**MULTI-CLASS PREDICTION OF TOBACCO-RELATED**

**DISEASES USING ARTIFICIAL INTELLIGENCE**

**TECHNIQUES**

A Dissertation

Submitted to the

University of Bradford School of Management

In Partial Fulfilment

Of the Requirements for the Degree

Master of Science in Applied Artificial Intelligence and Data Analytics

**Olubukola O. Ojagbemi**

UOB 20029277

March 2022

WORD COUNT: 11979

# **STATEMENT OF AUTHENTICITY**

I have read the University Regulations relating to plagiarism. In submitting this I certify that this research is all my own work and does not contain any unacknowledged work from any other sources.

# **ABSTRACT**

Keywords (Artificial intelligence, Algorithms, Classifiers, Tobacco-related diseases, Multiclass)

Using tobacco products predispose users to the risk of developing respiratory disease such as chronic obstructive pulmonary disease, Asthma, bronchitis etc. cardiovascular diseases such as coronary heart disease, angina and heart attack etc, cerebrovascular diseases like stroke and cancers most significantly, lung cancer. This research was motivated by the need to give the growing number of tobacco users worldwide the opportunity to stay abreast of the diseases they may likely develop, due to their risk characteristics, and also serve as decision support for clinicians, by making a multiclass prediction of tobacco-related diseases using artificial intelligence techniques. In addition, it aimed to answer the question, to what extent can Artificial intelligence techniques be used for multiclass prediction of tobacco-related diseases among users based on their risk characteristics? Drawing from the data value chain theory, the study used available features present in the open-source Behavioural Risk Factor Surveillance System (BRFSS) survey from 2016 to 2020 collected and made available by the Centre for Disease Control (CDC). Random Forest, K nearest neighbour, Support Vector Machines and Naive Bayes algorithms were applied. Python and its libraries were used in the implementation. Performance was measured using precision, F score, and recall. Performance results were poor across all applied algorithms with Random Forest being the best performer with a precision of 40%. Thus, implying that these algorithms could not reliably classify tobacco users into their risk groups based on the features available in the BRFSS datasets. As such, these algorithms cannot at this time be recommended to make predictions of tobacco-related diseases using these features.

# **ACKNOWLEDGEMENT**

I am grateful to God for seeing me through this phase of learning and also to my family and friends for all of their support.

**Table of Content**

[**STATEMENT OF AUTHENTICITY** ii](#_Toc99013566)

[**ABSTRACT** iii](#_Toc99013567)

[**ACKNOWLEDGEMENT** iv](#_Toc99013568)

[**LIST OF FIGURES** vii](#_Toc99013569)

[**LIST OF TABLES** viii](#_Toc99013570)

[**LIST OF EQUATION** ix](#_Toc99013571)

[**CHAPTER 1** 1](#_Toc99013572)

[**INTRODUCTION** 1](#_Toc99013573)

[**1.1 RESEARCH QUESTION** 2](#_Toc99013574)

[**1.2 RESEARCH AIM** 3](#_Toc99013575)

[**1.3 RESEARCH OBJECTIVES** 3](#_Toc99013576)

[**1.4 JUSTIFICATION** 3](#_Toc99013577)

[**1.5 METHODOLOGY** 3](#_Toc99013578)

[**1.6 RESEARCH OUTLINE** 4](#_Toc99013579)

[**CHAPTER 2** 5](#_Toc99013580)

[**METHODOLOGY** 5](#_Toc99013581)

[**2.1 RESEARCH PHILOSOPHY** 5](#_Toc99013582)

[**2.2 RESEARCH STRATEGY** 7](#_Toc99013583)

[**2.3 RESEARCH DESIGN** 8](#_Toc99013584)

[**2.3.1 DATA SOURCE** 8](#_Toc99013585)

[**2.3.2 SAMPLING** 8](#_Toc99013586)

[**2.3.3 DEPENDENT VARIABLES** 8](#_Toc99013587)

[**2.3.4 INDEPENDENT VARIABLES (FEATURES)** 9](#_Toc99013588)

[**2.4 RESEARCH ADMINISTRATION** 9](#_Toc99013589)

[**2.4.1 DATA PROCESSING** 9](#_Toc99013590)

[**2.4.2 EXPLORATORY ANALYSIS** 10](#_Toc99013591)

[**2.4.3 FEATURE SELECTION** 10](#_Toc99013592)

[**2.4.4 LABEL ENCODING** 11](#_Toc99013593)

[**2.4.5 DATA VALIDATION AND SPLITTING** 11](#_Toc99013594)

[**2.4.6 MULTI-CLASS MODELS** 12](#_Toc99013595)

[**2.4.7 SUPPORT VECTOR MACHINES (SVM)** 12](#_Toc99013596)

[**2.4.8 K-NEAREST NEIGHBOUR (KNN)** 12](#_Toc99013597)

[**2.4.9 NAIVE BAYES** 13](#_Toc99013598)

[**2.4.11 EVALUATION** 14](#_Toc99013599)

[**2.4.12 IMPLEMENTATION** 14](#_Toc99013600)

[**2.5 RESEARCH ETHICS** 16](#_Toc99013601)

[**2.6 METHODOLOGY LIMITATIONS** 16](#_Toc99013602)

[**2.7 CHAPTER SUMMARY** 17](#_Toc99013603)

[**CHAPTER 3** 18](#_Toc99013604)

[**LITERATURE REVIEW** 18](#_Toc99013605)

[**3.1 TOBACCO-RELATED DISEASE RISK** 18](#_Toc99013606)

[**3.2 MULTI-CLASS CLASSIFIERS** 22](#_Toc99013607)

[**3.3 ALGORITHM PERFORMANCE** 27](#_Toc99013608)

[**3.4 RESEARCH GAP** 29](#_Toc99013609)

[**3.5 THEORETICAL FRAMEWORK** 30](#_Toc99013610)

[**3.6 SUMMARY** 31](#_Toc99013611)

[**CHAPTER 4** 32](#_Toc99013612)

[**FINDINGS** 32](#_Toc99013613)

[**4.1 DATA PROCESSING** 32](#_Toc99013614)

[**4.2 EXPLORATORY ANALYSIS** 33](#_Toc99013615)

[**4.3 FEATURE SELECTION** 38](#_Toc99013616)

[**CHAPTER 5** 49](#_Toc99013617)

[**ANALYSIS AND CONCLUSION** 49](#_Toc99013618)

[**5.1 ADOPTED APPROACH** 49](#_Toc99013619)

[**5.2 ANALYSIS** 50](#_Toc99013620)

[**5.3 CONCLUSION** 52](#_Toc99013621)

[**5.4** **LIMITATIONS** 53](#_Toc99013622)

[**5.5 FURTHER RESEARCH** 53](#_Toc99013623)

[**5.6 REFLECTION** 54](#_Toc99013624)

[**REFERENCES** 56](#_Toc99013625)

[**APPENDIX A RESEARCH PROPOSAL** 63](#_Toc99013626)

[**APPENDIX B APPROVED ETHICS FORM** 76](#_Toc99013627)

# **LIST OF FIGURES**

[**Figure 1. 1 Machine learning techniques (Rashidi et al. 2019: 9) 2**](#_Toc98939344)

[**Figure 2. 1 Research onion (Saunders et al. 2019) 5**](#_Toc98939685)

[**Figure 2. 2 Implementation framework 16**](#_Toc98939686)

[**Figure 3. 1 Shows predictors used in clinical models for cardiovascular disease adapted from (Damen et al. 2016) 20**](#_Toc98940513)

[**Figure 3. 2 Approaches for Data Imbalance (Kaur et al. 2019: 13) 27**](#_Toc98940514)

[**Figure 3. 3 Cross validation (Jiao and Du 2016: 323) 28**](#_Toc98940515)

[**Figure 3. 4 Confusion matrix adapted from (Tharwat 2018: 170) 28**](#_Toc98940516)

[**Figure 3. 5 Data Value Chain framework for data based value creation adapted from (Lim et al. 2018: 126) 30**](#_Toc98940517)

[**Figure 4. 1 Imbalance distribution of dependent variables** 35](#_Toc98941544)

[**Figure 4. 2 Distribution of disease class among age groups** 36](#_Toc98941545)

[**Figure 4. 3 Distribution of disease class by weight** 36](#_Toc98941546)

[**Figure 4. 4 Python implementation of Chi-square** 37](#_Toc98941547)

[**Figure 4. 5 Associated variables from Chi-Square test** 38](#_Toc98941548)

[**Figure 4. 6 Python implementation of LASSO** 38](#_Toc98941549)

[**Figure 4. 7 SMOTE implementation to balance classes** 40](#_Toc98941550)

[**Figure 4. 8 Confusion matrix Random Forest** 43](#_Toc98941551)

[**Figure 4. 9 Confusion matrix K-nearest neighbour (KNN)** 44](#_Toc98941552)

[**Figure 4. 10 Confusion matrix Support vector machine** 45](#_Toc98941553)

[**Figure 4. 11 Confusion matrix Naive Bayes** 46](#_Toc98941554)

[**Figure 4. 12 Performance of Predictive algorithms** 47](#_Toc98941555)

# **LIST OF TABLES**

[**Table 3. 1 Predictive models for cardiovascular disease adapted from (Yun et al. 2017)** 28](#_Toc98878127)

[**Table 4. 1 Missing values** 32](#_Toc98942543)

[**Table 4. 2 Variable description** 33](#_Toc98942544)

[**Table 4. 3 Disease prevalence in the sample population** 35](#_Toc98942545)

[**Table 4. 4 Least absolute shrinkage** 39](#_Toc98942546)

[**Table 4. 5 Result for Random Forest** 41](#_Toc98942547)

[**Table 4. 6 Result for K-nearest neighbour (KNN)** 41](#_Toc98942548)

[**Table 4. 7 Result for SVM** 42](#_Toc98942549)

[**Table 4. 8 Result for Naive Bayes** 42](#_Toc98942550)

[**Table 4. 9 Comparison of algorithm performance** 47](#_Toc98942551)

# **LIST OF EQUATION**

[**Equation 2. 1 Least absolute shrinkage (LASSO) (Awais et al. 2021: 3413)** 11](#_Toc98877489)

[**Equation 2. 2 Support vector machines (SVM) (Cavallaro et al. 2015: 4636)** 12](#_Toc98877490)

[**Equation 2. 3 Euclidean distance K-nearest neighbour (Moeez M. Subhani and Ashiq Anjum 2020: 23)** 13](#_Toc98877491)

[**Equation 2. 4 Gaussian Naive Bayes (Mandal and Jana 2019)** 13](#_Toc98877492)

[**Equation 2. 5 Precision, Recall, F-score** 14](#_Toc98877493)

[**Equation 2. 6 Prevalence rate (Bhopal 2016: 238)** 15](#_Toc98877494)

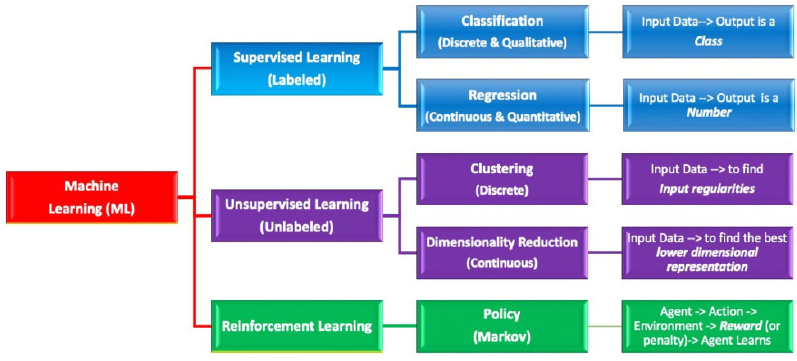
[**Equation 4. 1 Chis-square formula** 44](#_Toc98879857)

# **CHAPTER 1**

# **INTRODUCTION**

Tobacco use has been identified as a major public health concern, with it being associated with about 6 million deaths annually, of which, 30% are cancers (Perez-Warnisher et al. 2018). It is the leading cause of preventable cancer according to (Al-Obaide et al. 2018). Worldwide, tobacco has been identified as a well-known risk factor for early morbidity and mortality (Holipah et al. 2020). Tobacco use has been identified to have a causal relationship with diseases such as cancer of the lung, oral-nasal cavities, stomach, liver, urinary tract and also, leukaemia. As well as, cardiovascular diseases (CVD) and chronic obstructive pulmonary disease (COPD) (Taghizadeh et al. 2016). Even with the introduction of tobacco cessation protocols, due to the addictive nature of nicotine, according to (Perez-Warnisher et al. 2018) there is a projected increase in tobacco users worldwide by 1.6 billion in the next 25 years. Thus, more deaths associated with tobacco use are to be expected. According to (Vineis et al. 2004) 50% of ever smokers are at risk of early mortality, with a majority of these at middle age. And if these smoking trends remain, a projected 1 billion deaths may be related to smoking in the 21st century. Tobacco products may be in form of smoked products such as cigars, cigarettes, electronic cigarettes, pipes, or non-smoked products such as chewed tobacco, snuffs, and dips (Onor et al. 2017). (MacArthur et al. 2016) reported the use of such products in individuals as young as 16 in the United Kingdom. Thus, increasing their lifetime disease risk. The use of these products carries a lot of health risks of which users are mostly aware, but find it difficult to cease using these products due to their addictive nature. There is also high financial implication from the management of these diseases (Rezaei et al. 2016), as well as its ability to generally reduce the quality of life (Feirman et al. 2016). As such, early follow-up of disease risk may help mitigate some of these effects.

Artificial intelligence (AI) refers to technologies that show human-like intelligence and learning (Kilic 2020). Machine learning is the branch of artificial intelligence that trains and learns underlying patterns from large amounts of data with algorithms, through mathematical and statistical approaches and produces results that will be used for decision-making and predictions (Huang et al. 2020). Broadly, these can be either supervised learning, unsupervised learning, or reinforcement learning techniques. Supervised algorithms are trained on datasets, and provide an output based on the prior training (Rashidi et al. 2019). Classification and regression techniques are based on supervised learning. Unsupervised learning produces results by aggregating and clustering unlabelled data based on shared similarities with the absence of prior data training (Rashidi et al. 2019). An example is clustering algorithms. Reinforcement learning, on the other hand learns by optimal mapping i.e. policy-based, with multiple trial and reward attempts to achieve the best result (Nian et al. 2020).



**Figure 1. 1 Machine learning techniques (Rashidi et al. 2019: 9)**

Artificial intelligence techniques have been used with success to predict diseases based on disease characteristics (Huang et al. 2020). Classifier algorithms such as Random Forest, Neural networks, support vector machines (SVM), K-nearest neighbour (KNN), Naive Bayes, have been used for disease risk prediction in various studies. A similar approach could be used to predict different tobacco-related diseases a tobacco user might come down with based on their characteristics. This is a Multiclass problem because several risk features will be used to predict several diseases. Disease prediction will be beneficial for monitoring and early follow-up and has become important for personalized care (Steyerberg and Vergouwe 2014). Such predictive models when accessible to patients, is of value in enabling them to play an active role in making decisions about their health, and also to the society at large for better planning and management of these risk group (Lee et al. 2016). (Lim et al. 2018) proposed how such benefits could be provided by a data-intensive service such as healthcare service, using the data value chain framework.

# **1.1 RESEARCH QUESTION**

Artificial intelligence algorithms have been applied successfully to predict heart disease (Vembandasamy et al. 2015), lung cancer (Toumazis et al. 2020), Chronic obstructive pulmonary diseases (Matheson et al. 2018) among others, using a range of predictors. With tobacco users as a target population, this research will apply similar artificial intelligence techniques to answer the question; To what extent can Artificial intelligence techniques be used for multiclass prediction of tobacco-related diseases among users based on their risk characteristics?

# **1.2 RESEARCH AIM**

This research aims to use Artificial intelligence classifier algorithms for multiclass prediction of tobacco-related diseases of high burden (morbidity and mortality) among tobacco users based on their risk characteristics.

# **1.3 RESEARCH OBJECTIVES**

The objective of this research is outlined as follows;

1. To outline tobacco-related diseases of highest burden among the study population.
2. To identify risk characteristics of tobacco users that could be used to make predictions.
3. To apply artificial intelligence algorithms to predict diseases in objective 1 using risk features in objective 2.
4. To evaluate the performance of algorithms used in objective 3.
5. Compare the performance of the artificial intelligence algorithms used in objective 3

# **1.4 JUSTIFICATION**

Early identification of diseases such as lung cancer is an important factor in reducing morbidity and mortality (Rankin et al. 2020). Early initiation of treatment measures in Chronic obstructive pulmonary disease improves the quality of life of patients and slow down disease progression (Welte et al. 2015). Disease prediction among tobacco users will allow for early monitoring so as to subsequently reduce disease morbidity and mortality and thus improve health outcomes.

# **1.5 METHODOLOGY**

This research is a quantitative cross-sectional study of the population which will use a secondary dataset from a cross-section of the population to predict tobacco-related diseases among users using their high-risk characteristics. The participants for this study are tobacco users. The dataset containing the participants is a secondary dataset that has been collected by the Centre for Disease Control via landline and telephone call to identify individuals with modifiable risk factors for chronic diseases and other leading causes of death from various states across the United States of America from 2011 till date and contains 2.7 million entries (Centers for disease control and prevention 2021). The data will be uploaded, and data wrangling done using various python libraries. To ensure a logical and quality predictive model is applied to meet research objectives, the framework outlined by (Steyerberg and Vergouwe 2014) for developing predictive models is adopted for this research. It involves, problem definition and data inspection, coding the predictors, model specification, model estimation, model performance

# **1.6 RESEARCH OUTLINE**

This research has five chapters, they include, Introduction, Methodology, Literature review, Findings and Analysis and overall conclusion. In chapter one a brief introduction about tobacco burden, the research aim, objectives, justification and how the methodology is approached is highlighted. Chapter two highlights the methodological approach to apply some Artificial intelligence algorithms such as Naive Bayes, Random Forest, K nearest neighbour and Support Vector Machine to predict tobacco-related diseases and measure their performance. Chapter three contains a literature review on the diseases of which tobacco use poses a high risk, risk characteristics that have been used in predictions of tobacco-related diseases, Artificial intelligence techniques and Classifier algorithms that have been used in literature for disease predictions, performance measurement techniques, theoretical framework and research gap. Chapter 4 contains the result and findings, Chapter five analyses and discusses the result of prediction, assess how the classifiers compare with each other, as well as their limitations, and draw conclusions based on the research objectives.

This chapter thus identifies tobacco use as a burden and briefly highlights the benefits prediction of tobacco-related disease among these users will provide by applying artificial intelligence algorithms. It also provides an insight into the content of the subsequent chapters in this research, as well as highlights the aims, objectives, and justification of this research.

# **CHAPTER 2**

# **METHODOLOGY**

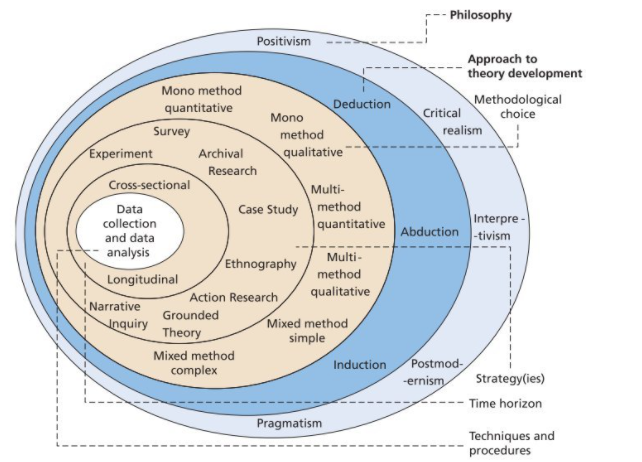
This chapter outlines the research philosophy and approach of choice for this research, as well as highlights the strategy to accomplish the research objectives, and also hights research limitations.

# **2.1 RESEARCH PHILOSOPHY**

Research philosophy could be described as beliefs and assumptions that result in the development of knowledge. Regarding this research, it involves how knowledge about the use of Artificial intelligence techniques to predict tobacco-related diseases could be developed. This process will involve making a number of assumptions (Burrell and Morgan 1979). The nature of research assumptions differs and are used to describe the various research ideologies, they include;

1. Epistemology: Assumptions about knowing
2. Ontology: Assumptions about realities
3. Axiological: Assumptions based on values

The choice of research philosophy determines the research strategy and methodological design (Saunders et al. 2019).



**Figure 2. 1 Research onion (Saunders et al. 2019)**

In epistemology, basic questions about knowledge are asked. It describes the nature and bases of human knowledge (Hofer and Pintrich 1997). How do we know what we know? Although knowledge has been described in various ways, epistemology argues that our perception of valid knowledge can result from a combination of belief, confidence, truth and a good basis for claiming research findings are a fact (Kelly 2021). It thus implies a somewhat measurable and objective way to seek knowledge, which is less abstract. Another Epistemological belief is that of interpretivism, i.e., developing knowledge from subjective interpretation (Hovorka and Lee 2010). Finally, some believe that knowledge could be gained by both objective measurements and subjective interpretations combined (Žukauskas et al. 2018). Ontology questions reality and perception of what is currently in existence in the social environment (Saunders et al. 2019). It seeks to answer questions like, what is the nature of reality and what is the world like? Ontological beliefs vary from a singular view of what reality is, to having multiple realities that are dependent on the experience of people involved. Axiology makes a judgement about values. Personal values affect how the research is being conducted (Saunders et al. 2019).

This research is not founded on the Axiological assumptions which are based on personal values. It however seeks to understand the truth about reality and the existence of a particular knowledge. This thus narrows research philosophical focus between epistemological and Ontological ideologies. The combination of these views is used in the development of research philosophies. Positivism combines the ontological view of a singular reality and the epistemological view that this reality can be assessed objectively. Thus, there is only one knowledge and this can be measured using reliable techniques (Mackenzie and Knipe 2006). Constructivism combines the ontological assumptions of the existence of multiple realties and the epistemological assumption that these can be assessed by interpretation (Žukauskas et al. 2018). Pragmatism is of a view that there might be one or more than one reality that could be interpreted as well as objectively measured (Mackenzie and Knipe 2006). This research is quantitative research that plans to use performance metrics of artificial intelligence algorithms that would be applied on data sets to meet research objectives. Conclusions would be drawn based on these findings without the input of the researcher’s personal values, reality or experience. As such, it leans towards a positivism research philosophy. Deduction, induction, and abduction are cognitive approaches applied to research to draw a conclusion. Deductive reasoning draws conclusions by a stepwise and logical approach based on previous theories, while inductive reasoning makes conclusions from observation, and finally abductive reasoning draws a conclusion from what is previously known (Ketokivi and Mantere 2010). The research suggests a deductive approach will be the best approach to make conclusions. Conclusions will be drawn by applying similar performance metrics used in literature to assess the ability of Artificial intelligence algorithms to make predictions. As opposed to an inductive approach that would suggest new ways to draw conclusions about the ability of Artificial intelligence algorithms to make similar predictions.

# **2.2 RESEARCH STRATEGY**

Broadly, research strategies could be qualitative, quantitative or mixed methods. However, based on the positivism strategy adopted, a quantitative research strategy will be applied to this research (Schutt 2019). This is because quantitative research can be used to describe, explain, predicts, or investigates relationships between variables and assess their outcome. It analyses the correlation between variables and uses numerical summaries to make a generalization about the population. The goal of quantitative research is to apply methodological techniques for empirical evaluation without bias (Schutt 2019). To answer the research question, the researcher will examine existing pieces of literature that applied Artificial intelligence to predict diseases, especially those of which tobacco use pose a high risk. The literature review is vital to explaining and understanding the topic under study (Vom Brocke et al. 2015). An iterative literature review process that involves searching, reading, learning and back to searching until the researcher is satisfied, could be applied (Vom Brocke et al. 2015). Alternatively, a sequential step of searching, acquiring, analysing and interpreting could be adopted (Boell and Cecez-Kecmanovic 2014). In this work, the researcher adopts the latter. Contemporary peer-reviewed journal articles from 2016 till date were sourced from Google scholar, University of Bradford repository(summons), PubMed, and Web of science using key search words. Identifying and applying search parameters can affect search results, this may require some experimenting initially to access the effectiveness of chosen search words and combining operators (Kitchenham 2004). The key search terms for this study are divided broadly into the following groups;

1. Classification problems: “multiclass classification” “classification”
2. Artificial Intelligence algorithms: “Artificial intelligence”, “Machine learning”, “Supervised learning”, “Algorithms”, “Disease prediction models”, “algorithm performance metrics”
3. Tobacco use burden: “Tobacco use”,” Smoking”, “Risk factors for tobacco-related diseases”, “Tobacco-related diseases of high burden”

The collection and utilisation of secondary data are vital to meet these research objectives. Secondary data refers to data that has been collected and used for a different purpose (Kimberlin and Winterstein 2008). It is usually data that has been collected by another person or institution (Yin 2009). Secondary data that contain information about tobacco users, as well as their associated diseases, was sourced. The dataset used in this research has been collected by the (Centers for disease control and prevention 2021) as part of an ad-hoc survey and is available in their database.

# **2.3 RESEARCH DESIGN**

The literature search will be focused on highlighting tobacco-related diseases of high burden, identifying risk characteristics that have been applied in the literature to predict tobacco-related diseases, predictive models, and performance evaluation techniques. This information will be sourced from peer-reviewed journal articles on subjects about epidemiology, clinical predictive models, and predictive algorithms.

## **2.3.1 DATA SOURCE**

The Dataset for this research was sourced from the Behavioural risk factor surveillance system (BRFSS) survey from 2016 to 2020. The Data is based on a telephone survey conducted across 50 states in the United States including the district of Columbia, Guam, and Puerto Rico by states health departments in partnership with the centre of disease control (CDC) (Centers for disease control and prevention 2021). It contains self-reported responses about health-related risk behaviours, chronic health conditions, and other information from noninstitutionalized adult population, age 18 or older. This is obtained from the public repository (Centers for disease control and prevention 2021).

## **2.3.2 SAMPLING**

The research is a cross-sectional study that targets the tobacco user population that took part in the BRFSS survey between 2016 and 2020. It adopts a convenience sampling technique and includes all respondents who answered “Yes” to the use of either smokeless tobacco products or smoked at least 100 Cigarettes in their lifetime. The respondents in these categories were 2,079,386.

## **2.3.3 DEPENDENT VARIABLES**

The dependent variables to be used are responses to questions related to diagnosis of chronic health conditions that tobacco use poses a high risk, as shown in the literature, and are similarly present in the BRFSS dataset. The questions on chronic health conditions in the data set includes

1. Have you ever been diagnosed with a Heart attack? Response categories were “Yes”,” No”, “Not sure”, “Refused”
2. Have you ever been diagnosed with Angina or coronary heart disease? Response categories were “Yes”,” No”, “Not sure”, “Refused”
3. Have you ever been diagnosed with Asthma? Response categories were “Yes”,” No”, “Not sure”, “Refused”
4. Have you ever been diagnosed with any type of cancer? Response categories were “Yes”,” No”, “Not sure”, “Refused”
5. Have you ever been diagnosed with chronic obstructive pulmonary disease, (COPD) emphysema or chronic bronchitis? Response categories were “Yes”,” No”, “Not sure”, “Refused”

While coding the outcome variables, those who answered “No”, “Not sure”, “Refused” values would be dropped. Those who answered “Yes” would be selected.

## **2.3.4 INDEPENDENT VARIABLES (FEATURES)**

The predictors used will be based on what is available in literature as well as similar variables available in the BRFSS dataset. Responses to questions about Age, Gender, General health wellness, Body mass index (BMI), Exercise, smoked at least 100 cigarettes, Frequency of days now smoking, stopped smoking in the past 12 months, interval since last smoked, use of smokeless tobacco, Alcohol consumptions are potential risk predictors present in the BRFSS dataset.

# **2.4 RESEARCH ADMINISTRATION**

In multiclass classification, each observation could belong to a number of classes (Amarappa and Sathyanarayana 2014). With the framework proposed by (Steyerberg and Vergouwe 2014) in mind, the research objectives will be met as follows;

## **2.4.1 DATA PROCESSING**

Data processing and cleaning task is required to inspect the data to remove noise and have the data in a suitable format for model training, in order to improve model quality (Nakhle and Harfouche 2021). For this research these involve;

* Extracting participants for this research from the dataset in each year i.e., those that responded Yes to the question; Have you smoked at least 100 cigarettes in your entire life? Or Yes to the question; Do you currently use chewing tobacco, snuff, or snus every day, some days, or not at all?
* Aggregating the dataset with study sample from 2016 to 2018
* Removal of missing details and duplicated entries
* Dropping columns that would not be relevant to the research
* Dropping entries that answered I don’t know or did not provide responses to the diagnosis of the dependent variables
* Dropping entries who reported they had not been diagnosed with the dependent variables
* Randomization and reshuffling entries to ensure entries are not in any ordered format.
* Data transformation to have all the dependent variables in a column
* Renaming of columns to be more descriptive

## **2.4.2 EXPLORATORY ANALYSIS**

The exploratory analysis aims to examine the data for features distribution pattern, detect outliers and data anomalies using descriptive statistical techniques, univariant, bivariant and multi-variant analysis of variables with graphical representation (Komorowski et al. 2016). This is vital to derive insight from the dataset, visualize relationships in both direction and magnitude between the dependent and independent variables (Komorowski et al. 2016).

## **2.4.3 FEATURE SELECTION**

Feature selection is the process of identifying the best variables to be applied in the model. This is important to improve the performance of the model and minimize computational cost (Cunningham et al. 2021). Findings from exploratory analysis, domain knowledge and also least absolute shrinkage (LASSO) technique will be used to select the best features. The least absolute shrinkage (LASSO) is a technique used to select the best variables that will improve predictive accuracy. It eliminates insignificant predictive features by shrinking their data values, it adds a penalty that is equal to the absolute value of the coefficient (Awais et al. 2021). The tuning parameter lambda (λ) determines the degree of shrinkage. Higher values of lambda result in more shrinkage. Variables that have values of zero or less are eliminated.

**Equation 2. 1 Least absolute shrinkage (LASSO) (Awais et al. 2021: 3413)**

= Cost function (sum of squared residuals)

| = Penalty

|| = Absolute value of coefficient

λ = Tuning parameter

## **2.4.4 LABEL ENCODING**

Label encoding allows the categorical variables to be mapped into numerical values that the algorithms understand (Hancock and Khoshgoftaar 2020). This can be achieved with either one hot encoder or Label encoder from the scikit learn python library to encode the dependent variable.

## **2.4.5 DATA VALIDATION AND SPLITTING**

There is a need to split the data into training and testing sets. This is important as it helps to prevent overfitting while also ensuring the model learns from enough samples. Overfitting occurs when the model can make predictions on the trained dataset with high accuracy, but it is unreliable on novel samples as because it has learnt the training dataset too well (Rice et al. 2021). The models first learn from the training set and make predictions from the test set (Strieth-Kalthoff et al. 2020). Although no actual proportion exists that must be utilized, it is essential the training set significantly outnumber the testing set for best result. 80% of the data will be used for training and 20% for testing. Cross-validation is also done assess model performance prior to testing.

## **2.4.6 MULTI-CLASS MODELS**

Multi-classification is a multioutput kind of learning that predicts each instance concurrently into multiple outputs (Xu et al. 2019). From the broad collections of classification algorithms available, the researchers choose to apply K-nearest neighbour (KNN), Naive Bayes and Support Vector Machine and Random Forest to meet objective 3 due to their computational efficiency. The literature review will highlight the application of the various algorithms along with their strength and weaknesses. Performance measures such as precision, recall, F-score and accuracy will be applied to meet objective 4, and their performance compared as stated in objective 5.

## **2.4.7 SUPPORT VECTOR MACHINES (SVM)**

Support vector machines (SVM) is a supervised classification algorithm that has the advantage of requiring a fewer number of features to make predictions of high-dimensional data. It is also memory efficient. It identifies the best hyperplane (decision boundary) that separates the classes (Amarappa and Sathyanarayana 2014).

**Equation 2. 2 Support vector machines (SVM) (Cavallaro et al. 2015: 4636)**

= Minimum distance of margins of hyper plane from support vectors

C = Hyperparameter, controls the shape of the solution of the decision boundary.

ξ = Slack variable that penalizes classification error

## **2.4.8 K-NEAREST NEIGHBOUR (KNN)**

KNN classifies by placing data points based on their identified nearest neighbours. It has the advantage of being memory efficient with good run time performance. The nearest neighbours are determined either by majority voting, or distance weighted voting (Cunningham and Delany 2021).

**Equation 2. 3 Euclidean distance K-nearest neighbour (Moeez M. Subhani and Ashiq Anjum 2020: 23)**

D(x,y) = Distance between data points

xiyi : ith feature of the two data points

## **2.4.9 NAIVE BAYES**

The Naive Bayes algorithm applies probabilities to determine class using Bayes theorem. Prediction is made by determining the probability of the class instance and by selecting the those with the highest probability (Lamba et al. 2015). Each feature is conditionally independent. Gaussian Naive Bayes classifier is the most commonly used Bayesian classifiers multi-class classifier (Mandal and Jana 2019).

**Equation 2. 4 Gaussian Naive Bayes (Mandal and Jana 2019)**

= Features

Y = Classes

= Standard deviation

= Mean

**2.4.10 RANDOM FOREST**

Random forest is an ensemble decision tree classifier that produces a randomly selected subset from training samples in a bagging process, and make predictions with randomly selected variables. This tree building cycle is done on a user-defined number of times. In creating the training subset, some samples can be selected several times, while others may not be selected at all. Prediction is made by evaluating new data points against all decision trees created by voting on class membership (Belgiu and Drăguţ 2016).It handles high dimension and multicollinear data effectively at low computational cost with high variance and low bias.

## **2.4.11 EVALUATION**

The model performance and efficiency will be measured by applying a confusion matrix. The precision score, F score and recall will be assessed. A confusion matrix is a statistical contingency table that contains four attributes i.e. True Positive (TP), False Positive, True Negative (TN), and False Negative (FN. An ideal performance will be one with no false positives and negatives (Salih and Abdulazeez 2021). Precision is also known as specificity. It represents the proportion of predictive results that are actually true. Recall, also known as sensitivity, represents actual true values that are predicted. F score is referred to as the harmonic mean of precision and recall, combining both metrics (Jain and Singh 2018).

Precision =

Recall =

F-Score =

**Equation 2. 5 Precision, Recall, F-score**

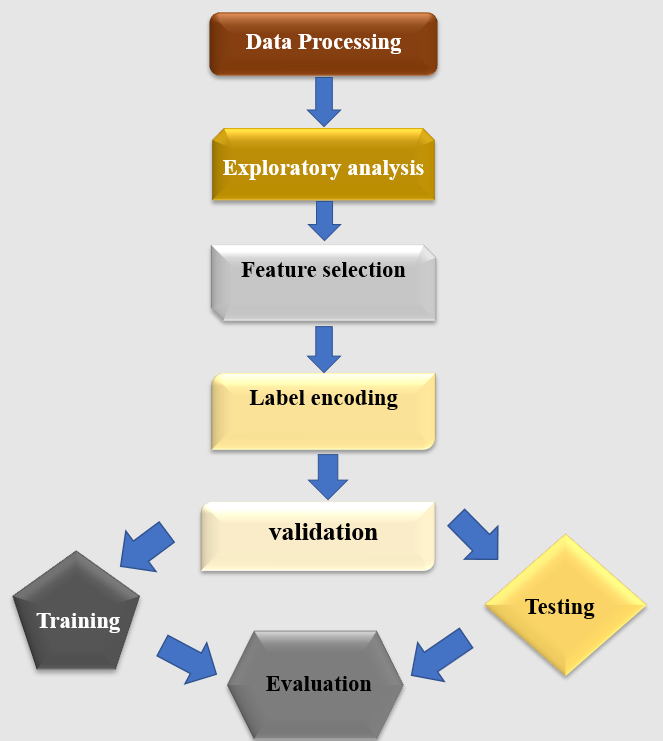
## **2.4.12 IMPLEMENTATION**

The data will be loaded and wrangled using python and its libraries for data cleaning, analysing, as well as visual reporting. Python is a high-performance, object-oriented programming language, which is popular among developers due to its simplicity as well as its efficiency, because of its very interactive libraries. It has the ability to synthesise a large amount of both structured and unstructured data, and meet analytical needs. Some of the libraries to be used in this research include, NumPy, Pandas, SciPy and Sci-kit-learn, Matplotlib and Seaborn. NumPy is a numerical python library used to create multi-dimension numerical arrays as well-performing numeric calculations. The Pandas library creates data structures that are useful for data analysis and manipulations in form of data frames and series. SciPy is a scientific library for mathematics, and engineering. The Scikit-learn library for various analytical algorithms both supervised and unsupervised (Dhawan 2019). Matplotlib and Seaborn are visualization libraries for python. The computational environment used is Google pro laboratory, Tensor Processing unit. Objective 1 will be achieved by outlining all tobacco-related diseases that have been identified in the literature and similarly available in the dataset and calculating their disease prevalence.The prevalence rate is one of the most common measures of disease burden. It is the proportion of the study population that is affected by a disease of interest at a given time (Bhopal 2016).

Prevalence (%) =

**Equation 2. 6 Prevalence rate (Bhopal 2016: 238)**

Objective 2 will be achieved by applying the Least absolute shrinkage (LASSO) technique to select the best predictive variables for the model. Objective 3 will be achieved by applying the multi-class models to make predictions of diseases with the highest prevalence in study population, as well as those identified as having high morbidity and mortality. These will be used as dependent variables. Selected features from objective 2 will be used as independent variables. Objective 4 will be achieved by assessing model performance using a confusion matrix and assessing the average precision, F1 score and Recall. Objective 5 will be achieved by making a comparison of model performance.

****

**Figure 2. 2 Implementation framework**

# **2.5 RESEARCH ETHICS**

Using secondary data can raise privacy and confidentiality issues. However, the scientific benefit of archiving and sharing data cannot be overlooked (Morrow et al. 2014). Ethical guidelines are based on deontological ethical principles which provide rules that direct research conduct (Saunders et al. 2019). All appropriate ethical guidelines are adhered to during the course of this research. This research is intended for the benefit of the public good, will do no harm and intends to maximize benefit of the dataset which will be gotten from (Centers for disease control and prevention 2021). The data is de-identified and the process of analysis will not result in re-identifying participants, thus anonymity will be maintained. All sources of information for this research are acknowledged through referencing and this research will adhere to international standards for ethics in research.

# **2.6 METHODOLOGY LIMITATIONS**

Although using a secondary dataset is convenient as it provides easy access to a large sample size with the required features, the data was not collected for the sole purpose of this research, the data collection process was not directed by the researcher. Specifically, information such as chronic disease status are based on respondent responses and not on clinical information thus it may be difficult to verify as it will be difficult to ascertain if respondents were truthful in their responses, if questions were taken seriously by respondents, if respondents had issues with recall and not fully understanding the question. The sampling technique may not be truly representative of the population as adults who are institutionalized, no access to telephone having been excluded in the BRFSS survey. The dataset may not contain all possible predictive risk characteristics that are used in the literature. The researcher is not able to apply and compare all the available multiclassification models due to limitations with computing resources and models selected are those that require the lowest computing cost and are take the least time to train thus, keeping the scope of research narrow.

# **2.7 CHAPTER SUMMARY**

This chapter highlights the various research ideologies and determined positivism as the research philosophy of choice for this study, with conclusions been made by the deductive approach. The research is a quantitative, cross-sectional study that uses data from the Behavioural risk factor surveillance system (BRFSS) survey. It identified the algorithms of choice for this research and also outlined the limitations of the study.

# **CHAPTER 3**

# **LITERATURE REVIEW**

This chapter provides a comprehensive overview of the literature relating to the research topic. It covers tobacco-related diseases, their disease burden as well as their predictive risk characteristics. It also explores disease predictive models of tobacco-related disease in literature, and classifier algorithms. Performance measurement techniques for these algorithms are also assessed. At the end of this chapter, the theoretical framework is described and identified research gaps are highlighted.

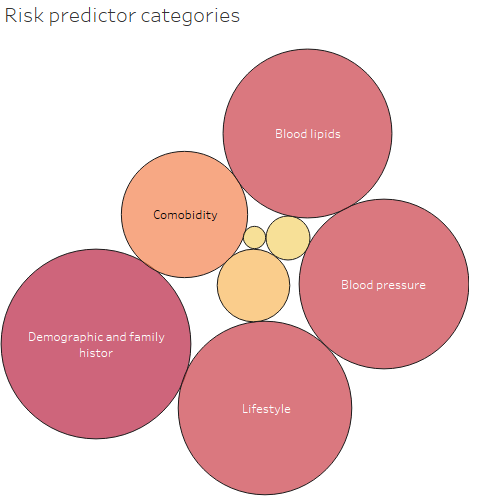
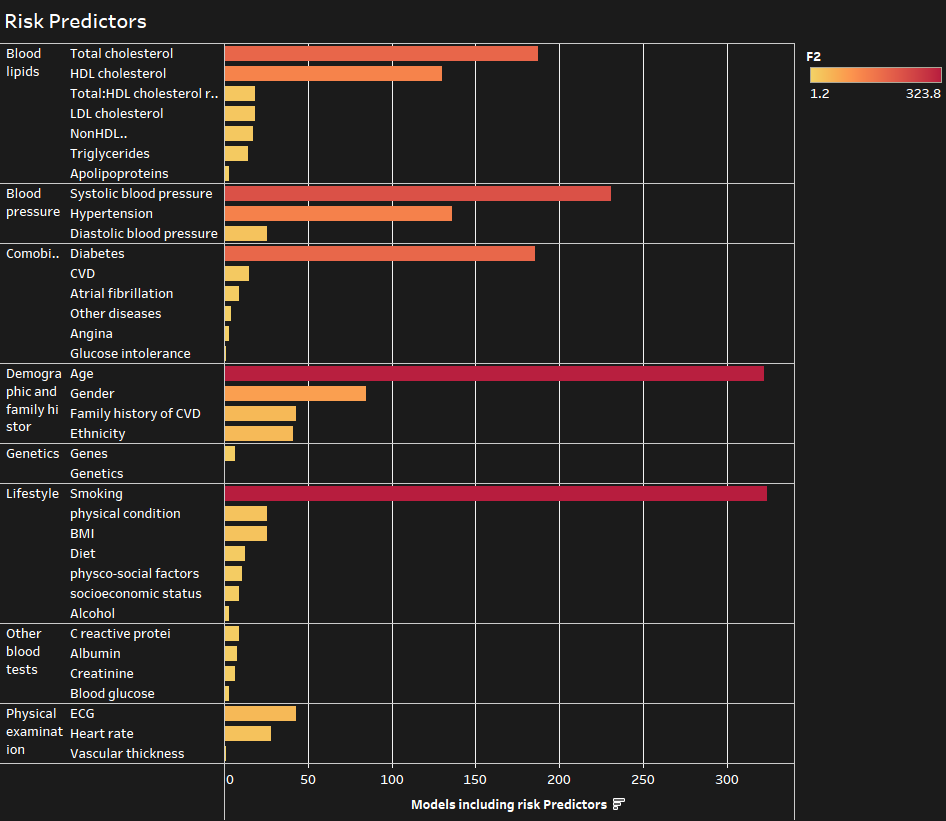
# **3.1 TOBACCO-RELATED DISEASE RISK**

Tobacco-related diseases can be described as chronic diseases for which tobacco use pose a high relative risk, as well as excess death rates i.e., the deathrate in smokers exceed the death rate in never smokers (Jha 2020). Tobacco use is a public health burden in terms of death, disability, and financial implication. It has been identified as an important risk factor for cardiovascular disease, chronic pulmonary disease, lung cancer, as well as other cancer, accounting for 41 million deaths in the US, UK, and Canada cumulatively (Jha 2020). Smoking has been associated with 82% of lung cancer mortality (Torre et al. 2016). According to (West 2017) tobacco-related deaths are from

1. Cardiovascular diseases especially coronary heart disease
2. Cancers mainly lung cancer.
3. Respiratory disease significantly Chronic obstructive pulmonary disease
4. Other diseases such as maternal and foetal conditions, digestive diseases, diabetes mellitus, diseases of the genitourinary system, neuropsychiatric conditions and musculoskeletal diseases.

The disease burden among tobacco users is significant, as these health implications apply to millions of users worldwide due to the addictive nature of tobacco. These diseases can remain symptomless in early stages, and when diagnosed at later stages can be associated with poor prognosis, with lung cancer for example having only 5 years survival rate post-diagnosis (Toumazis et al. 2020). Late diagnosis adversely affects survival from lung cancer, and picking up on the disease early can increase 5 years survival by up to 75% (Weller et al. 2019). According to (Li et al. 2021), although smoking has been shown to pose an increased risk of ischemic stroke in smokers, it has no impact on the prognostic outcome. In terms of disease prevalence, Asthma is considered the most prevalent respiratory disease in the United States with increasing incidence (Kuruvilla et al. 2019). Globally, its prevalence varies from population to population. (Underner et al. 2021) noted that among smokers, its prevalence either equals or is greater than that of the general population. In some population, in middle eastern cities, its prevalence was particularly high among smokers (Alavinezhad and Boskabady 2018). Prevention and treatment success of these diseases largely depends on early intervention. This can be achieved with good disease prediction. A good predictive model can be useful for a more accurate selection of clinical trials participants (Tammemägi 2018). It can also be of benefit in the surveillance of disease incidence and prevalence, as well as public health management in general (Cichosz et al. 2016). Overall, a good prediction of these diseases can potentially result in improved disease outcomes and reduce the cost of treatment. These diseases can thus be used as risk classes when developing a predictive model.

Risk predictors such as gender, age, smoking, blood pressure, blood lipids measurements and body mass index have been used for cardiovascular disease prediction (Van Bussel et al. 2020). In addition, other factors such as ethnicity, family history of the disease, and presence of other chronic diseases and social deprivation, asbestos exposure, duration of smoking, average amount smoked per day, and duration of abstinence from smoking for former smokers, have also been used (Gray et al. 2016). Documented risk predictors for chronic obstructive pulmonary disease includes tobacco smoking, asthma, environmental exposures to toxic fumes and alpha-1-antitrypsin deficiency (Matheson et al. 2018). Feature engineering is an important technique for selecting the best variables from the dataset to be used as predictors, it also provides useful information about the dataset (Waring et al. 2020). Good feature engineering traditionally depends on good domain knowledge combined with trial-and-error attempts. This manual endeavour is a time-consuming process that can be challenging (Waring et al. 2020). (Katz et al. 2016) proposed the use of explore-kit, an automated feature selection approach, which was shown to be effective on tested datasets. A similar automated approach such as the One Button Machine according to (Lam et al. 2017) improved efficiency by more than 20% when compared with the traditional process.

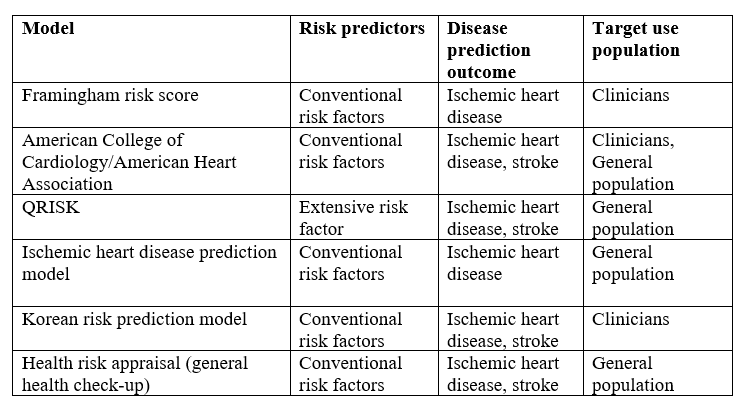
 

**Figure 3. 1 Shows predictors used in clinical models for cardiovascular disease adapted from (Damen et al. 2016)**

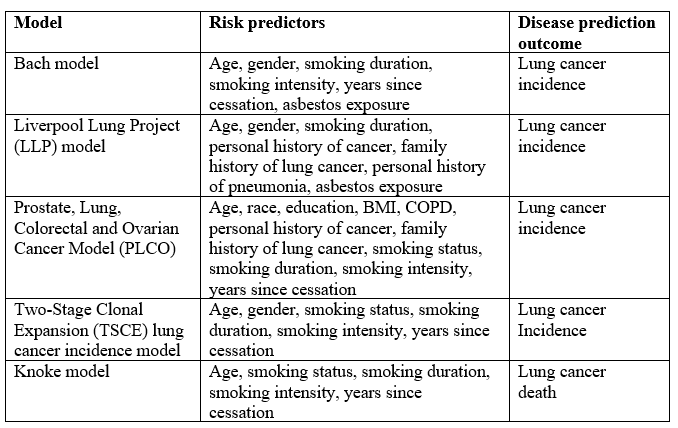
The need for early disease mitigation has resulted in the development of several clinical predictive models over the years to be applied in decision support systems and as predictive calculators for clinicians. Thus, allowing for early follow-up and intervention among at-risk patients. Predictive models for cardiovascular disease have been worked on extensively in literature. Successful models such as Framingham risk models and QRISK have been developed, validated and deployed in clinical practice to predict 10 years risk of cardiovascular disease and thus used to provide early intervention (Ofori and Odia 2016). However, according to (Riley et al. 2016), only a handful of models developed for cardiovascular diseases have undergone external validation. Various lung cancer risk predictors have been developed, especially to screen for the disease (Dubey et al. 2016). In contrast to cardiovascular disease and cancers, only a few predictive models for chronic obstructive pulmonary disease have been developed and according (Bellou et al. 2019) even though the models performed fairly well on their study population, they had methodological insufficiencies, low rate of external validation and assessment for clinical usefulness. (Damen et al. 2016) highlighted the paucity of models developed with datasets from the African population. With a majority of models being developed and validated in European, Northern American and Asian populations. Asthma Predictive Index (API), Asthma Prediction Tool (APT), Asthma Detection and Monitoring (ADM) represents some of the predictive models developed for asthma, using features such as using age, gender, wheezing allergy, family history etc as predictive features (Castro-Rodriguez et al. 2019).

Evidently, multiple binary classifier predictive models exist in the literature for the individual diseases for which tobacco use pose a high risk, most significantly for cardiovascular diseases. These models consider a single task i.e., risk or no risk, disease present or absence, but do not consider other diseases together. Considering diseases of similar risks is of practical importance, as an individual’s characteristic can predispose them to many other diseases. (Wang et al. 2014) highlights this need and states more of such models may be more practical and beneficial for patients. (Bayati et al. 2018) proposed a computationally effective way for multi-disease prediction using disease biomarkers, but not with other characteristics. Thus, there is yet to be at this time, an integrated multiclass model which attempts to predict these diseases using identified risk characteristics, that would be of benefit to tobacco users.

**Table 3. 1 Predictive models for cardiovascular disease adapted from (Yun et al. 2017)**



**Table 3. 2 Predictive models for lung cancer adapted from (Gray et al. 2016: 97)**



# **3.2 MULTI-CLASS CLASSIFIERS**

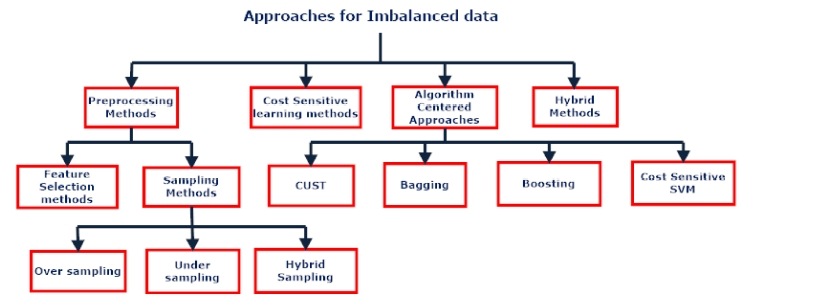
Classification is a common artificial intelligence strategy used to make predictions about observations by assigning them into classes based on pre-assigned rules (Lyu et al. 2020). It is a supervised artificial intelligence technique that makes predictions about new inputs based on learnings from the training dataset (Rocha and Klein Goldenstein 2014). It is suited for the prediction of discrete variables, unlike regression which is adapted for predicting continuous variables. Essentially, the assigned classes might be two, where a binary classifier will be applied or more than two classes which will require a multi-classifier (Lyu et al. 2020). Although most of the vastly used classifiers are binary classifiers, realistically, there exist more multiclassification problems (Grandini et al. 2020). Thus, oftentimes, binary classifiers are optimized to solve multiclass problems by class binarization. However, this might result in inefficiency in the model, especially when dealing with a large number of classes (Zhang et al. 2018). Class binarization essentially breaks down the multiclass problem into several binary classification problems, i.e., using a divide and conquer strategy and combining outcomes by a voting scheme (Fernández et al. 2017). The approaches used for class binarization includes One vs All (OVA), One vs One (OVO), Error correcting output code (ECOC), a combination of OVO and OVA (Fernández et al. 2017).

K-nearest neighbour classifies based on its k nearest neighbours’ points, which is the distance between the test samples and training set (Saadatfar et al. 2020). Majority voting or weighed distance voting is applied to resolve entries with similar nearest neighbours (Cunningham and Delany 2021). It has a fast execution time as it has short training periods (Soofi and Awan 2017). Implementation is easy and it’s not as complicated as other algorithms (Abdulrahman et al. 2020). Its computational cost is large for high dimensional datasets (Saadatfar et al. 2020). Artificial Neural Network (ANN) consists of layers of processing units with weighted connections (Bertolaccini et al. 2017). It has high predictive accuracy and is efficient with high dimensional datasets (Bertolaccini et al. 2017). Predictive outcomes are however vague and difficult to describe (Zajmi et al. 2018). Decision tree consists of nodes, branches and leaves, that are derived from the training dataset, each node is split to form branches with an associated value for the splitting, this is further split into leaves which are associated with the class labels (Bellou et al. 2019). It is simple to understand and easy to implement, with a good interpretation of the relationship between predictor and dataset (Speiser et al. 2019). It however has a problem with overfitting and low accuracy with high dimensional datasets. Random Forest is a type of decision tree classifier that combines tree predictors such that, each tree are made of random independent samples (Abdulkareem and Abdulazeez 2021). It is good with high dimensional datasets, and has better accuracy than Decision tree. It also reduces overfitting and does not require much pre-processing as they are efficient with missing values, outliers (Speiser et al. 2019). It, however, requires higher computational cost, a long training period and poor result performance with unbalanced data (Brieuc et al. 2018). Naive Bayes estimates the probability of class membership by assuming that the observations are conditionally independent. It is efficient in handling noisy data and does not require a large training set, and computational speed is fast (Wickramasinghe and Kalutarage 2020). Model assumption of feature independence by the algorithm can affect predictive accuracy (Wickramasinghe and Kalutarage 2020). Support vector machines classify by generating a separator plane between data points to assign classes in a multidimensional environment. The hyperplane separator concept uses the largest distance analysis to identify classes and reduce error. It is easy to train with no local optimum and scalability to high dimensional data is good (Huang et al. 2018). It requires a good kernel for optimal performance (Huang et al. 2018).

(Moeez M. Subhani and Ashiq Anjum 2020) applied K-nearest neighbour for prediction of diseases from an integrated clinical and genomic dataset, KNN was found to have up to 75% accuracy in making predictions of 75 possible classes. One of the main successful clinical applications of artificial neural networks is in PAPNET, a clinical screening model deployed to assist in the screening of cervical cancer (Thrall 2019). Artificial neural network has been applied in imaging of lung cancer with high accuracy in providing disease prognosis (Huang et al. 2020). (Maheswari and Pitchai 2019) deployed a user-friendly web application based on a model for predicting cardiovascular disease using user risk characteristics with Naive Bayes. They concluded that Naive Bayes was the most effective for predicting heart disease. Similarly, (Wickramasinghe and Kalutarage 2020) reported Naive Bayes in heart disease prediction and found the algorithm performed well when compared with other techniques having high predictive accuracy with high computational speed. (Kour et al. 2020) reported the application of multiclass support vector machine to hyperthyroid detection and classification and found that the one against all support vector machine (OAASVM) had superior accuracy to Decision tree algorithm on the dataset. (Uddin et al. 2019) stated that support vector machines (SVM) were the most widely used algorithm for medical datasets and also concluded that SVM produced better percentage accuracy in classification. However, (Christodoulou et al. 2019) identified the most commonly used artificial intelligence algorithms for disease prediction as Classification tree, Random forest, Support vector machine, Artificial neural network (ANN), and others algorithms such as K-nearest neighbour (KNN), Naive Bayes, genetic algorithm. As such, it is evident that results varied across studies and datasets, with no single algorithm outperforming the others across board. Multiple algorithms have thus been applied to predict diseases in literature. This, for some, raises concern of overdiagnosis, an effect that may arise on the opposite spectrum of early detection of disease and thus may run the risk of eventually increasing health costs (Delpino et al. 2022).

The ideal algorithm to apply does not exist, and the choice and relative superiority of one model over the other relies on factors such as the type of data, the performance criteria and the expertise of the researcher using the chosen model, the time available for development and training, and the nature of the classification problems (Rácz et al. 2019). High diagnostic accuracy, efficiency in dealing with missing and noisy data, ease of understanding and explaining generated results while using a minimum amount of data, were identified as some of the desirable characteristics of artificial intelligence algorithms for clinical use (Alsharqi et al. 2018). (Christodoulou et al. 2019) concluded that most research that applied artificial intelligence to predict clinical conditions had suboptimal model performance validation, model development methodology, clinical utility assessment and head-to-head comparison. (Palazón‐Bru et al. 2020) also identified bias in the handling of missing data, sampling and analysis in models developed to predict Oro-pharyngeal cancers. There is an overlap between traditional statistical modelling and artificial intelligence techniques. Sometimes artificial intelligence models only provide theoretical superiority over traditional statistical modelling methods, as when applied in real-world situations the difference in performance is only marginal (Kagiyama et al. 2019). He justified his arguments by citing several examples including the fact that the benefit of fitting a dataset diminishes as the process becomes more complex, as is the case with most artificial intelligence models. Thus, whatever predictive superiority of artificial intelligence models was lost when applied to the population due to the changing characteristics of the population. This, among other reasons, was said to weaken the claim for the superiority of artificial intelligence over conventional statistical modelling for risk prediction. As such, this has created some doubt among health professionals and policymakers on the usefulness of their adoption. Even though studies such as (Weng et al. 2017) prospective cohort study to assess if artificial intelligence can improve cardiovascular risk prediction using clinical features in data, applying binary classifier algorithms with random forest, logistic regression, gradient boosting and neural network, showed improved accuracy of cardiovascular risk prediction when compared to the American College of Cardiology guidelines for predicting the first cardiovascular event over 10 years. Thus, this greatly increases the number of patients that would benefit from preventive care. (Alaa et al. 2019) also performed a similar study and applied support vector machines (SVM), random forest, neural networks and gradient boosting machines to predict cardiovascular disease, and found the model improves the accuracy of cardiovascular risk prediction when compared with currently recommended clinical guidelines. (Seetharam et al. 2019) concluded that artificial intelligence algorithms generally performed better than conventional techniques in predicting adverse outcomes in cardiovascular medicine. Essentially, applying easily interpretable and reproducible methodology when developing and reporting predictive models might help improve their level of acceptance within the health community.

Data imbalance is frequently encountered in most real-world datasets, especially in medical datasets (Boughorbel et al. 2017). Imbalanced data occurs when the members of one class grossly outnumber the members of the other classes, thus making it difficult for the algorithm to learn from the minority set, and it thus shows bias towards the majority set (Grandini et al. 2020). Thus, algorithm performance is determined by majority classes. According to (Leevy et al. 2018), class imbalance beyond a ratio of 50: 1 of the majority to minority classes is considered a high-level class imbalance. Different approaches have been reported to handle the problem of data imbalance. They include, either applying pre-processing techniques, algorithmic learning modification with cost-sensitive solutions by applying a penalty to miss-classification, or a combination of various classifiers (Fernández et al. 2017). Pre-processing involves implementing change directly in the dataset to balance the ratio of the classes. It may involve applying sampling manipulation of the dataset such as over-sampling of the minority class or under-sampling of the majority class on the training dataset (Kaur et al. 2019). The algorithm learning technique modifies the algorithm to accommodate the skewed data by removing bias towards majority classes. The cost-sensitive technique applies misclassification costs to each class, while ensemble classifiers are a combination of multiple classifier systems, that aims to improve performance (Tanha et al. 2020). Bagging and Boosting are the most common ensemble techniques (Krawczyk 2016). (Fotouhi et al. 2019) applied over-sampling techniques such as Synthetic minority over-sampling technique (SMOTE), Adaptive synthetic Sampling (ADASYN), and Selective Pre-processing of Imbalanced Data as well as under-sampling techniques such as Condensed Nearest neighbour(CNN), Neighbourhood Cleaning Rule (NCL), One Sided selection( OSS), Random under-sampling (RUS) etc. on 18 imbalanced datasets and found that there was a general improvement in the performance of the classifier with the application of these techniques with over-sampling pre-processing techniques producing the best result. According to (Fernández et al. 2017) RUS and SMOTE are the most commonly used techniques in research. Cost-sensitive techniques when compared with over-sampling techniques, showed no significant difference in performance (Leevy et al. 2018).



**Figure 3. 2 Approaches for Data Imbalance (Kaur et al. 2019: 13)**

For parametric algorithms such as Naive Bayes, Support vector machines and least absolute shrinkage (LASSO), hyperparameter tuning is an important optimization approach to get the best result from the chosen model. It optimizes the objective function on the training set while learning (Raschka 2018). Thus, balanced data and optimized algorithms are essential to good performing algorithms.

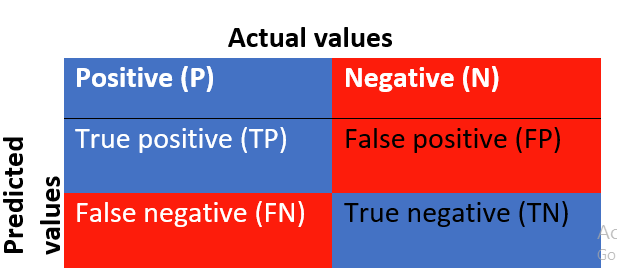
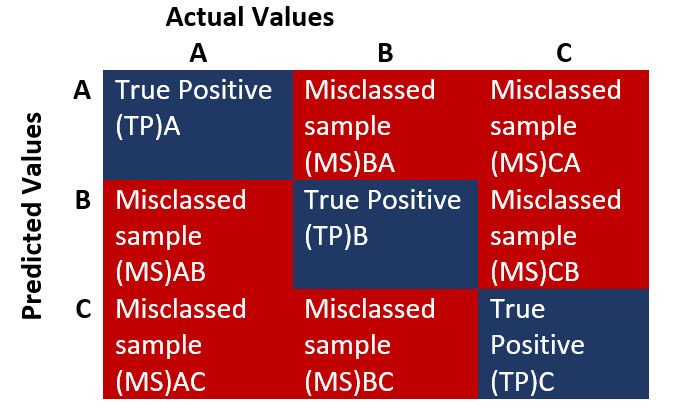
# **3.3 ALGORITHM PERFORMANCE**

The output of predictive models needs to the assessed and interpreted to evaluate their performance (Tharwat 2018). This will help determine how far results from the model are from actual results and make a generalisation about how the model will perform on unseen data. The performance measure is also used to compare algorithms and select the best-suited model for the classification problem (Raschka 2018). Cross-validation is an important technique to assess model generalizability i.e., how it can fit other samples from the same population and thus ensures the trained model is not overfitted. This essentially involves splitting the data set into a training set and then, fitting the training set to the model to obtain parameters, this is then followed by applying the model to make predictions on the test set and results aggregated (Song et al. 2021). *k*-fold cross-validation and stratified *k*-cross-validation are the two most common cross-validation techniques. In the *k*-fold technique, the dataset is split randomly into *k* equal parts. Each time one of the subsets is used as a test set and the rest for training. The training and testing would be carried out for *k* rounds (Jiao and Du 2016). Stratified cross-validation is similar to *k* fold cross validation however predefined proportion of data points are present in each subset (Song et al. 2021).



**Figure 3. 3 Cross validation (Jiao and Du 2016: 323)**

The parameters available for assessing the quality of a classification model can often give conflicting results (Rácz et al. 2019). Matrices such as accuracy, sensitivity and specificity can be calculated from a confusion matrix which represents the values of predictive outputs (Tharwat 2018).

**Figure 3. 4 Confusion matrix adapted from (Tharwat 2018: 170)**

True positive (TP) represents the correctly classified positive output, False negative (FN) or type II error represents a wrongly classified positive output, True negative (TN) represent a correctly classified negative output, False positive (FP), or Type I error represents a wrongly classified negative output.

Even though these are easy to use, it runs the risk of being sensitive to imbalanced data and might not represent the true performance of some classes. Data imbalance may affect the choice of metrics to be used. When making predictions the cost of misclassification is an important factor to consider. Imbalance classes can contribute significantly to misclassification. A false negative classification which states the absence of disease risk when it is actually present could be more harmful to the patient in the long run as opposed to a false positive misclassification which would mean only more medical investigations for the patient (Haixiang et al. 2017).

Performance can be assessed in terms of overall performance, discrimination or calibration.

Accuracy is a commonly adopted metrices, it measures overall performance by determining the number of correctly classified observations to the total observations. This means the high performance of the majority class can mask the poor performance of the minority class and thus not suitable for imbalanced data (Fernández et al. 2017).

Discrimination describes how well the model differentiates event from non-event, while calibration refers to the agreement between outcome and the prediction and are used for external validation of models (Alba et al. 2017). Graphical performance methods such as Receiver operating characteristics (ROC) and precision-recall curve can be applied for discrimination performance measures (Pak and Oh 2016). (Rácz et al. 2019) concluded in their comparison of performances metrics for artificial intelligence models that diagnostic odd ratio, Area under the ROC curve and markedness (MK) as the best options, showing the best consistency in performance. According to (Fleuren et al. 2020), the area under the receiver operating characteristics (AUROC) curve was used in most studies to assess the performance of clinical predictive models. Essentially, performance metrics are important tools used to evaluate the quality of the model developed. Most studies apply a combination of the above-mentioned metrics in assessing model success, and also in comparative research.

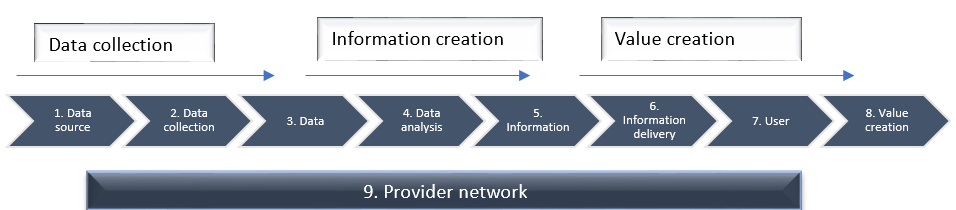
# **3.4 RESEARCH GAP**

The significant research gaps include;

1. Lack of multi-disease classification models for tobacco-related diseases that uses patients’ characteristics which could be of benefit to tobacco users to identify and stay abreast of their disease risk class allowing for early follow-up, and possible intervention.
2. Paucity of models developed from datasets from African populations.
3. Need for improved research methodology, reporting as well as model performance comparison.

# **3.5 THEORETICAL FRAMEWORK**

Predictive models for diseases are of high value, not just to patients but also to clinicians, lawmakers and the larger society. The value co-creation theory explains how creating value is a shared responsibility of all stakeholders (Vargo and Lusch 2007). As such, in disease management both patients and clinicians share information and knowledge to arrive at the right conclusion and thus make the best decision. Patients are now more involved and interested in taking ownership of their health. The shared decision-making theory in health (SDM) is built on a similar philosophy, allowing for patient involvement in making decisions after evaluating benefit and also based on their values and belief. In healthcare, this is achieved by providing decision aids that assist patients to understand information about the risks and benefits (Lillie et al. 2014). (Lim et al. 2018) on the bases of the concept of value co-creation by (Vargo and Lusch 2007), proposed a 9 step framework, applying data value chain to maximise benefit from data generated from various sources in an information-intensive service. The framework highlights activities and resources to be managed to create value from data, with the dynamic interaction between providers and users an essence in meeting that objective.



**Figure 3. 5 Data Value Chain framework for data based value creation adapted from (Lim et al. 2018: 126)**

# **3.6 SUMMARY**

This literature review has highlighted the disease burden associated with tobacco use, which will be useful in making class labels. It also shows the commonly used predictors in literature for these diseases as this can be applied during the process of feature selection.

Furthermore, it identifies the various multi-classifiers, their strength and weaknesses. This is important when choosing an algorithm, and ensures comprehensive comparisons can be made about results from them.

# **CHAPTER 4**

# **FINDINGS**

This Chapter presents the research findings based on the methodology described in chapter 2. Results from data processing and cleaning, exploratory analysis, feature selection, and model implementation are highlighted.

# **4.1 DATA PROCESSING**

The combined datasets of smokers in the BRFSS survey from 2016 to 2019 was 2,079386 with 571 columns of Variables. Initial data processing to drop irrelevant variables not related to this research was done. Further data cleaning was performed and missing values, those without a reported diagnosis of target diseases were dropped. This reduced the sample to 79,4378 with 19 columns. The variables were mostly nominal and ordinal categorical variables. Table 4.1 shows the percentage of missing values in the dataset. Missing values in Gender made up more than half of the dataset (62.8%), although it is an important variable, the column was dropped so that it does not significantly reduce the sample size. Body mass index (BMI), which is a binned numerical variable, had the second-highest percentage of missing data, these missing values were replaced with the median BMI. Other missing entries were dropped.

**Table 4. 1 Missing values**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Missing Values** | **Percentage (%)** |
| Gender | 1306273 | 62.8 |
| BMI | 137420 | 6.6 |
| Exercise | 21097 | 1 |
| Coronary heart disease | 18442 | 0.9 |
| Current smokeless tobacco | 2212 | 0.1 |
| Check-up | 828 | 0 |
| Marital status | 37 | 0 |
| General Health | 35 | 0 |
| Employment | 8 | 0 |
| Heart Attack | 7 | 0 |
| Memory impaired | 6 | 0 |
| Home type | 5 | 0 |
| Disability (independent activity) | 4 | 0 |
| Ever smoked | 4 | 0 |
| Disability (Self-care) | 3 | 0 |
| Stroke | 3 | 0 |
| Race | 3 | 0 |
| Disability(walking) | 2 | 0 |

# **4.2 EXPLORATORY ANALYSIS**

Descriptive statistics showing the final columns selected, their number of unique entries and their count are shown in Table 4.2.

**Table 4. 2 Variable description**

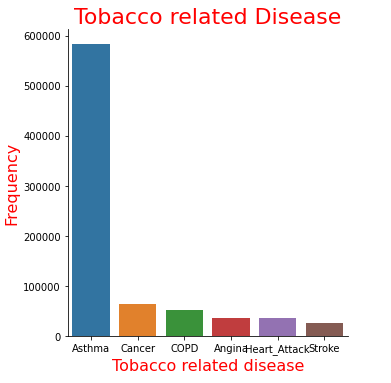
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variables** | **Count** | **Unique** | **Top entries** | **Frequency** |
| General Health | 19583 | 5 | Good | 7972 |
| Check-up | 19583 | 5 | Less than 1 year | 14780 |
| Exercise | 19583 | 2 | Yes | 12192 |
| Marital status | 19583 | 6 | Married | 6828 |
| Employment | 19583 | 8 | Employed | 7317 |
| Memory impaired | 19583 | 2 | No | 14604 |
| Disability(walking) | 19583 | 2 | No | 13804 |
| Disability (Self-care) | 19583 | 2 | No | 18375 |
| Disability (independent activity) | 19583 | 2 | No | 16703 |
| Health state | 19583 | 2 | Good | 12399 |
| Physical health | 19583 | 1 | >13 days of good health | 19583 |
| Mental Health | 19583 | 3 | Zero days with no good MH | 8074 |
| Race | 19583 | 5 | White | 15507 |
| Age | 19583 | 6 | 55 to 64 | 5315 |
| BMI | 19583 | 4 | Normal weight | 6437 |
| Education | 19583 | 4 | Graduate high school | 7433 |
| Smoking status | 19583 | 1 | Current smoker everyday | 19583 |
| Current smoker | 19583 | 1 | Yes | 19583 |
| Diseases | 19583 | 6 | Asthma | 12228 |

Table 4.3 shows the prevalence rate of each disease class. Asthma had the highest prevalence rate of 73% among tobacco users. Also significant was that class membership increases with age group and weight.

**Table 4. 3 Disease prevalence in the sample population**

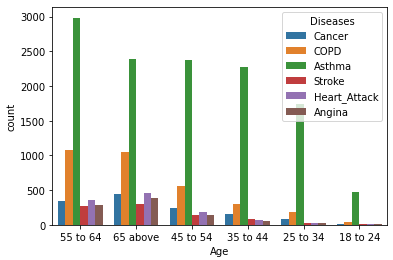
|  |  |
| --- | --- |
| **Diseases** | **Prevalence (%)** |
| Asthma | 73 |
| Cancer | 7.9 |
| COPD | 6.4 |
| Angina | 4.6 |
| Heart Attack | 4.4 |
| Stroke | 3.2 |

Figure 4.1 shows the imbalanced distribution of the data set. Assessing the distribution of variables shows the dataset to be imbalanced, with Asthma as the majority class and had an imbalance proportion of 10:1 compared with other classes.

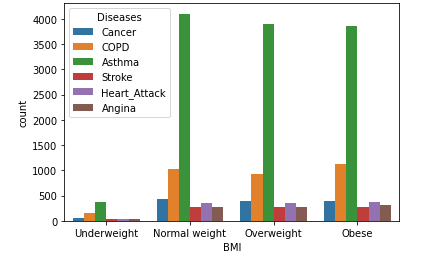


**Figure 4. 1 Imbalance distribution of dependent variables**

The grouped column chart in Figure 4.2 and Figure 4.3 shows that the number of individuals with these diseases increased with age group and BMI class.



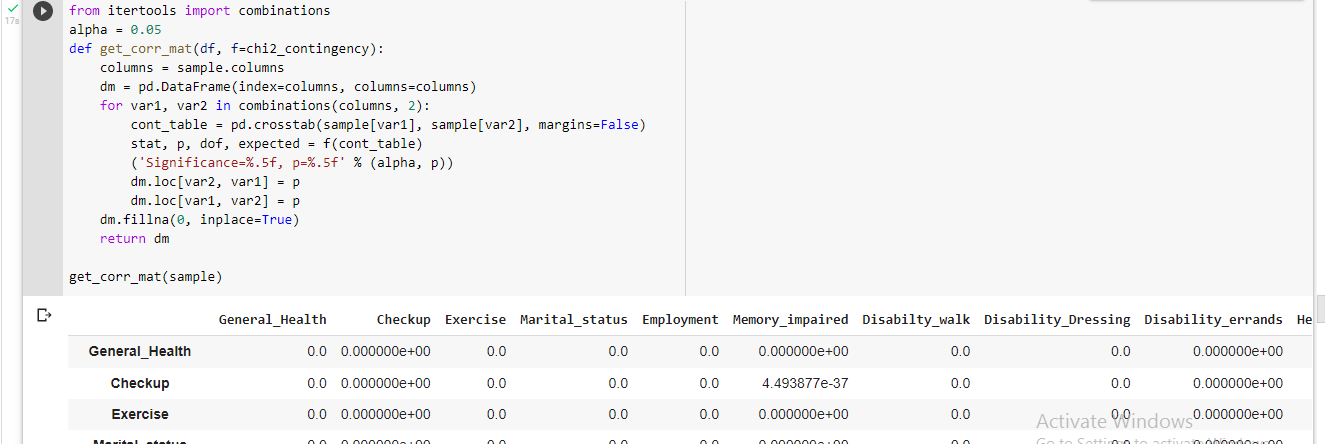
**Figure 4. 2 Distribution of disease class among age groups**



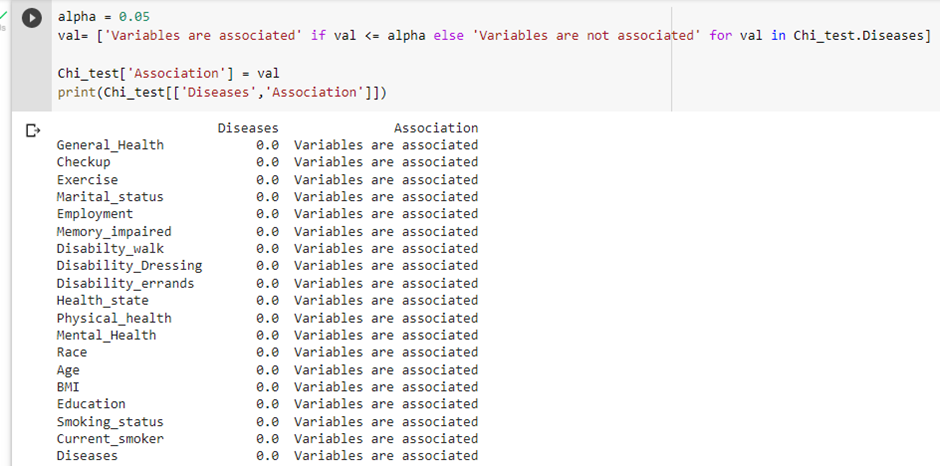
**Figure 4. 3 Distribution of disease class by weight**

Bivarent analysis with Pearson Chi-square was done as shown in Figure 4.4. Chi-square tests the association between categorical variables by determining if there is a significant difference between expected frequencies and the observed frequencies among categories. It is used to select features based on association with the target classes. The null hypothesis(H0) is that there is no association between the features and the target classes. The alternate hypothesis is that there is an association between the features and the target variable (H1). The p-value indicates the level of significance of the result. P-value less than alpha 0.05 indicates that we can reject the null hypothesis with a confidence interval of 95%. The Chi-square results in Figure 4.5 shows an association between features and the target variable.

**Equation 4. 1 Chis-square formula**



**Figure 4. 4 Python implementation of Chi-square**



**Figure 4. 5 Associated variables from Chi-Square test**

# **4.3 FEATURE SELECTION**

Feature selection was done by applying the least absolute shrinkage (LASSO) as shown in Figure 4.6. LASSO applies a penalty to non-significant variables and shrinks their coefficients to zero. A grid search is performed to find the best value for the parameter alpha. An alpha value of 0.1 was determined and applied in the LASSO equation. Non zero coefficients are features are selected as the best features to apply to the model as shown in Table 4.4.

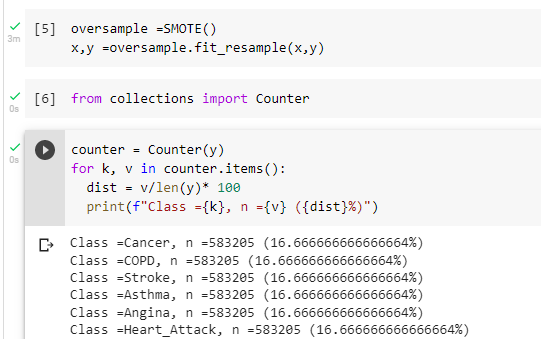


**Figure 4. 6 Python implementation of LASSO**

**Table 4. 4 Least absolute shrinkage**

|  |  |
| --- | --- |
| **LASSO Coefficient = 0** | **Features** |
| FALSE | General Health |
| FALSE | Check-up |
| FALSE | Exercise |
| FALSE | Marital status |
| TRUE | Employment |
| TRUE | Memory impaired |
| TRUE | Disability(walking) |
| FALSE | Disability (Self-care) |
| TRUE | Disability (independent activity) |
| FALSE | Health state |
| FALSE | Physical health |
| FALSE | Mental Health |
| FALSE | Race |
| TRUE | Age |
| FALSE | BMI |
| TRUE | Education |
| FALSE | Smoking status |
| FALSE | Current smoker |

Synthetic minority over-sampling technique (SMOTE) was applied to add synthetic data points to minority classes, thus balancing the data as shown in Figure 4.7. This brings all classes to have a proportionate distribution of 16.6%.



**Figure 4. 7 SMOTE implementations to balance classes**

**4.4 EVALUATION**

The confusion matrix is adopted to evaluate the results of the models. Precision score, F score and Recall are calculated. The models were trained with both features selected by LASSO and also with features segmented into categories of Socio-demographic factors (Age, Marital status, employment, BMI, Education), comorbidity (Disabilities, Memory impairment), Health state (General Health, Physical health, Mental health), lifestyle factors (Exercising, Check-up, smoking status). Similar results were obtained. Across the applied classifier algorithms, the predictive performance was poor with all algorithms as shown in the Tables below. Random Forest had the best result as shown in Table 4.5 with an average F score (0.40), Precision (0.40), Recall (0.28). Navies Bayes has the worse result as shown in Table 4.8. F-score (0.25), Precision (0.27), Recall (0.28). The order of performance in decreasing order are Random Forest, K-nearest neighbour (KNN) Support Vector Machine, Naive Bayes.

**Table 4. 5 Result for Random Forest**

|  |  |  |  |
| --- | --- | --- | --- |
| **Classes** | **Precision** | **Recall** | **F-score** |
| Angina | 0.30 | 0.40 | 0.35 |
| Asthma | 0.59 | 0.46 | 0.52 |
| COPD | 0.44 | 0.44 | 0.44 |
| Cancer | 0.33 | 0.40 | 0.36 |
| Heart Attack | 0.37 | 0.29 | 0.32 |
| Stroke | 0.43 | 0.42 | 0.43 |
| Accuracy |  |  | 0.40 |
| Macro average | 0.40 | 0.40 | 0.40 |
| Weighted average | 0.40 | 0.40 | 0.40 |

**Table 4. 6 Result for K-nearest neighbour (KNN)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Classes** | **Precision** | **Recall** | **F-score** |
| Angina | 0.27 | 0.39 | 0.32 |
| Asthma | 0.45 | 0.44 | 0.44 |
| COPD | 0.38 | 0.4 | 0.39 |
| Cancer | 0.3 | 0.27 | 0.29 |
| Heart Attack | 0.31 | 0.26 | 0.29 |
| Stroke | 0.39 | 0.31 | 0.35 |
| Accuracy |  |  | 0.35 |
| Macro average | 0.35 | 0.35 | 0.34 |
| Weighted average | 0.35 | 0.35 | 0.34 |

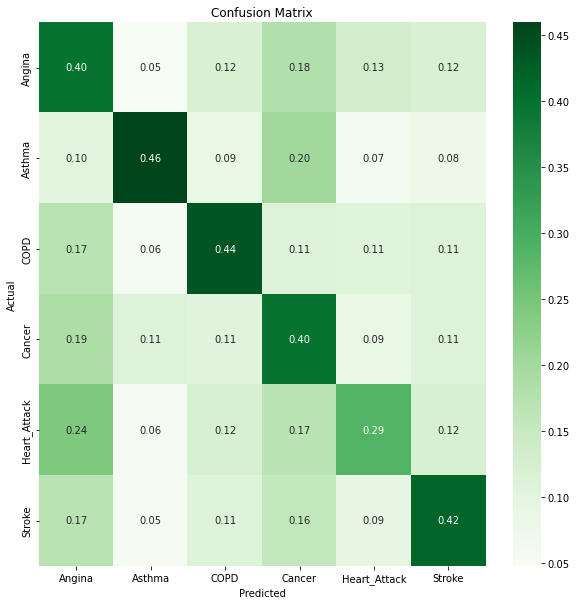
**Table 4. 7 Result for SVM**

|  |  |  |  |
| --- | --- | --- | --- |
| **Classes** | **Precision** | **Recall** | **F-score** |
| Angina | 0.23 | 0.26 | 0.25 |
| Asthma | 0.38 | 0.61 | 0.47 |
| COPD | 0.29 | 0.39 | 0.33 |
| Cancer | 0.24 | 0.3 | 0.27 |
| Heart Attack | 0.22 | 0.05 | 0.08 |
| Stroke | 0.26 | 0.11 | 0.16 |
| Accuracy |  |  | 0.29 |
| Macro average | 0.27 | 0.29 | 0.26 |
| Weighted average | 0.27 | 0.29 | 0.26 |

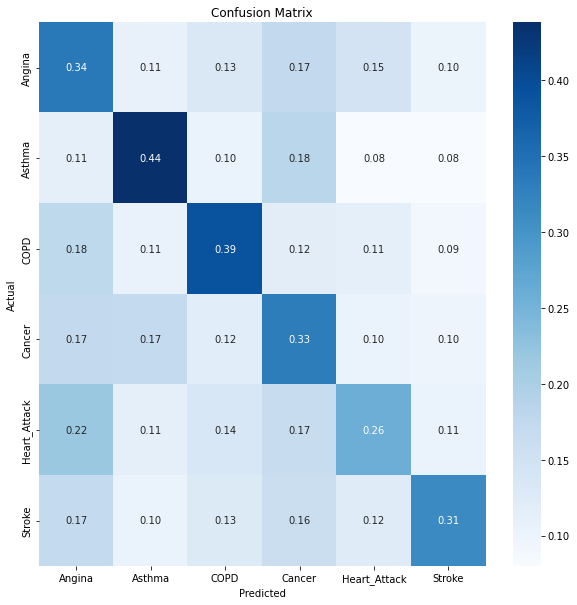
**Table 4. 8 Result for Naive Bayes**

|  |  |  |  |
| --- | --- | --- | --- |
| **Classes** | **Precision** | **Recall** | **F-score** |
| Angina | 0.22 | 0.25 | 0.25 |
| Asthma | 0.41 | 0.5 | 0.47 |
| COPD | 0.3 | 0.34 | 0.33 |
| Cancer | 0.22 | 0.45 | 0.27 |
| Heart Attack | 0.23 | 0.02 | 0.08 |
| Stroke | 0.25 | 0.11 | 0.16 |
| Accuracy |  |  | 0.28 |
| Macro average | 0.27 | 0.28 | 0.25 |
| Weighted average | 0.27 | 0.28 | 0.25 |

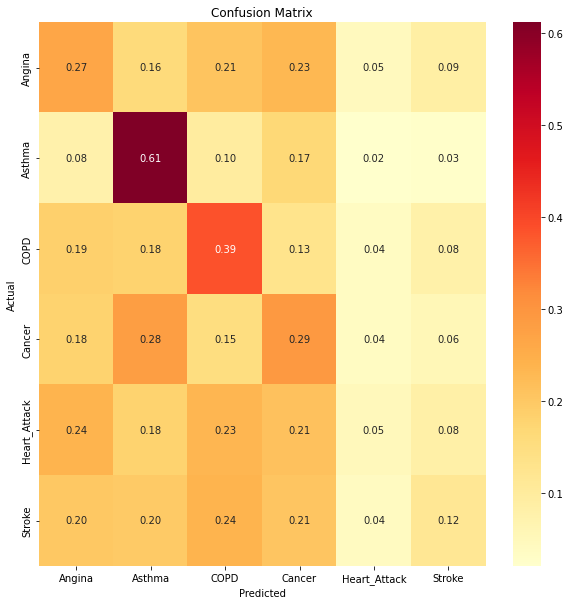
The confusion matrix plots show the proportion of true positive results on the diagonal axis i.e. The proportion of predicted labels that were actually correctly classified, with the other segments showing the proportion of misclassified entries. It can be seen that the Random Forest classifier in Figure 4.8 correctly classified Asthma, COPD, Stroke and Angina best, and Heart attack the least. Figure 4.9 showed KNN classified Asthma most accurately and Heart attack the least. A similar observation is made of Naive Bayes in Figure 4.11.



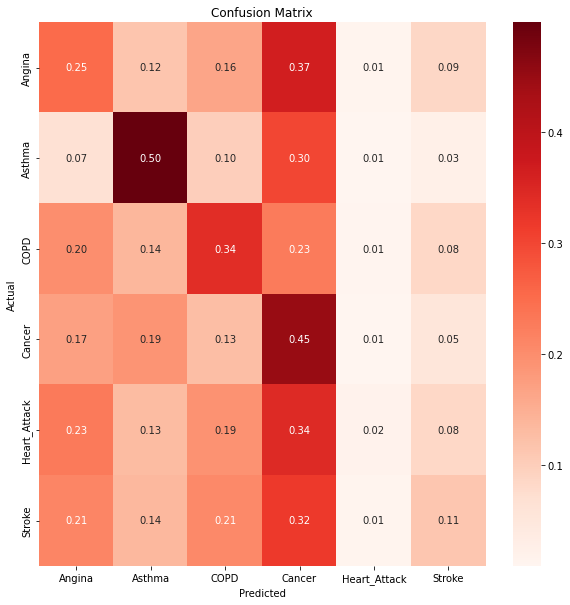
**Figure 4. 8 Confusion matrix Random Forest**



**Figure 4. 9 Confusion matrix K-nearest neighbour (KNN)**



**Figure 4. 10 Confusion matrix Support vector machine**



**Figure 4. 11 Confusion matrix Naive Bayes**

**Table 4. 9 Comparison of algorithm performance**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Performance Metrics** | **Random Forest** | **K-nearest neighbour** | **Naive Bayes** | **Support vector machine** |
| Precision (%) | 40 | 35 | 27 | 27 |
| Recall (%) | 40 | 35 | 28 | 29 |
| F-score | 40 | 34 | 25 | 26 |
| Accuracy | 40 | 35 | 28 | 29 |
| Training time(hours) | 0.3 | 0.3 | 0.02 | 1 |

As shown in Figure 4.12. Random Forest was the best performer across the metrics, closely followed by K-nearest neighbour, with Support vector machine taking the longest to test and train.

**Figure 4. 12 Performance of Predictive algorithms**

**4.5 SUMMARY**

The research findings have been presented in this chapter. The result has shown that the dataset had a large percentage of missing features resulting in potentially useful predictors being dropped. The features applied to the classifier algorithms produced poor predictive outcomes in all algorithms. Precision, Recall and F scores were applied as performance evaluation metrics and showed that Random Forest was the best multi-classifier and Naive Bayes was the worst on this dataset.

# **CHAPTER 5**

# **ANALYSIS AND CONCLUSION**

This chapter discusses the findings from applying selected features to make multiclass predictions of tobacco-related diseases using different artificial intelligence algorithms. It analyses the research results and how it meets the research objectives, presents the challenges and limitations of this research, and draws a conclusion that answers the research question and meets the aim of this research with appropriate recommendations.

# **5.1 ADOPTED APPROACH**

The research was based on a quantitative cross-sectional study on a secondary dataset. The computational nature of this research made the quantitative approach the best fit. This approach offered the opportunity to provide new insights following a laid down framework of analysis. Thus, the result can be quantifiable, with the ability to make a generalisation about the population of study. The research made use of secondary data. This previously collected open-source data provided an opportunity to utilise a large sample for this study while saving on time and resources. However, because the data was not collected primarily for the purpose of this research, not all the ideal features were present in the dataset. Even with its limitations, according to (Johnston 2017), secondary data remains viable for research inquiry. The validated guidelines for making disease prediction using artificial intelligence algorithms proposed by (Steyerberg and Vergouwe 2014) was applied. The data processing step highlighted a significant number of missing values and non-respondents, which affected the final sample size obtained, and features used. Missing data affect analytics results, and consequently, leads to findings that could be different from an otherwise complete dataset (Sidi and Harel 2018). Although the dataset contained tobacco-related diseases present in literature, such as Cardiovascular disease, Stroke, Asthma, Cancer, Chronic obstructive pulmonary disease (COPD) (West 2017), the dataset was extremely imbalanced with the prevalence of Asthma in the study sample highest (73%). Data imbalance can be problematic as this can result in bias of findings towards the dominant class (Kaur et al. 2019). Synthetic minority over-sampling technique (SMOTE) was applied to mitigate this problem by synthetically balancing out the classes (Prusty et al. 2017). The Dataset was mostly nominal categorical entries. This then raises the need to encode labels. The Label encoding technique was adopted. Although a One-hot encoder was the best option, because it replaces each value with a unique numerical value thus, allowing the algorithm to learn each class member as an independent value rather than in a hierarchy, while the Label encoder replaces values with hierarchical numeric values thus the machine see values of each class member in hierarchy. Computational limitation and runtime complexity, due to the size of the dataset made adopting the One-hot encoder an impractical choice (Hancock and Khoshgoftaar 2020). Feature selection techniques used were Chi-square and Least absolute shrinkage (LASSO). Chi-square is the bivariant analysis of choice for categorical variables and can determine the association between features while being insensitive to the nature of data distribution (Kamala and Thangaiah 2019). LASSO was also applied as it helps to further eliminate various irrelevant features (Khaire and Dhanalakshmi 2019). Google Pro-lab tensor processing unit was the adopted cloud computing environment with the average training and testing time for each algorithm was about 0.4 hours.

# **5.2 ANALYSIS**

The study was directed towards answering the research question, to what extent can artificial intelligence techniques be used for multiclass prediction of tobacco-related diseases among users based on their risk characteristics? With the aim of applying these algorithms for the prediction of tobacco-related diseases among tobacco users. The algorithms applied to achieve this were Random Forest, Support vector machines, Naive Bayes and K nearest neighbour. The result showed a high prevalence rate of Asthma among the study population (73%), showing an increasing trend with age group. (Stern et al. 2020) highlighted the varying trend of Asthma prevalence among varying populations globally, however, concluded that the highest prevalence was noted in children and young adults in the United States. (Alavinezhad and Boskabady 2018) reported a high prevalence of asthma among smokers in middle eastern countries. (Zacharasiewicz 2016) reported the effects of smoking in the development and severity of Asthma in children and adolescents, which was primarily attributed to the effect of nicotine on the developing lungs. Similarly, (Xian and Chen 2021) found a significant association between the use of electronic cigarettes and asthma especially in young adults, with similar findings among those who combined with traditional cigarettes. Thus, this research findings on asthma prevalence are not completely inconsistent with those found in literature as asthma’s prevalence has been shown to vary among populations even though it has been found to be more common in the younger population globally which is not similar to the findings in the research sample. Although asthma may be more prevalent among smokers, mainly due to allergic reactions of the lungs to tobacco products, it isn’t the major cause of death or morbidity among these populations. Cancer, Cardiovascular disease and Chronic Obstructive pulmonary disease were reported to with the highest morbidity and mortality rates in these populations (Fleuren et al. 2020). As such, the high prevalence rate of asthma combined with the reported morbidity and mortality of other tobacco-related diseases made them selected as significant predictive outcomes for the study.

From the available features in the dataset, Chi-square result showed an association between age, body mass index, state of general health, check-up frequency, Exercise frequency, Marital status, Employment status, memory impairment and other disabilities such as walking, self-care, independent activity, physical and mental health state, race, education, and smoking status. Out of these features, those selected by Least absolute shrinkage (LASSO) included employment, memory impairment, disability (walking), disability (independent activity), age, body mass index. And these were applied to make predictions. These features show a slight variation from those used in other successful models that predicted tobacco-related diseases. (Van Bussel et al. 2020) highlighted features such as gender, age, smoking, blood pressure, blood lipids measurements and body mass index for cardiovascular disease prediction. (Gray et al. 2016) applied features such as ethnicity, and family history of the disease, and presence of other chronic diseases, social deprivation, asbestos exposure, duration of smoking. Tobacco smoking, asthma, environmental exposures to toxic fumes and alpha-1-antitrypsin deficiency average amount smoked per day, and duration of abstinence from smoking for former smokers were features used to predict chronic obstructive pulmonary disease (Matheson et al. 2018). Thus, no definitive set of features exists that has been used as different studies have adopted different features. Some studies have combined both user characteristics, clinical findings as well as investigative findings to achieve results.

The predicative performance of Random Forest, K-nearest neighbour (KNN), Support Vector Machine, Naive Bayes showed poor results in Precision, Recall and F-scores. Random Forest showed the best performance of these algorithms with 40% accuracy. Thus, these algorithms could not reliably predict each user to their actual classes using the selected features. These findings are in contrast to those obtained in literature where good predictive results were reported especially for cardiovascular disease. Most of these models, however, were binary classifiers on clinical trial participants (Yun et al. 2017). (Haq et al. 2018) achieved good results when applying Decision tree, KNN and ANN, SVM and Naive Bayes as a binary classifier on public repository dataset from Cleveland Heart Disease database, to predict presence or absence of heart disease, with SVM as best classifier closely followed by Naive Bayes using features such as age, gender echocardiography and fluoroscopy results as predictive features. (Mustaqeem et al. 2018) reported good multiclass results with SVM while predicting different types of cardiac arrhythmias using electrocardiogram (ECG) with an accuracy of 81%. (Alam et al. 2018) applied binary SVM classifier for multiclass prediction of lung cancer using image investigations with 97% accuracy. While comparing algorithms used for disease prediction (Uddin et al. 2019) found that although SVM was the most frequently employed, Random forest gave the best results. Unsupervised learning especially deep learning convolution neural networks has also be applied for complex classification problems involving medical imaging (Litjens et al. 2017). Notably, these results are based on predictions of individual diseases, working on one disease at a time and not on multi-disease prediction. They also applied a combination of both patients’ characteristics and investigative findings. (Arumugam et al. 2021) indicated that decision tree models produce better results than SVM and Naive Bayes while making multi-disease predictions when applied to make predictions of diabetes-related heart diseases on the Cleveland Heart Disease database, which is an open-source dataset, with good results obtained. Essentially, multiple success stories exist for models making multiclass classification with good performance. However, most of these were based on single disease classification, clinical trial dataset, and utilised features that were a combination of patient’s characteristics as well as investigative findings. Multi disease, multiclassification results obtained from the open-source Cleveland Heart Disease database also had good results as the dataset had more extensive features than was obtainable in the BRFSS dataset.

The results from applying artificial intelligence to predict tobacco-related disease using features in the BRFSS are not good enough to be proposed for external validation, and possible application to make predictions that could be added into decision support systems, predictive calculators for clinicians or to provide patients with insights on their health based on their characteristics.

# **5.3 CONCLUSION**

This research was intended to determine the extent to which artificial intelligence algorithms can make predictions of tobacco-related diseases using risked characteristics. This research was based on the value co-creation framework (Lim et al. 2018), that utilises a data value chain from data-intensive services like health, to create benefits for both clinicians and tobacco users. For clinicians, it could serve as a predictive calculator and support decision making, and for tobacco users to allow them to stay informed about diseases they might be at risk of due to their characteristics. All in all, it can be concluded that these algorithms performed poorly when applied to make multiclass predictions of tobacco-related diseases on the open-source BRFSS dataset. This implies the need for further studies with more robust features in a high-quality dataset needs to be carried out before these techniques can be recommended for practical application. Such high-quality data as described by (Kodra et al. 2017) refers to the available features present in the dataset that are able to meet the needs of the intended research. Stating that some of the desirable characteristics of such data include its accessibility, completeness, validity, coherence and timeliness. The research objectives were met as Asthma was identified as the most prevalent tobacco-related disease in the study population. (Kuruvilla et al. 2019) considers Asthma to be the most prevalent respiratory disease in the United States, with an increasing incidence rate. The other tobacco-related diseases presented in the dataset were also used as predictive outcomes as they were identified to have high morbidity and mortality rates from literature (Fleuren et al. 2020). The risk characteristics for prediction were selected using a combination of the Chi-square test and LASSO. Support Vector Machine, Random Forest, Naive Bayes and K-nearest neighbour were trained and tested with the selected features. The Precision, Recall, and F-score showed poor performance of these algorithms’ ability to make multiclass predictions of these diseases with the available features in the dataset. Comparison of the performance showed random forest had the best performance of these algorithms and a good option to be adopted in multiclassification problems (Uddin et al. 2019).

# **LIMITATIONS**

The limitations of this study include the amount of missing and non-respondents in the dataset which prevented the use of some valuable features. Although applying secondary data enabled access to a large sample of participants, the data was not initially collected for this research, thus contributing to the amount of missing data as well as, a limited number of features. Diseases used as the dependent variables were based on respondents’ responses, with no way to validate if these responses were definitive clinical diagnoses. In addition, there is no way to be certain how truthful respondents were on other responses. The sampling technique adopted was based on convivence to the available samples and may not be truly representative of the population. Computational limitations prevented some best practices to be applied e.g., using Label encoder instead of one hot encoder. Computational limitation was also a significant factor in the choice of algorithms applied.

# **5.5 FURTHER RESEARCH**

A high-quality primary dataset that would represent the study population more accurately could be used for future studies, with features that are more closely related to each of the outcome variables collected. Future research may also include the understudied African population. Other supervised learning algorithms such as Decision trees, Xgboost, Gradient boosting as well as unsupervised algorithms like deep learning could be applied, and a comparison of their performance made, with several iterations and hyperparameter tuning performed.

# **5.6 REFLECTION**

The study has highlighted some limitations of artificial intelligence in making multiclass predictions of different diseases using tobacco users’ characteristics on the BRFSS dataset. It also showed that Random Forest is the best option for disease classification. Highly accurate predictions depend both on the quality and quantity of the data provided to make them. It is thus not practical at this time to provide this as a solution for tobacco users to use to predict the disease they might come down with, based on their characteristics or be used by clinicians. Further research with better quality datasets and more advanced algorithms may show promising results in the future.

Reflecting on my personal journey with this research, choosing this topic was based on my experience living in the UK. I was brought closer to the smoking culture as I had never been before. It is a habit that cuts across all age groups and I knew the health implication, especially starting from a young age. I thus had the motivation of combining my learning from this degree with a need to reduce the health risk among smokers. This research was a lot of hard work, with moments of anxiety and worry about how it will be completed. Because, even though I had an idea what to write, research and writing were skills I was yet to conquer. I also did not have the full range of technical skills required for the implementation of this study at the start of this research. (Mumford 1995) described various learning styles, most relatable is the activist learning. Which involved learning from new experiences. Drawing from this, with a lot of work, I learnt what I needed to know at each stage of this research. Thus, I have had the opportunity to study a lot of materials around the research topic and it has deepened my understanding of machine learning and how to apply it to real-life problems. I have also experienced an improvement in my literature and writing skills as a result of this research. Most of my learning has been from actively applying myself at each stage of the study and this has served not just as an avenue for knowledge acquisition but for all-round development. Beyond this work, I have been able to apply these skills in securing a placement. I plan to consolidate this learning by continuing with this momentum to take part in other studies by joining a research mentorship program which will provide me with the opportunity to continue to work with others and self-improve.

# **REFERENCES**

Abdulkareem, N. M. and Abdulazeez, A. M. (2021) Machine learning classification based on Radom Forest Algorithm: A review. *International Journal of Science and Business* 5 (2), 128-142.

Abdulrahman, S. A., Khalifa, W., Roushdy, M. and Salem, A.-B. M. (2020) Comparative study for 8 computational intelligence algorithms for human identification. *Computer science review* 36, 100237.

Al-Obaide, M. A. I., Ibrahim, B. A., Al-Humaish, S. and Abdel-Salam, A.-S. G. (2018) Genomic and Bioinformatics Approaches for Analysis of Genes Associated With Cancer Risks Following Exposure to Tobacco Smoking. *Frontiers in public health* 6, 84-84.

Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H. F. and van der Schaar, M. (2019) Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PloS one* 14 (5), e0213653-e0213653.

Alam, J., Alam, S. and Hossan, A. (2018) Multi-stage lung cancer detection and prediction using multi-class svm classifie. *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2).* IEEE.

Alavinezhad, A. and Boskabady, M. H. (2018) The prevalence of asthma and related symptoms in Middle East countries. *The Clinical Respiratory Journal* 12 (3), 865-877.

Alba, A. C., Agoritsas, T., Walsh, M., Hanna, S., Iorio, A., Devereaux, P., McGinn, T. and Guyatt, G. (2017) Discrimination and calibration of clinical prediction models: users’ guides to the medical literature. *Jama* 318 (14), 1377-1384.

Alsharqi, M., Woodward, W., Mumith, J., Markham, D., Upton, R. and Leeson, P. (2018) Artificial intelligence and echocardiography. *Echo research and practice* 5 (4), R115-R125.

Amarappa, S. and Sathyanarayana, S. (2014) Data classification using Support vector Machine (SVM), a simplified approach. *Int. J. Electron. Comput. Sci. Eng* 3, 435-445.

Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A. and Gonzales-Yanac, T. (2021) Multiple disease prediction using Machine learning algorithms. *Materials Today: Proceedings*.

Awais, M., Munwar, I., Zahid, M., Aun, I., Marriam, N., Tahira, N. and Momina, M. (2021) Prediction of Heart Disease Using Deep Convolutional Neural Networks. *The Arabian Journal for Science and Engineering. Section B, Engineering* 46 (4), 3409-3422.

Bayati, M., Bhaskar, S. and Montanari, A. (2018) Statistical analysis of a low cost method for multiple disease prediction. *Statistical methods in medical research* 27 (8), 2312-2328.

Belgiu, M. and Drăguţ, L. (2016) Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing* 114, 24-31.

Bellou, V., Belbasis, L., Konstantinidis, A. K., Tzoulaki, I. and Evangelou, E. (2019) Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal. *BMJ* 367, l5358.

Bertolaccini, L., Solli, P., Pardolesi, A. and Pasini, A. (2017) An overview of the use of artificial neural networks in lung cancer research. *Journal of thoracic disease* 9 (4), 924.

Bhopal, R. S. (2016) *Concepts of epidemiology: integrating the ideas, theories, principles, and methods of epidemiology.* Oxford University Press.

Boell, S. K. and Cecez-Kecmanovic, D. (2014) A hermeneutic approach for conducting literature reviews and literature searches. *Communications of the Association for information Systems* 34 (1), 12.

Boughorbel, S., Jarray, F. and El-Anbari, M. (2017) Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PloS one* 12 (6), e0177678-e0177678.

Brieuc, M. S. O., Waters, C. D., Drinan, D. P. and Naish, K. A. (2018) A practical introduction to Random Forest for genetic association studies in ecology and evolution. *Mol Ecol Resour* 18 (4), 755-766.

Burrell, G. and Morgan, G. (1979) *Sociological paradigms and organisational analysis: elements of the sociology of corporate life.* Aldershot: Ashgate.

Castro-Rodriguez, J. A., Cifuentes, L. and Martinez, F. D. (2019) Predicting asthma using clinical indexes. *Frontiers in Pediatrics*, 320.

Cavallaro, G., Riedel, M., Richerzhagen, M., Benediktsson, J. A. and Plaza, A. (2015) On understanding big data impacts in remotely sensed image classification using support vector machine methods. *IEEE journal of selected topics in applied earth observations and remote sensing* 8 (10), 4634-4646.

Centers for disease control and prevention (2021) *Behavioral Risk Factor Surveillance System (BRFSS) Prevalence Data (2011 to present).* United states: Centers for disease control and prevention. [*https://chronicdata.cdc.gov/Behavioral-Risk-Factors/Behavioral-Risk-Factor-Surveillance-System-BRFSS-P/dttw-5yxu*](https://chronicdata.cdc.gov/Behavioral-Risk-Factors/Behavioral-Risk-Factor-Surveillance-System-BRFSS-P/dttw-5yxu) Accessed 21/6/2021.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y. and Van Calster, B. (2019) A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology* 110, 12-22.

Cichosz, S. L., Johansen, M. D. and Hejlesen, O. (2016) Toward big data analytics: review of predictive models in management of diabetes and its complications. *Journal of diabetes science and technology* 10 (1), 27-34.

Cunningham, P. and Delany, S. J. (2021) k-Nearest neighbour classifiers-A Tutorial. *ACM Computing Surveys (CSUR)* 54 (6), 1-25.

Cunningham, P., Kathirgamanathan, B. and Delany, S. J. (2021) Feature Selection Tutorial with Python Examples. *arXiv preprint arXiv:2106.06437*.

Damen, J. A., Hooft, L., Schuit, E., Debray, T. P., Collins, G. S., Tzoulaki, I., Lassale, C. M., Siontis, G. C., Chiocchia, V. and Roberts, C. (2016) Prediction models for cardiovascular disease risk in the general population: systematic review. *bmj* 353.

Delpino, F. M., Costa, Â. K., Farias, S. R., Chiavegatto Filho, A. D. P., Arcêncio, R. A. and Nunes, B. P. (2022) Machine learning for predicting chronic diseases: a systematic review. *Public health (London)* 205, 14-25.

Dhawan, R. (2019) A Path To Data Scientist: A Review. *Think India Journal* 22 (30), 1219-1225.

Dubey, A. K., Gupta, U. and Jain, S. (2016) Epidemiology of lung cancer and approaches for its prediction: a systematic review and analysis. *Chinese journal of cancer* 35 (1), 1-13.

Feirman, S. P., Glasser, A. M., Teplitskaya, L., Holtgrave, D. R., Abrams, D. B., Niaura, R. S. and Villanti, A. C. (2016) Medical costs and quality-adjusted life years associated with smoking: a systematic review. *BMC Public Health* 16 (1), 1-11.

Fernández, A., del Río, S., Chawla, N. V. and Herrera, F. (2017) An insight into imbalanced big data classification: outcomes and challenges. *Complex & Intelligent Systems* 3 (2), 105-120.

Fleuren, L. M., Klausch, T. L. T., Zwager, C. L., Schoonmade, L. J., Guo, T., Roggeveen, L. F., Swart, E. L., Girbes, A. R. J., Thoral, P., Ercole, A., Hoogendoorn, M. and Elbers, P. W. G. (2020) Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive care medicine* 46 (3), 383-400.

Fotouhi, S., Asadi, S. and Kattan, M. W. (2019) A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of biomedical informatics* 90, 103089.

Grandini, M., Bagli, E. and Visani, G. (2020) Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

Gray, E. P., Teare, M. D., Stevens, J. and Archer, R. (2016) Risk prediction models for lung cancer: a systematic review. *Clinical lung cancer* 17 (2), 95-106.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H. and Bing, G. (2017) Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications* 73, 220-239.

Hancock, J. T. and Khoshgoftaar, T. M. (2020) Survey on categorical data for neural networks. *Journal of Big Data* 7 (1), 1-41.

Haq, A. U., Li, J. P., Memon, M. H., Nazir, S. and Sun, R. (2018) A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems* 2018.

Hofer, B. K. and Pintrich, P. R. (1997) The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of educational research* 67 (1), 88-140.

Holipah, H., Sulistomo, H. W. and Maharani, A. (2020) Tobacco smoking and risk of all-cause mortality in Indonesia. *PloS one* 15 (12), e0242558.

Hovorka, D. S. and Lee, A. S. (2010) Reframing interpretivism and positivism as understanding and explanation: Consequences for information systems research.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y. and Xu, W. (2018) Applications of support vector machine (SVM) learning in cancer genomics. *Cancer genomics & proteomics* 15 (1), 41-51.

Huang, S., Yang, J., Fong, S. and Zhao, Q. (2020) Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Letters* 471, 61-71.

Jain, D. and Singh, V. (2018) Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Informatics Journal* 19 (3), 179-189.

Jha, P. (2020) The hazards of smoking and the benefits of cessation: a critical summation of the epidemiological evidence in high-income countries. *Elife* 9, e49979.

Jiao, Y. and Du, P. (2016) Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quantitative Biology* 4 (4), 320-330.

Johnston, M. P. (2017) Secondary data analysis: A method of which the time has come. *Qualitative and quantitative methods in libraries* 3 (3), 619-626.

Kagiyama, N., Shrestha, S., Farjo, P. D. and Sengupta, P. P. (2019) Artificial intelligence: practical primer for clinical research in cardiovascular disease. *Journal of the American Heart Association* 8 (17), e012788.

Kamala, R. and Thangaiah, R. J. (2019) An improved hybrid feature selection method for huge dimensional datasets. *IAES International Journal of Artificial Intelligence* 8 (1), 77.

Katz, G., Shin, E. C. R. and Song, D. (2016) Explorekit: Automatic feature generation and selection. *2016 IEEE 16th International Conference on Data Mining (ICDM).* IEEE.

Kaur, H., Pannu, H. S. and Malhi, A. K. (2019) A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys (CSUR)* 52 (4), 1-36.

Kelly, M. (2021) Epistemology, epistemic belief, personal epistemology, and epistemics: A review of concepts as they impact information behavior research. *Journal of the Association for Information Science and Technology* 72 (4), 507-519.

Ketokivi, M. and Mantere, S. (2010) Two strategies for inductive reasoning in organizational research. *Academy of management review* 35 (2), 315-333.

Khaire, U. M. and Dhanalakshmi, R. (2019) Stability of feature selection algorithm: A review. *Journal of King Saud University-Computer and Information Sciences*.

Kilic, A. (2020) Artificial Intelligence and Machine Learning in Cardiovascular Health Care. *The Annals of thoracic surgery* 109 (5), 1323-1329.

Kimberlin, C. L. and Winterstein, A. G. (2008) Validity and reliability of measurement instruments used in research. *American journal of health-system pharmacy* 65 (23), 2276-2284.

Kitchenham, B. (2004) Procedures for performing systematic reviews. *Keele, UK, Keele University* 33 (2004), 1-26.

Komorowski, M., Marshall, D. C., Salciccioli, J. D. and Crutain, Y. (2016) Exploratory data analysis. *Secondary analysis of electronic health records*, 185-203.

Kour, H., Manhas, J. and Sharma, V. (2020) Usage and implementation of neuro-fuzzy systems for classification and prediction in the diagnosis of different types of medical disorders: a decade review. *Artificial Intelligence Review* 53 (7), 4651-4706.

Krawczyk, B. (2016) Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence* 5 (4), 221-232.

Kuruvilla, M. E., Vanijcharoenkarn, K., Shih, J. A. and Lee, F. E.-H. (2019) Epidemiology and risk factors for asthma. *Respiratory Medicine* 149, 16-22.

Lam, H. T., Thiebaut, J.-M., Sinn, M., Chen, B., Mai, T. and Alkan, O. (2017) One button machine for automating feature engineering in relational databases. *arXiv preprint arXiv:1706.00327*.

Lamba, A., Singh, S., Bhardwaj, S., Dutta, N. and Rela, S. (2015) Uses of Artificial Intelligent Techniques to Build Accurate Models for Intrusion Detection System. *International Journal For Technological Research In Engineering* 2 (12).

Lee, Y.-h., Bang, H. and Kim, D. J. (2016) How to establish clinical prediction models. *Endocrinology and Metabolism* 31 (1), 38-44.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A. and Seliya, N. (2018) A survey on addressing high-class imbalance in big data. *Journal of Big Data* 5 (1), 1-30.

Li, B., Li, D., Liu, J.-F., Wang, L., Li, B.-Z., Yan, X.-J., Liu, W., Wu, K. and Xiang, R.-L. (2021) “Smoking paradox” is not true in patients with ischemic stroke: a systematic review and meta-analysis. *Journal of Neurology* 268 (6), 2042-2054.

Lillie, S. E., Partin, M. R., Rice, K., Fabbrini, A. E., Greer, N. L., Patel, S. S., MacDonald, R., Rutks, I. and Wilt, T. J. (2014) VA Evidence-based Synthesis Program Reports. *The Effects of Shared Decision Making on Cancer Screening – A Systematic Review.*  Washington (DC): Department of Veterans Affairs (US).

Lim, C., Kim, K.-H., Kim, M.-J., Heo, J.-Y., Kim, K.-J. and Maglio, P. P. (2018) From data to value: A nine-factor framework for data-based value creation in information-intensive services. *International journal of information management* 39, 121-135.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B. and Sánchez, C. I. (2017) A survey on deep learning in medical image analysis. *Medical image analysis* 42, 60-88.

Lyu, S., Tian, X., Li, Y., Jiang, B. and Chen, H. (2020) Multiclass Probabilistic Classification Vector Machine. *IEEE transaction on neural networks and learning systems* 31 (10), 3906-3919.

MacArthur, G. J., Harrison, S., Caldwell, D. M., Hickman, M. and Campbell, R. (2016) Peer‐led interventions to prevent tobacco, alcohol and/or drug use among young people aged 11–21 years: a systematic review and meta‐analysis. *Addiction* 111 (3), 391-407.

Mackenzie, N. and Knipe, S. (2006) Research dilemmas: Paradigms, methods and methodology. *Issues in educational research* 16 (2), 193-205.

Maheswari, S. and Pitchai, R. (2019) Heart disease prediction system using decision tree and naive bayes algorithm. *Current Medical Imaging* 15 (8), 712-717.

Mandal, L. and Jana, N. D. (2019) A comparative study of naive bayes and k-NN algorithm for multi-class drug molecule classification. *2019 IEEE 16th India Council International Conference (INDICON).* IEEE.

Matheson, M. C., Bowatte, G., Perret, J. L., Lowe, A. J., Senaratna, C. V., Hall, G. L., de Klerk, N., Keogh, L. A., McDonald, C. F., Waidyatillake, N. T., Sly, P. D., Jarvis, D., Abramson, M. J., Lodge, C. J. and Dharmage, S. C. (2018) Prediction models for the development of COPD: a systematic review. *Int J Chron Obstruct Pulmon Dis* 13, 1927-1935.

Moeez M. Subhani and Ashiq Anjum (2020) Multiclass Disease Predictions Based on Integrated Clinical and Genomics Datasets. *In Poceedings of The Eleventh International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies*.

Morrow, V., Boddy, J. and Lamb, R. (2014) The ethics of secondary data analysis.

Mumford, A. (1995) Learning styles and mentoring. *Industrial and Commercial Training*.

Mustaqeem, A., Anwar, S. M. and Majid, M. (2018) Multiclass classification of cardiac arrhythmia using improved feature selection and SVM invariants. *Computational and mathematical methods in medicine* 2018.

Nakhle, F. and Harfouche, A. L. (2021) Ready, Steady, Go AI: A practical tutorial on fundamentals of artificial intelligence and its applications in phenomics image analysis. *Patterns* 2 (9), 100323.

Nian, R., Liu, J. and Huang, B. (2020) A review on reinforcement learning: Introduction and applications in industrial process control. *Computers & Chemical Engineering* 139, 106886.

Ofori, S. N. and Odia, O. J. (2016) Risk assessment in the prevention of cardiovascular disease in low-resource settings. *Indian Heart Journal* 68 (3), 391-398.

Onor, I. O., Stirling, D. L., Williams, S. R., Bediako, D., Borghol, A., Harris, M. B., Darensburg, T. B., Clay, S. D., Okpechi, S. C. and Sarpong, D. F. (2017) Clinical effects of cigarette smoking: epidemiologic impact and review of pharmacotherapy options. *International journal of environmental research and public health* 14 (10), 1147.

Pak, S.-I. and Oh, T.-H. (2016) Application of receiver operating characteristic (ROC) curve for evaluation of diagnostic test performance. *Journal of Veterinary Clinics* 33 (2), 97-101.

Palazón‐Bru, A., Mares‐García, E., López‐Bru, D., Mares‐Arambul, E., Folgado‐de la Rosa, D. M., Carbonell‐Torregrosa, M. d. l. Á. and Gil‐Guillén, V. F. (2020) A critical appraisal of the clinical applicability and risk of bias of the predictive models for mortality and recurrence in patients with oropharyngeal cancer: Systematic review. *Head & neck* 42 (4), 763-773.

Perez-Warnisher, M. T., De Miguel, M. D. P. C. and Seijo, L. M. (2018) Tobacco Use Worldwide: Legislative Efforts to Curb Consumption. *Annals of global health* 84 (4), 571.

Prusty, M. R., Jayanthi, T. and Velusamy, K. (2017) Weighted-SMOTE: A modification to SMOTE for event classification in sodium cooled fast reactors. *Progress in Nuclear Energy* 100, 355-364.

Rácz, A., Bajusz, D. and Héberger, K. (2019) Multi-Level Comparison of Machine Learning Classifiers and Their Performance Metrics. *Molecules (Basel, Switzerland)* 24 (15), 2811.

Rankin, N. M., McWilliams, A. and Marshall, H. M. (2020) Lung cancer screening implementation: Complexities and priorities. *Respirology* 25, 5-23.

Raschka, S. (2018) Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.

Rashidi, H. H., Tran, N. K., Betts, E. V., Howell, L. P. and Green, R. (2019) Artificial intelligence and machine learning in pathology: the present landscape of supervised methods. *Academic pathology* 6, 2374289519873088.

Rezaei, S., Sari, A. A., Arab, M., Majdzadeh, R. and Poorasl, A. M. (2016) Economic burden of smoking: a systematic review of direct and indirect costs. *Medical journal of the Islamic Republic of Iran* 30, 397.

Rice, P., Pugh, M., Geraghty, R., Hameed, B. Z., Shah, M. and Somani, B. K. (2021) Machine Learning Models for Predicting Stone-Free Status after Shockwave Lithotripsy: A Systematic Review and Meta-Analysis. *Urology* 156, 16-22.

Riley, R. D., Ensor, J., Snell, K. I., Debray, T. P., Altman, D. G., Moons, K. G. and Collins, G. S. (2016) External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *bmj* 353.

Rocha, A. and Klein Goldenstein, S. (2014) Multiclass from binary: expanding one-versus-all, one-versus-one and ECOC-based approaches. *IEEE transaction on neural networks and learning systems* 25 (2), 278.

Saadatfar, H., Khosravi, S., Joloudari, J. H., Mosavi, A. and Shamshirband, S. (2020) A New K-Nearest Neighbors Classifier for Big Data Based on Efficient Data Pruning. *Mathematics (Basel)* 8 (2), 286.

Salih, A. A. and Abdulazeez, A. M. (2021) Evaluation of classification algorithms for intrusion detection system: A review. *Journal of Soft Computing and Data Mining* 2 (1), 31-40.

Saunders, M. N. K. P., Lewis, P. and Thornhill, A. (2019) *Research Methods for Business Students.* Boston: Pearson Education limited.

Schutt, R. K. (2019) Quantitative methods. *The Wiley Blackwell Companion to Sociology*, 39-56.

Seetharam, K., Shrestha, S. and Sengupta, P. P. (2019) Artificial Intelligence in Cardiovascular Medicine. *Current treatment options in cardiovascular medicine* 21 (5), 1-14.

Sidi, Y. and Harel, O. (2018) The treatment of incomplete data: reporting, analysis, reproducibility, and replicability. *Social Science & Medicine* 209, 169-173.

Song, Q. C., Tang, C. and Wee, S. (2021) Making sense of model generalizability: A tutorial on cross-validation in R and Shiny. *Advances in Methods and Practices in Psychological Science* 4 (1), 2515245920947067.

Soofi, A. A. and Awan, A. (2017) Classification techniques in machine learning: applications and issues. *Journal of Basic and Applied Sciences* 13, 459-465.

Speiser, J. L., Miller, M. E., Tooze, J. and Ip, E. (2019) A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications* 134, 93-101.

Stern, J., Pier, J. and Litonjua, A. A. (2020) Asthma epidemiology and risk factors. *Seminars in immunopathology.* Vol. 42. Springer.

Steyerberg, E. W. and Vergouwe, Y. (2014) Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 35 (29), 1925-31.

Strieth-Kalthoff, F., Sandfort, F., Segler, M. H. and Glorius, F. (2020) Machine learning the ropes: principles, applications and directions in synthetic chemistry. *Chemical Society Reviews* 49 (17), 6154-6168.

Taghizadeh, N., Vonk, J. M. and Boezen, H. M. (2016) Lifetime Smoking History and Cause-Specific Mortality in a Cohort Study with 43 Years of Follow-Up. *PloS one* 11 (4), e0153310-e0153310.

Tammemägi, M. C. (2018) Selecting lung cancer screenees using risk prediction models—where do we go from here. *Translational lung cancer research* 7 (3), 243.

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N. and Asadpour, M. (2020) Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data* 7 (1), 1-47.

Tharwat, A. (2018) Classification assessment methods. *Applied computing & informatics*.

Thrall, M. J. (2019) Automated screening of Papanicolaou tests: a review of the literature. *Diagnostic cytopathology* 47 (1), 20-27.

Torre, L. A., Siegel, R. L. and Jemal, A. (2016) Lung cancer statistics. *Lung cancer and personalized medicine*, 1-19.

Toumazis, I., Bastani, M., Han, S. S. and Plevritis, S. K. (2020) Risk-Based lung cancer screening: A systematic review. *Lung cancer (Amsterdam, Netherlands)* 147, 154-186.

Uddin, S., Khan, A., Hossain, M. E. and Moni, M. A. (2019) Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making* 19 (1), 1-16.

Underner, M., Peiffer, G., Perriot, J. and Jaafari, N. (2021) Smoking cessation in asthmatic patients and its impact. *Revue des Maladies Respiratoires*.

Van Bussel, E., Hoevenaar-Blom, M., Poortvliet, R., Gussekloo, J., van Dalen, J., Van Gool, W., Richard, E. and van Charante, E. M. (2020) Predictive value of traditional risk factors for cardiovascular disease in older people: a systematic review. *Preventive medicine* 132, 105986.

Vargo, S. L. and Lusch, R. F. (2007) Service-dominant logic: continuing the evolution. *Journal of the Academy of Marketing Science* 36 (1), 1-10.

Vembandasamy, K., Sasipriya, R. and Deepa, E. (2015) Heart Diseases Detection Using Naive Bayes Algorithm. *International Journal of Innovative Science, Engineering & Technology* Vol. 2 (Issue 9).

Vineis, P., Alavanja, M., Buffler, P., Fontham, E., Franceschi, S., Gao, Y. T., Gupta, P. C., Hackshaw, A., Matos, E., Samet, J., Sitas, F., Smith, J., Stayner, L., Straif, K., Thun, M. J., Wichmann, H. E., Wu, A. H., Zaridze, D., Peto, R. and Doll, R. (2004) Tobacco and cancer: recent epidemiological evidence. *JNCI : Journal of the National Cancer Institute* 96 (2), 99-106.

Vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R. and Cleven, A. (2015) Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. *Communications of the association for information systems* 37 (1), 9.

Wang, X., Wang, F. and Hu, J. (2014) A multi-task learning framework for joint disease risk prediction and comorbidity discovery. *2014 22nd International Conference on Pattern Recognition.* IEEE.

Waring, J., Lindvall, C. and Umeton, R. (2020) Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial Intelligence in Medicine* 104, 101822.

Weller, D., Peake, M. and Field, J. (2019) Presentation of lung cancer in primary care. *NPJ primary care respiratory medicine* 29 (1), 1-5.

Welte, T., Vogelmeier, C. and Papi, A. (2015) COPD: early diagnosis and treatment to slow disease progression. *International journal of clinical practice* 69 (3), 336-349.

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M. and Qureshi, N. (2017) Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one* 12 (4), e0174944-e0174944.

West, R. (2017) Tobacco smoking: Health impact, prevalence, correlates and interventions. *Psychology & health* 32 (8), 1018-1036.

Wickramasinghe, I. and Kalutarage, H. (2020) Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft computing (Berlin, Germany)* 25 (3), 2277-2293.

Xian, S. and Chen, Y. (2021) E‐cigarette users are associated with asthma disease: A meta‐analysis. *The Clinical Respiratory Journal* 15 (5), 457-466.

Xu, D., Shi, Y., Tsang, I. W., Ong, Y.-S., Gong, C. and Shen, X. (2019) Survey on multi-output learning. *IEEE transactions on neural networks and learning systems* 31 (7), 2409-2429.

Yin, R. K. (2009) *Case study research: Design and methods.* Vol. 5. sage.

Yun, J.-M., Yoo, T., Oh, S.-W., Cho, B., Kim, E. and Hwang, I. (2017) Prediction of cardiovascular disease in Korean population: Based on health risk appraisal of national health screening program. *Journal of the Korean Medical Association* 60, 746.

Zacharasiewicz, A. (2016) Maternal smoking in pregnancy and its influence on childhood asthma. *ERJ open research* 2 (3).

Zajmi, L., Ahmed, F. Y. and Jaharadak, A. A. (2018) Concepts, methods, and performances of particle swarm optimization, backpropagation, and neural networks. *Applied Computational Intelligence and Soft Computing* 2018.

Zhang, M.-L., Li, Y.-K., Liu, X.-Y. and Geng, X. (2018) Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science* 12 (2), 191-202.

Žukauskas, P., Vveinhardt, J. and Andriukaitienė, R. (2018) Philosophy and paradigm of scientific research. *Management Culture and Corporate Social Responsibility* 121.

# **APPENDIX A RESEARCH PROPOSAL**

Text

Description automatically generated

Text, letter

Description automatically generated

Text, letter

Description automatically generated

Text, letter

Description automatically generated

Text

Description automatically generated

Text, letter

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Text, application

Description automatically generated

Text, letter

Description automatically generated

Text, letter

Description automatically generated

Text

Description automatically generated

Text, letter

Description automatically generated

Text, letter

Description automatically generated

# **APPENDIX B APPROVED ETHICS FORM**

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Timeline, calendar

Description automatically generated

Table

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated