
Medical School
University

E-mail: 2606612@students.wits.ac.za

CC: Supervisor: Mr M Mapundu and Dr WC Chen
Michael.Mapundu@wits.ac.za
and <HREC-Medical Research Office@wits.ac.za>

FROM: Mr Iain Burns
Human Research Ethics Committee (Medical)
Tel: 011 717 1252

E-mail: Iain.Burns@wits.ac.za

DATE: 5 February 2024

REF: R14/49

PROTOCOL NO: **M231114** (This is your ethics application reference number. Please quote it in all enquiries, oral or written, relating to this study.)

PROJECT TITLE: *Identifying high-risk prostate cancer risk factors and morbidity using unsupervised machine learning algorithms*

Please find attached the Clearance Certificate for the above project. I hope it goes well and that an article in a recognized publication comes out of it. This will reflect well on your professional standing and contribute to Government funding of the University.

MSWorks2000/Iain0007/Clearscan.wps

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG	HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

Office of the Deputy Vice-Chancellor (Research and Innovation)

TO: Drs M Joffe and M Muchengeti
Wits Health Consortium
Non-communicable Diseases Research Division

E-mail: mjoffe@witshealth.co.za

CC: Supervisor: Not applicable
<>
and <HREC-Medical Research Office@wits.ac.za>

FROM: Mr Iain Burns
Human Research Ethics Committee (Medical)
Tel: 011 717 1252

E-mail: Iain.Burns@wits.ac.za

DATE: 2022/07/01

REF: R14/49

PROTOCOL NO: **M220673** (This is your ethics application reference number. Please quote it in all enquiries, oral or written, relating to this study.)

PROJECT TITLE: *Men of African descent with cancer of the Prostate*

Please find attached the Clearance Certificate for the above project. I hope it goes well and that an article in a recognized publication comes out of it. This will reflect well on your professional standing and contribute to Government funding of the University.

MSWorks2000/Iain0007/Clearscan.wps

Appendix B

Plagiarism Declaration

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I _____TANAKA, RWAFA__ (Student number: _2606612____) am a student registered for the degree of MSc EPIDEMIOLOGY IN PUBLIC HEALTH INFORMATICS in the academic year _2025_.

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: _____11/05/2025____

Appendix C

TurnItIn Report

Tanaka_Rwafa_2606612 (17)-1.pdf

ORIGINALITY REPORT

Michael Mapundu:

11%	3%	3%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Witwatersrand Student Paper	7%
2	David J. Hernandez, Matthew E. Nielsen, Misop Han, Alan W. Partin. "Contemporary Evaluation of the D'Amico Risk Classification of Prostate Cancer", Urology, 2007 Publication	<1%
3	www.ncbi.nlm.nih.gov Internet Source	<1%
4	Witness Mapanga, Shane A. Norris, Ashleigh Craig, Yoanna Pumpalova et al. "Prevalence of multimorbidity in men of African descent with and without prostate cancer in Soweto, South Africa", PLOS ONE, 2022 Publication	<1%
5	Francesco Chierigo, Rocco Simone Flammia, Gabriele Sorce, Benedikt Hoeh et al. "Racial/ethnic disparities in the distribution and effect of type and number of high-risk criteria on mortality in prostate cancer patients treated with radiotherapy", Arab Journal of Urology, 2022 Publication	<1%
6	Submitted to Glyndwr University Student Paper	<1%
7	www.mdpi.com Internet Source	<1%
8	www.researchgate.net Internet Source	<1%
	sajr.org.za	

Appendix D

Training Certificate

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

TANAKA RWAFA

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: 2022/01/30
CID : SHprJlyVe3

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 Programms
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]

Appendix E

Analysis Codes

In [3]:
```
import pandas as pd
import numpy as np
import warnings
import pandas as pd
import os
import plotly.express as px
import seaborn as sns
import numpy as np
import tqdm
import datetime
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('ggplot')
from sklearn.decomposition import PCA
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.ensemble import RandomForestClassifier
from sklearn.cluster import KMeans
from IPython.display import clear_output
from sklearn.preprocessing import StandardScaler, normalize
from sklearn.metrics import silhouette_score
from sklearn.preprocessing import LabelEncoder
# 3D Visualization
import plotly as py
import plotly.graph_objs as go
from numpy import unique
from numpy import where
from sklearn.model_selection import train_test_split
```

In [4]:
```
pd.set_option('display.float_format', lambda x: '%.2f' % x)
pd.set_option('display.max_colwidth', None)
pd.set_option('display.max_rows', None)

#import plotly.io as pio
#pio.templates.default = 'simple_white'

warnings.filterwarnings('ignore')
```

In [7]:
```
df = pd.read_stata('C:/Users/rwafa/OneDrive/Documentos/Aka/CommentsDoc/Research Dat
```

Data Overview

In [9]:
```
df.shape
```

Out[9]: (1096, 47)

In [12]:
```
df.head(5)
```

Appendix F

Data Dictionary

	548	[mr_radiaxbeamdose] Show the field ONLY if: [mr_radiationtreat] = '1'	X-Beam treatment / radiation dose	notes
	549	[mr_radiaprotondate] Show the field ONLY if: [mr_radiationtreat] = '2'	Proton therapy date started	text (date_ymd)
	550	[mr_radiaprotondesc] Show the field ONLY if: [mr_radiationtreat] = '2'	Proton therapy description	notes
	551	[mr_radiabrachythedate] Show the field ONLY if: [mr_radiationtreat] = '3'	Low dose brachytherapy(seeds) date started	text (date_ymd)
	552	[mr_radiabrachythedose] Show the field ONLY if: [mr_radiationtreat] = '3'	Low dose brachytherapy(seeds) dose	notes
	553	[mr_radiabrachytheseed] Show the field ONLY if: [mr_radiationtreat] = '3'	Low dose brachytherapy(seeds) number of seeds	text (integer)
	554	[mr_radiaotherdate] Show the field ONLY if: [mr_radiationtreat] = '4'	Other treatment date started	text (date_ymd)
	555	[mr_radiaotherdesc] Show the field ONLY if: [mr_radiationtreat] = '4'	Other treatment description	notes
	556	[mr_finalstagingt]	Section Header: *Final staging* Final staging - T	text
	557	[mr_finalstagingn]	Final staging - N	text
	558	[mr_finalstagingm]	Final staging - M	text
	559	[medical_record_complete]	Section Header: *Form Status* Complete?	dropdown 0 Incomplete 1 Unverified 2 Complete
Instrument: PSA summary (psa_summary)				
	560	[psa_psareferraldate]	Section Header: *PSA* PSA referral	text (date_ymd)
	561	[psa_psareferrallevel]	Referral PSA level (ng/ml)	text (number_2dp)
	562	[psa_psareferralleveldesc]	Specify referral PSA level	dropdown 1 Normal 2 Abnormal/suspicious 3 Unknown
	563	[psa_psadiagndate]	Diagnostic PSA	text (date_ymd)
	564	[psa_psadiagnlevel]	Specify diagnostic PSA level (ng/ml)	text (number_2dp)
	565	[psa_diagnoproscar]	Was patient on Proscar when PSA was performed?	radio 1 Yes 2 No 3 Unknown Custom alignment: RH
	566	[psa_followupdate1]	Section Header: *PSA follow up* PSA follow up 1	text (date_ymd)
	567	[psa_followupvalue1]	Follow up PSA level (ng/ml) 1	text (number_2dp)
	568	[psa_followupdate2]	PSA follow up 2	text (date_ymd)
	569	[psa_followupvalue2]	Follow up PSA level (ng/ml) 2	text (number_2dp)
	570	[psa_followupdate3]	PSA follow up 3	text (date_ymd)
	571	[psa_followupvalue3]	Follow up PSA level (ng/ml) 3	text (number_2dp)
	572	[psa_followupdate4]	PSA follow up 4	text (date_ymd)
	573	[psa_followupvalue4]	Follow up PSA level (ng/ml) 4	text (number_2dp)