QATAR UNIVERSITY

COLLEGE OF HEALTH SCIENCES

# MACHINE LEARNING PREDICTION OF CANCER FROM THE PUBLICLY AVAILABLE DATASET

BY

Halah Noor Nasir

A Capstone Project Submitted to

the College of Health Sciences

in Partial Fulfillment of the Requirements for the Degree of

Masters of Science  in Biomedical Sciences

January   2024

# COMMITTEE PAGE

The members of the Committee approve the Capstone Project of Halah Noor Nasir defended on 06/12/2023.

Dr Rozaimi Razali
Thesis/Dissertation Supervisor

Dr Maha Al Asmakh
Committee Member

Dr Abdulaziz Khalid A M Al-Ali
Committee Member

# ABSTRACT

NOOR NASIR, HALAH, Masters of Science:

January: [2024], Biomedical Sciences

Title: Machine Learning Prediction of Cancer from the Publicly Available Dataset

Supervisor of Capstone Project: Dr Rozaimi Razali

Prostate cancer in the second most common cause of cancer in men around the world and in Qatar with a high incidence rate worldwide. This has resulted in an increased mortality rate, making prostate cancer a healthcare burden. Early detection of prostate cancer is crucial in reducing mortality; however, the current detection procedures are invasive with prostate cancer screening test not being easily accessible. This has led to the development of machine learning approaches in detection of cancer with aims to improve healthcare accuracy and patient outcomes. This study examines the efficacy of machine learning model in prediction of prostate cancer using publicly available healthcare dataset. It aims to determine the best classifier algorithm and to develop a standard operating procedure (SOP) that can be used in a machine learning model for prostate cancer prediction. Lastly, this study examines the main feature class based on machine learning model that can increase the risk of developing prostate cancer.

# DEDICATION

*"This capstone project is dedicated to my loving parents for their unconditional support throughout my educational and professional endeavors.”*

# ACKNOWLEDGMENTS

“I would like to acknowledge the support of Qatar University for providing all the needs to achieve the requirements of this study. I would also like to extend my sincere gratitude towards the supervisor of the capstone project Dr Rozaimi Razali for his immense guidance and commitment.”

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1: INTRODUCTION

Prostate cancer is the second most common type of cancer in men around the world resulting in significant morbidity and mortality rate. It accounts for an annual 1.2 million newly diagnosed cases and 350,000 deaths.(Sung et al., 2021) On a global scale, the incident cases, deaths, and disability-adjusted life years of prostate cancer have increased by just over 100% over the last three decades making prostate cancer a global health burden(Buskin et al., 2021). In Qatar alone, prostate cancer is the second most common type of cancer in men with 7% percent of new cases in 2020. Early detection of prostate cancer with effective treatment can reduce the mortality rate. (Buskin et al., 2021; Sung et al., 2021). Currently, prostate cancer metastasis detection relies on procedures such as biopsies of affected distant organs, radiological examinations, imaging, and evaluation of serum tumor markers. Prostate screening such as prostate specific antigen (PSA) test is also implemented by many countries; however, the interpretation can be challenging which leads to false positives. These methods are frequently incompetent in detection of metastasis at an early stage. Many of these diagnostic procedures are also invasive. These disadvantages have led to the development of various machine learning (ML) approaches. (Bray et al., 2018; W. Zhang et al., 2023)

Machine learning is a branch of artificial intelligence that uses algorithms to analyze data and it can be classified based on the type of label and feature. Artificial Intelligence Prediction Models (AIPM) can be an essential tool due to drawbacks of current procedures. (Bini, 2018; S. L. Goldenberg et al., 2019). Men with likelihood of prostate cancer can be detected at any early stage using non-invasive prediction models without the need of undergoing invasive procedures. It will also result in facilitating clinical management of patients as well as in classification of cancer patients into high and low risk. It is essential for the clinician to detect cancer in time for more favorable outcomes such as survival, diagnosis at an early stage, and an improvement in the quality of life. If the clinician could not predict the cancer on time it can lead to spreading of the cancer to other organs resulting in a higher chance of treatment not being effective which eventually leads to lower, chances of survival. (Bini, 2018).

Many of the ML techniques have been applied on these prediction models giving accurate and effective results. However, an assessment framework of the machine learning prediction experiments is currently lacking. (S. L. Goldenberg et al., 2019; Welch & Albertsen, 2009). This has resulted in many conflicting results and experiments that are not replicable. It is necessary to have a standard framework to safely integrate the usage of artificial intelligence models in healthcare through careful assessment of the entire AIPM workflow such as data preparation, model training, model validation and implementation into the healthcare practice. (S. L. Goldenberg et al., 2019; Toivanen & Shen, 2017). AI plays a pivotal role in the development of algorithms that can help urologists to diagnose prostate cancer antigen without requesting unnecessary prostate biopsies.

## 1.1 Study Hypothesis, Aim and Objectives.

The primary aim of this research endeavor is to assess the efficacy of machine learning, trained on comprehensive health-related surveys, for the prediction of prostate cancer occurrence. To attain this aim, the following research objectives are delineated:

1. Identification of the optimal machine learning classifier algorithm for application in prostate cancer diagnosis.
2. Identification of features that contribute most to the machine learning model.
3. Formulation of a specialized protocol, akin to best practice guidelines or standard operating procedures (SOP).

# CHAPTER 2: LITERATURE REVIEW

## 2.1 Overview of Prostate Cancer

Prostate cancer, a prevalent malignancy among men, originates in the prostate gland—a vital component of the male reproductive system. This gland, roughly the size of a walnut, resides just below the bladder and surrounds the urethra, the tube responsible for carrying urine and semen out of the body. The prostate's primary function is to produce seminal fluid, a clear and slightly alkaline liquid that constitutes a significant portion of semen. This fluid serves as a nourishing medium for sperm and aids in their transport during ejaculation. Without the prostate's contributions, the journey of sperm would be far more challenging, as the acidic environment of the female reproductive tract could impair their mobility. (Toivanen & Shen, 2017). Consequently, the prostate gland's role is indispensable for successful reproduction. Furthermore, most prostate cancers are low-grade, have low risk and result in limited aggressiveness. Prostate cancer does not cause initial symptoms, however, some of the late symptoms include anemia, pain in the bone, spinal metastases resulting in paralysis as well as urethral obstruction resulting in renal failure. (Roberts et al., 2000).

The primary method of prostate cancer diagnosis is prostate specific antigen (PSA) testing as well as tissue biopsies. Recent diagnostic methods include PCA3 urine testing, exosome testing, genomic analysis we well as free and total PSA levels. (Sivaraman & Bhat, 2017). As with many cancers, the exact causes of prostate cancer remain a subject of ongoing research. It is believed to result from a complex interplay of genetic, environmental, and hormonal factors. Some studies have suggested a potential link between dietary habits and prostate cancer risk, with diets high in red meat and low in fruits and vegetables potentially increasing the likelihood of developing the disease. Family history, old age, obesity, hypertension, and ethnicity are also some of the risk factors.  However, the precise mechanisms through which these factors influence prostate cancer development are still being explored.

(Kaiser et al., 2019; Mullins & Loeb, 2012; Rhoden & Averbeck, 2009).

In its early stages, prostate cancer is often asymptomatic, meaning it does not produce noticeable symptoms. This feature highlights the importance of routine screening for early detection, especially in men at higher risk. As the cancer progresses, it can manifest symptoms such as urinary difficulties (including increased frequency, urgency, or a weak urinary stream), blood in the urine or semen, discomfort, or pain in the pelvic region, and, in advanced cases, bone pain. (Sivaraman & Bhat, 2017). However, it's important to note that these symptoms can also be caused by various non-cancerous conditions, making a thorough medical evaluation crucial to determine the cause and initiate appropriate treatment. If the cancer is localized, it is potentially curable, however, if the disease spread out the prostate gland into bones or other organs, pain medication, immunotherapy, chemotherapy, radiation, hormonal treatment, and other targeted therapy must be used depending upon the stage of cancer. (Loriot et al., 2012).

## 2.2 Screening and Diagnosis of Prostate Cancer

### *2.2.1: PSA Test*

The current screening and detection techniques of prostate cancer have a two-step approach with firstly undergoing a PSA blood test, and, secondly by confirming the suspected diagnosis by biopsy. Some of the main methods for screening prostate cancer include serum PSA levels especially at mid-life from the age of 50-70 years. A reduction in mortality rate of prostate cancer can result due to PSA screening test. The mortality rate of prostate cancer serves as quality control in population-based screening. However, there is a consensus that despite the existing evidence of early screening reducing prostate cancer mortality, population-based screening should not be implemented due to lack of evidence. (Attard et al., 2016).

It is important to note that PSA test is a continuous variable and does not have a cut-off. Therefore, very low levels don’t eliminate the risk of prostate cancer and very high levels increase the likelihood of detecting prostate cancer. Increase in serum PSA levels can be attributed to several conditions with the most common being benign prostatic hyperplasia. In addition, there are many factors that affect PSA measurements including prostatic infection and aging. (Schröder et al., 2009; Vickers et al., 2010)

### *2.2.2: Digital Rectal Examination:*

DRE is another screening test for prostate cancer with up to 50% of palpable prostate masses can be attributed to prostate cancer. Even though DRE is not a reliable method for diagnosis prostate cancer especially if it is performed by non-urologists, it is still considered to have an essential diagnostic role. This is because DRE can independently detect patients at risk of prostate cancer and the detection should be considered clinically significant irrespective of serum PSA levels. (Ellis et al., 1994).

### *2.2.3: PCA3 Test:*

PCA3 or Prostate cancer antigen is a gene that is only expressed in human tissue and high expression of the gene is present in prostate cancer. The expression of this gene is also used as a tumor marker with urinary PCA3 being evaluated as a diagnostic tool in detection of prostate cancer. To perform the test, PCA3 RNA is measured using PCR from a urine specimen that is collected immediately after a rectal examination. The role of urine PCA3 test is significant in identifying serious malignancies in comparison with total PSA levels, however, its role in prostate cancer screening still needs to be established. Moreover, the cut-off value of urinary PCS3 is controversial with some studies establishing urinary PSA3 cut-off value at 35 as this cut-off avoided more biopsies but also missed diagnosis of 28% of prostate cancers. Nevertheless, the cut-off value of 35 is used in clinical practice till date. (Roobol et al., 2010).

### *2.2.4: Magnetic Resonance Imaging:*

Another method for diagnosis of prostate cancer is MRI or magnetic resonance imaging technique which is used as an alternative to transrectal ultrasonography guided biopsy. According to a randomized, multicentered and noninferiority trial, men who had clinical suspicion of prostate cancer and didn’t undergo biopsy in the past underwent MRI with or without transrectal ultrasonography-guided biopsy. Similarly, if the MRI was suggestive of prostate cancer, the participants underwent a targeted biopsy, whereas, for participants whose MRI was not suggestive of prostate cancer did not underwent a biopsy. The results of the study proved that performing MRI before biopsy and MRI- targeted biopsy was a superior method to the standard transrectal ultrasonography. With 38% of the participants had clinically significant cancer detected in the MRI-targeted biopsy group when compared with standard biopsy at 26%. (Kasivisvanathan et al., 2018).

## 2.3 Prevalence and Incidence of Prostate Cancer in Qatar

Prostate cancer ranks as one of the most frequently diagnosed cancers in men worldwide. While it can affect men of all ages, its incidence notably increases with advancing age, making it more prevalent among older populations. Typically, prostate cancer is diagnosed in men aged 50 and older, though it can occur in younger individuals as well. Various factors contribute to an individual's risk of developing prostate cancer, including genetics, family history, race, and lifestyle choices. (Bray et al., 2018). For instance, African American men have a notably higher risk of developing prostate cancer than men of other racial backgrounds. Additionally, individuals with a family history of the disease are at an increased risk. The incidence rate of prostate cancer varies across different regions and populations. For instance, Africa and Asia have low incidence rates compared to developed countries. (Ferlay et al., 2021; Panigrahi et al., 2019).

In Qatar, in 2020, the incidence of prostate cancer amongst the male population is ranked the second behind colorectal cancer with an incidence rate of 7% .(Bray et al., 2018). Furthermore, prostate cancer is also the second most frequent cancer in Qatar and amongst the top five most frequent cancers in the MENA region .(W. Zhang et al., 2023). Several factors can be contributed to the incidence of the disease such as dietary factors, androgenic factors as well as socioeconomic factors. Screening for prostate cancer in the Arab countries is not very common, however, many studies have demonstrated that the benefits of early detection outweigh the risks and have also resulted in reduced mortality in the western countries. Due to lack of knowledge and hesitancy towards prostate cancer screening, it is still a challenge in the MENA region which could be the main cause of less disease detection. (Daher et al., 2021).

## 2.4 Prostate Cancer Etiologies

The etiologies of prostate cancer are multifactorial with many risk factors with some that are modifiable and others that are not. Some risk factors include but are not limited to age, family history and ancestry, however, there are other risk factors involved as well according to several epidemiological studies conducted. (Culp et al., 2020; Page et al., 2019). These risk factors include environmental factors such as lifestyle and diet that can greatly influence the risk of developing prostate cancer as well as its progression. Moreover, there is a noticeably increase in the use of prostate specific antigen screening as well as better diagnostic techniques which can lead to a high incidence rate of prostate cancer. By understanding the etiology, history, and pathophysiology of prostate cancer, it can be better managed as well as aid in its diagnosis. (Bratt et al., 2016) (Brookman-May et al., 2019; Campi et al., 2019; Krstev & Knutsson, 2019).

### *2.4.1 Age*

One of the well-established risk factors includes age as there is a high incidence rate of prostate cancer with an increase in age with a low risk below the age of 40. The correlation between age and prostate cancer development is consistent in both the developed and the developing world and screening of prostate specific antigen has led to the detection of prostate cancer almost a decade before symptoms can be seen. (Bray et al., 2018). The probability of developing prostate cancer increases in men as they age with the highest probability seen in men above the age of 60 being 13.7%. Furthermore, histological diagnosis and malignancy is mainly seen in men between the age of 70 and 80 years, however, majority of the histologically diagnosed cases follows an indolent course without any risk of mortality. (Scardino, 1989).

### *2.4.2 Family History and Genetics*

Men who have a family history such as a brother or father diagnosed with prostate cancer have a two to four-fold higher risk of developing prostate cancer with a high risk that can be attributed to genetic factors. (Gallagher & Fleshner, 1998). For instance, according to a study done by Nordic Twin Study of Cancer estimated that genetic variation relating to inheritance of prostate cancer amongst twins was 57% which makes prostate cancer inheritable. There has also been a correlation seen in risk of prostate cancer in families that also have a family history breast cancer according to a large prospective study done in USA that identified that there is a 21% greater risk of developing prostate cancer in families with familial prostate cancer with an overall elevated risk of both cancers .(Bruner et al., 2003; Hemminki, 2012). The reason for this increased risk of both cancer could be due to the BRCA gene mutation as there is a link between BRCA and breast cancer as well as a high risk of developing prostate cancer in men that are carriers of the BRCA gene. The inheritance of this gene can provide a biological pathway for familial inheritance and increased risk of developing breast and prostate cancer. This has confirmed genetic predisposition and led to GWAS or genome wide association studies in prostate cancer. More than 180 SNPs are associated with prostate cancer risk. The largest prostate cancer GWAS and meta-analysis reported identified 63 new susceptibility loci related to prostate cancer making the total number of loci to 167. This can lead to utilization of genetic identification kits to form screening programs for individuals at a significantly higher risk. (Nyberg et al., 2020; Nyberg et al., 2019).

### *2.4.3 Ethnicity*

Ethnic and geographic variations are also related in the incidence of prostate cancer. For instance, the incidence and mortality rate amongst men of black African decent is much higher compared to other ethnic groups with 2.4x higher mortality rate when compared with white men in the USA .(Wu & Modlin, 2012). The reason could be due to the increased prevalence of genetic risk loci related to prostate cancer in ethnic groups. For example, African American men have chromosome 8q24 variants that are associated with increased prostate cancer risk, furthermore, they have high variations in genes that suppress tumors or regulate cell apoptosis. Other factors that contribute to high incidence rate in African American men are due to low-quality healthcare and less access to PSA screening tests. (Hatcher et al., 2009; Robbins et al., 2011).

### *2.4.4 Smoking and Alcohol*

Smoking has been proven to have a high association with incidence and mortality related to prostate cancer. In a meta-analysis done by Huncharek et al of 24 cohort studies there was increase in risk of prostate cancer with an increase in the amount smoked with ex-smokers being at higher risk as well as heavy smokers had an association with prostate cancer related mortalities. Furthermore, a meta-analysis of 340 studies found a relationship between alcohol dosage and prostate cancer risk which increases with an increase in alcohol volume consumption as compared to non-drinkers. (Huncharek et al., 2010).

### *2.4.5 Obesity*

A high body mass index and obesity has an association with prostate with adiposity leading to increased mortality risk. According to a study, as the body mass index increases by 5kg/m2, the risk of prostate. Cancer mortality also increases by 20%. The three main reasons that could lead to an increased risk include IGF-1, sex hormones as well as adipokines which are chemokines secreted by adipocytes into plasma. (Cao & Ma, 2011). Adiponectin is an adipokine that has been linked to prostate cancer development and progression. As an individual becomes more obese, the concentration of plasma adiponectin reduces in men which leads to higher risk of developing prostate cancer. (Liao et al., 2015).

### *2.4.6 Physical Activity, Diet and Nutrition*

Many studies have proven that the risk of physical activity and development of prostate cancer is inversely proportional. A study done on 2705 men that had prostate cancer had a reduction of 61% in risk of prostate cancer-specific mortality due to minimum three hours of exercise per week when compared to men who only had one hour of exercise.(Kenfield et al., 2011). Furthermore, dietary habits are also significant when it comes to risk associated with many cancers including prostate cancer and according to a study done, diet with highly processed food leads to an increase in the risk of developing prostate cancer with unprocessed food leading to a reduced risk. (Trudeau et al., 2020). When it comes to specific type of foods that are associated with prostate cancer risk, foods containing soy are linked with lower incidence rate of prostate cancer with numerous studies done on soy. (Applegate et al., 2018). Furthermore, dairy products and in particular those with calcium have a positive association with prostate cancer as increased calcium suppresses calcitriol levels. Calcitriol is an active form of vitamin D which affects cell cycle by inducing apoptosis and inhibiting the growth of normal epithelial cells in the prostate gland. (Feldman et al., 2000). According to World Cancer Research Fund on Diet and Cancer, calcium intake can be considered as a “probable” risk factor for developing prostate cancer. (Wiseman, 2008).

## 2.5 Clinical Diagnosis of Prostate Cancer

There are several ways prostate cancer is diagnosed. The first is looking at the medical history and physical examination. This involves a thorough medical history review and a physical examination. The healthcare provider will inquire about any urinary symptoms, family history of prostate cancer, and other relevant medical conditions. Another method is using the Digital Rectal Examination (DRE). DRE is a medical procedure in which a healthcare provider inserts a lubricated, gloved finger into the rectum to examine the rectal and prostate area. This examination is often performed to check for abnormalities, such as prostate cancer, hemorrhoids, or other rectal conditions. It can help assess the size, shape, and texture of the prostate gland and detect any unusual growths or abnormalities in the rectal area. (Castillejos-Molina & Gabilondo-Navarro, 2016). Other than DRE, performing the Prostate-Specific Antigen (PSA) Testing is another common method to diagnose prostate cancer. The PSA test measures the level of a protein called prostate-specific antigen in the blood. Elevated PSA levels can be a sign of prostate cancer, although other conditions, such as benign prostatic hyperplasia (BPH) or inflammation, can also cause increased PSA levels. If the PSA levels are elevated or if there are abnormalities detected during the DRE, a prostate biopsy may be recommended. (Stamey et al., 1987). During a biopsy, small tissue samples are taken from different areas of the prostate gland using a thin needle. These samples are then examined under a microscope to determine if cancer is present. After a biopsy, the pathologist assigns a Gleason score to the tissue samples. The Gleason score grades the aggressiveness of the cancer based on the appearance of cancer cells under the microscope. Scores range from 2 to 10, with higher scores indicating more aggressive cancer .(Chen & Zhou, 2016).

Increasingly, advanced imaging tests like transrectal ultrasound (TRUS), magnetic resonance imaging (MRI), or computed tomography (CT) scans are now being used to help visualize the prostate and surrounding tissues. These imaging techniques can help determine the extent and stage of the cancer. Other new methods such as staging and genetic testing are also being utilized more nowadays. Staging is the process of determining the extent of cancer and whether it has spread to other parts of the body. The most used staging system for prostate cancer is the TNM system, which considers the tumor size, lymph node involvement, and distant metastasis. With regards to genetic testing, it may be recommended to identify specific genetic mutations or alterations that can affect treatment decisions, particularly in cases of advanced or metastatic prostate cancer. (Benafif & Eeles, 2016; Thalgott et al., 2018).

## 2.6 Overview of Machine Learning and Artificial Intelligence

The definition of artificial intelligence is based upon the ability of a computer to make decisions that are like human intellect based on the surrounding to achieve a certain goal. Machine learning is a subset of artificial intelligence that utilized algorithms for the purpose of data analysis. Based on the type of label and feature, machine learning techniques are classified. Machine learning is mainly classified into three models based on the labeling of data such as supervised, unsupervised and reinforcement learning. (S. Larry Goldenberg et al., 2019). A subset of machine learning is deep learning that learns from experience and understanding of environment. Recently, deep convolutional neural networks have an application in computer aided diagnosis of prostate cancer using imaging features. This leads to artificial intelligence and machine learning being considered an area of development when it comes to cancer prediction and diagnosis. (Song et al., 2018).

Machine learning uses data that has been collected or stored to make future predictions. This can also include determining if an individual has the risk of developing cancer or not. Machine learning modeling techniques are trained using existing data samples. The first category of machine learning which is supervised learning in which the algorithm predicts the outcome using data that is labelled. When it comes to prediction of cancer, it is mainly a supervised problem as it includes risk factors diet, ethnicity, alcohol, or drug abuse to determine if these factors lead to the development of cancer or not. Some of the main machine learning supervised models include random forest, support vector machine, naïve bayes, logistic regression and artificial neural network. (S. L. Goldenberg et al., 2019; Hussain et al., 2018).

### *2.6.1 Random Forest Classifier*

Random forest is a supervised machine learning model in which several decision trees are built for classification and regression trees which are binary splits on predictor variable for prediction of outcome. For classification task, the output depends upon the class which is selected by majority of the decision trees. Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. (Breiman et al., 1984). This results in a wide diversity that generally results in a better model. Therefore, in a random forest classifier, only a random subset of the features is taken into consideration by the algorithm for splitting a node. The random forest algorithm randomly selects observations and features to build several decision trees and then averages the results. Random forests prevent overfitting by creating random subsets of the features and building smaller trees using those subsets. Afterwards, it combines the subtrees. (Speiser et al., 2015).

### *2.6.2 Naïve Bayes Classifier*

Naïve Bayes is a supervised machine learning algorithm that utilizes Bayes' rule. It makes a strong assumption that the features are conditionally independent given the class. Naïve Bayes nonetheless often delivers competitive classification accuracy. Naïve Bayes is widely applied in practice when coupled with its computational efficiency and many other desirable features. Naïve Bayes provides a mechanism for using the information in sample data to estimate the posterior probability P (y | x) of each class y given an object x. Once we have such estimates, we can use them for classification or other decision support applications. (Frank et al., 2000). Naïve Bayes' features include many features such as computational efficiency in which training time is linear with respect to both the number of training examples and the number of features and classification time is linear with respect to the number of features and unaffected by the number of training examples. Key features of Naïve Bayes include robustness due to its use of probabilities and insensitivity to noise in the training data as well as another important feature includes handling missing values. Moreover, Naïve Bayes utilizes all the features for predictions, therefore, in case of any missing features, it[i] will not affect performance as other features will still be used. This results in a graceful degradation in performance and makes Naïve Bayes insensitive to missing features values in training set due to its probabilistic framework. (Nigsch et al., 2008).

### *2.6.3 Logistic Regression Classifier*

Statistical models in which a logistic curve is fitted to the dataset. This technique is applied when the dependent variable or target variable is dichotomous. Unlike Decision Trees or SVM's, there is nice probabilistic interpretation and model can be updated to take new data easily (using online gradient descent method). Since it returns probability, the classification thresholds can be easily adjusted. The logistic model can be an alternative for Discriminant Analysis. It has fewer assumptions - no assumption on the distribution of the independent variables, no linear relationship between the predictors and target variable must be assumed. It can handle interaction effect, nonlinear effect, and power terms. However, it requires a large sample size to achieve stable results. (B. Zhang et al., 2023).

### *2.6.4 Support Vector Machines Classifier*

Support Vector Machines (SVM) is another class of supervised machine learning algorithms used for classification. The map data into a high-dimensional space by using a non-linear function called kernel function. SVMs are particularly effective when dealing with high-dimensional data and are known for their ability to find a clear separation boundary, called a hyperplane, between different classes. The key idea behind SVM is to identify the hyperplane that maximizes the margin, which is the distance between the hyperplane and the nearest data points of each class. Some strengths of using SVM include their versatility and robustness against overfitting as they find the optimal hyperplane. This maximizes the margin rather than fitting the training data exactly which makes SVM a popular choice in real-world applications like image classification, text categorization, and bioinformatics. While SVMs offer many advantages, they also come with some computational complexity, especially when dealing with large datasets, and require careful tuning of hyperparameters. (Cannon et al., 2007).

## 2.7 Applications of Machine Learning in Prostate Cancer Prediction

Numerous studies have explored the utilization of artificial intelligence (AI) in the classification of lesions in prostate cancer detection. AI models categorize lesions outlined by radiologists into various groups, including cancer or benign, clinically significant cancer or benign, and distinct Gleason grade groups. The approach to lesion classification with AI often involves traditional machine learning methods, which entail extracting manually designed features from the region of interest and subsequently employing a classifier to determine the category to which the lesion belongs .(B. Zhang et al., 2023). These hand-crafted features encompass an assessment of factors like texture, shape, volume, and image-based radiomic characteristics. They can be broadly categorized into two main groups based on their algorithm type – traditional machine learning and Deep Learning. Moreover, there exist AI models that employ complete sets of images from a prostate magnetic resonance (MR) examination as their inputs. These models aim to identify, pinpoint, and potentially assess the aggressiveness of cancer across the entirety of the prostate MRI, primarily for lesion detection. These methods for lesion detection offer a granular, pixel-level estimation of cancer distribution within the prostate, illuminating regions with a high likelihood of cancer presence. (Johnson et al., 2019; van der Leest et al., 2019).

## 2.8 The Behavioral Risk Factor Surveillance System (BRFSS)

The Behavioral Risk Factor Surveillance System (BRFSS) is the main telephone survey system that is related to healthcare in the United States.(Hsia et al., 2020). It is mainly used for health-related telephone surveys to assess the health conditions of the local population. It was established in 1984 and now collects data in all US states and three US territories as well as the district of Columbia. Through this data collection, BRFSS is a very impactful tool in building activities that lead to health promotion. BRFSS has a wide range of sponsorships such as CDC (Centers for Disease Control), other federal agencies such as Administrations of Health Resources, Aging, as well as Services of Substance Abuse and Mental Health. Many countries such as Australia, Canada, China, Brazil, Egypt, Mexico, Jordan have requested staff from BRFSS to develop similar healthcare surveillance systems .(Hsia et al., 2020).

### *2.8.1 History and Background of BRFSS*

Early research has proven that personal health behaviors have a major role to play in the morbidity and mortality of many diseases. There are other entities in the US such as the National Center for Health Statistics that provide information regarding the national estimates of health risk behaviors, however, there is less availability of data on a state specific basis. This has been proven as a deficiency for state health agencies that are trying to find target resources that can reduce behavioral risks and the illnesses related to them. To achieve national health goals, state and local agency participation is required. As personal health behaviors were getting recognition with their link to the morbidity and mortality of many chronic conditions, another method that emerged was telephone surveys that could be helpful to determine the prevalence of many health risk behaviors. There are many advantages of conducting a telephone survey such as the cost advantage, as well as at the state level where there is a chance local expertise not being available to hold in-person interviews. (53). (Pierannunzi et al., 2012). This resulted in the development of surveys that could monitor prevalence of major behavioral risk factors amongst the adult population at the level of the state to find its association with premature morbidity and mortality. This is achieved by collecting data on actual behaviors which would help in planning, initiation and evaluation of health promotion and disease prevention programs .(Hsia et al., 2020; Pierannunzi et al., 2012).

To understand the practicality of behavioral surveillance, point-in-time state surveys were conducted in 29 states across the United States from 1981-1983. This led to the establishment of a monthly data collection system called Behavioral Risk Factor Surveillance System (BRFSS) by the Center for Disease Control and Prevention in 1984. A standard core questionnaire was developed that could collect data to be compared across states which topic including smoking, alcohol, physical inactivity, diet, hypertension etc. In 1988, further modules were implemented with question on specific topics.

Moreover, in 1993, BRFSS became a nationwide surveillance system with the questionnaire redesigned to include fixed core, rotating core, and emerging core questions by completion of approximately 100,000 interviews. The first panel meeting for BRFSS was held in 2002 which included many survey statisticians, methodologists, and operational experts where a discussion was held on the challenges and their implications related to survey research by BRFSS. Many states have used BRFSS survey to address urgent and emerging health issues such as monitoring of influenza vaccine shortage by BRFSS during the 2005 flu season or assessing the impact of hurricane Katrina and Rita in the same year. In 2009, modules for influenza like illness were also included due to H1N1 flu pandemic. (Esser et al., 2020).

Furthermore, in 2009 cell phones were also added as part of the BRFSS surveys to include population that had cell phones but not landline. This resulted in production of much higher quality data and cell phone surveys were publicly released in the beginning of 2011 with more than 500,000 telephone interviews conducted. This made BRFSS the largest telephone survey in the world.

### *2.8.2 Management of Survey by BRFSS*

A BRFSS survey, also known as the Behavioral Risk Factor Surveillance System survey, is managed through a well-structured process to gather important health-related information from individuals across the United States. The survey begins with careful planning and designing of the questionnaire. Experts work together to determine the key health topics to be covered, the sample size, and the sampling methodology to ensure it represents the population accurately. A representative sample of individuals is selected from the population using various statistical techniques. This ensures that the survey results can be generalized to the larger population. Trained interviewers or survey administrators reach out to the selected participants either via phone, mail, or even through web-based surveys. They collect responses to the survey questions while ensuring confidentiality and privacy.

Rigorous quality control measures are implemented throughout the data collection process. This includes regular training and monitoring of interviewers, data validation checks, and consistent adherence to survey protocols. Once the data collection phase is complete, statisticians and researchers analyze the collected information. They use advanced statistical methods to draw meaningful conclusions from the data and identify trends or patterns related to the surveyed health behaviors and risk factors. The findings and insights from the BRFSS survey are compiled into reports or publications. These reports often serve as valuable resources for policymakers, public health officials, and researchers, helping them make informed decisions and develop effective health interventions.

### *2.8.3 Components and Collection of the BRFSS Survey*

The BRFSS survey consists of six key components that work together to gather important health-related information. The questionnaire is a crucial component of the BRFSS survey. It is carefully designed to collect data on various health behaviors, risk factors, and chronic conditions. The questions cover topics such as tobacco use, physical activity, diet, alcohol consumption, chronic diseases, mental health, and access to healthcare. A scientifically valid sampling methodology is employed to select a representative sample of individuals from the target population. This ensures that the survey results can be generalized to the broader population. Trained interviewers or survey administrators collect data from the selected participants using various methods, such as telephone interviews, mail surveys, or web-based surveys. They follow standardized protocols to ensure consistency in data collection.

Rigorous quality control measures are implemented to maintain the integrity of the data. This includes interviewer training, data validation checks, and adherence to survey protocols. Regular monitoring and supervision are conducted to ensure data accuracy. Once the data collection phase is completed, statisticians and researchers analyze the collected data. The findings from the BRFSS survey are compiled into reports, publications, and data sets. These reports are often made publicly available and serve as valuable resources for policymakers, researchers, and public health officials. The data collected may also be used for further analysis and research. These six components work together to provide a comprehensive picture of the health behaviors, risk factors, and chronic conditions within a population, which policymakers and public health professionals can make informed decisions and develop targeted interventions to improve public health efforts.

### *2.8.4 Participants of the BRFSS Survey*

The participants of the BRFSS (Behavioral Risk Factor Surveillance System) survey are adults aged 18 years and older residing in the United States. The survey aims to collect information on various health behaviors, risk factors, and chronic conditions that can impact public health. The BRFSS survey utilizes a sampling methodology to select a representative sample of individuals from the target population. The sampling can be conducted at the state level or at a smaller geographic level, depending on the survey's objectives. The participants are selected using scientifically valid sampling techniques, such as random digit dialing for telephone surveys or address-based sampling for mail surveys.

The survey is designed to capture data from a diverse range of individuals, including people from different demographic groups (e.g., age, gender, race/ethnicity), geographic locations, and socioeconomic backgrounds. This diversity ensures that the survey results can be generalized to the broader population. It is important to note that the BRFSS survey focuses on collecting data from non-institutionalized individuals. This means that individuals residing in institutions such as nursing homes, correctional facilities, or military installations are not included in the survey.

Participation in the BRFSS survey is voluntary, and participants can decline or withdraw from it at any time. Confidentiality and privacy of the participants' responses are maintained throughout the data collection process to encourage honest and accurate reporting. Overall, the BRFSS survey aims to gather data from a representative sample of adults in the United States to monitor health behaviors, risk factors, and chronic conditions. The collected data helps inform public health policies and interventions to improve the health and well-being of the population.

### *2.8.5 Clinical Significance of BRFSS*

The BRFSS survey has significant clinical implications for public health. The first is in terms of identifying risk factors. The BRFSS collects data on various risk factors such as smoking, excessive alcohol consumption, physical inactivity, poor nutrition, and obesity. This information helps healthcare professionals and policymakers understand the prevalence and distribution of these risk factors in the population. By identifying these risk factors, interventions and targeted programs can be developed to address them, leading to improved health outcomes. The BRFSS collects data on chronic conditions such as diabetes, hypertension, asthma, and heart disease for monitoring purposes. This data collection process allows healthcare professionals to monitor the prevalence and trends of these conditions, identify high-risk populations, and allocate resources accordingly. It also helps in evaluating the effectiveness of interventions and healthcare policies aimed at preventing and managing chronic diseases.

The BRFSS provides valuable information on health behaviors such as tobacco use, physical activity, and diet. This data helps in understanding the prevalence of unhealthy behaviors and their association with chronic diseases. This information is usually used by health professionals to develop targeted interventions and educational campaigns to promote healthier behaviors and prevent disease. The BRFSS data allows for the examination of health disparities among different populations. It helps identify groups that are at a higher risk for certain health conditions or behaviors due to factors such as race, ethnicity, socioeconomic status, or geographic location. This information allows to address health inequities and to implement targeted interventions to reduce disparities in healthcare access and outcomes. Finally, the BRFSS data is used to evaluate the impact and effectiveness of public health programs and interventions.

# CHAPTER 3: METHODOLOGY

## 3.1 Ethical Approval

Ethical approval for this study was obtained from the institutional review boards of Qatar University (QU-IRB 1666-E/22). The study was conducted according to the guidelines of the Declaration of Helsinki. Since the BRFFS dataset is a publicly available dataset, informed consent/assent was not required to be obtained from all study subjects. All data collection sheets including codebooks, questionnaire and answers were stored as in a digitalized format in the highly secured Qatar University-Microsoft Azure server environment.

## 3.2 Study Subjects and Data Collection

The study subjects were selected from the US Behavioral Risk Factor Surveillance System telephone survey (Pierannunzi et al., 2013), the US premier system for health-related telephone surveys. The BRFSS annual survey dataset for 2020 was downloaded from the BRFSS portal - https://www.cdc.gov/brfss/annual_data/annual_data.htm. The BRFSS codebook, questionnaire and survey data were downloaded. The survey data were downloaded in the SAS transport format. We then converted the SAS transport format to the TSV (tab separated values) format using in-house script written in the R language. The survey data for 2020 consisted of 401,959 participants. In total, there were 280 questions in the survey for the year 2020.

## 3.3 Inclusion/Exclusion Criteria

Inclusion and exclusion criteria were based on the chosen questions. We went through a laborious exercise to identify which question should be included and excluded. Questions included are questions directly or indirectly related to prostate cancer. In total, we excluded 249 questions, reducing them to 31 questions. The exclusion of questions was determined by reviewing them as features. Each feature must contribute to the study of prostate cancer.

Questions showing a correlation to prostate cancer were included, while those unrelated to the study's objective were excluded. The 249 questions were excluded because including them would introduce noise to the machine learning model. In machine learning, any dataset that does not provide a correlation or rationale for linking it to prostate cancer is considered wasteful. For the remaining 31 questions, we created a hypothesis for each of them on how they are related to prostate cancer from our exhaustive search in online health related databases and registries.

### 3.4 Class Identification and Justification

In our feature class selection process, we employed the 2020 BRFSS codebook to train our machine learning model for predicting the risk of cancer development from a randomly sampled dataset. We conducted a thorough search for the term "cancer" across all the codebooks and specifically focused on questions pertaining to prostate cancer. Subsequently, we meticulously reviewed all the questions related to prostate cancer and pinpointed those that could effectively discriminate between individuals with or without the condition. This comprehensive analysis led us to the determination that questions related to prostate cancer would be the most informative for distinguishing cases from non-cases. Our selected feature, denoted by the code "PCPSARS1," corresponds to the question: "What was the MAIN reason you had this PSA test?" Within this feature, individuals who responded with "Because they were told that they had prostate cancer" or "Because of family history of prostate cancer" were categorized as cases, while all other responses were categorized as non-cases. This specific feature choice was instrumental in our efforts to create a robust differentiator for identifying individuals with prostate cancer, enabling our machine learning model to make more precise predictions in this context.

## 3.5 Features Identification and Justification

For identification of our features, we went through the codebooks to identify the risk factors that will most likely result in the development of prostate cancer or any type of cancer. These risk factors will be the features for our machine learning model. There were many risk factor questions listed in the codebooks but not all of them were present in all three codebooks so we had to find the questions that were common in all three and that could increase the risk of developing prostate cancer. After scanning the codebooks, we shortlisted nine main features that could most likely contribute to prostate cancer which includes fitness level, disability, smoking, HIV, diabetes, hepatitis, mental health/abuse, drug/alcohol as well as the socioeconomic status. These features are the risk factors that are mainly associated with the development of prostate cancer.

In the developing world, the most diagnosed cancer is prostate cancer in men with a survival rate of 95%. There are a wide range of treatment options ranging from radical prostatectomy, radiation therapy and chemotherapy. However, the treatments lead to serious effects in the patient such as a significant decline in quality of life and physical activity, reduced bone density as well as an alteration in body composition. These adverse effects are mainly due to cancer treatment. (Bray et al., 2018).

### *3.5.1 Demographics:*

Prostate cancer is the most common type of cancer in elderly males and there is an increase incidence rate of prostate cancer especially in senior citizens which could be due to increased average life expectancy as well as increased prostate cancer screening. (Bray et al., 2018). According to a study, it was observed that the risk of developing prostate cancer increases in white men after 50 years of age even in those individuals that have no family history. Whereas, in Black men or men with a history of prostate cancer, it increases after the age of 40. Moreover, another study reported that 30% of men above the age of 50 that died for various causes other than prostate cancer were found to have prostate cancer upon histological tests done during autopsy. (Scardino, 1989).

### *3.5.2 Fitness*

Prostate cancer treatment can have many detrimental effects on the health of the individual including a reduced quality of life, reduced bone density as well as physical function and altered body composition such as a gain in fat mass and reduced lean mass. These adverse effects can either be directly related to treatment or indirectly due to a decline in physical activity that resulted during treatment. (Gardner et al., 2014; Taylor et al., 2009). Therefore, it is essential to understand the role of physical fitness and exercise in reducing mortality after a prostate cancer diagnosis. According to a study done by randomized control trials that were carried out in 1891 men that were receiving different types of treatments for prostate cancer. There was a positive outcome reported with 75% of cases reporting a statistically significant outcome. However, a limitation to the study was that there was no information provided on the dosage of exercise required to get the desired outcome. (Farris et al., 2017; Vashistha et al., 2016).

Secondly, another study evaluated the use of androgen deprivation therapy as there is a high increase in use of this therapy for the treatment of men with prostate cancer.  Men who do not die of prostate cancer can die due to cardiovascular disease, however, data on the effect of androgen deprivation therapy in men that are receiving the therapy is very limited. (Friedenreich et al., 2016; Keilani et al., 2017). A cohort study was carried out that calculated the risk of cardiovascular morbidity in men that are diagnosed with prostate cancer and receiving androgen deprivation therapy. The results proved that patients that are newly diagnosed with prostate cancer that received androgen deprivation therapy for minimum of a year have at least 20% more risk of death due to cardiovascular disease compared with men that did not receive the treatment. This therapy can significantly increase cardiovascular morbidity in men with prostate cancer, however, cardiac risks can be reduced through diet and exercise. (Saigal et al., 2007).

### *3.5.3 Disability*

Androgen deprivation therapy is a common treatment used for prostate cancer and this treatment can lead to several side effects in a patient with prostate cancer. A study investigated event-based prospective memory and time-based prospective memory in patients diagnosed with prostate cancer that have cognitive impairment due to androgen-deprivation therapy. The study included prostate cancer patients that had undergone androgen deprivation therapy as well as those that didn’t undergo the therapy along with healthy control that had the same age and education. All the participants were tested on various neuropsychological tasks which also included EBPM and TBPM. The results of the study demonstrated that the androgen deprivation therapy group received much lower scores on the event based prospective memory tasks compared to those individuals with prostate cancer that are not on androgen deprivation therapy as well as the healthy control group. Furthermore, no significant differences were observed in all the three groups for time-based memory tasks. In addition, the group that received ADT presented extremely low scores on other cognitive tasks related to attention, memory and processing information compared with non-ADT and control group. The results of the study demonstrated that prostate cancer patients receiving androgen deprivation therapy suffer from reduction in event-based prospective memory which may result due to changes in structure and function of the pre-frontal cortex .(Yang et al., 2015).

Another study investigated the effect of cancer treatment and factors associated with self- reported fall, balance as well as difficulty in walking. This study was a cross-sectional study that aimed to examine factors that are associated with falls, imbalance, and walking difficulty in four major cancer survivors. They analyzed population-based data from the Medicare Health Outcomes survey and extracted data from cohorts 9 to 14. The duration of the period begins from January 2006 to December 2013. Inclusion criteria included individuals with an age greater than or equal to 65 at cancer diagnosis with the first survey done within five years duration of cancer diagnosis. They examined four cancers including prostate cancer along with staging information of each cancer. These four cancer types were chosen as they are the most prevalent forms of cancer in adults aged 65 or above. The sample size was 9,540 survivors with 4,245 survivors of prostate cancer. Furthermore, they constructed logistic regression for each of the cancer types to analyze and identify independent factors that could result in falls, imbalance, and difficulty in walking. The results indicated that in all cancer types, age, and dependence in daily activities at the time of cancer diagnosis were factors of great significance with an increased odds of reporting falls, imbalance, and difficulty in walking. In addition, depression was another independent factor linked to falls and sensory impairment was an independent factor linked with imbalance, and difficulty in walking for all cancer types including prostate cancer. The main finding includes screening for individuals that are cancer survivors including prostate cancer for imbalance, risk of falls and difficulty in walking. (Huang et al., 2018).

### *3.5.4 Smoking*

It is widely accepted amongst the medical community that cigarette smoking is a major cause of mortality and a report according to surgeons general found that smoking has an increased association with advanced stage prostate cancer or death from prostate cancer, however, there is no significant association with the overall incidence of the disease. The extent to which other tobacco products can cause harm to have also not been identified. For example, snus, which is a moist smokeless tobacco product, is a source of nicotine. According to the study conducted by where information using tobacco use was collected within a cohort of Swedish construction workers which included people categorized into never users of tobacco, exclusive snus users, exclusive smokers, and users that use both snus and smoking. The results showed that exclusive snus users were at a higher risk of prostate cancer mortality with a confidence interval between (1.03-1.49). The study proved that tobacco related carcinogenic products can cause cancer progression irrespective of tobacco’s combustion. (Wilson et al., 2016).

Another study examined the association between smoking at the time of diagnosis with an increased risk of mortality due to prostate cancer in a cohort study of men with prostate cancer. Data was collected from 752 prostate cancer patients between the ages of 40-64. The cases were enrolled in a case-control study with a long-term follow up. Hazard ratios and 95% confidence intervals were estimated using cox proportional hazard models to determine an association between smoking and mortality due to prostate cancer. In the results, it was found that smoking at the time of diagnosis leads to mortality due to prostate cancer significantly with a hazard ratio of 2.66 and 95% confidence interval between 1.1-6.43 leading to a significant association. (Gong et al., 2008).

### *3.5.5 HIV*

HIV patients have a high life expectancy due to advancements in viral treatments, however, these patients are at risk of secondary treatments. True incidence of prostate cancer in HIV-positive men is unknown. The cases that have been presented, it appears to behave in the same way in both HIV-positive and HIV-negative men. As prostate cancer is the most common malignancy in men with approximately a million men in the US which are HIV positive, there is not adequate literature about prostate cancer in HIV-positive patients which no consensus on screening or treatment of this patient population. However, a review identified prostate cancer to be a common malignancy in HIV-positive men. This can lead to the development of therapies for HIV which can test and screen for prostate cancer. (Biggar et al., 2004; Levinson et al., 2005; Manfredi et al., 2006).

### *3.5.6 Diabetes*

According to a study, pre-existing diabetes is linked to worse overall mortality in cancer patients, however, its impact varied depending upon the type of cancer. (Barone et al., 2008). As discussed, prostate cancer is the second most common malignancy diagnosed in adult men with androgens being considered as primary growth factors for normal and prostate cancer cells. However, there are other non-androgenic growth factors that are also linked to prostate cancer cells growth regulation. There is a well-established association between IGF1 and prostate cancer risk, however, there is a lack of evidence that IGF-1 measurement results in an enhanced specificity of detection of prostate cancer that is beyond the detection level achieved by prostate specific antigen (PSA) levels. Furthermore, the study indicates that high insulin levels can be associated with prostatic tumors although it is not well established. (DiGiovanni et al., 2000; Greenberg et al., 1995).

Although there are several studies done to study the risk factors related to prostate cancer, more research is required. There are many well established risk factors such as age, family history and ethnicity, however, many other risk-factors such as androgens, high fat intake, along with the role of insulin in development of prostate cancer remains unclear. (Suba & Ujpál, 2006). Another study focuses on evidence that relates insulin to pathogenesis of prostate cancer. One of the risk factors in the development of malignancies is insulin resistance resulting in hyperglycemia and tumor genesis. This is due to increased DNA synthesis in tumor cells caused by high glucose levels. Further detrimental effects of hyperglycemia include nonenzymatic glycation of proteins along with deliberation of free-radical and growth factors. Hyperinsulinemia is considered on the top risk factors for development of several malignancies including benign prostatic hyperplasia. (Hussain et al., 2003; Nandeesha, 2008; Okumura et al., 2002; Salahudeen et al., 1997). Some studies reported hyperinsulinemia to be considered a risk factor in the development of prostate cancer and insulin could also be used as a marker for prognosis and tumor aggressiveness of prostate cancer. In conclusion, hyperinsulinemia in association with insulin resistance may play a role in prostate cancer pathogenesis. (Nandeesha et al., 2008).

### *3.5.7. Drug/Alcohol*

Other potential risk factors for prostate cancer include drug/alcohol. The effects of alcohol consumption on prostate cancer remain unclear due to which a study investigated the genetic variants that are present in genes that metabolize alcohol and their association with the incidence and survival rate of prostate cancer. Data analysis was done from 25 studies consisting of 23,868 men with prostate cancer and 23,051 control. The study found an association between 68 SNPs in eight genes that metabolize alcohol and prostate cancer mortality rate using logistic and cox regression models. The study performed a meta-analysis of 25 studies and there was no association found between prostate cancer diagnosis and variants in alcohol metabolizing genes. The results of the meta-analysis concluded that alcohol consumption is less likely to influence prostate cancer incidence rate, however, it can contribute to disease progression. (Brunner et al., 2017).

Furthermore, there was another study conducted that examined the risk of developing low- or high-grade prostate cancer in association with the type of alcoholic beverage and drinking pattern. In this study, data was collected from 2,129 participants that had cancer detected and 8,791 participants that were cancer-free by the end of the trials. The trials were known as Prostate Cancer Prevention Trials which ran for a period of 7 years. Relative risks were calculated using Poisson regression with 95% confidence intervals to determine the link between the risk of developing prostate cancer and alcohol intake. The study found that less heavy drinking had no association with risk of developing prostate cancer, however, regular, and heavy alcohol consumption have an association with increased risk of developing high-grade prostate cancer. Moreover, the results also demonstrate that heavy drinking also makes the treatment of finasteride ineffective. (Gong et al., 2009).

In addition, there is not a clear investigation done on the association between tumor stage and alcohol consumption. Another study investigated the relation between prostate cancer and current or lifetime intake of alcohol. The study took place in Canada and was a population-based case-control study. The number of cases was 947 that had stage T2, or higher prostate cancer and the number of controls was 1,039. Cases were classified based on cancer stage and severity. Interviews were conducted to assess the current and lifetime history of prostate cancer and it was found that the risk of prostate cancer did not increase in current alcohol intake. However, lifetime intake of alcohol resulted in an increased risk of prostate cancer for both aggressive and non-aggressive cases. (McGregor et al., 2013).

Finally, men that consume drugs maybe at a high risk of death due to prostate cancer. A study in Sweden was conducted to investigate prostate cancer mortality, stage, and incidence rate in men with drug use disorders with general male population. The study was carried out on 1.3 million men above the age of 50 out of whom 9,259 had drug use disorders. Prostate cancer stage at the time of diagnosis, incidence and mortality of prostate cancer cases registered with DUD were analyzed using cox regression analysis. The results demonstrate that drug use disorder was significantly linked to fatal prostate cancer with a slightly increased risk linked to incidence of the disease. The study found that there is an increased risk of death due to prostate cancer in men that suffer from drug use disorders due to several factors such as a delay in diagnosis or insufficient treatment. (Dahlman et al., 2022).

### *3.5.8. Mental Health/Abuse:*

The incidence of depression in prostate cancer patients is high compared to those without. However, there is little information about the incidence rate of depression subtypes. To further investigate this, a survey questionnaire was completed by over 500 prostate cancer patients. Various factors relating to depression as well as prostate cancer were examined such as symptoms of depression, as well as stressors related to prostate cancer. Amongst the patients, a score was given for each of the common subtypes of depression based on the depressive symptomology. The results suggested that nearly 50% of the patients had scores that were considered clinically significant for at least one of the depression subtypes with some patients demonstrating clinically significant score for various of those subtypes. The results found an association with prostate cancer related stressor and different subtypes of depression. Lastly, the study concluded that the treatment of prostate cancer patients differs depending on the subtype of depression presented by the patient. (Sharpley et al., 2013).

According to a study, one in five men with prostate cancer can become depressed resulting is high chances of suicide compared to those men without prostate cancer.  As a high number of prostate cancer patients experience severe levels of depression, it can have an immense negative impact on their treatment and disease course. (Watts et al., 2014). Furthermore, certain prostate cancer treatments such as anhedonia and erectile dysfunction might lead to an increase in depression severity. According to a review of 26 studies done on depression in men with prostate cancer with a sample size of 4,494 patients aged between 57-73 years which assessed the prevalence of depression. It was found that 17.27% had depression before, 14.7% during and 18.44% after treatment which was much higher in comparison to men of similar age group. No statistical comparisons were, however, high prevalence of depression after treatment might be linked to anxiety due to outcome. (Watts et al., 2014). Prevalence of depression in individuals with prostate cancer is 2-3 times higher compared to those without regardless of the time point of assessment with regards to diagnosis and treatment. (Caruso et al., 2017).

Lastly, forms of abuse such as adverse childhood experiences have been linked to higher odds of developing cancer in adulthood. This study examined the association between adverse childhood experiences with the compliance with screening of prostate, breast, cervical and colorectal cancer. The study utilized data from 2014 Kansas Behavioral Risk Factor Surveillance System with a sample size of 11,794. Odds of cancer screening behaviors were calculated from nine different ACE using logistic regression. For PSA screening, clinical breast exam and pap test guidelines, there were low odds of compliance from individual ACEs. Whereas, certain ACE had an increased odds of compliance association. Physical abuse had a common association with cancer screening with specific ACEs linked with lower odds of cancer screening. This proves that extra effort should be made to promote screening of prostate cancer in individuals with a history of adverse childhood experiences. (Alcalá et al., 2018).

## 3.6 WEKA: Machine Learning Tool

Weka, the open-source software tool developed at the University of Waikato in New Zealand, has gained popularity in the field of machine learning for several compelling reasons. Its user-friendly graphical interface, extensive array of machine learning algorithms, and robust data preprocessing capabilities have solidified its position as a top choice for both newcomers and experienced data scientists. With Weka, a diverse range of machine learning experiments becomes possible. These experiments include tasks such as supervised learning, for instance, classification and regression, along with unsupervised learning for discovering data patterns through clustering. Weka also provides features for handling tasks like feature selection and dimensionality reduction, simplifying high-dimensional data. Additionally, Weka serves as a valuable tool for conducting model evaluations, parameter tuning, and the comparison of multiple algorithms, offering a comprehensive platform for the development of effective machine learning solutions.

## 3.7 Handling Missing Data

In the BRFSS dataset, missing data is denoted as 'NA' (not available), indicating that the data is absent because the study subjects did not answer the respective question. We transform these missing values into zeros. This conversion is necessary because Weka, the software we are using, requires numerical data for its algorithms to function correctly. Utilizing a non-numeric placeholder, such as 'NAN,' would lead to errors during dataset operations. To handle this transformation, we employed the NumericalCleaner filter function in Weka. We specified that all features containing a value of 0 should be treated as missing and disregarded in subsequent analyses. We set the minThreshold option to 1E-10 and assigned minDefault as NaN (Not A Number). Setting the minThreshold close to 0 ensures that we retain features with very low variance in the dataset. This is particularly valuable as low-variance features may contain essential information for training the model. During the model training process, any values below the 1E-10 threshold are replaced with NaN, a special symbol recognized by Weka as indicating missing values.

## 3.8 Test Strategy

We selected 10-fold Cross Validation (CV) option for assessing model performance. CV is a technique in machine learning to assess the performance ability of a predictive model. It involves systematically splitting the dataset into multiple "folds" which in our case is 10. For each fold, the model iteratively trained on a portion of the data and tested on the remaining unseen data. The key advantage of cross-validation is that it provides a more reliable estimate of a model's performance compared to a simple train-test split (the Percentage Split option in WEKA). CV is preferred over a simple percentage split because of its robustness and reliable performance assessments. This is because in percentage split, the model's performance is determined by a single random split of the data, which can be highly influenced by the random seed used for the split. In contrast, CV systematically tests the model on different “folds” of the data, providing a more comprehensive and stable evaluation. It helps ensure that the model's performance estimates are less sensitive to the specific composition of the training and testing data. Also, CV efficiently utilizes the available data by maximizing its use for both training and testing, reducing the risk of data partitioning influencing the results.

## 3.9 Model Training

We selected the four most common supervised-based classification algorithms: i) Random Forest, ii) Logistic Regression, iii) Naïve Bayes, and iv) Sequential Minimal Optimization (SMO), which is a more efficient implementation for training SVM. Running these algorithms in Weka involves several steps. In Weka, once we have loaded and filtered our dataset, we go to the 'Classify' tab. In the 'Classify' panel, we select our algorithm of interest from the list of classifiers. For all four classifiers, we use the default parameters.

## 3.10 Test Strategy Selection

Weka offers a range of performance assessment measures for evaluating the effectiveness of classifiers utilized in the analysis. Two commonly employed techniques are cross-validation and percentage split. Cross-validation is a widely used technique in machine learning that involves dividing the dataset into multiple subsets, typically referred to as "folds," and iteratively training the model on a subset while using the remaining data for testing. This process is repeated multiple times, each time with a different fold as the test set and the remaining data as the training set. The results are then aggregated to evaluate the model's performance in a way that helps mitigate issues like overfitting and provides a more robust estimation of the model's predictive accuracy. Common types of cross-validation include k-fold cross-validation, stratified cross-validation, and leave-one-out cross-validation, with the choice of method depending on the specific modeling problem and dataset.

In contrast, the percentage split strategy involves dividing the dataset into two portions: a training set and a test set, typically with a specified percentage split. The model is trained on the larger training set, and its performance is assessed by how well it predicts outcomes on the separate test set. This strategy allows for a straightforward assessment of model accuracy and generalizability, and it's particularly useful when working with large datasets where the computational resources required for techniques like cross-validation may be prohibitive.

In our experiment, we opted for cross validation because it offers a more robust assessment of model performance due to its repeated training and testing cycles on different subsets of data. In contrast, percentage splits, while simpler and computationally less demanding, may not capture the full variability and complexity of the data. This is because our dataset might have variations and hidden patterns that are not apparent in a single training-test split. Additionally, cross-validation helps in addressing the risk of overfitting by providing a more realistic estimate of a model's predictive accuracy.

## 3.11 Classifiers Evaluation

### Confusion Matrix

Confusion matrix provides a comparison between model’s predictions in respect to the class labels. It is represented as a table in which instances in class are represented by rows and instances in a predicted class are represented by columns. This matrix is used to determine the accuracy of the model by comparing between correct and incorrect predictions. Key metrics provided by confusion matrix include true positives, true negatives, false positives and false negatives. Other performance metrics that can be derived from confusion matrix include true positive, false positive, true negative, false negative, precision, recall and F-measure, all of which can be used to evaluate the performance of a model.

True positive rate

The true positive rate, which is also known as sensitivity or recall, is used to represent the actual number of positive instances that have been correctly classified by the model. True positive rate is equal to true positive divided by true positive plus false negative. False negative is the number of instances that are positive but are wrongly classified as negative.

TPR= TP/ (TP+FN)

TP= True Positive (instances correctly predicted as positive)

FN= False Negative (instances incorrectly predicted as negative)

False positive rate

False positive rate in Weka is a performance metric which helps in evaluating the performance of a classification model and it measures the incorrect predictions made by the model when it comes to positive class.

FPR= FP/ (FP+TN)

FP= False Positive (instances incorrectly predicted as positive)

TN= True Negative (instances correctly predicted as negative)

Precision

Another performance metric Weka that can be used to evaluate a classifier's accuracy is precision. It is a measure of correctly predicted positive instances out of the total number of positive instances that are predicted. The calculation of precision is done by true positive divided by the sum of true positive and false positives. The higher the value for precision, the better it indicates that the classifier is better at identifying positive instances. Alternatively, a low precision value indicates that the classifier is more prone to making incorrect positive predictions.

Precision= TP/ (TP+FP)

TP= True Positive (instances correctly predicted as positive)

FP= False Positive (instances incorrectly predicted as positive)

### Recall

Recall is another performance measure which is used in evaluation of a classification model in its ability to correctly identify positive instances from a dataset. Recall can be measured as the ratio of true positive divided by the sum of true positive and false negative instances. Recall value for a classifier is directly proportional to a model’s effectiveness to correctly identify positive instances.

Recall= TP/ (TP + FN).

TP= True Positive (instances correctly predicted as positive)

FN= False Negative (instances incorrectly predicted as negative)

### F-Measure

F-measure, which is also known as F1 score, is a performance measure that combines precision and recall into a single metric. As described above, precision is the model's ability to correctly predict positive instances out of the total number of instances, whereas recall, also known as sensitivity, measures the correctly predicted true positive out of the total number of actual positives.  F-measure is more sensitive to imbalance between precision and recall.

F-measure= 2* (precision*recall)/ (precision + recall)

F-measure ranges from 0-1 and the closer a value is to 1 the better the classification accuracy. It is used when the false positives and false negatives need to be balanced where precision and recall both are important.

### ROC Curve

ROC, which stands for Receiver Operating Characteristic curve is a graphical representation of classification algorithms. It is an illustration of true positive rate which is sensitivity and false positive rate (1-specificity) for different classification threshold. ROC curve plots TPR true positive rate on y axis against FPR false positive rate on x-axis for different classification models. The AUC Area under the ROC curve can be used to evaluate classifier performance with a perfect classifier having an AUC of 1. ROC curve can be constructed on Weka using the following steps:

1) Train classification model using training dataset.
2) Apply the trained model on a test set to obtain predicted class labels by using evaluation models such as cross validation or percentage split.
3) Utilize "Visualize Threshold Curve" option in Weka's classifier output window to generate a ROC curve.

### 3.12 Identification of Top Risk Factors

GainAttributeEval, CorrelationAttributeEval and InfoGainAttributeEval are feature selection methods in WEKA, each offering a distinct approach to assessing the relevance of features in a dataset. GainAttributeEval quantifies information gain, measuring the ability of features to reduce uncertainty and enhance predictive accuracy, making it a versatile choice. It is widely applicable in a range of scenarios where improving model accuracy is the primary goal. CorrelationAttributeEval, in contrast, evaluates the linear correlation between features and the class, focusing on the strength of linear relationships, which can be valuable when linearity is crucial. It is useful when understanding and leveraging linear relationships between features and the class features is important. InfoGainAttributeEval also calculates information gain but is particularly suitable for discrete data. It assesses the knowledge provided by features about the class variable, making it a good choice for non-numeric features. The method selected should align with the data type and the specific objectives of feature selection.

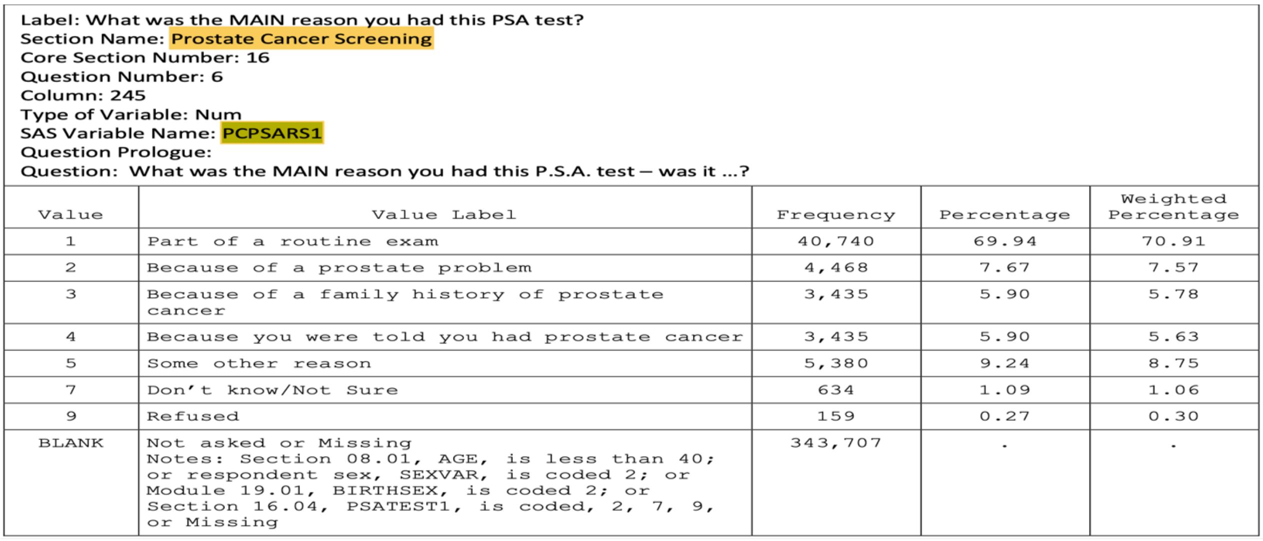
GainAttributeEval assesses features based on their information gain, measuring their ability to reduce uncertainty and improve predictive accuracy. It's well-suited for both numeric and categorical data. It emphasizes feature selection based on information gain, aiming to find features that contribute the most to predictive accuracy. Similarly, InfoGainAttributeEval evaluates information gain but is particularly suitable for discrete and categorical data, which some of our features consist of. It focuses on identifying features with strong linear relationships with the class, which can be valuable for specific modeling situations. For CorrelationAttributeEval, it focuses on the linear correlation between features and the class, emphasizing the strength of linear relationships, which is especially relevant for numeric data, which represents most features.

In our experiments, we opted to use all three methods. For each, we identified the top and bottom 10 features respectively. We then chose the best performing training model and re-trained our model based on the exclusion of the top 10 features and compared the model accuracy result with all features included. Similarly, we did the same experiment for the bottom 10 features and compared the model accuracy result.

# CHAPTER 4: RESULTS

## 4.1 Background of Classes

The feature class used for our experiment was taken from the BRFSS codebook 2020. The question that was chosen was based on the answers that would help us most differentiate between case and non-case. The feature class question is as follows: what was the main reason you had this PSA test? The answers provided would clearly help us differentiate between people who were either diagnosed with prostate cancer or had a family history of prostate cancer. These answers were considered as case while the rest of the options were considered as non-case. This was the only feature class that could be selected from all the questions as the questions on prostate cancer were limited.



**Label: What was the MAIN reason you had this PSA test?**
**Section Name: Prostate Cancer Screening**
**Core Section Number: 16**
**Question Number: 6**
**Column: 245**
**Type of Variable: Num**
**SAS Variable Name: PCPSARS1**
**Question Prologue:**
**Question: What was the MAIN reason you had this P.S.A. test – was it ...?**

| Value | Value Label | Frequency | Percentage | Weighted Percentage |
|---|---|---|---|---|
| 1 | Part of a routine exam | 40,740 | 69.94 | 70.91 |
| 2 | Because of a prostate problem | 4,468 | 7.67 | 7.57 |
| 3 | Because of a family history of prostate cancer | 3,435 | 5.90 | 5.78 |
| 4 | Because you were told you had prostate cancer | 3,435 | 5.90 | 5.63 |
| 5 | Some other reason | 5,380 | 9.24 | 8.75 |
| 7 | Don't know/Not Sure | 634 | 1.09 | 1.06 |
| 9 | Refused | 159 | 0.27 | 0.30 |
| BLANK | Not asked or Missing<br>Notes: Section 08.01, AGE, is less than 40; or respondent sex, SEXVAR, is coded 2; or Module 19.01, BIRTHSEX, is coded 2; or Section 16.04, PSATEST1, is coded, 2, 7, 9, or Missing | 343,707 | . | . |

Figure 1: Feature class from BRFSS 2020 codebook

## 4.2 Background of Features

The feature class is divided into nine main categories based on questions that were found in the BRFSS codebook 2020 and how those questions can be associated with diagnosis of prostate cancer. The categories were namely fitness, disability, smoking, HIV, diabetes, PSA Test, drug/alcohol, mental health/abuse, and socioeconomic status. Our features questions were selected and divided under these categories. Literature review was done to prove association between each category and the risk of developing prostate cancer.

Table 1: Groups and feature using BRFSS Codebook2020

| | GROUP | Feature | Label |
|---|---|---|---|
| 1) | DEMOGRAPHIC | _SEX | Gender |
| 2) | DEMOGRAPHIC | _AGE_G | Imputed age in 6 groups |
| 3) | FITNESS | EXERANY2 | Exercise in Past 30 Days |
| 4) | FITNESS | CVDINFR4 | Ever Diagnosed with Heart Attack |
| 5) | DISABILITY | DECIDE | Difficulty Concentrating or Remembering |
| 6) | DISABILITY | DIFFWALK | Difficulty Walking or Climbing Stairs |
| 7) | DISABILITY | DIFFDRES | Difficulty Dressing or Bathing |
| 8) | SMOKING | USENOW3 | Use of Smokeless Tobacco Products |
| 9) | SMOKING | LCSLAST | How old when you last smoked? |
| 10) | SMOKING | LCSNUMCG | On average, how many cigarettes do you smoke each day? |
| 11) | SMOKING | SMOKE100 | Smoked at Least 100 Cigarettes |
| 12) | HIV | HIVTST7 | Ever tested H.I.V. |
| 13) | DIABETES | PDIABTST | Had a test for high blood sugar or diabetes in the past three years? |
| 14) | DIABETES | PREDIAB1 | Ever been told by a doctor or other health professional that you have pre-diabetes or borderline diabetes? |
| 15) | DIABETES | INSULIN1 | Now Taking Insulin |
| 16) | PSA Test | PSATEST1 | Have you ever had a P.S.A. test? |
| 17) | PSA Test | PSATIME | How long has it been since you had your last P.S.A. test? |
| 18) | DRUG/ALCOHOL | ACEDRINK | Live With a Problem Drinker/Alcoholic? |
| 19) | DRUG/ALCOHOL | ALCDAY5 | Days in past 30 had alcoholic beverage. |
| 20) | DRUG/ALCOHOL | ACEDRUGS | Live With Anyone Who Used Illegal Drugs or Abused Prescriptions? |
| 21) | MENTAL HEALTH/ABUSE | ADDEPEV3 | (Ever told) you had a depressive disorder |

| | | | |
|---|---|---|---|
| 22) | MENTAL HEALTH/ABUSE | ACEDEPRS | Live With Anyone Depressed, Mentally Ill, Or Suicidal? |
| 23) | MENTAL HEALTH/ABUSE | ACEPRISN | Live With Anyone Who Served TIme in Prison or Jail? |
| 24) | MENTAL HEALTH/ABUSE | ACEPUNCH | How Often Did Your Parents Beat Each Other Up? |
| 25) | MENTAL HEALTH/ABUSE | ACEHURT1 | How Often Did A Parent Physically Hurt You In Any Way? |
| 26) | MENTAL HEALTH/ABUSE | ACESWEAR | How Often Did A Parent Swear At You? |
| 27) | DRUG/ALCOHOL | ACEDRINK | Live With a Problem Drinker/Alcoholic? |
| 28) | SOCIOECONOMCS | _URBSTAT | Urban/Rural Status |
| 29) | SOCIOECONOMCS | EMPLOY1 | Employment Status |
| 30) | SOCIOECONOMCS | ACEDIVRC | Were Your Parents Divorced/Separated? |

## 4.3 Pre-Processing of the Data

### *4.3.1 Recoding the class question.*

To recode the class question is extremely essential in differentiating between case and non-case. Our class question was chosen based on its ability to help us differentiate the most between individuals who either have prostate cancer or with a history of prostate cancer with individuals that have not been diagnosed or had a history of the disease. The only question that could explicitly separate the two was with the variable name PCPSARS1, “what was the MAIN reason you had this PSA test? The question had nine options from 1-9 as well as blank which was coded as zero. To recode the class question, the options from 0-9 had to be divided into binary with only “CASE” and “CONTROL” being our only two choices. The class question was re-coded by using option 2,3 and 4 as “CASE” as these answers were clear in identifying individuals that were either diagnosed or had a history of prostate cancer. Furthermore, options 0,1,5,7 and 9 were used as our “CONTROL” group. The pcpsars1 column was then recoded into a CLASS column with only a CASE and CONTROL group.

### *4.3.2 Loading the data and removing missing values:*

Data can be loaded onto Weka by using the software and opening Weka Explorer. Once

Weka explorer has been opened, the data file can be loaded onto Weka using the “Open file” icon. When the file has been loaded, many of the features will have “missing values”. These are frequency of answers to the features questions in codebook 2020 that were left blank. It is crucial to remove such missing values from the data as they can affect the accuracy and reliability of results.

Table 2: Features and missing values:

| **Feature** | **MISSING VALUES** |
|---|---|
| _SEX | 0 (0%) |
| _AGE_G | 0 (0%) |
| EXERANY2 | 0 (0%) |
| CVDINFR4 | 0 (0%) |
| DECIDE | 227 (1%) |
| DIFFWALK | 252 (2%) |
| DIFFDRES | 260 (2%) |
| USENOW3 | 322 (2%) |
| LCSLAST | 15174 (96%) |
| LCSNUMCG | 15176 (96%) |
| SMOKE100 | 311 (2%) |
| HIVTST7 | 781 (5%) |
| PDIABTST | 7392 (47%) |
| PREDIAB1 | 7392 (47%) |
| INSULIN1 | 15180 (96%) |
| PSATEST1 | 5459 (34%) |
| PSATIME | 6854 (43%) |
| ACEDRINK | 10541 (66%) |
| ALCDAY5 | 391 (2%) |
| ACEDRUGS | 10542 (66%) |
| ADDEPEV3 | 0 (0%) |
| ACEDEPRS | 10536 (66%) |
| ACEPRISN | 10546 (66%) |
| ACEPUNCH | 10549 (66%) |
| ACEHURT1 | 10554 (66%) |
| ACESWEAR | 10559 (67%) |
| ACEDRINK | 10541 (66%) |
| _URBSTAT | 258 (2%) |
| EMPLOY1 | 81 (1%) |
| ACEDIVRC | 10546 (66%) |

### 4.4 Model Training and Comparison

After the careful selection of features from the 2020 codebook, we curated a dataset comprising 30 features and 1 class, bringing the total number of features to 31. We derived the SAS variable names from each feature in the codebook and transformed the codebook data into a CSV format using Linux software. The class features were converted into a binary format to distinctly distinguish between "case" and "control" instances, preparing our data for processing in Weka software. To ensure a balanced dataset, we randomly selected instances based on class features, resulting in a total of 15,874 instances, with an equal number of cases and controls. This balanced dataset allows us to train various classifiers and construct machine learning models. The selection of the best classifier is based on the results obtained and subsequently validated using a test set.

Table 1 exhibits the results obtained by running the training set on the classifier logistic regression. The logistic regression model demonstrates strong performance in binary classification with the given evaluation metrics. In the "CASE" class, it achieves a high true positive rate (TP Rate) of 0.863, indicating its ability to correctly classify the positive instances, with an excellent precision of 1.000, reflecting the low rate of false positives. The recall, which is also 0.863, suggests that it captures a substantial proportion of positive instances. The F-measure stands at 0.926, highlighting a well-balanced trade-off between precision and recall. Overall, the Matthews Correlation Coefficient (MCC) is 0.871, and the Receiver Operating Characteristic (ROC) Area is 0.938, showcasing the model's overall discriminative ability. The "CONTROL" class shows an even more impressive performance, with a perfect precision of 1.000 and a recall of 1.000, leading to an outstanding F-measure of 0.936. The average (AVG.) of these two classes is also impressive, with a high ROC Area of 0.938, indicating the model's generalization ability.

Table 3: Training set on Logistic Regression

| TP Rate | FP Rate | Precision | Recall | F-measure | MCC | ROC Area | PRC Area | CLASS |
|---|---|---|---|---|---|---|---|---|
| 0.863 | 0.000 | 1.000 | 0.863 | 0.926 | 0.871 | 0.938 | 0.960 | CASE |
| 1.000 | 0.137 | 0.880 | 1.000 | 0.936 | 0.871 | 0.938 | 0.893 | CONTROL |
| 0.931 | 0.069 | 0.940 | 0.931 | 0.931 | 0.871 | 0.938 | 0.926 | AVG. |

The random forest model demonstrates strong performance in binary classification, as evident from the evaluation metrics. In the "CASE" class, it achieves a robust TP Rate of 0.860, indicating its ability to correctly classify positive instances, with a commendable precision of 0.970. The recall, standing at 0.860, suggests that it captures a significant proportion of positive instances, resulting in a solid F-measure of 0.912. The MCC is 0.839, and the ROC Area is 0.950, underlining the model's strong discriminative ability. In the "CONTROL" class, the model performs even better, with a remarkable recall of 0.973 and an F-measure of 0.921. The average of these two classes is also impressive, with a high ROC Area of 0.950, showcasing the model's strong generalization ability.

Table 4: Training set on Random Forest

| TP Rate | FP Rate | Precision | Recall | F-measure | MCC | ROC Area | PRC Area | CLASS |
|---|---|---|---|---|---|---|---|---|
| 0.860 | 0.027 | 0.970 | 0.860 | 0.912 | 0.839 | 0.950 | 0.965 | CASE |
| 0.973 | 0.140 | 0.875 | 0.973 | 0.921 | 0.839 | 0.950 | 0.922 | CONTROL |
| 0.917 | 0.083 | 0.922 | 0.917 | 0.917 | 0.839 | 0.950 | 0.943 | AVG. |

The Naïve Bayes model provides respectable performance in binary classification, as observed from the evaluation metrics. In the "CASE" class, it demonstrates a reasonably strong true positive rate (TP Rate) of 0.847, successfully classifying positive instances, accompanied by a precision of 0.926, reflecting a satisfactory trade-off between true positives and false positives. The recall, at 0.847, implies that it captures a substantial portion of positive instances, resulting in an F-measure of 0.885. The Matthews Correlation Coefficient (MCC) stands at 0.783, and the Receiver Operating Characteristic (ROC) Area is 0.924, indicating moderate discriminative ability. In the "CONTROL" class, the model shows a high recall of 0.933 and an F-measure of 0.895, emphasizing its ability to correctly classify negative instances. The average (AVG.) performance, with an ROC Area of 0.924, suggests the model's reasonable generalization ability.

Table 5: Training set on Naïve Bayes

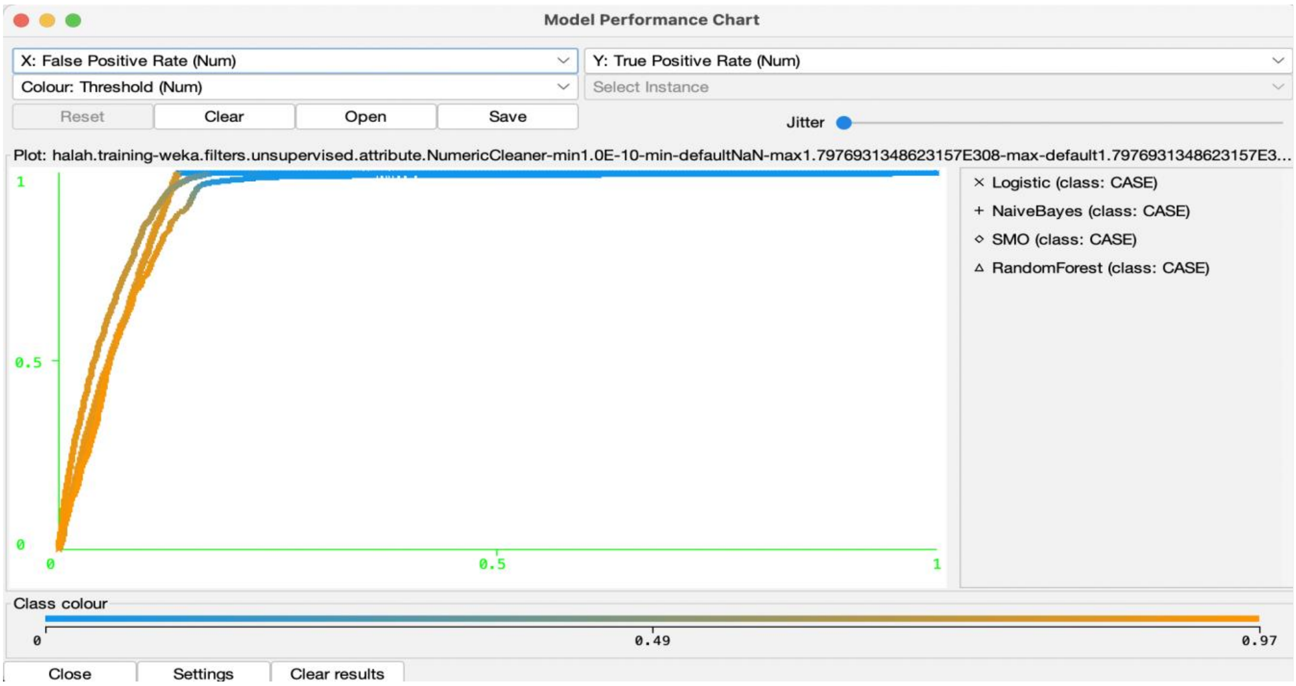
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-measure** | **MCC** | **ROC Area** | **PRC Area** | **CLASS** |
|---|---|---|---|---|---|---|---|---|
| 0.847 | 0.067 | 0.926 | 0.847 | 0.885 | 0.783 | 0.924 | 0.934 | CASE |
| 0.933 | 0.153 | 0.859 | 0.933 | 0.895 | 0.783 | 0.924 | 0.878 | CONTROL |
| 0.890 | 0.110 | 0.893 | 0.890 | 0.890 | 0.783 | 0.924 | 0.906 | AVG. |

The SMO/SVM (Sequential Minimal Optimization Support Vector Machine) model delivers a robust performance in binary classification, evident from the evaluation metrics. In the "CASE" class, it achieves a strong true positive rate (TP Rate) of 0.840, successfully identifying positive instances, and an exceptionally high precision of 0.998, highlighting an extremely low rate of false positives. The recall, standing at 0.840, indicates that it captures a substantial proportion of positive instances, leading to a remarkable F-measure of 0.912. The Matthews Correlation Coefficient (MCC) is 0.849, and the Receiver Operating Characteristic (ROC) Area is 0.919, showcasing the model's exceptional discriminative ability. In the "CONTROL" class, the model performs even better, with a recall of 0.998 and an F-measure of 0.925, emphasizing its ability to correctly classify negative instances. The average (AVG.) of these two classes is outstanding, with a high ROC Area of 0.919, illustrating the model's robust generalization ability.

Table 6: Training set on SMO

| TP Rate | FP Rate | Precision | Recall | F-measure | MCC | ROC Area | PRC Area | CLASS |
|---|---|---|---|---|---|---|---|---|
| 0.840 | 0.002 | 0.998 | 0.840 | 0.912 | 0.849 | 0.919 | 0.918 | CASE |
| 0.998 | 0.160 | 0.862 | 0.998 | 0.925 | 0.849 | 0.919 | 0.861 | CONTROL |
| 0.919 | 0.081 | 0.930 | 0.919 | 0.919 | 0.849 | 0.919 | 0.890 | AVG. |

In conclusion, based on the metrics and the focus on the "CASE" class, Logistic Regression is the most suitable model for this specific classification task. It achieves a balance between precision and recall while maintaining a competitive ROC Area, making it the best choice for effectively classifying positive instances.


Model Performance Chart

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Plot: halah.training-weka.filters.unsupervised.attribute.NumericCleaner-min1.0E-10-min-defaultNaN-max1.7976931348623157E308-max-default1.7976931348623157E3...

Figure 2: Using Knowledge Flow in Weka to build a ROC Curve

## 4.5 Model Validation

The performance of the Logistic Regression model on the test set demonstrates its robust and consistent ability to classify instances effectively. In the "CASE" class, the model achieves an impressive true positive rate (TP Rate) of 0.866, which indicates its capability to correctly classify positive instances with a recall of 0.866. Additionally, it maintains an exceptional precision of 1.000, highlighting its ability to minimize false positives, which is crucial in many real-world applications. The F-measure of 0.928 signifies a balanced trade-off between precision and recall, and the Matthews Correlation Coefficient (MCC) of 0.874 demonstrates its strong discriminative ability. The Receiver Operating Characteristic (ROC) Area of 0.947 further underlines the model's capacity to distinguish between classes. In the "CONTROL" class, the model exhibits perfect precision (1.000) and a recall of 1.000, leading to a remarkable F-measure of 0.937.

The overall average performance (AVG.) in the test set is also outstanding, with a high ROC Area of 0.947, emphasizing the model's robust generalization ability. The classifier evaluation results on the test set further support the excellence of the Logistic Regression model. With 93.2814% of instances correctly classified and a low rate of 6.7186% incorrectly classified instances, the model achieves a high level of accuracy, reflecting its reliability in real-world applications. The Kappa Statistic of 0.8656 suggests substantial agreement between the model's predictions and the actual class labels. The Relative Absolute Error of 23.675% and the Root Relative Squared Error of 48.4004% indicate that the model's predictions are close to the actual values, further emphasizing its effectiveness. These results underscore the Logistic Regression model's strong performance and make a compelling case for its selection as the best model for this classification task, both in terms of classification metrics and overall predictive accuracy.

Table 7: Test set on Logistic Regression

| TP Rate | FP Rate | Precision | Recall | F-measure | MCC | ROC Area | PRC Area | CLASS |
|---|---|---|---|---|---|---|---|---|
| 0.866 | 0.000 | 1.000 | 0.866 | 0.928 | 0.874 | 0.947 | 0.964 | CASE |
| 1.000 | 0.134 | 0.882 | 1.000 | 0.937 | 0.874 | 0.947 | 0.914 | CONTROL |
| 0.933 | 0.067 | 0.941 | 0.933 | 0.933 | 0.874 | 0.947 | 0.939 | AVG. |

Table 8: Classifier Evaluation Test Set

| | Logistic Regression |
|---|---|
| 1. Correctly classified instances | 93.2814% |
| 2. Incorrectly Classified Instances | 6.7186% |
| 3. Kappa Statistic | 0.8656 |
| 4. Relative Absolute Error | 23.675% |
| 5. Root relative squared error | 48.4004% |

## 4.6 Identification of Top Risk Factors for Prostate Cancer.

To rank our features, we used three different types of FeaturesEvaluators which are (I) CorrelationAttributeEval, (II) GainAttributeEval and (III) InfoGainAttributeEval. The search method used was Ranker for all the featuresevaluators. The results for the evaluators are as follows:

Table 9: Features ranking based on CorrelationAttributeEval

| Average Merit | Average Rank | Features |
|---|---|---|
| 0.613 +- 0.001 | 1 +- 0 | 29 x.sex |
| 0.449 +- 0.001 | 2 +- 0 | 30 x.age.g |
| 0.326 +- 0.001 | 3 +- 0 | 12 psatest1 |
| 0.203 +- 0.002 | 4 +- 0 | 4 employ1 |
| 0.082 +- 0.004 | 5 +- 0 | 2 cvdinfr4 |
| 0.07 +- 0.002 | 6.2 +- 0.4 | 8 smoke100 |
| 0.067 +- 0.003 | 6.8 +- 0.4 | 16 prediab1 |
| 0.048 +- 0.002 | 8.7 +- 1 | 15 pdiabtst |
| 0.045 +- 0.003 | 9.3 +- 0.78 | 25 acepunch |
| 0.046 +- 0.003 | 9.6 +- 1.2 | 3 addepev3 |

| | | |
|---|---|---|
| 0.044 +- 0.002 | 10.4 +- 0.66 | 14 hivtst7 |
| 0.039 +- 0.002 | 12 +- 0 | 6 diffwalk |
| 0.031 +- 0.002 | 13.4 +- 0.49 | 7 diffdres |
| 0.029 +- 0.003 | 13.8 +- 0.87 | 27 aceswear |
| 0.022 +- 0.002 | 16.5 +- 1.28 | 18 lcslast |
| 0.022 +- 0.003 | 16.6 +- 1.74 | 11 avedrnk3 |
| 0.022 +- 0.002 | 16.9 +- 1.37 | 13 psatime |
| 0.019 +- 0.002 | 19 +- 2.24 | 5 decide |
| 0.018 +- 0.003 | 19.6 +- 2.46 | 22 acedrugs |
| 0.018 +- 0.004 | 19.7 +- 2.9 | 20 acedeprs |
| 0.017 +- 0.003 | 20.9 +- 2.62 | 19 lcsnumcg |
| 0.017 +- 0.002 | 21.1 +- 1.87 | 9 usenow3 |
| 0.015 +- 0.002 | 22.9 +- 2.59 | 28 x.urbstat |
| 0.015 +- 0.002 | 23 +- 1.48 | 23 aceprisn |
| 0.012 +- 0.002 | 24.7 +- 1.62 | 10 alcday5 |
| 0.01 +- 0.002 | 25.9 +- 0.94 | 24 acedivrc |
| 0.009 +- 0.003 | 27.1 +- 1.87 | 26 acehurt1 |
| 0.007 +- 0.003 | 27.9 +- 1.22 | 21 acedrink |
| 0.006 +- 0.002 | 28.4 +- 1.02 | 1 exerany2 |
| 0.002 +- 0.002 | 29.6 +- 0.8 | 17 insulin1 |

Table 10: Features ranking based on GainAttributeEval:

| **Average Merit** | **Average Rank** | **Features** |
|---|---|---|
| 0.411 +- 0.001 | 1 +- 0 | 29 x.sex |
| 0.262 +- 0.001 | 2 +- 0 | 12 psatest1 |
| 0.093 +- 0.001 | 3 +- 0 | 30 x.age.g |
| 0.046 +- 0.001 | 4 +- 0 | 4 employ1 |
| 0.03 +- 0.001 | 5.7 +- 0.64 | 2 cvdinfr4 |
| 0.007 +- 0.001 | 7.4 +- 0.8 | 3 addepev3 |
| 0.007 +- 0 | 7.5 +- 0.67 | 14 hivtst7 |
| 0.028 +- 0.013 | 7.6 +- 6.2 | 9 usenow3 |
| 0.006 +- 0 | 8.9 +- 0.7 | 8 smoke100 |
| 0.005 +- 0 | 10.2 +- 0.6 | 15 pdiabtst |
| 0.004 +- 0 | 10.8 +- 0.6 | 16 prediab1 |
| 0.004 +- 0 | 12.5 +- 0.92 | 6 diffwalk |
| 0.003 +- 0 | 12.9 +- 0.54 | 10 alcday5 |
| 0.003 +- 0 | 13.6 +- 0.92 | 5 decide |
| 0.003 +- 0 | 15.2 +- 0.6 | 23 aceprisn |
| 0.002 +- 0 | 16.7 +- 1 | 22 acedrugs |
| 0.002 +- 0 | 17.4 +- 1.02 | 24 acedivrc |
| 0.002 +- 0 | 17.5 +- 0.81 | 20 acedeprs |
| 0.001 +- 0 | 19.3 +- 0.64 | 25 acepunch |
| 0.001 +- 0 | 20.3 +- 0.64 | 11 avedrnk3 |
| 0.004 +- 0.011 | 21.3 +- 6.9 | 7 diffdres |
| 0 +- 0 | 21.7 +- 0.9 | 27 aceswear |

|  |  |  |
|---|---|---|
| 0 +- 0 | 21.9 +- 0.7 | 13 psatime |
| 0 +- 0 | 24 +- 1 | 26 acehurt1 |
| 0 +- 0 | 24.5 +- 1.02 | 28 x.urbstat |
| 0 +- 0 | 25.6 +- 2.11 | 21 acedrink |
| 0 +- 0 | 27.3 +- 1.1 | 17 insulin1 |
| 0 +- 0 | 27.4 +- 1.36 | 18 lcslast |
| 0 +- 0 | 27.8 +- 1.6 | 19 lcsnumcg |
| 0 +- 0 | 30 +- 0 | 1 exerany2 |

Table 11: Features ranking based on InfoGainAttributeEval:

| **Average merit** | **Average rank** | **Features** |
|---|---|---|
| 0.347 +- 0.001 | 1 +- 0 | 29 x.sex |
| 0.18 +- 0.001 | 2 +- 0 | 30 x.age.g |
| 0.149 +- 0.001 | 3 +- 0 | 12 psatest1 |
| 0.095 +- 0.001 | 4 +- 0 | 4 employ1 |
| 0.013 +- 0.001 | 5 +- 0 | 2 cvdinfr4 |
| 0.007 +- 0 | 6.1 +- 0.3 | 10 alcday5 |
| 0.007 +- 0 | 7 +- 0.45 | 14 hivtst7 |
| 0.006 +- 0 | 8.4 +- 0.8 | 8 smoke100 |
| 0.005 +- 0 | 9.1 +- 0.7 | 15 pdiabtst |
| 0.005 +- 0 | 9.4 +- 0.8 | 3 addepev3 |
| 0.003 +- 0 | 11 +- 0 | 16 prediab1 |
| 0.002 +- 0 | 12 +- 0 | 6 diffwalk |
| 0.002 +- 0 | 13.1 +- 0.3 | 24 acedivrc |
| 0.002 +- 0 | 14 +- 0.45 | 5 decide |
| 0.001 +- 0 | 15.5 +- 1.2 | 11 avedrnk3 |
| 0.001 +- 0 | 16 +- 0.45 | 20 acedeprs |
| 0.001 +- 0 | 17.5 +- 0.92 | 25 acepunch |
| 0.001 +- 0 | 18.2 +- 0.87 | 23 aceprisn |
| 0.001 +- 0 | 18.7 +- 0.9 | 22 acedrugs |
| 0.001 +- 0 | 20.3 +- 2.72 | 9 usenow3 |
| 0 +- 0 | 21.7 +- 0.9 | 27 aceswear |
| 0 +- 0 | 21.8 +- 0.87 | 13 psatime |
| 0 +- 0 | 23.1 +- 3.65 | 7 diffdres |
| 0 +- 0 | 24.6 +- 1.36 | 28 x.urbstat |
| 0 +- 0 | 25.1 +- 2.02 | 21 acedrink |
| 0 +- 0 | 26.3 +- 1.35 | 19 lcsnumcg |
| 0 +- 0 | 26.6 +- 1.69 | 26 acehurt1 |
| 0 +- 0 | 26.7 +- 1.73 | 18 lcslast |
| 0 +- 0 | 27.8 +- 1.33 | 17 insulin1 |
| 0 +- 0 | 30 +- 0 | 1 exerany2 |

Table 12: Top 10 features based on ranking from three evaluators:

| Index | CorrelationAttributeEval | GainAttributeEval | InfoGainAttributeEval |
|---|---|---|---|
| 1 | 29 x.sex | 29 x.sex | 29 x.sex |
| 2 | 30 x.age.g | 12 psatest1 | 30 x.age.g |
| 3 | 12 psatest1 | 30 x.age.g | 12 psatest1 |
| 4 | 4 employ1 | 4 employ1 | 4 employ1 |
| 5 | 2 cvdinfr4 | 2 cvdinfr4 | 2 cvdinfr4 |
| 6 | 8 smoke100 | 3 addepev3 | 10 alcday5 |
| 7 | 16 prediab1 | 14 hivtst7 | 14 hivtst7 |
| 8 | 15 pdiabtst | 9 usenow3 | 8 smoke100 |
| 9 | 25 acepunch | 8 smoke100 | 15 pdiabtst |
| 10 | 3 addepev3 | 15 pdiabtst | 3 addepev3 |

The feature ranking experiment reveals variations in the importance of features as risk factors for prostate cancer. In terms of common features, there is a degree of consistency across the methods, with some features appearing in the top 10 across all three evaluations. These common features include "x.sex" (gender), "x.age.g" (age groups), "psatest1" (PSA testing), and "employ1" (employment status). These features consistently show a strong association with prostate cancer risk. In terms of differences in ranking, while some features are consistently ranked highly across methods, the exact ranking order varies. For instance, "x.sex" consistently ranks as the most influential featuresin all three methods. However, the ranking of "x.age.g" and "psatest1" differs between CorrelationAttributeEval, GainAttributeEval, and InfoGainAttributeEval, indicating variations in the perceived importance of age and PSA testing. Additionally, each method introduces unique features into the top 10. For CorrelationAttributeEval, "smoke100" (smoking) and "prediab1" (pre-diabetic conditions) are among the top features. In GainAttributeEval, "alcday5" (alcohol consumption) and "hivtst7" (HIV testing) make the top 10. In InfoGainAttributeEval, "alcday5" also appears, along with "addepev3" (psychological or emotional factors). Furthermore, each method introduces unique features into the top 10. For CorrelationAttributeEval, "smoke100" (smoking) and "prediab1" (pre-diabetic conditions) are among the top features. In GainAttributeEval, "alcday5" (alcohol consumption) and "hivtst7" (HIV testing) make the top 10. In InfoGainAttributeEval, "alcday5" also appears, along with "addepev3" (psychological or emotional factors).

The results of the logistic regression model trained after removing the top 10 features (sex, age, psatest1, employ1, cvdinfr4, smoke100, hivtst7, addepev3, pdiabtst, and prediab1) indicate a notable change in model performance. In terms of precision for both "CASE" and "CONTROL" is relatively low at 55.5% and 54.5%, respectively, indicating that the model has a high rate of false positives in both classes. The recall is also moderate for both classes, with values of 49.8% for "CASE" and 60% for "CONTROL.". The F-measure is 52.5% for "CASE" and 57.1% for "CONTROL," suggesting an average balance between precision and recall. The ROC and PRC (Precision-Recall Curve) areas both have values of 0.570, which are mediocre and imply that the model's discrimination ability is modest at best.

In terms of error rate, the relative absolute error rate is extremely high at 99.0877%, highlighting a substantial level of inaccuracy in the model's predictions and the root relative squared error is also very high at 99.5967%, indicating that the model's predictions differ significantly from the actual values. This experiment showed that the removal of the top 10 features has led to a significant deterioration in the logistic regression model's performance. The reduced true positive rate and precision, along with a higher false positive rate, suggesting that the model's ability to distinguish between cases and controls has been compromised. The MCC and Kappa statistics both indicate very weak model performance, with limited agreement between predictions and actual outcomes. Based on the comparison, we selected the top 10 features to be x.sex (gender), x.age (gender), psatest1, employ1, cvdinfr4, smoke100, hivtst7, addepev3, pdiabtst and prediab1.

Table 13: Training Set excluding top 10 features on Logistic Regression:

| TP Rate | FP Rate | Precision | Recall | F-measure | MCC | ROC Area | PRC Area | CLASS |
|---|---|---|---|---|---|---|---|---|
| 0.498 | 0.400 | 0.555 | 0.498 | 0.525 | 0.099 | 0.570 | 0.547 | CASE |
| 0.600 | 0.502 | 0.545 | 0.600 | 0.571 | 0.099 | 0.570 | 0.561 | CONTROL |
| 0.549 | 0.451 | 0.550 | 0.549 | 0.548 | 0.099 | 0.570 | 0.554 | AVG. |

Table 14: Classifier Evaluation on training set removing top 10 features:

| | Logistic Regression |
|---|---|
| 1. Correctly classified instances | 54.9011% |
| 2. Incorrectly Classified Instances | 45.0989% |
| 3. Kappa Statistic | 0.098 |
| 4. Relative Absolute Error | 99.0877% |
| 5. Root relative squared error | 99.5967% |

The results of the logistic regression model after removing the bottom 10 features demonstrate a strikingly different outcome compared to the previous experiment. The precision is high for both "CASE" and "CONTROL" at 100% and 88%, respectively, indicating that the model has significantly reduced false positives for both classes. The recall, or the model's ability to correctly identify true positives, is also high for both classes, with values of 86.3% for "CASE" and 100% for "CONTROL". The F-measure, which combines precision and recall, is also high, suggesting an excellent balance between precision and recall. Both the ROC and PRC areas have values of 0.936, indicating a marked improvement in the model's discrimination ability. In terms of the error rate, the relative absolute error has decreased to 24.0363%, signifying a significant reduction in the inaccuracy of the model's predictions while the root relative squared error is also lower at 49.0975%, indicating that the model's predictions now align more closely with the actual values. The Kappa statistic indicates a high level of agreement between the model's predictions and actual values, and both relative absolute error and root relative squared error have substantially decreased, signifying a considerable improvement in the model's predictive accuracy.

These results suggest that the bottom 10 features removed in this experiment were less important for the logistic regression model's predictive power. The true positive rate, precision, recall, F-measure, MCC, and areas under the ROC and PRC curves have all significantly increased. In this context, it appears that the initial model may have contained redundant or less informative features that were affecting its performance negatively. By removing these less important features, the model has been streamlined, resulting in an improved prostate cancer risk assessment model.

Table 15: Bottom 10 features based on ranking from three evaluators:

| **CorrelationAttributeEval** | **GainAttributeEval** | **InfoGainAttributeEval** |
|---|---|---|
| 19 lcsnumcg | 7 diffdres | 27 aceswear |
| 9 usenow3 | 27 aceswear | 13 psatime |
| 28 x.urbstat | 13 psatime | 7 diffdres |
| 23 aceprisn | 26 acehurt1 | 28 x.urbstat |
| 10 alcday5 | 28 x.urbstat | 21 acedrink |
| 24 acedivrc | 21 acedrink | 19 lcsnumcg |
| 26 acehurt1 | 17 insulin1 | 26 acehurt1 |
| 21 acedrink | 18 lcslast | 18 lcslast |
| 1 exerany2 | 19 lcsnumcg | 17 insulin1 |
| 17 insulin1 | 1 exerany2 | 1 exerany2 |

Table 16: Training set excluding bottom 10 features on Logistic Regression:

| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-measure** | **MCC** | **ROC Area** | **PRC Area** | **CLASS** |
|---|---|---|---|---|---|---|---|---|
| 0.863 | 0.000 | 1.000 | 0.863 | 0.926 | 0.871 | 0.936 | 0.959 | CASE |
| 1.000 | 0.137 | 0.880 | 1.000 | 0.936 | 0.871 | 0.936 | 0.886 | CONTROL |
| 0.932 | 0.068 | 0.940 | 0.932 | 0.931 | 0.871 | 0.936 | 0.922 | AVG. |

Table 17: Classifier Evaluation on training set removing bottom 10 features:

| | **Logistic Regression** |
|---|---|
| 1. Correctly classified instances | 93.1523% |
| 2. Incorrectly Classified Instances | 6.8477% |
| 3. Kappa Statistic | 0.863 |
| 4. Relative Absolute Error | 24.0363% |
| 5. Root relative squared error | 49.0975% |

# CHAPTER 5: DISCUSSION

## 5.1 Algorithm Selection for Prediction of Prostate Cancer

Among the four models, Logistic Regression consistently demonstrates strong performance in the "CASE" class, with a high True Positive Rate (TP Rate) of 0.863, precision of 1.000, and an F-measure of 0.926, indicating its ability to correctly classify positive instances with high precision while maintaining a balanced trade-off between precision and recall. Moreover, its ROC Area of 0.938 is also competitive. While Random Forest performs well with a high ROC Area of 0.950, its precision and F-measure in the "CASE" class are slightly lower compared to Logistic Regression. Naïve Bayes and SMO/SVM also exhibit good performance, but their precision, recall, and F-measure in the "CASE" class are not as strong as Logistic Regression. (Faradmal et al., 2014; Witteveen et al., 2018).

The reasons why Logistic regression works the best could be due to several reasons. First is in terms of linear separability. Logistic regression is inherently designed to handle linear relationships between features and the target variable. If the relationships in the dataset are primarily linear, logistic regression can capture them effectively. In cases where the decision boundary between classes is relatively linear, logistic regression is known to perform well. It is particularly well-suited to our small dataset, which contains highly informative features that have a strong linear relationship with the class variable. The ability to identify and leverage the importance of individual features could explain the strong performance of this algorithm compared to the rest. For example, in our datasets, we have features which have a linear relationship to prostate cancer. (Faradmal et al., 2014; Witteveen et al., 2018).

Furthermore, in terms of balancing trade-off between precision and recall, it is one of the key strengths of logistic regression. In binary classification problems where both false positives and false negatives are costly, logistic regression's ability to fine-tune this trade-off is beneficial. It might have achieved a better balance compared to the other models.

Additionally, Logistic regression is less prone to overfitting, especially when the dataset is not very large. If the dataset size is moderate and the data distribution is relatively stable, logistic regression's simplicity and regularization can help prevent overfitting and improve generalization. In our dataset, we only used one dataset from BRFSS i.e., Year 2020 instead of combining from different dataset from several years.

## 5.2 Top Risk Factors for Prostate Cancer:

One of the objectives of this study is to identify the top features (questions) that can act as a risk factor for prostate cancer. We identified that sex, age, psatest1, employ1, cvdinfr4, smoke100, hivtst7, addepev3, pdiabtst and prediab1 can be an excellent risk factors for prostate cancer. Despite using three different evaluation methods to rank the features, the features are pretty much standard throughout. The differences in the rankings between these methods arise from their distinct methodologies and objectives. For example, while CorrelationAttributeEval primarily captures linear associations, GainAttributeEval and InfoGainAttributeEval offer more flexibility by considering non-linear relationships and interactions. These methods are particularly useful when complex, non-linear, or multivariate relationships exist within the dataset.

When it comes to defining the top 10 risk factors and their contribution in prostate cancer development, age is ranked as one of the top risk factors with 67% of the " CASE" fall under the category above the age range of 65, whereas, more than 50% of "CONTROL" fall below the age of 40 proving that age is a very important risk factor. Additionally, employment is another important factor in the prediction of prostate cancer as it determines the socioeconomic status of a person. Retired individuals are mostly above the age of 60 and are expected to live on pension or personal funding. Moreover, men employed on wages fall under low-income bracket proving association between low-income and development of prostate cancer. This is due to lack of awareness or resources available for screening of prostate cancer and age is a crucial factor in the case of retirees. (Coughlin, 2020; Stangelberger et al., 2008).

Furthermore, PSA test1 was also amongst the top features or risk factors which proves the importance of this test. Although PSA test doesn’t confirm the diagnosis and is used just as a screening test, high serum PSA levels are linked to detection of prostate cancer. It is important to note that PSA test doesn’t distinguish between individuals that do or do not have prostate cancer and high serum PSA levels are also linked to other malignancies. Nevertheless, this doesn’t diminish the importance of PSA as a screening test for the disease. (Schröder et al., 2009; Vickers et al., 2010).

# CHAPTER 6: Standard Operating Procedure (SOP) of Performing Machine Learning for Prostate Cancer using WEKA:

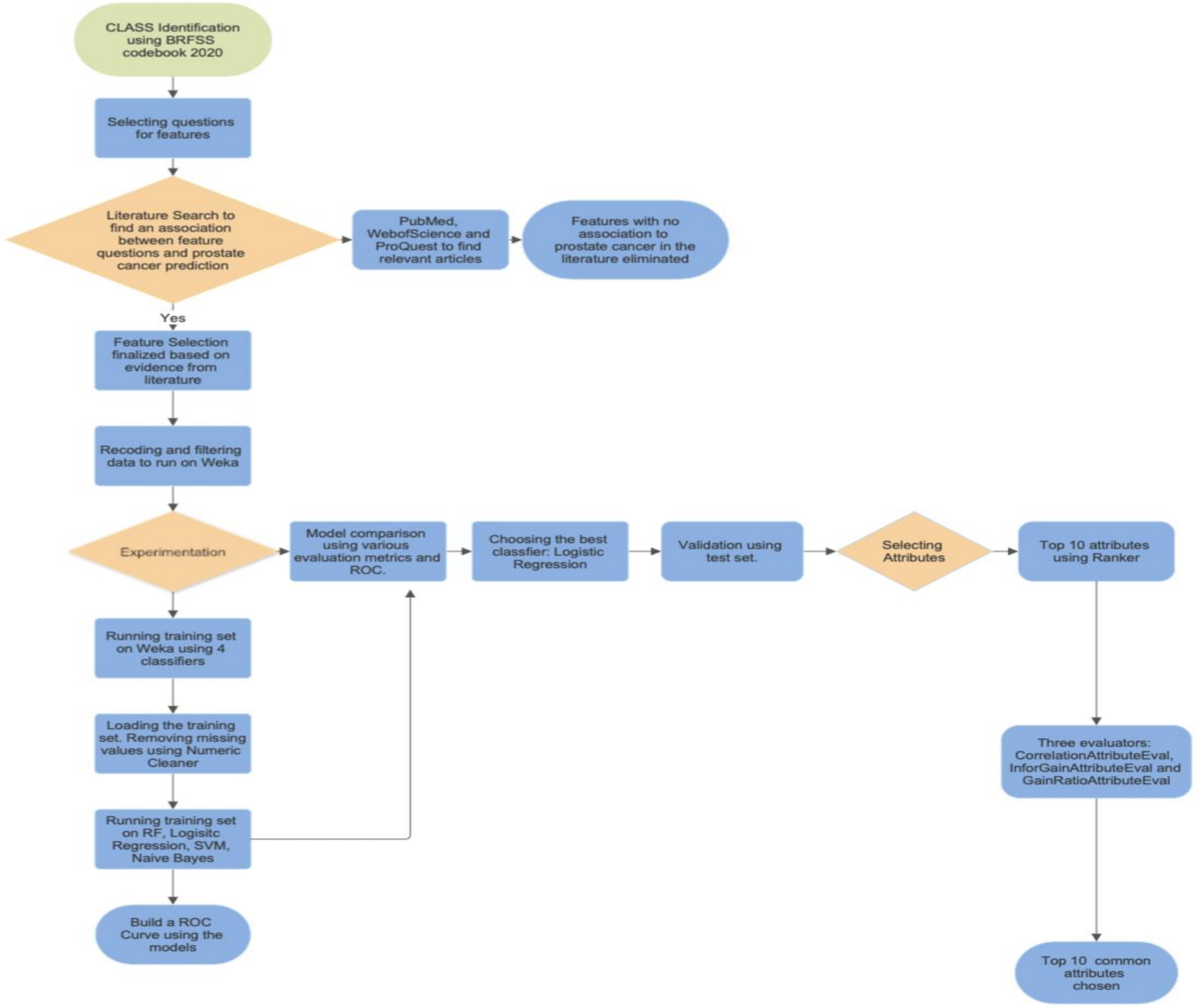

CLASS Identification using BRFSS codebook 2020
Selecting questions for features
Literature Search to find an association between feature questions and prostate cancer prediction
PubMed, WebofScience and ProQuest to find relevant articles
Features with no association to prostate cancer in the literature eliminated
Yes
Feature Selection finalized based on evidence from literature
Recoding and filtering data to run on Weka
Experimentation
Model comparison using various evaluation metrics and ROC.
Choosing the best classfier: Logistic Regression
Validation using test set.
Selecting Attributes
Top 10 attributes using Ranker
Running training set on Weka using 4 classifiers
Loading the training set. Removing missing values using Numeric Cleaner
Running training set on RF, Logisitc Regression, SVM, Naive Bayes
Build a ROC Curve using the models
Three evaluators: CorrelationAttributeEval, InforGainAttributeEval and GainRatioAttributeEval
Top 10 common attributes chosen

Figure 3: Machine Learning for Prostate Cancer Workflow

### 6.1. Class Identification:

In WEKA, the target or class variable must be specified, which in this case is likely whether an individual has prostate cancer or not. Load the dataset into WEKA and specify the class. The CLASS variable should be the last column in the dataset and must be binary. In this case, the CLASS variable consisted of CASE and CONTROL.

### 6.2. Features Identification and Selection:

Use WEKA's feature selection and features evaluation methods to identify relevant features. You can explore different feature selection techniques, such as information gain or correlation-based feature selection, to choose the most relevant features for your prediction task. The feature selection must be based on relevance to the class that has been selected. This can be achieved by carrying out thorough literature review on feature questions and their significance in association with the class. For example, association of diabetes in development of prostate cancer.

### 6.3. Data Filtering and Cleaning:

Preprocess the dataset to handle missing values, outliers, and noisy data. WEKA provides various preprocessing options, including imputation, removing outliers, and filtering. The missing values can be removed by going to the Preprocess tab on WEKA and clicking “choose” under Filter. Then click on the dropdown menu on ‘filters’ and then ‘unsupervised’. Select Numeric Cleaner on WEKA and choose the min and max Threshold value and then click on ‘apply’. Down sampling the data can help solve the issue of class imbalance and result in an improved performance of classifiers. It also reduces computation time and leads to faster computation and training time especially when dealing with large datasets. Finally, down sampling improves model performance and prevents the model from being biased towards majority class.

On the contrary, some cons of down sampling in WEKA includes loss of information resulting in the dataset not fully representing the complexity of the original data. It can also lead to loss of valuable information and the sampling strategy can yield different results; therefore, it is crucial to choose the most efficient approach based on a specific dataset.

### 6.4. Select Test Strategy:

Decide on your evaluation strategy, whether you want to use a simple split percentage (e.g., 70/30) for training and testing or cross-validation. Cross-validation is a robust approach that helps in obtaining a more reliable model. It splits the data into multiple subsets depending on the number of “folds” chosen. It repeatedly trains and tests the model on various combinations of these folds and then provides an estimate of the model’s performance by taking an average of results across all folds. Cross-validation is useful in handling limited amounts of data. This experiment used a 10-fold cross-validation. Moreover, percentage split is another technique which divides the data into a training and test set based on the percentage specified. It is best suited for a large dataset. The choice between the two methods depends on the size of dataset, complexity of the model and specific goals of analysis.

### 6.5. Training the Model using Classification Algorithm:

Once the training set has been finalized using a certain number of features and classes, various algorithms such as Random Forest, Logistic Regression, SVM and Naïve Bayes can be trained. This can be done by going to the classify tab in Weka and choosing the relevant classifier by clicking on the dropdown menu of various classifiers. Click on the START button to run the experiment for each classifier.

### 6.6. Comparing the Best Model using Classifier Evaluator:

The performance of classifiers can be compared using the confusion matrix. The best evaluation metrics that must be considered are Precision, Recall, F-measure, and ROC curve. The closer their value is to 1, the more accurate the model’s prediction is. After the model on each classifier has been built using training set, a ROC curve for all the classifiers must be generated using Knowledge Flow on Weka. Comparing the ROC curve will help to identify the best classifier. Other metrics that must be considered include kappa statistic, relative absolute error, and root relative squared error.

### 6.7. Validating the Model with Unseen Data:

Once the best classifier has been identified based on comparison between all the evaluation metrics, the model can be validated on the test set. Logistic Regression proved to be the best model for prostate cancer prediction. Therefore, the test set was applied to the model and the result comparison was made between the training and test set.

### 6.8. Ranking the Features in terms of Contribution to the Model:

After training the models, you can use feature ranking techniques in WEKA to determine which features contributed the most to the model's predictions. This can help identify the most important factors that contribute to prostate cancer. The search method used for our dataset was Ranker and the three evaluators used were CorrelationAttributeEval, GainRatioAttributeEval and InfoGainAttributeEval. Based on these search evaluators, the top ten features were identified.

# CHAPTER 7: CHALLENGES AND LIMITATIONS

## 7.1 Challenges:

Some of the challenges faced were to justify class and features and to select relevant feature questions from the codebook that will help in prostate cancer prediction. Initially, our feature and class selection were done on codebooks from 2018, 2019 and 2020, however, due to lack of common features in each of these codebooks, the final feature selection was limited to BRFSS codebook 2020. Selecting feature class from the codebooks was also one of the challenges as the codebooks are composed of general questions related to patient health, lifestyle, and dietary habits. Each feature question could greatly impact our machine learning model, therefore, thorough research into the risk factors associated with prostate cancer was done before finalizing the feature class questions. Furthermore, recoding of data and converting into binary options was also critical as it could greatly impact the machine learning. “CASE” and “CONTROL” had to be very carefully chosen based on the individual answer. Selecting feature and class required the most time and effort due to their crucial role in model training.

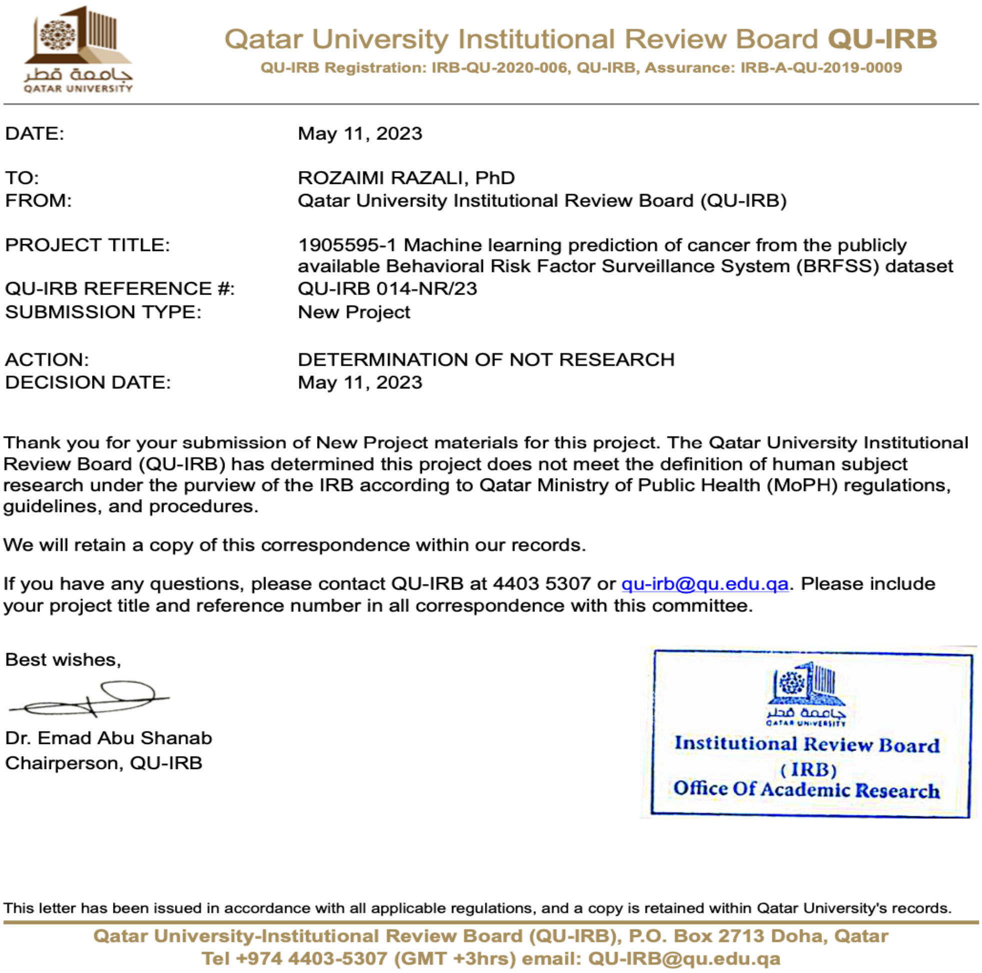
## 7.2 Limitations:

Some of the limitations of the study include the imbalance between case and control as the number of controls was very large compared to the case and to overcome this imbalance, down sampling was done. Due to this reason the analysis of our data was only limited to classification algorithms. Lastly, the data source used which is the BRFSS codebook 2020 was taken from USA as there was no such publicly available healthcare surveillance information in Qatar. Other BRFSS codebooks were not used due to lack of common questions present in each codebook.

# CHAPTER 8: CONCLUSION and FUTURE DIRECTION

In conclusion, this study assessed the ability of a machine learning model to predict prostate cancer using a publicly available healthcare survey. Machine learning in prediction of cancer is much less invasive compared to current procedures and it can help identify early onset of cancer. Secondly, doing a healthcare survey is much more convenient as it is faced with much less resistance as opposed to doing a screening test for prostate cancer. This model proved effective in predicting the top features that might contribute to prostate cancer as well in identifying important questions that can be implemented by local healthcare providers such as HMC for prediction of prostate cancer in the local population. For the future direction of this study, machine learning can be implemented on individual risk factors and their association with the severity of the disease can be studied. Machine learning is an important technique for studying various chronic illnesses and this model can be implemented in the study of other types of cancer from locally and internationally available dataset.

# APPENDIX: Ethical Approval



جامعة قطر
QATAR UNIVERSITY

Qatar University Institutional Review Board **QU-IRB**

**QU-IRB Registration: IRB-QU-2020-006, QU-IRB, Assurance: IRB-A-QU-2019-0009**

|  |  |
|---|---|
| DATE: | May 11, 2023 |
| TO: | ROZAIMI RAZALI, PhD |
| FROM: | Qatar University Institutional Review Board (QU-IRB) |
| PROJECT TITLE: | 1905595-1 Machine learning prediction of cancer from the publicly available Behavioral Risk Factor Surveillance System (BRFSS) dataset |
| QU-IRB REFERENCE #: | QU-IRB 014-NR/23 |
| SUBMISSION TYPE: | New Project |
| ACTION: | DETERMINATION OF NOT RESEARCH |
| DECISION DATE: | May 11, 2023 |

Thank you for your submission of New Project materials for this project. The Qatar University Institutional Review Board (QU-IRB) has determined this project does not meet the definition of human subject research under the purview of the IRB according to Qatar Ministry of Public Health (MoPH) regulations, guidelines, and procedures.

We will retain a copy of this correspondence within our records.

If you have any questions, please contact QU-IRB at 4403 5307 or qu-irb@qu.edu.qa. Please include your project title and reference number in all correspondence with this committee.

Best wishes,

Dr. Emad Abu Shanab
Chairperson, QU-IRB

جامعة قطر
QATAR UNIVERSITY
Institutional Review Board
( IRB)
Office Of Academic Research

This letter has been issued in accordance with all applicable regulations, and a copy is retained within Qatar University's records.

**Qatar University-Institutional Review Board (QU-IRB), P.O. Box 2713 Doha, Qatar**
**Tel +974 4403-5307 (GMT +3hrs) email: QU-IRB@qu.edu.qa**

# REFERENCES

Alcalá, H. E., Mitchell, E. M., & Keim-Malpass, J. (2018). Heterogeneous impacts: adverse childhood experiences and cancer screening. *Cancer Causes Control*, *29*(3), 343-351. https://doi.org/10.1007/s10552-018-1007-2

Applegate, C. C., Rowles, J. L., Ranard, K. M., Jeon, S., & Erdman, J. W. (2018). Soy Consumption and the Risk of Prostate Cancer: An Updated Systematic Review and Meta-Analysis. *Nutrients*, *10*(1). https://doi.org/10.3390/nu10010040

Attard, G., Parker, C., Eeles, R. A., Schröder, F., Tomlins, S. A., Tannock, I., Drake, C. G., & de Bono, J. S. (2016). Prostate cancer. *The Lancet*, *387*(10013), 70-82. https://doi.org/https://doi.org/10.1016/S0140-6736(14)61947-4

Barone, B. B., Yeh, H. C., Snyder, C. F., Peairs, K. S., Stein, K. B., Derr, R. L., Wolff, A. C., & Brancati, F. L. (2008). Long-term all-cause mortality in cancer patients with preexisting diabetes mellitus: a systematic review and meta-analysis. *Jama*, *300*(23), 2754-2764. https://doi.org/10.1001/jama.2008.824

Benafif, S., & Eeles, R. (2016). Genetic predisposition to prostate cancer. *Br Med Bull*, *120*(1), 75-89. https://doi.org/10.1093/bmb/ldw039

Biggar, R. J., Kirby, K. A., Atkinson, J., McNeel, T. S., & Engels, E. (2004). Cancer risk in elderly persons with HIV/AIDS. *J Acquir Immune Defic Syndr*, *36*(3), 861-868. https://doi.org/10.1097/00126334-200407010-00014

Bini, S. A. (2018). Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? *J Arthroplasty*, *33*(8), 2358-2361. https://doi.org/10.1016/j.arth.2018.02.067

Bratt, O., Drevin, L., Akre, O., Garmo, H., & Stattin, P. (2016). Family History and Probability of Prostate Cancer, Differentiated by Risk Category: A Nationwide Population-Based Study. *J Natl Cancer Inst*, *108*(10). https://doi.org/10.1093/jnci/djw110

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, *68*(6), 394-424. https://doi.org/10.3322/caac.21492

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. *Biometrics*, *40*, 874.

Brookman-May, S. D., Campi, R., Henríquez, J. D. S., Klatte, T., Langenhuijsen, J. F., Brausi, M., Linares-Espinós, E., Volpe, A., Marszalek, M., Akdogan, B., Roll, C., Stief, C. G., Rodriguez-Faba, O., & Minervini, A. (2019). Latest Evidence on the Impact of Smoking, Sports, and Sexual Activity as Modifiable Lifestyle Risk Factors for Prostate Cancer Incidence, Recurrence, and Progression: A Systematic Review of the Literature by the European Association of Urology Section of Oncological Urology (ESOU). *Eur Urol Focus*, *5*(5), 756-787. https://doi.org/10.1016/j.euf.2018.02.007

Bruner, D. W., Moore, D., Parlanti, A., Dorgan, J., & Engstrom, P. (2003). Relative risk of prostate cancer for men with affected relatives: systematic review and meta-analysis. *Int J Cancer*, *107*(5), 797-803. https://doi.org/10.1002/ijc.11466

Brunner, C., Davies, N. M., Martin, R. M., Eeles, R., Easton, D., Kote-Jarai, Z., Al Olama, A. A., Benlloch, S., Muir, K., Giles, G., Wiklund, F., Gronberg, H., Haiman, C. A., Schleutker, J., Nordestgaard, B. G., Travis, R. C., Neal, D., Donovan, J., Hamdy, F. C., . . . Zuccolo, L. (2017). Alcohol consumption and prostate cancer incidence and progression: A Mendelian randomisation study. *Int J Cancer*, *140*(1), 75-85. https://doi.org/10.1002/ijc.30436

Buskin, A., Singh, P., Lorenz, O., Robson, C., Strand, D. W., & Heer, R. (2021). A Review of Prostate Organogenesis and a Role for iPSC-Derived Prostate Organoids to Study Prostate Development and Disease. *Int J Mol Sci*, *22*(23). https://doi.org/10.3390/ijms222313097

Campi, R., Brookman-May, S. D., Subiela Henríquez, J. D., Akdoğan, B., Brausi, M., Klatte, T., Langenhuijsen, J. F., Linares-Espinos, E., Marszalek, M., Roupret, M., Stief, C. G., Volpe, A., Minervini, A., & Rodriguez-Faba, O. (2019). Impact of Metabolic Diseases, Drugs, and Dietary Factors on Prostate Cancer Risk, Recurrence, and Survival: A Systematic Review by the European Association of Urology Section of Oncological Urology. *Eur Urol Focus*, *5*(6), 1029-1057. https://doi.org/10.1016/j.euf.2018.04.001

Cannon, E. O., Amini, A., Bender, A., Sternberg, M. J., Muggleton, S. H., Glen, R. C., & Mitchell, J. B. (2007). Support vector inductive logic programming outperforms the naive Bayes classifier and inductive logic programming for the classification of bioactive chemical compounds. *J Comput Aided Mol Des*, *21*(5), 269-280. https://doi.org/10.1007/s10822-007-9113-3

Cao, Y., & Ma, J. (2011). Body mass index, prostate cancer-specific mortality, and biochemical recurrence: a systematic review and meta-analysis. *Cancer Prev Res (Phila)*, *4*(4), 486-501. https://doi.org/10.1158/1940-6207.Capr-10-0229

Caruso, R., Nanni, M. G., Riba, M., Sabato, S., Mitchell, A. J., Croce, E., & Grassi, L. (2017). Depressive spectrum disorders in cancer: prevalence, risk factors and screening for depression: a critical review. *Acta Oncol*, *56*(2), 146-155. https://doi.org/10.1080/0284186x.2016.1266090

Castillejos-Molina, R. A., & Gabilondo-Navarro, F. B. (2016). Prostate cancer. *Salud Publica Mex*, *58*(2), 279-284. https://doi.org/10.21149/spm.v58i2.7797

Chen, N., & Zhou, Q. (2016). The evolving Gleason grading system. *Chin J Cancer Res, 28*(1), 58-64. https://doi.org/10.3978/j.issn.1000-9604.2016.02.04

Coughlin, S. S. (2020). A review of social determinants of prostate cancer risk, stage, and survival. *Prostate Int*, *8*(2), 49-54. https://doi.org/10.1016/j.prnil.2019.08.001

Culp, M. B., Soerjomataram, I., Efstathiou, J. A., Bray, F., & Jemal, A. (2020). Recent Global Patterns in Prostate Cancer Incidence and Mortality Rates. *Eur Urol*, *77*(1), 38-52. https://doi.org/10.1016/j.eururo.2019.08.005

Daher, M., Telvizian, T., Dagher, C., Abdul Sater, Z., Massih, S., Chediak, A., Charafeddine, M., Shahait, M., Alameddine, R., Temraz, S., Geara, F., Youssef, B., Hajj, A., Nasr, R., Wazzan, W., Bulbul, M., Khauli, R., Shamseddin, A., & Mukherji, D. (2021). High rates of advanced prostate cancer in the Middle East: Analysis from a tertiary care center. *Urology Annals*, *13*. https://doi.org/10.4103/UA.UA_47_20

Dahlman, D., Li, X., Crump, C., Sundquist, J., & Sundquist, K. (2022). Drug use disorder and risk of incident and fatal prostate cancer among Swedish men: a nationwide epidemiological study. *Cancer Causes Control*, *33*(2), 213-222. https://doi.org/10.1007/s10552-021-01513-2

DiGiovanni, J., Kiguchi, K., Frijhoff, A., Wilker, E., Bol, D. K., Beltrán, L., Moats, S., Ramirez, A., Jorcano, J., & Conti, C. (2000). Deregulated expression of insulin-like growth factor 1 in prostate epithelium leads to neoplasia in transgenic mice. *Proc Natl Acad Sci U S A, 97*(7), 3455-3460. https://doi.org/10.1073/pnas.97.7.3455

Ellis, W. J., Chetner, M. P., Preston, S. D., & Brawer, M. K. (1994). Diagnosis of prostatic carcinoma: the yield of serum prostate specific antigen, digital rectal examination and transrectal ultrasonography. *J Urol*, *152*(5 Pt 1), 1520-1525. https://doi.org/10.1016/s0022-5347(17)32460-6

Esser, M. B., Sacks, J. J., Sherk, A., Karriker-Jaffe, K. J., Greenfield, T. K., Pierannunzi, C., & Brewer, R. D. (2020). Distribution of Drinks Consumed by U.S. Adults by Average Daily Alcohol Consumption: A Comparison of 2 Nationwide Surveys. *Am J Prev Med*, *59*(5), 669-677. https://doi.org/10.1016/j.amepre.2020.04.018

Faradmal, J., Soltanian, A. R., Roshanaei, G., Khodabakhshi, R., & Kasaeian, A. (2014). Comparison of the performance of log-logistic regression and artificial neural networks for predicting breast cancer relapse. *Asian Pac J Cancer Prev*, *15*(14), 5883-5888. https://doi.org/10.7314/apjcp.2014.15.14.5883

Farris, M. S., Kopciuk, K. A., Courneya, K. S., McGregor, S. E., Wang, Q., & Friedenreich, C. M. (2017). Associations of Postdiagnosis Physical Activity and Change from Prediagnosis Physical Activity with Quality of Life in Prostate Cancer Survivors. *Cancer Epidemiol Biomarkers Prev*, *26*(2), 179-187. https://doi.org/10.1158/1055-9965.Epi-16-0465

Feldman, D., Zhao, X. Y., & Krishnan, A. V. (2000). Vitamin D and prostate cancer. *Endocrinology*, *141*(1), 5-9. https://doi.org/10.1210/endo.141.1.7341

Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., & Bray, F. (2021). Cancer statistics for the year 2020: An overview. *Int J Cancer*. https://doi.org/10.1002/ijc.33588

Frank, E., Trigg, L., Holmes, G., & Witten, I. H. (2000). Technical Note: Naive Bayes for Regression. *Machine Learning*, *41*(1), 5-25. https://doi.org/10.1023/A:1007670802811

Friedenreich, C. M., Wang, Q., Neilson, H. K., Kopciuk, K. A., McGregor, S. E., & Courneya, K. S. (2016). Physical Activity and Survival After Prostate Cancer. *Eur Urol*, *70*(4), 576-585. https://doi.org/10.1016/j.eururo.2015.12.032

Gallagher, R. P., & Fleshner, N. (1998). Prostate cancer: 3. Individual risk factors. *Cmaj*, *159*(7), 807-813.

Gardner, J. R., Livingston, P. M., & Fraser, S. F. (2014). Effects of exercise on treatment-related adverse effects for patients with prostate cancer receiving androgen-deprivation therapy: a systematic review. *J Clin Oncol*, *32*(4), 335-346. https://doi.org/10.1200/jco.2013.49.5523

Goldenberg, S. L., Nir, G., & Salcudean, S. E. (2019). A new era: artificial intelligence and machine learning in prostate cancer. *Nature Reviews Urology*, *16*(7), 391-403. https://doi.org/10.1038/s41585-019-0193-3

Goldenberg, S. L., Nir, G., & Salcudean, S. E. (2019). A new era: artificial intelligence and machine learning in prostate cancer. *Nat Rev Urol*, *16*(7), 391-403. https://doi.org/10.1038/s41585-019-0193-3

Gong, Z., Agalliu, I., Lin, D. W., Stanford, J. L., & Kristal, A. R. (2008). Cigarette smoking and prostate cancer-specific mortality following diagnosis in middle-aged men. *Cancer Causes Control*, *19*(1), 25-31. https://doi.org/10.1007/s10552-007-9066-9

Gong, Z., Kristal, A. R., Schenk, J. M., Tangen, C. M., Goodman, P. J., & Thompson, I. M. (2009). Alcohol consumption, finasteride, and prostate cancer risk: results from the Prostate Cancer Prevention Trial. *Cancer*, *115*(16), 3661-3669. https://doi.org/10.1002/cncr.24423

Greenberg, N. M., DeMayo, F., Finegold, M. J., Medina, D., Tilley, W. D., Aspinall, J. O., Cunha, G. R., Donjacour, A. A., Matusik, R. J., & Rosen, J. M. (1995). Prostate cancer in a transgenic mouse. *Proc Natl Acad Sci U S A*, *92*(8), 3439-3443. https://doi.org/10.1073/pnas.92.8.3439

Hatcher, D., Daniels, G., Osman, I., & Lee, P. (2009). Molecular mechanisms involving prostate cancer racial disparity. *Am J Transl Res*, *1*(3), 235-248.

Hemminki, K. (2012). Familial risk and familial survival in prostate cancer. *World J Urol*, *30*(2), 143-148. https://doi.org/10.1007/s00345-011-0801-1

Hsia, J., Zhao, G., Town, M., Ren, J., Okoro, C. A., Pierannunzi, C., & Garvin, W. (2020). Comparisons of Estimates From the Behavioral Risk Factor Surveillance System and Other National Health Surveys, 2011-2016. *Am J Prev Med*, *58*(6), e181-e190. https://doi.org/10.1016/j.amepre.2020.01.025

Huang, M. H., Blackwood, J., Godoshian, M., & Pfalzer, L. (2018). Factors associated with self-reported falls, balance or walking difficulty in older survivors of breast, colorectal, lung, or prostate cancer: Results from Surveillance, Epidemiology, and End Results-Medicare Health Outcomes Survey linkage. *Plos One*, *13*(12), e0208573. https://doi.org/10.1371/journal.pone.0208573

Huncharek, M., Haddock, K. S., Reid, R., & Kupelnick, B. (2010). Smoking as a risk factor for prostate cancer: a meta-analysis of 24 prospective cohort studies. *Am J Public Health*, *100*(4), 693-701. https://doi.org/10.2105/ajph.2008.150508

Hussain, L., Ahmed, A., Saeed, S., Rathore, S., Awan, I. A., Shah, S. A., Majid, A., Idris, A., & Awan, A. A. (2018). Prostate cancer detection using machine learning techniques by employing combination of features extracting strategies. *Cancer Biomark*, *21*(2), 393-413. https://doi.org/10.3233/cbm-170643

Hussain, S. P., Hofseth, L. J., & Harris, C. C. (2003). Radical causes of cancer. *Nat Rev Cancer*, *3*(4), 276-285. https://doi.org/10.1038/nrc1046

Johnson, D. C., Raman, S. S., Mirak, S. A., Kwan, L., Bajgiran, A. M., Hsu, W., Maehara, C. K., Ahuja, P., Faiena, I., Pooli, A., Salmasi, A., Sisk, A., Felker, E. R., Lu, D. S. K., & Reiter, R. E. (2019). Detection of Individual Prostate Cancer Foci via Multiparametric Magnetic Resonance Imaging. *Eur Urol*, *75*(5), 712-720. https://doi.org/10.1016/j.eururo.2018.11.031

Kaiser, A., Haskins, C., Siddiqui, M. M., Hussain, A., & D'Adamo, C. (2019). The evolving role of diet in prostate cancer risk and progression. *Curr Opin Oncol*, *31*(3), 222-229. https://doi.org/10.1097/cco.0000000000000519

Kasivisvanathan, V., Rannikko, A. S., Borghi, M., Panebianco, V., Mynderse, L. A., Vaarala, M. H., Briganti, A., Budäus, L., Hellawell, G., Hindley, R. G., Roobol, M. J., Eggener, S., Ghei, M., Villers, A., Bladou, F., Villeirs, G. M., Virdi, J., Boxler, S., Robert, G., . . . Moore, C. M. (2018). MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis. *New England Journal of Medicine*, *378*(19), 1767-1777. https://doi.org/10.1056/NEJMoa1801993

Keilani, M., Hasenoehrl, T., Baumann, L., Ristl, R., Schwarz, M., Marhold, M., Sedghi Komandj, T., & Crevenna, R. (2017). Effects of resistance exercise in prostate cancer patients: a meta-analysis. *Support Care Cancer*, *25*(9), 2953-2968. https://doi.org/10.1007/s00520-017-3771-z

Kenfield, S. A., Stampfer, M. J., Giovannucci, E., & Chan, J. M. (2011). Physical activity and survival after prostate cancer diagnosis in the health professionals follow-up study. *J Clin Oncol*, *29*(6), 726-732. https://doi.org/10.1200/jco.2010.31.5226

Krstev, S., & Knutsson, A. (2019). Occupational Risk Factors for Prostate Cancer: A Meta-analysis. *J Cancer Prev*, *24*(2), 91-111. https://doi.org/10.15430/jcp.2019.24.2.91

Levinson, A., Nagler, E. A., & Lowe, F. C. (2005). Approach to management of clinically localized prostate cancer in patients with human immunodeficiency virus. *Urology*, *65*(1), 91-94. https://doi.org/10.1016/j.urology.2004.08.053

Liao, Q., Long, C., Deng, Z., Bi, X., & Hu, J. (2015). The role of circulating adiponectin in prostate cancer: a meta-analysis. *Int J Biol Markers*, *30*(1), e22-31. https://doi.org/10.5301/jbm.5000124

Loriot, Y., Massard, C., & Fizazi, K. (2012). Recent developments in treatments targeting castration-resistant prostate cancer bone metastases. *Ann Oncol*, *23*(5), 1085-1094. https://doi.org/10.1093/annonc/mdr573

Manfredi, R., Fulgaro, C., Sabbatani, S., Dentale, N., & Legnani, G. (2006). Disseminated, lethal prostate cancer during human immunodeficiency virus infection presenting with non-specific features. Open questions for urologists, oncologists, and infectious disease specialists. *Cancer Detect Prev*, *30*(1), 20-23. https://doi.org/10.1016/j.cdp.2005.10.002

McGregor, S. E., Courneya, K. S., Kopciuk, K. A., Tosevski, C., & Friedenreich, C. M. (2013). Case–control study of lifetime alcohol intake and prostate cancer risk. *Cancer Causes & Control*, *24*(3), 451-461. https://doi.org/10.1007/s10552-012-0131-7

Mullins, J. K., & Loeb, S. (2012). Environmental exposures and prostate cancer. *Urol Oncol*, *30*(2), 216-219. https://doi.org/10.1016/j.urolonc.2011.11.014

Nandeesha, H. (2008). Benign prostatic hyperplasia: dietary and metabolic risk factors. *Int Urol Nephrol*, *40*(3), 649-656. https://doi.org/10.1007/s11255-008-9333-z

Nandeesha, H., Koner, B. C., & Dorairajan, L. N. (2008). Altered insulin sensitivity, insulin secretion and lipid profile in non-diabetic prostate carcinoma. *Acta Physiol Hung*, *95*(1), 97-105. https://doi.org/10.1556/APhysiol.95.2008.1.7

Nigsch, F., Bender, A., Jenkins, J. L., & Mitchell, J. B. O. (2008). Ligand-Target Prediction Using Winnow and Naive Bayesian Algorithms and the Implications of Overall Performance Statistics. *Journal of Chemical Information and Modeling*, *48*(12), 2313-2325. https://doi.org/10.1021/ci800079x

Nyberg, T., Frost, D., Barrowdale, D., Evans, D. G., Bancroft, E., Adlard, J., Ahmed, M., Barwell, J., Brady, A. F., Brewer, C., Cook, J., Davidson, R., Donaldson, A., Eason, J., Gregory, H., Henderson, A., Izatt, L., Kennedy, M. J., Miller, C., . . . Antoniou, A. C.

(2020). Prostate Cancer Risks for Male BRCA1 and BRCA2 Mutation Carriers: A Prospective Cohort Study. *Eur Urol*, *77*(1), 24-35. https://doi.org/10.1016/j.eururo.2019.08.025

Nyberg, T., Govindasami, K., Leslie, G., Dadaev, T., Bancroft, E., Ni Raghallaigh, H., Brook, M. N., Hussain, N., Keating, D., Lee, A., McMahon, R., Morgan, A., Mullen, A., Osborne, A., Rageevakumar, R., Kote-Jarai, Z., Eeles, R., & Antoniou, A. C. (2019). Homeobox B13 G84E Mutation and Prostate Cancer Risk. *Eur Urol*, *75*(5), 834-845. https://doi.org/10.1016/j.eururo.2018.11.015

Okumura, M., Yamamoto, M., Sakuma, H., Kojima, T., Maruyama, T., Jamali, M., Cooper, D. R., & Yasuda, K. (2002). Leptin and high glucose stimulate cell proliferation in MCF-7 human breast cancer cells: reciprocal involvement of PKC-alpha and PPAR expression. *Biochim Biophys Acta*, *1592*(2), 107-116. https://doi.org/10.1016/s0167-4889(02)00276-8

Page, E. C., Bancroft, E. K., Brook, M. N., Assel, M., Hassan Al Battat, M., Thomas, S., Taylor, N., Chamberlain, A., Pope, J., Raghallaigh, H. N., Evans, D. G., Rothwell, J., Maehle, L., Grindedal, E. M., James, P., Mascarenhas, L., McKinley, J., Side, L., Thomas, T., . . . Eeles, R. A. (2019). Interim Results from the IMPACT Study: Evidence for Prostate-specific Antigen Screening in BRCA2 Mutation Carriers. *Eur Urol*, *76*(6), 831-842. https://doi.org/10.1016/j.eururo.2019.08.019

Panigrahi, G. K., Praharaj, P. P., Kittaka, H., Mridha, A. R., Black, O. M., Singh, R., Mercer, R., van Bokhoven, A., Torkko, K. C., Agarwal, C., Agarwal, R., Abd Elmageed, Z. Y., Yadav, H., Mishra, S. K., & Deep, G. (2019). Exosome proteomic analyses identify inflammatory phenotype and novel biomarkers in African American prostate cancer patients. *Cancer Med*, *8*(3), 1110-1123. https://doi.org/10.1002/cam4.1885

Pierannunzi, C., Town, M., Garvin, W., Shaw, F. E., & Balluz, L. (2012). Methodologic Changes in the Behavioral Risk Factor Surveillance System in 2011 and Potential Effects on Prevalence Estimates. *Morbidity and Mortality Weekly Report*, *61*, 410-413.

Rhoden, E. L., & Averbeck, M. A. (2009). [Prostate carcinoma and testosterone: risks and controversies]. *Arq Bras Endocrinol Metabol*, *53*(8), 956-962. https://doi.org/10.1590/s0004-27302009000800008 (Câncer de próstata e testosterona: riscos e controvérsias.)

Robbins, C. M., Hooker, S., Kittles, R. A., & Carpten, J. D. (2011). EphB2 SNPs and sporadic prostate cancer risk in African American men. *Plos One*, *6*(5), e19494. https://doi.org/10.1371/journal.pone.0019494

Roberts, M. J., Teloken, P., Chambers, S. K., Williams, S. G., Yaxley, J., Samaratunga, H., Frydenberg, M., & Gardiner, R. A. (2000). Prostate Cancer Detection. In K. R. Feingold, B. Anawalt, M. R. Blackman, A. Boyce, G. Chrousos, E. Corpas, W. W. de Herder, K. Dhatariya, K. Dungan, J. Hofland, S. Kalra, G. Kaltsas, N. Kapoor, C. Koch, P. Kopp, M. Korbonits, C. S. Kovacs, W. Kuohung, B. Laferrère, M. Levy, E. A. McGee, R. McLachlan, M. New, J. Purnell, R. Sahay, A. S. Shah, F. Singer, M. A. Sperling, C. A. Stratakis, D. L. Trence, & D. P. Wilson (Eds.), *Endotext*. MDText.com, Inc.

Copyright © 2000-2023, MDText.com, Inc.

Roobol, M. J., Schröder, F. H., van Leeuwen, P., Wolters, T., van den Bergh, R. C., van Leenders, G. J., & Hessels, D. (2010). Performance of the prostate cancer antigen 3 (PCA3) gene and prostate-specific antigen in prescreened men: exploring the value of PCA3 for a first-line diagnostic test. *Eur Urol*, *58*(4), 475-481. https://doi.org/10.1016/j.eururo.2010.06.039

Saigal, C. S., Gore, J. L., Krupski, T. L., Hanley, J., Schonlau, M., & Litwin, M. S. (2007). Androgen deprivation therapy increases cardiovascular morbidity in men with prostate cancer. *Cancer, 110*(7), 1493-1500. https://doi.org/10.1002/cncr.22933

Salahudeen, A. K., Kanji, V., Reckelhoff, J. F., & Schmidt, A. M. (1997). Pathogenesis of diabetic nephropathy: a radical approach. *Nephrol Dial Transplant, 12*(4), 664-668. https://doi.org/10.1093/ndt/12.4.664

Scardino, P. T. (1989). Early detection of prostate cancer. *Urol Clin North Am*, *16*(4), 635-655.

Schröder, F. H., Hugosson, J., Roobol, M. J., Tammela, T. L., Ciatto, S., Nelen, V., Kwiatkowski, M., Lujan, M., Lilja, H., Zappa, M., Denis, L. J., Recker, F., Berenguer, A., Määttänen, L., Bangma, C. H., Aus, G., Villers, A., Rebillard, X., van der Kwast, T., . . . Auvinen, A. (2009). Screening and prostate-cancer mortality in a randomized European study. *N Engl J Med, 360*(13), 1320-1328. https://doi.org/10.1056/NEJMoa0810084

Sharpley, C. F., Bitsika, V., & Christie, D. R. (2013). The incidence and causes of different subtypes of depression in prostate cancer patients: implications for cancer care. *Eur J Cancer Care (Engl)*, *22*(6), 815-823. https://doi.org/10.1111/ecc.12090

Sivaraman, A., & Bhat, K. R. S. (2017). Screening and Detection of Prostate Cancer-Review of Literature and Current Perspective. *Indian J Surg Oncol, 8*(2), 160-168. https://doi.org/10.1007/s13193-016-0584-3

Song, Y., Zhang, Y. D., Yan, X., Liu, H., Zhou, M., Hu, B., & Yang, G. (2018). Computer-aided diagnosis of prostate cancer using a deep convolutional neural network from multiparametric MRI. *J Magn Reson Imaging, 48*(6), 1570-1577. https://doi.org/10.1002/jmri.26047

Speiser, J. L., Durkalski, V. L., & Lee, W. M. (2015). Random forest classification of etiologies for an orphan disease. *Stat Med*, *34*(5), 887-899. https://doi.org/10.1002/sim.6351

Stamey, T. A., Yang, N., Hay, A. R., McNeal, J. E., Freiha, F. S., & Redwine, E. (1987). Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate. *N Engl J Med*, *317*(15), 909-916. https://doi.org/10.1056/nejm198710083171501

Stangelberger, A., Waldert, M., & Djavan, B. (2008). Prostate cancer in elderly men. *Rev Urol*, *10*(2), 111-119.

Suba, Z., & Ujpál, M. (2006). [Correlations of insulin resistance and neoplasms]. *Magy Onkol*, *50*(2), 127-135. (Az insulinresistentia és a daganat összefüggései.)

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*, *71*(3), 209-249. https://doi.org/10.3322/caac.21660

Taylor, L. G., Canfield, S. E., & Du, X. L. (2009). Review of major adverse effects of androgen-deprivation therapy in men with prostate cancer. *Cancer*, *115*(11), 2388-2399. https://doi.org/10.1002/cncr.24283

Thalgott, M., Kron, M., Brath, J. M., Ankerst, D. P., Thompson, I. M., Gschwend, J. E., & Herkommer, K. (2018). Men with family history of prostate cancer have a higher risk of disease recurrence after radical prostatectomy. *World J Urol*, *36*(2), 177-185. https://doi.org/10.1007/s00345-017-2122-5

Toivanen, R., & Shen, M. M. (2017). Prostate organogenesis: tissue induction, hormonal regulation and cell type specification. *Development*, *144*(8), 1382-1398. https://doi.org/10.1242/dev.148270

Trudeau, K., Rousseau, M. C., & Parent, M. (2020). Extent of Food Processing and Risk of Prostate Cancer: The PROtEuS Study in Montreal, Canada. *Nutrients*, *12*(3). https://doi.org/10.3390/nu12030637 van der Leest, M., Cornel, E., Israël, B., Hendriks, R., Padhani, A. R., Hoogenboom, M., Zamecnik, P., Bakker, D., Setiasti, A. Y., Veltman, J., van den Hout, H., van der Lelij, H., van Oort, I., Klaver, S., Debruyne, F., Sedelaar, M., Hannink, G., Rovers, M., Hulsbergen-van de Kaa, C., & Barentsz, J. O. (2019). Head-to-head Comparison of Transrectal Ultrasound-guided Prostate Biopsy Versus Multiparametric Prostate Resonance Imaging with Subsequent Magnetic Resonance-guided Biopsy in Biopsy-naïve Men with Elevated Prostate-specific Antigen: A Large Prospective Multicenter Clinical Study. *Eur Urol*, *75*(4), 570-578. https://doi.org/10.1016/j.eururo.2018.11.023

Vashistha, V., Singh, B., Kaur, S., Prokop, L. J., & Kaushik, D. (2016). The Effects of Exercise on Fatigue, Quality of Life, and Psychological Function for Men with Prostate Cancer: Systematic Review and Meta-analyses. *Eur Urol Focus*, *2*(3), 284-295. https://doi.org/10.1016/j.euf.2016.02.011

Vickers, A. J., Cronin, A. M., Björk, T., Manjer, J., Nilsson, P. M., Dahlin, A., Bjartell, A., Scardino, P. T., Ulmert, D., & Lilja, H. (2010). Prostate specific antigen concentration at age 60 and death or metastasis from prostate cancer: case-control study. *Bmj*, *341*, c4521. https://doi.org/10.1136/bmj.c4521

Watts, S., Leydon, G., Birch, B., Prescott, P., Lai, L., Eardley, S., & Lewith, G. (2014). Depression and anxiety in prostate cancer: a systematic review and meta-analysis of prevalence rates. *Bmj Open*, *4*(3), e003901. https://doi.org/10.1136/bmjopen-2013-003901

Welch, H. G., & Albertsen, P. C. (2009). Prostate cancer diagnosis and treatment after the introduction of prostate-specific antigen screening: 1986-2005. *J Natl Cancer Inst*, *101*(19), 1325-1329. https://doi.org/10.1093/jnci/djp278

Wilson, K. M., Markt, S. C., Fang, F., Nordenvall, C., Rider, J. R., Ye, W., Adami, H. O., Stattin, P., Nyrén, O., & Mucci, L. A. (2016). Snus use, smoking and survival among prostate cancer patients. *Int J Cancer*, *139*(12), 2753-2759. https://doi.org/10.1002/ijc.30411

Wiseman, M. (2008). The second World Cancer Research Fund/American Institute for Cancer Research expert report. Food, nutrition, physical activity, and the prevention of cancer: a global perspective. *Proc Nutr Soc*, *67*(3), 253-256. https://doi.org/10.1017/s002966510800712x

Witteveen, A., Nane, G. F., Vliegen, I. M. H., Siesling, S., & MJ, I. J. (2018). Comparison of Logistic Regression and Bayesian Networks for Risk Prediction of Breast Cancer Recurrence. *Med Decis Making*, *38*(7), 822-833. https://doi.org/10.1177/0272989x18790963

Wu, I., & Modlin, C. S. (2012). Disparities in prostate cancer in African American men: what primary care physicians can do. *Cleve Clin J Med*, *79*(5), 313-320. https://doi.org/10.3949/ccjm.79a.11001

Yang, J., Zhong, F., Qiu, J., Cheng, H., & Wang, K. (2015). Dissociation of event-based prospective memory and time-based prospective memory in patients with prostate cancer receiving androgen-deprivation therapy: a neuropsychological study. *Eur J Cancer Care (Engl)*, *24*(2), 198-204. https://doi.org/10.1111/ecc.12299

Zhang, B., Shi, H., & Wang, H. (2023). Machine Learning and AI in Cancer Prognosis, Prediction, and Treatment Selection: A Critical Approach. *J Multidiscip Healthc*, *16*, 1779-1791. https://doi.org/10.2147/jmdh.S410301

Zhang, W., Cao, G., Wu, F., Wang, Y., Liu, Z., Hu, H., & Xu, K. (2023). Global Burden of Prostate Cancer and Association with Socioeconomic Status, 1990-2019: A Systematic Analysis from the Global Burden of Disease Study. *J Epidemiol Glob Health*, *13*(3), 407-421. https://doi.org/10.1007/s44197-023-00103-6