# PREDICTIVE MODELS FOR COLORECTAL CANCER: A UNITED KINGDOM STUDY

OIDAMAE KEN TOBIKO, 096983

**Submitted in partial fulfillment of the requirements for the Degree of Bachelor of Business Science in Actuarial Science at Strathmore University**

**Strathmore Institute of Mathematical Sciences**
**Strathmore University**
**Nairobi, Kenya**

**February 2021**

## DECLARATION

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the Research Project contains no material previously published or written by another person except where due reference is made in the Research Project itself.



Oidamae Ken Tobiko

February 10th, 2021

This Research Project has been submitted for examination with my approval as the Supervisor.

Dr. Evans Otieno Omondi

February 10th, 2021

Strathmore Institute of Mathematical Sciences
Strathmore University

# Table of Contents

# LIST OF FIGURES.

# LIST OF TABLES.

## LIST OF ABBREVIATIONS.

ML – Machine Learning.

AI- Artificial Intelligence.

EHR- Electronic Health Record.

LR- Logistic Regression.

SVM- Support Vector Machine.

DT- Decision Tree.

RF- Random Forest.

KNN- k-Nearest Neighbor.

NB- Naïve Bayes.

ANN- Advanced Neural Network.

DR- Dimensionality Reduction.

PCA- Principal Component Analysis

UK – United Kingdom

OGL – Open Government License

## ABSTRACT.

The role of Machine Learning (ML) and Artificial Intelligence (AI) in healthcare and other aspects of life is growing every day. The purpose of this study was to build predictive ML models to determine whether patients with colorectal cancer live or die and to draw insights on them.

This study was of a causal design. Patients from the UK of various ages, different times of diagnosis and with different grades and stages of cancer were the focus of the study. These coupled with their vital status – alive or dead - were used to build the models.

Three models were built – regression model, decision tree and extreme gradient boosting ensemble trees. The three models had various accuracies in predicting the survival – with the regression model performing the worst followed by the decision tree and then the ensemble trees. Apart from the models, feature importance showed the significance of attributes like the stage of cancer and grade of tumor differentiation had on the likelihood of a patient surviving or dying.

# CHAPTER ONE: INTRODUCTION.

## 1.1 Background to the study.

Machine learning is an implementation of artificial intelligence (AI) that equips computer systems with the capacity to learn and master tasks from previous encounters without straightforward programming (Expert System, 2017). The focus of machine learning is the structuring of computer software that can retrieve data and from the data, master its workings (Expert System, 2017). (Mitchell, 1997) defined the general learning problem as "a computer is said to learn from past experience $E$ with respect to some class of tasks $T$ and performance measure $P$, if its performance in said task $T$, as measured by $P$, improves with $E$." Take for instance, a computer program designed to solve maze puzzles may refine its execution and solve future puzzles more accurately from its encounters with previous maze puzzles. Therefore, fundamentally a great number of ML problems can be designed as an optimization problem as regards to a certain data set. The goal of ML is to develop methods that are capable of automatically detecting trends in certain data, using the discovered trends in the data to forecast particular outcomes of interest (Murphy, 2012).

ML draws on ideas from other disciplines including data mining, probability and statistics and cognitive science (Murphy, 2012). ML algorithms are then applied in various fields for various purposes. For instance, in speech processing and recognition, determining financial risk and even determining threat of a terrorist attack just to name a few. This demonstrates the capability of ML to robustly solve complex tasks producing faster, more accurate results identifying profitable opportunities and dangerous risks while relying on real-world data not by intuition. It is also adaptable to new and upcoming situations, while collecting more and more data. Despite this, to train ML properly two things are required – a great number of resources and time. To be able to increase its success and its ability to operate with vast amounts of data, ML is coupled with AI. ML as a discipline, has seen a great number of victories recently and will continue to make an impression. This has largely been attributed to the accessibility of large amounts of data. An example is in the field of computer vision that deals with image-related tasks. The ability of ML algorithms to successfully recognize images has equaled if not surpassed that of people. This has been

made possible by the existence of large image databases for example 'ImageNet' that has millions of images that provide the training data (Wiens and Shenoy, 2017).

As ML becomes more prevalent in today's workplaces and in day-to-day living, it has also been used in the health sector. ML methods possess the ability to change different features of patient care. Precision medicine is the most familiar ML method applied. This involves predicting which treatment conventions would be best for a particular patient focusing on the various patient characteristics. ("The potential for artificial intelligence in healthcare," 2019). A different but more intricate method of machine learning is the *neural network*. It sees different problems in terms of inputs, outputs and weights of variables that associate inputs with outputs. This can then be used for classification where it looks at different patient characteristics and dictates whether they would get a certain disease or if they already are sick, how the disease progresses.

## 1.2 Types of ML methods.

ML methods may fall into two categories; supervised and unsupervised.

### 1.2.1 Unsupervised Learning.

This is where only the input data is present and the output is not present or is unknown. Data is unlabeled. The main aim for unsupervised learning is to get more information about the data input by modelling its structure (Brownlee, 2019). Here, the algorithms are allowed to work, discover by themselves and then show how the data is structured. According to ("Supervised vs Unsupervised Learning: Key Differences, 2020") unsupervised machine learning not only reveals trends in data but also it aids in the discovery of features relevant for classification. A simple illustration given in ("Supervised vs Unsupervised Learning: Key Differences," 2020) of unsupervised learning is a child with their family dog. The child knows and recognizes their dog. Should a new dog be brought by a visitor, even though the child has not seen it before, but because it has the same characteristics as theirs i.e., it is furry walks on all fours – the child can identify it to be a dog not a mouse. Unsupervised learning works in the same way- no teaching involved, just learning by itself.

Two groups of unsupervised learning problems arise – association and clustering. A clustering problem is presented as wanting to find groups in a particular dataset for

example grouping loan applicants by their respective bank balances. The algorithms will look at the available data and if they are present, uncover any natural clusters. The number of the clusters to be uncovered by the algorithms can also be preset. This makes it possible to adjust the scale of detail needed. An association rule problem is presented as wanting to uncover orders which characterize the dataset, for example most people who purchase a toothbrush also purchase toothpaste for it. Two examples of such algorithms are the k-means that deal with clustering and for association the Apriori algorithm is also used (Brownlee, 2019).

### 1.2.2 Supervised learning

Supervised ML algorithms learn from labeled data, forecasting outcomes for unanticipated data (Brownlee, 2019). Using more technical terms, "supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output -Y=f (X)" (Brownlee, 2019). The objective is to estimate the mapping function properly so much so that when totally new dataset is introduced it becomes possible to come up with the appropriate output variables as regards to the new dataset (Brownlee, 2019). This means that for some of the data, the correct answer has been given. A similarity can be made to learning in the company of an instructor. When the outcomes are known, the algorithm repetitively makes forecasts on the data and where it is wrong, the instructor corrects it. This process will only come to a halt when the algorithm has achieved a certain required level of success. An illustration given in ("Supervised vs Unsupervised Learning: Key Differences," 2020) is that someone wants to use his computer to help him predict how long it may take him to get home from work. The man feeds his computer with some data including the time he gets off work, how is the weather at that time and is it on a public holiday or a normal working day. The desired output for the man is how long it takes to get to his house. As a person, you know that should there be rain, then there will be traffic jams and it you may get to the house later than usual. The computer does not know this, but if given relevant data it could. Starting with a training data set containing how long it takes to get home and the prevalent weather conditions and maybe time one left the office, the computer could deduce that there exists a direct relationship between how much it rains and how long one takes to get to the house. Using this, it concludes that the heavier the rain, the more time

a person takes to get home. Also, it may uncover a relationship between the time one leaves the office and when they get home. For example, if you leave the office towards 5pm it takes more time to get to the house. This demonstrates how the computer is capable of finding relationships and connections within a set of labeled data making it possible for one to source data or draw an inference from some past experiences. With more experience this makes the algorithm perform even better making it easier to be applied in various types of real-world computation issues ("Supervised vs Unsupervised Learning: Key Differences," 2020).

There are two groups of supervised learning problems – classification and regression. A classification problem occurs where the output variable is to be a definitive class, for example "passed exam" and "failed exam" ("Supervised vs Unsupervised Learning: Key Differences," 2020). This brings the concept of binary classification. In binary classification, there are only two distinct categories. Should there be more than two categories it is now multiclass classification ("Supervised vs Unsupervised Learning: Key Differences," 2020). An example is determining whether or not an applicant will pay out their mortgage completely or default on it. Some types of classification algorithms include; Naïve Bayes Classifiers and the Decision Trees.

A regression problem is presented as the output variable being a real value, for example "height" and "prices" (Brownlee, 2019). The regression method will predict a singular output value. For instance, regression can be used to predict the value of a piece of land. The input variables could be where the land is located or the size of the land et cetera. There is also logistic regression method that estimates values based on a certain group of independent variables ("Supervised vs Unsupervised Learning: Key Differences," 2020). It helps you to forecast the possibility of a certain phenomenon happening by placing the data into a logit function and because it is a probability value, the result can only be a value between 0 and 1. ("Supervised vs Unsupervised Learning: Key Differences," 2020) gives the strengths and weaknesses.

One strength is that the algorithm can be adjusted to deal with overfitting. On the flipside, it lacks flexibility. Moreover, logistic regression may perform poorly if the decision margins are non-linear or if they are many.

## 1.3 Applications of ML.

Looking at the health sector, ML methods are applied in pinpointing and diagnosis of ailments which are otherwise thought of to be hard to spot (Flat World Solutions, 2020). This can include anything from various cancers to genetic disorders. An example is IBM Watson Genomics that uses cognitive computing to be able to come up with a quick spot of genetic disease (Flat World Solutions, 2020). Moreover, it is used in early-stage drug discovery process (Flat World Solutions, 2020). This also includes Research and Design technologies such as next-generation sequencing (Flat World Solutions, 2020). For example, Project Hanover developed by Microsoft is using ML-based technologies for multiple initiatives including developing AI-based technology for cancer treatment and personalizing drug combination for AML (Acute Myeloid Leukemia) (Flat World Solutions, 2020). Moreover, Medical Imaging Diagnosis through Computer Vision and an example of this is Microsoft's Inner Eye that uses image analysis to come up with instruments for diagnosis.

Also, by coupling personal health with predictive analytics it becomes easier to come up with treatments that are specifically designed for a certain patient (Flat World Solutions, 2020). At the moment, doctors are bound to a certain set of diagnoses (Wiens, Jenna & Shenoy, Erica, 2017) and for this, IBM Watson Oncology is also using individual patient medical history to assist in coming up with different treatment avenues to be used (Flat World Solutions, 2020). Also, in clinical trial and research, using ML methods for prediction analysis can help in finding potential clinical trial candidates from a pool of a large variety characteristics like medical history, past hospital trips et cetera which goes to reduce the overall cost and time for these clinical trials.

ML methods are also used in observe and forecast disease outbreaks all over the planet (Wiens, Jenna & Shenoy, Erica, 2017). With availability of vast amount of data this information can be used to predict everything from Ebola to other serious deadly ailments (Wiens, Jenna & Shenoy, Erica, 2017). Being able to predict the outbreaks is essential in developing countries which may not have strong health systems to be able to combat these diseases.

## 1.4 Problem Statement.

As already seen, ML offers exciting opportunities to augment the healthcare sector. In the UK today, ML and AI is increasingly being employed in the healthcare sector. There has been a recent growth in startups that use ML algorithms. (Cambridge Wireless, 2019) gives examples of Hear Angel – that created an app for headphones that helps to protect against hearing damage by learning the listening patterns of the user and Granta Innovation that seeks to predict prostate cancer. However, application in the NHS is still relatively limited (Marr, 2019).

There are around 367,000 new cancer cases in the UK every year, that translates to around 1,000 every day which is a new diagnosis every two minutes (Cancer Research UK, 2020). Of these cancers, colorectal cancer is the second most-deadliest after only lung cancer (Cancer Research UK, 2020). Although cancer treatment is free, because of the NHS, it ends up costing the government in excess of 1.4 billion pounds a year (Cancer Research UK, 2020). Treatment in private care is faster and more efficient but costs for private care can run up to excess of 30,000 pounds. Also, with a fast-growing and ageing population, evolving healthcare needs such as the increases in obesity and medical costs that keep running up, the situation looks like it is bound to get worse (The Medic Portal, 2019).

Considering this, early detection is important because when abnormal tissue or cancer is found early, it may be easier to treat. By the time symptoms appear, cancer may have begun to spread and be harder to treat. Therefore, this brings the need for building meaningful machine learning models for disease prediction especially for detection of cancer, to be employed in the UK health sector.

This project advances the use of machine learning to develop a model of disease prognosis prediction especially for colorectal cancer that can be used for either private healthcare services or by the public healthcare sector thereby assisting in decision-making and enhancing delivery of cancer services.

## 1.5 Research objectives.

### 1.5.1 The general objective.

The objective of this project is to examine machine learning as an approach to develop a meaningful model for disease prognosis prediction that could be applied and used in the UK health sector, private and public.

### 1.5.2 Specific objectives

i. To formulate machine learning models for disease prognosis prediction
ii. To assess the success of the machine learning models in disease prognosis prediction and insights drawn from the models.
iii. To examine the applicability of the ML model in the healthcare sector in the UK.

## 1.6 Significance of the Research.

The healthcare system is home to various stakeholders who are essentially involved in the healthcare system and would be greatly influenced by any and all changes to the status quo. These include Patients, Physicians, Private Hospital Owners and Employers, Insurance companies, Pharmaceutical companies, and the Government.

These stakeholders would benefit from results of these projects in several ways that include:

i. Patients would benefit from early and credible detection of disease. This is desirable because, it may serve to increase the effectiveness of treatment and in turn, improve their chances of survival. On a cost basis, treatment from an early stage may be cheaper than from later stages of disease progression. This assists in their personal financial management and planning of their futures accordingly.
ii. Physicians play a central role in making sure that patients under their care will receive adequate healthcare. Systems that can help in prediction and detection of disease serve to augment the physicians' abilities. It eases their workload and goes on to add onto their precision. They can then give the best possible care to the patients entrusted to them and save more lives.

iii. Private hospital owners and employers seek to provide the best possible services as institutions to remain competitive as businesses. Patients always want the best possible care. Employing a meaningful model in their establishment, serves to boost and grow their business, while at the same time meeting the first priority- giving proper service to patients.

iv. Insurance companies will be in a better position to tailor their insurance plans. With a way of forecasting the future, they are better placed to make decisions on premiums and reserves which will bode well in their bid to minimize uncertainty and risk while at the same time maximizing their profitability.

v. Pharmaceutical companies will benefit from the model because having a forecast of future outcomes will help them make informed decisions on their business models. What the demand for their products would be like, what operational costs they should incur and so on. This can only do them better than harm.

vi. One of the government's main agendas is to protect the nation's health. A healthy nation is a working nation. A working nation is a productive nation. Having such a model and such technology can only do the government good. First, it will help in subsidizing the poor. Proper planning could go a long way into making sure healthcare provisions are accessible even at the lowest levels of the public sector where even the poorest can access it. It will also help in budget formulation, coming up with the relevant policies for the healthcare system and overall service delivery and patient experience.

# CHAPTER 2: LITERATURE REVIEW

## 2.1 Introduction.

As health data becomes more available and accessible and as the current generation of computers keep evolving to have greater computational ability, researchers are now being involved in unearthing and formulating algorithms that aid in formulation of clinical predictions. ML has the opportunity to come up with strong instruments to enhance medical outcomes and to subsidize costs. The greater clinical fraternity should be involved in coming up with and assessing these instruments (Nevin, 2018). The reason for this is because ML algorithms can come up with predictor models, bringing in diverse classes of predictor characteristics without at all reducing the accuracy of forecasts (Konerman, 2019).

With early detection of ailments, there can be room for proper disease management, revamped involvement, and better resource distribution in the sector. To achieve this end, a number of ML methods have been advanced that use Electronic Health Records (EHRs). Most of the past attempts use structured data to achieve this purpose. This is not the best way to go because a large amount of information is found in unstructured data. (Liu, Zhang & Zaravian, 2018).

## 2.2 Machine learning models for disease prediction.

Prediction models in healthcare through ML can be designed through various methods. The machine learning algorithm, also called model, is a mathematical expression that represents data in the context of a problem (Castañón, 2019).

### 2.2.1 Regression.

The regression ML methods will be under the umbrella of supervised learning. Regression methods assist in the forecasting or explanation of a certain value founded on some previous data. An example is predicting the price of a new iPhone coming out based on how previous iPhones were priced or phones with related functionality.

The most basic way is linear regression where the equation of a line ($y = m * x + b$) is used for modelling a certain group of data. Should the linear regression have numerous

data pairs, training happens by calculating a slope of a line that would reduce the distance between all the data points and said line. (Castañón, 2019).

### 2.2.2 Logistic regression.

Still under supervised methods, logistic regression (LR) is a potent instrument. LR can be seen almost as a supplement to ordinary regression only that it models a binary variable that constitutes either an incident happening or not. Therefore, LR aids in determining the probability that a certain incident belongs to a said category (Uddin et al., 2019). Being a probability, the value given by an LR could only realistically be part of the closed interval [0,1]. To then utilize a LR as a dichotomous classifier, a boundary should be set in place to distinguish the two categories. For instance, probability scores greater than 0.50 for an incident it will categorize it as 'category X'; otherwise, 'category Y'. Should there be need to accommodate a variable with more than two values, the LR model can be generalized to be a multinomial logistic regression.

### 2.2.3 Support Vector Machine.

According to (Uddin et al., 2019) support vector machine (SVM) algorithms are used to categorize linear and non-linear data. The first step is mapping data items into a n-dimensional attribute space, n being the number of attributes. Next, a hyperplane that would do two things – increase the marginal distance between two categories while simultaneously reducing any errors in categories – is identified. This marginal distance for a certain category is the distance between the hyperplane and the closest occurrence for which is an instance in that category. Then singular data points are charted onto the n-dimension space and the value of each attribute becomes a particular coordinate point. For classification, determine the hyperplane that separates the categories by the largest margin possible.

### 2.2.4 Decision Tree.

A decision tree (DT) will model any decision logic evaluations and will match-up the results for categorizing the data units into a tree-like construct (Uddin et al., 2019). The nodes of a DT usually have several levels for which the highest or first node is termed as the root-node. The other internal nodes constitute tests on qualities. Contingent on what outcome is desired, the DT algorithm will keep branching towards the most significant

child node. This procedure of testing and then branching keeps on going repetitively until it hits the terminal node. This terminal node, also called a leaf node, will correlate to the decision end result. For many diagnostic and prognostic conventions, DTs are familiar elements because of their ease of learning and also being simple to understand (Uddin et al., 2019).

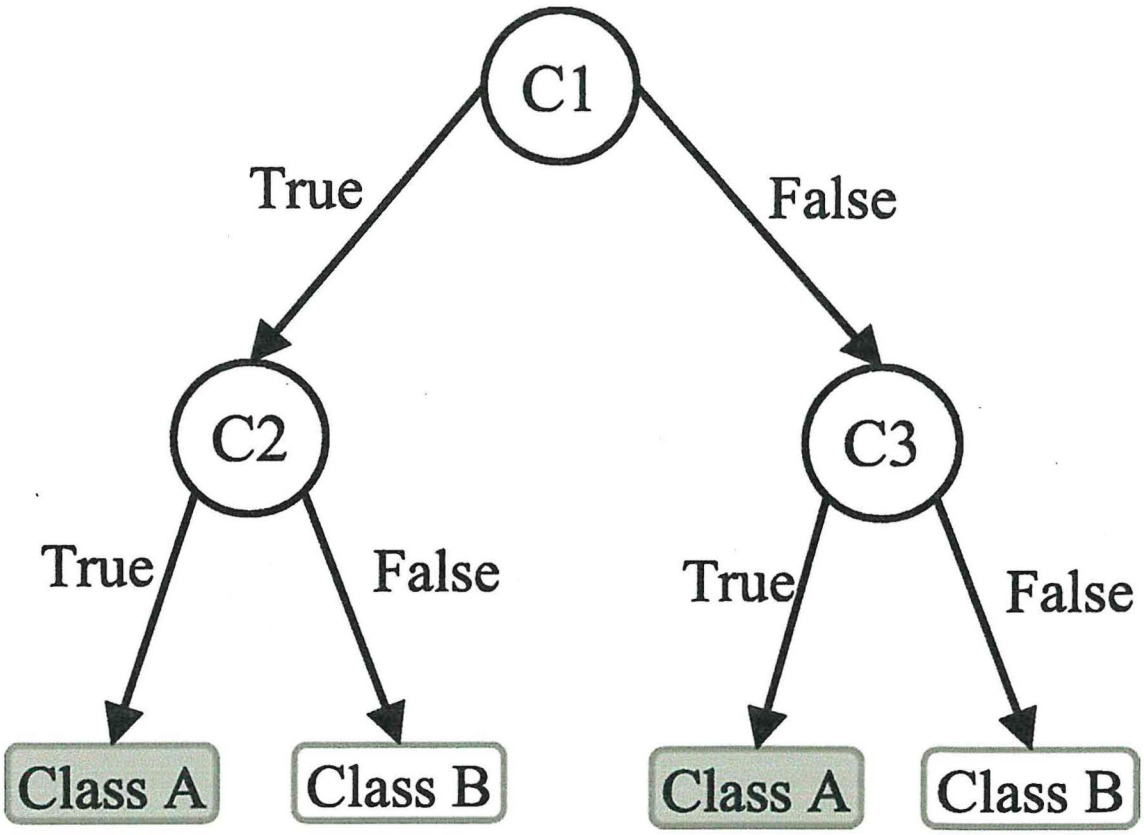



*Figure 2.1: Example of a Decision Tree 1*

### 2.2.5 Random Forest

Random forests (RFs) are group classifiers made up of a number of different decision trees in the same way that forests are groups of numerous trees (Uddin et al., 2019). It happens that DTs can become very extensive and this brings up the issue of overfitting the data. Overfitting will then cause a large discrepancy in classification end result for just a small variation in the data. This shows that they are susceptible to changes in the training data which causes them to be more likely to have errors.

Having various DTs for one RF means that they have to be trained in kind by various parts of the training data. Categorizing a new sample means that for that sample, the input has

to go through each and every DT in the RF. This means that each DT will now consider a separate part of the sample and will give its own end result. With a large number of classification end results, the RF chooses the result that has appeared the most in the case of discrete classification and should it be numerical, then the mean of the result of the DTs is taken. This consideration of results from numerous DTs means an RF can reduce the variance that would have otherwise resulted from considering just a singular DT for the particular data.

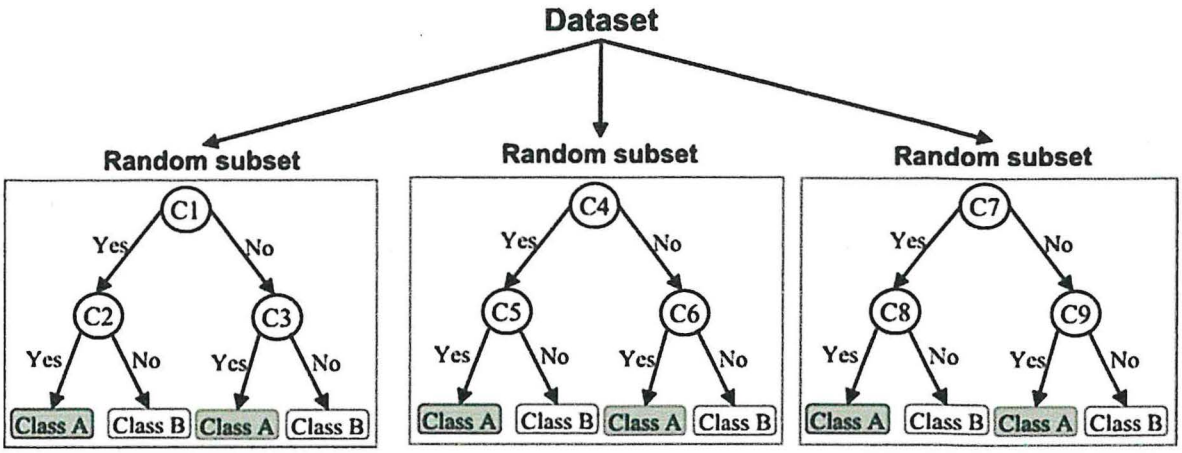



*Figure 2.2: Example of an RF.*

### 2.2.6 Naïve Bayes.

This method is anchored on the Bayes' Theorem. The theorem itself expresses the probability of an incident happening based on any past knowledge of circumstances related to that particular incident. The method is under the assumption that an attribute in a class is not at all directly related to any other attribute. However, these attributes in this particular class could display interdependence (Uddin et al., 2019).

### 2.2.7 k-nearest Neighbor.

According to (Uddin et al., 2019) this algorithm is one of the most basic and preliminary methods developed for classification. The KNN can be seen as an unostentatious version of the Naïve Bayes method. However, different from the Naïve Bayes method, it does not need to take probability into considerations. In the name, the k stands for the number of closest neighbors considered. With each different choice of a value for k, different classification outcomes can be realized for a particular sample.

### 2.2.8 Artificial Neural Network.

These are a group of ML algorithms and methods that seek to mimic the operations of the neural networks present in the brain of a human. It is possible to make these algorithms to appear as an intertwined group of several nodes. From node x, the output of it moves as input for node y, so on and so forth making for continuous processing of data in conformity with the interconnection. Matrices called layers are groups for these particular nodes. Nodes of similar functionality and transformation are placed in the same matrix. There may also exist hidden matrices in the ANN apart from the input and output matrix (Uddin et al., 2019). To maximize success and return from ANNs, large amounts of data are required in addition to machines that possess great computational power (Castañón, 2019).

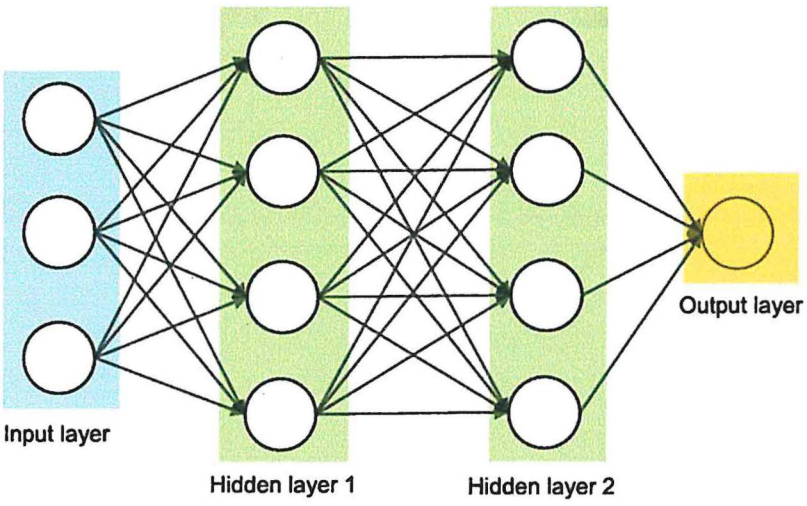



*Figure 2.3: Example of ANN.*

### 2.2.9 Dimensionality Reduction.

What this method does is that it discards of the least relevant information from a particular dataset (Castañón, 2019). This is largely essential because it helps in making the data set to be used practicable and results easily attainable.

One of the methods used for this is the Principal Component Analysis (PCA). This also happens to be the sought-after method of DR. How PCA works is that it decreases the dimensions of the attribute space by looking for brand new vectors that would increase the linear variations in the dataset to its highest possible value. This method can decrease the dimension of the dataset so greatly without compromising the amount of information should there be robust linear correlations in the data.

For their research, (Qianfan et.al, 2020) applied PCA while using genomic data to be able to predict disease. They argued that most genomic data that is produced to be utilized in disease and ailment prediction suffers from high dimensionality. This causes a challenge because it brings up the issue of over-fitting and not only that, but it also makes computation largely unsuitable. To combat this, their model used PCA to remove dimensionality present in the input attributes using the now low-dimensionality attributes for the prediction process.

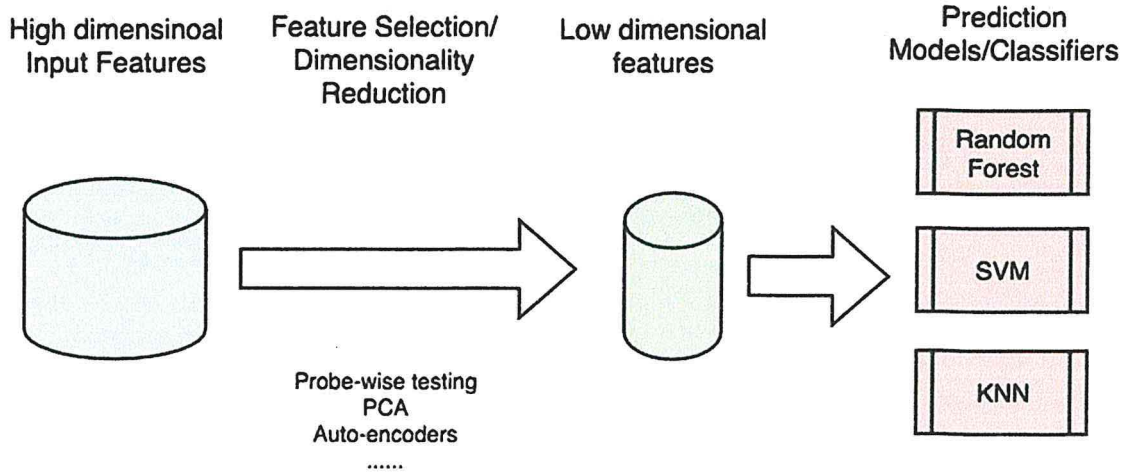



*Figure 2.4: Using Dimensionality Reduction.*

## 2.3 Performance of ML models in disease prediction.

Over the years, different machine learning models have been developed to assist in disease prediction. It is necessary to gauge the execution of the models to see which method of designing these models is best for prediction, in terms of accuracy.

According to (Uddin et al., 2019) the use of supervised machine learning algorithms has been the most sought-after method. In the study, they analyzed different ML models to identify key trends and gauge the execution, overall performance, and applicability in building predictor models. They analyzed 48 different models for predicting 49 different types of diseases from asthma, breast cancer to heart disease.

The study showed that the SVM method the most popular besting the Naïve Bayes that came second. In terms of overall accuracy of the models, the RF method exhibited senior accuracy to all other methods. The SVM algorithm was second in the category of accuracy.

In testing for the validation of model results, the 10-fold and 5-fold validation methods were employed. Upon validation, SVM was discovered to have greater accuracy more often than not. Conversely, where validation took place, the ANN algorithm showed more accuracy. This only goes to show the superiority of the Support Vector Machine method.

*Table 2.1: Comparing the application frequency and overall accuracy of independent ML methods.*

| ML method | Application frequency (x/49) | Instances it showed more accuracy (%) |
|---|---|---|
| Artificial Neural Network (ANN) | 20 | 6 (30%) |
| Decision Tree (DT) | 21 | 7 (33%) |
| k-nearest Neighbor (KNN) | 13 | 4 (31%) |
| Logistic Regression (LR) | 20 | 5 (25%) |
| Naïve Bayes (NB) | 23 | 7 (30%) |
| Random Forest (RF) | 17 | 9 (53%) |
| Support Vector Machine (SVM) | 29 | 13 (41%) |

Specifically looking into the method used:

In the Artificial Neural Network (ANN) the programmer cannot influence the required decision-making protocol, all is left to the machine. This means that it becomes tedious to train the network for composite classification problems. Also, the variables- independent and predictor - need pre-processing before they can be used which may take time (Uddin et al., 2019).

For the Decision tree (DT) the requirement is that classification categories must be mutually exclusive. Also, an absent characteristic for a non-leaf node means branching is impossible which means a result cannot be obtained. The DT is largely dependent on the order of the characteristics and it may also be functionally inferior to other methods like the ANN (Atlas et al., 1990).

In the K-nearest Neighbor (KNN) method, since model characteristics are all held to the same importance level, classification may not be optimal because no information is given about which characteristics are the most relevant towards making a proper classification. The Logistic regression method shows a deficiency in accuracy especially for input variables that have composite relationships. This means that they are more susceptible to overconfidence and may balloon the accuracy of the prediction because of sampling bias present (Uddin et al., 2019).

For Naïve Bayes (NB), it is mandatory for the classes to be mutually exclusive. Any existence of the characteristics being dependent on each other will cause the classification to be poor. In, Random Forests (RFs) it is necessary to define the number of base classifiers correctly. The method is also open to overfitting that can arise very easily. Also, RFs are more difficult to work with and computationally dear (Uddin et al., 2019).

The SVM also suffers from being computationally expensive in the case where huge and composite datasets are employed. If there is presence of noise in the data, the performance is going to be compromised and a common SVM cannot organize more than two categories unless it is extended (Uddin et al., 2019).

## 2.4 Applicability of ML models in the Healthcare sector in the UK.

In the UK healthcare sector, ML models for disease prediction could be applied in various ways. In the current, COVID-19 pandemic for example. As of February 3rd, 2020, the UK has had over 3.87 million confirmed cases of coronavirus. The efforts to test and to employ contact-tracing have increased coupled with an increasing vaccine drive. To be better prepared, and successfully manage the pandemic, understanding how the disease has progressed and continues to progress and whether the attenuative measures put in place are successful is necessary.

One model used to predict infectious disease models is the SEIR model. The model focuses on four groups of people – the Susceptible, the Exposed, the Infectious and the Recovered. This model was only recently applied for COVID-19 in the Chinese city Wuhan, Hubei Province where the disease was first reported. It aided in giving a layout of necessary interventions needed to be put in place in an effort to control and reduce the number of people at risk of contracting the virus. The same SEIR models were also applied to forecast and devise necessary interventions during the 2009 influenza pandemic in the United States of America (Nanyingi, 2020).

The SEIR model focuses on the change of people between four groups: the susceptible (S), the exposed (E), the infected (I), and the recovered or the removed (R). It also evaluates the speed at which people go through the groups. This is important as it augments public health readiness and proper response planning by upping surveillance, providing a signal of how the practicality of containing the virus and in which places and in what quantities resources should be given first concern should a pandemic break out (Nanyingi, 2020).

There have been two models that have also been newly developed – the QCOVID and the 4C mortality model. The QCOVID model predicts the risk of catching and dying from coronavirus in the general population while the 4C model is used to calculate the risk of mortality upon admission (Sperrin, 2020).

Moreover, Addenbrooke hospital in Cambridge is utilizing the Microsoft's InnerEye system to automatically process scans for patients with prostate cancer fast-tracking prostate cancer treatment. The hospital is also looking into using this for brain tumors (Marr, 2019).

The NHS is also employing the AI technology oh HeartFlow. This system looks at CT scans of patients who are suspected of having coronary heart disease and then creates a personalized 3D model of the heart that shows how their blood is flowing around it. This helps doctors find the places where blood flow is disrupted by clots and so on. This procedure is far cheaper and more intensive than the standard angiogram procedure, reducing the costs by almost 25% (Marr, 2019).

## 2.5 Conceptual Framework.

This research first gathers the relevant data required to undertake this research. The data is divided into two predictor classes that showed the occurrence or non-occurrence of breast cancer. Then using various variables like age in years and the Body Mass Index (BMI).

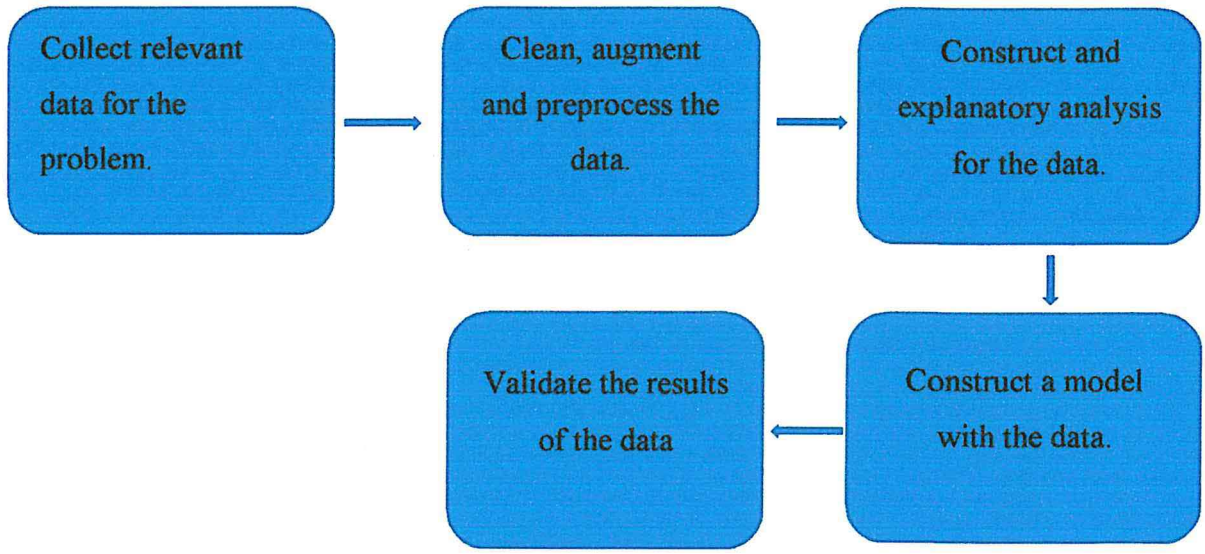



*Figure 2.5: The Conceptual Framework.*

# CHAPTER 3: RESEARCH METHODOLOGY.

## 3.1 Introduction.

This chapter sets forth the research methodology that was applied in this particular study. It binds the research design, population and sampling, data collection – instruments and procedure and also data analysis. There is also a section on data reliability, specification of the appropriate model to be used and the limitations of this particular study.

## 3.2 Research Design.

The explanatory study was based on a causal design. Causal effects happen where changes in a particular incident termed as the independent variable causes a change in another incident, termed as the dependent variable. This particular study seeks to assess the influences changes in attributes like insulin levels will ultimately have on breast cancer diagnosis.

### 3.3.1 Population.

A statistical population is a group of observations from which deductions are to be made, frequently based on a random sample drawn from the statistical population. An instance is where should generalizations need to be made for a certain characteristic, then the population is the group of all the observations where the characteristic is displayed at the moment, in the past or in times to come.

For this study, the population targeted is colorectal cancer patients. These subjects are in two groups – those alive and those who have passed away. The population targeted is multivariate, with different clinical features from which inferences can be drawn. The population for this study is 578717 people – some dead and some alive - with 12 features which serve as predictors.

### 3.3.2 Sampling Technique.

A sample is a mini class of a more sizable collection. It is a subdivision of a more sizable population, containing all the attributes of the population. containing the characteristics of a larger population. These samples are obtained, and relevant statistics are computed from them for the purpose of making deductions about the behavior of the larger population sample to the population. For our case, random sampling is used. A random sample is used when each independent observation possesses an equal probability of being selected for use in the sample.

A simple random sample of size n comprises of n observations from the larger population selected in a manner that all the observations possess an equal chance of being in the chosen sample. To determine the n sample size, first the Cochran's Sample Size Formula is used. The formula for this is $n_0 = \frac{z^2pq}{e^2}$ where Z is the Z value, p is the estimated proportion that has the desired characteristic, e is the amount of random sampling error and q is 1-p.

## 3.4 Data Collection – Instruments and Procedure.

The data used is cross-sectional data. It is provided by patients and then collected by the NHS as part of the patients' care. This data is secondary data. This data is raw statistical data obtained from the data bank of the government of the United Kingdom. This database publishes free datasets for analysis that are sourced from various government agencies and related bodies.

The data is the most suitable for this research question because using the various variables, it would be possible to build ML predictor model to predict colorectal cancer. On the question of validity, reliability and coverage, the data is reliable as the data is authentic, genuine, and consistent. The data is also valid and has proper coverage as it involves all necessary variables.

The variables used in this study include:

*Table 3.1: Variables of Interest in this study.*

| Variable | Unit of Measurement and Explanation |
|---|---|
| The age group | 5-year age bands |
| Sex | Coded as 1 for male, 2 for female |
| The site group of the tumor | Categorized as proximal colon, distal colon, rectum or overlapping |
| The stage of the cancer progression. | Stages rank from 1 to 4 |
| Grade of tumor differentiation | Grade ranges from undifferentiated to well differentiated – with some being unable to be accessed |
| Screening status | Whether the person has been screened or not |
| Days since diagnosis to current vital status | The number of days that have elapsed since the patient was diagnosed to their current vital status |

## 3.5 Data Analysis.

At the onset, the first step will be examining the data for completeness and undertaking quality control evaluations. The plan will involve;

1. Sorting the data.
2. Undertaking quality-control evaluations.
3. Processing the data.
4. Analyzing the data.

### 3.5.1 Understanding the Data.

The dataset applied in coming up with the model was sourced from the database of the United Kingdom government. The data collected included observations for different variables – age, grade of tumor differentiation, stage of cancer progression, sex, days since diagnosis, screening, underlying cause of death and site group of the tumor for 578717 patients. The patients are then classified into 2 groups; group A for those alive and group D for those who died from colorectal cancer.

### 3.5.2 Preparing the Data.

The data collected was very comprehensive and detailed. The data is loaded into R Studio for data preparation. Missing attributes and values were looked for using the is.na function. The relevant columns were labeled suitably awaiting use. After this transformation process, the data was stored in a new dataset for use.

### 3.5.3 Building the Model.

For this data, the first step would be to start with a regression model, then a Decision Tree, an Extreme Gradient Boosting Tree and then finally a Support Vector Machine (SVM).

#### *3.5.3.1 Regression Model.*

The caret package which is a condensed form Classification And REgression Training is a group of functions that try to ease the procedure of coming up with prediction models and algorithms. This package has instruments that aid in data splitting, pre-processing and feature selection. This is the package that will be utilized in making the first regression model.

The dataset is partitioned into training data and testing data. Then, a Generalized Linear Model of the training data is built. Predictions are then made on the GLM and using the testing data.

#### *3.5.3.2 Decision Tree.*

The next action is to formulate a classification decision tree. To be used is the rpart package which is a condensed form of Recursive Partitioning and Regression Trees which for partitions data recursively for classification, regression, and building survival trees. The training data is used to achieve this.

#### *3.5.3.3 Extreme Gradient Boosting Tree.*

An extreme gradient boosting tree is an ensemble tree model that is specifically designed to increase speed and precision (Morde, 2019). The xgbtree is still built by the caret package although it is necessary to define the method. After building the ensemble tree, it is necessary to cross-validate the results. Cross-validation is a method that is used for the evaluation of how the output of statistical analysis generalize to an independent data set.

Cross-validation is mainly employed in scenarios where the end-goal is prediction, and it is vital to evaluate the accuracy of the performance of a predictive model.

After cross-validation, testing for feature importance is next. Feature importance is a class of methods that give scores to input attributes to a predictive model that shows how important each feature is when used to make a forecast. These relative scores can show which attributes may be most apposite to the goal and the least apposite.

The next step is then prediction using the test data and coming up with the confusion matrix. The confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.

True Positive is the value predicted positive and it is true. True Negative is the value predicted negative and it is true. False Positive (Type 1 Error) is the value predicted positive and it is false. False Negative (Type 2 Error) is the value predicted negative and it is false.

Then the overall results of the model are tested and plotted to be examined in a visual form.

### 3.6 Limitations of the study.

First, models like SVM are computationally expensive for large and complex datasets. They may also not perform well if the data have noise. The resultant model, weight and impact of variables are often difficult to understand.

Also, there are many factors that influence the occurrence of cancer cells in people other than the ones used in the study that were not included. In the use of a random forest, there is the risk of overfitting.

# CHAPTER 4: ANALYSIS, RESULTS AND DISCUSSIONS

## 4.1 Sources of data

It is the task of government agencies to regularly collect and publish healthcare data. The National Health Service in the United Kingdom collects and publishes this data on their data.gov.uk website under an Open Government License (OGL) to be used when needed.

This analysis will be based on the Epidemiology of Colorectal Cancer from the Cancer Registry. This data has been collected from 2000 to 2017. This data is extensive, relevant, and collected over a long period. Moreover, the demographics of Kenya and the United Kingdom, particularly population-wise, is also similar.

The information that can be obtained in the epidemiological data includes:

- The pseudotumor ID - this is the project specific marker for the tumors.
- The age group – this is aggregated as under 40's then afterwards in 5-year age bands up to the age over 90 plus.
- The sex - coded as 1=male, 2=female.
- The year of diagnosis.
- The site group of the tumor – groups include the proximal colon, distal colon, rectum and overlapping and unspecified neoplasms of the colon.
- The stage of the cancer progression.
- The morph group of the tumors.
- The underlying cause of death – whether the patient died of the cancer or unrelated causes.
- The vital status of the patient.
- The days that have elapsed since diagnosis and current vital status.

• The grade of tumor – from differentiation that cannot be assessed, well differentiated, moderately differentiated, poorly differentiated and undifferentiated.

• Whether they have been screened or not.

• The basis of diagnosis – from histology, cytology to autopsy reports.

## 4.2 Description of the software used.

For the data analysis, R software is used. A number of packages are used for data visualizations and to train the models. To visualize the data – ggplot and ggpairs are used and for training mainly the caret package is used.

## 4.3 Assumptions, data modifications and data checks.

Checks for missing values and exaggerated values were done. Both were negative. Moreover, some data modifications were done. In the age group, the bands of ages were each given number from 1 all the way to 12 for the subjects over the age of 90.

The sites of the tumors were also given numbers 1-.4. Proximal colon being 1, distal colon being 2, rectum 3 and overlapping and unspecified neoplasms of the colon being 4. Grade of differentiation were also given 1-5 from well differentiated to cannot be assessed.

Whether screening or not happened, dummy variables 0 and 1 were chosen. 0 being unscreened and 1 being screened. Variables that were dropped and not used for the analysis were the pseudotumor ID, the year of diagnosis, the morph of the tumor and the basis of diagnosis.

## 4.4 Data Visualizations and Model Fitting

### 4.4.1 Data Visualization

The first plot made was a histogram of the age groups to see how the age groups were distributed in the data.

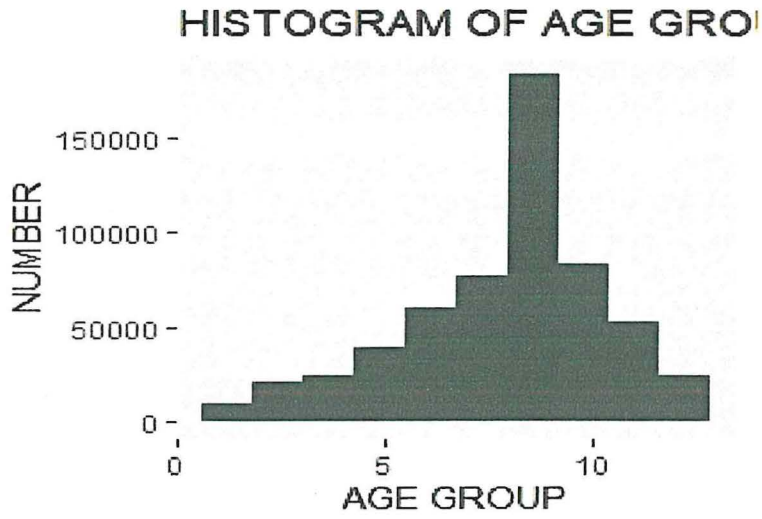



*Figure 4.1: Histogram of how the age groups are distributed.*

According to figure 4.1, the majority of the subjects in the study range from the ages of 60-79 with the number becoming progressively fewer in the subsequent age groups.

A bar plot of how the vital status varied with the age groups was also done.

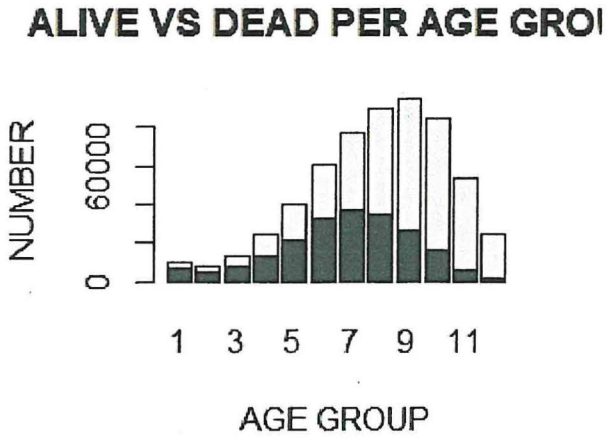



*Figure 4.2: Bar plot of vital status vs age group*

From figure 4.2, it can be seen that most of the alive patients are found in the younger ages while more patients that died were in the sunset years of their lives which is consistent with mortality assumptions as mortality increases with age.

It is also necessary to see how the stage of the cancer progression affects how the patient survives. A plot of this was made.

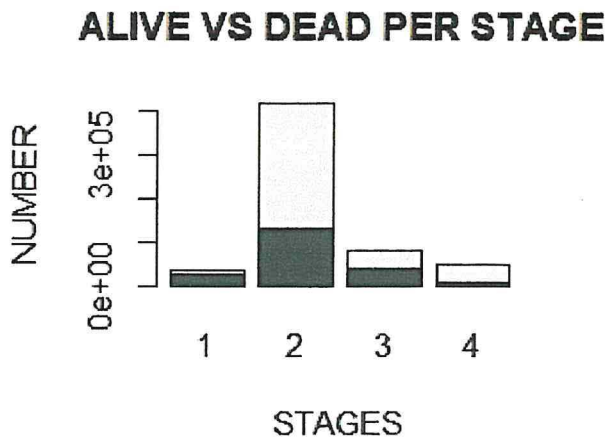



*Figure 4.3: Bar plot of alive and dead patients for each stage.*

From this figure 4.3, two observations can be made. First, more patients are found in the second stage, followed by stage 3, 4 then stage 1. Looking at the survival for each stage; in stage 1 of cancer progression, more patients are alive than dead. For stage 2, more patients die here than those who survive. The rate in stage 3 is almost the same while for stage 4, a large proportion are dead than alive - the chances of survival are too slim.

Next, a plot was made to seen to see how the grade of differentiation of the tumors affect the survival of patients.

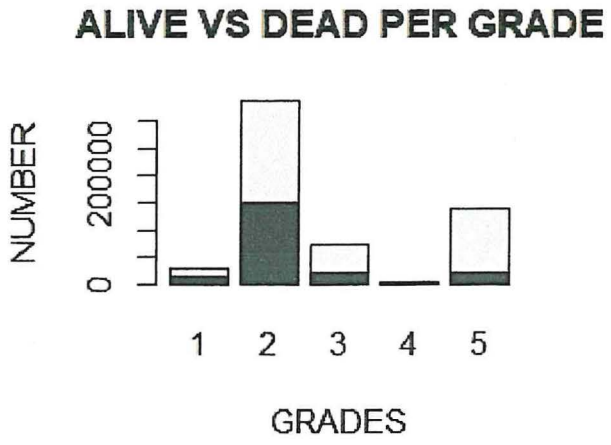



*Figure 4.4: Bar Plot of alive and dead patients per grade of differentiation.*

From figure 4.4, it can be seen that most patients have moderately differentiated tumors, followed by those that cannot be assessed, poorly differentiated tumors, well differentiated tumors and the least number of patients have undifferentiated tumors. For moderately differentiated tumors – the proportion of alive vs dead is almost the same although there are more dead patients, for poorly differentiated and not assessable tumors, there are more dead patients than those alive. For undifferentiated tumors, the chances of survival are high as almost all the patients here are alive.

There was a plot of every site group and how the alive and dead patients are distributed per site group.

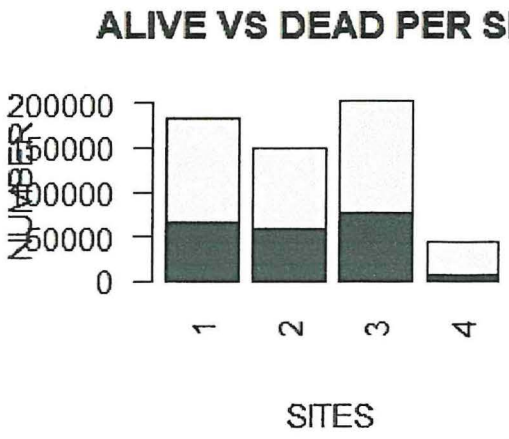



*Figure 4.5: Alive and Dead Patients per site group.*

From figure 4.5, it can be seen that most patients had tumors in their rectum followed by the proximal colon, distal colon then the overlapping tumors. For the overlapping tumors, the chance of survival is very low compared to the other tumor sites.

The next visualization is to see how the alive patients stack up against those who have passed away.

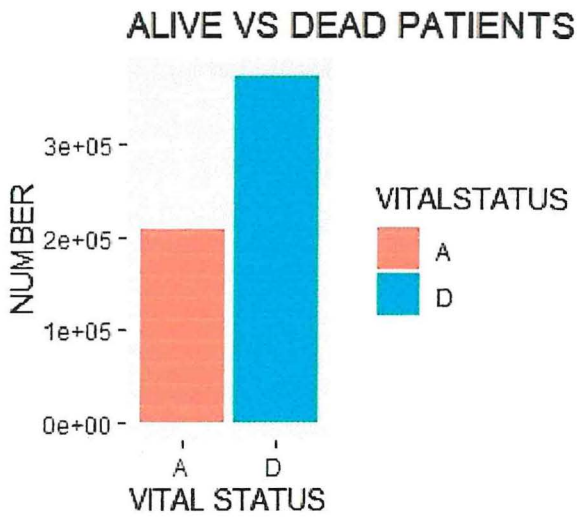



*Figure 4.6: Plot of Alive vs Dead Patients*

From figure 4.6, it can be seen that most subjects in the study are dead rather than alive - almost two times more. This shows that the dataset is imbalanced which could lead to problems with the prediction. To deal with this, Random Over-Sampling Examples (ROSE) was employed to seek to balance the data.

### 4.4.2 Regression Model.

To fit the regression model, the caret package in R is used. The data is loaded first and ROSE is applied to deal with the imbalance. The data is shuffled using an index to make sure that random observations are chosen for the training and testing data sets.

Training and testing data sets were chosen with 70% of the data used for training and 30% for testing. The model is trained with the generalized linear model method. 5 repeated cross validations are done while training.

After training the model, predictions are made on the testing data and residuals calculated. A graph of how the predictions vary from the actual is also plotted. The misclassification error is also calculated using a function. For our regression model, the misclassification error is 0.226 meaning the model is 77.4% accurate.

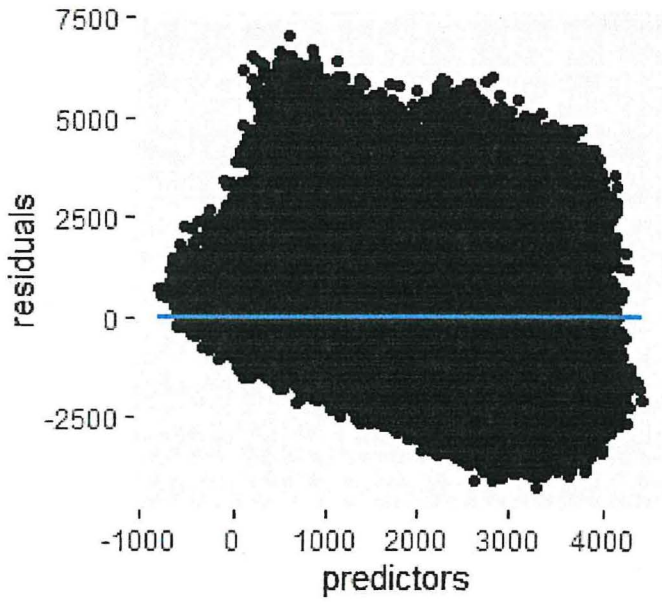



*Figure 4.7: Predictors vs Residuals*

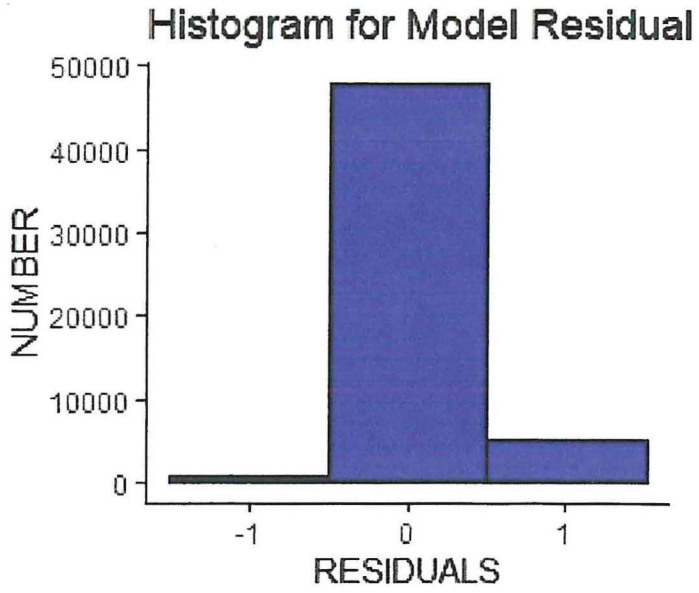



*Figure 4.8: Histogram for Model Residuals*

From figure 4.8, it can be seen that most of the residuals are centered around 0 which is a good thing. However, figure 4.7 has quite a number of points away from 0, meaning the regression might not have been that good at all.

### 4.4.3 Decision Tree

Using the same data, and same training and testing data sets, a decision tree is fitted to the data. In this case rpart is used for the training of the model. The training dataset has 401502 data points while the testing dataset has 173615 data points. The method used in this case is classification. Training on defaults results in a decision tree that has too many end nodes to be viewed. In this case, minbuckets are set to 5000 to get fewer end nodes that can be viewed.

Predictions are made on the testing data. A confusion matrix is called to see how the algorithm predicted the classes. The results, as shown in figure 4.10, were - it classified 72079 alive patients correctly and 67890 dead patients correctly. Conversely, it misclassified 14783 alive patients as dead and 18863 dead patients as alive. The accuracy for the model then is 0.806 which is 81%.

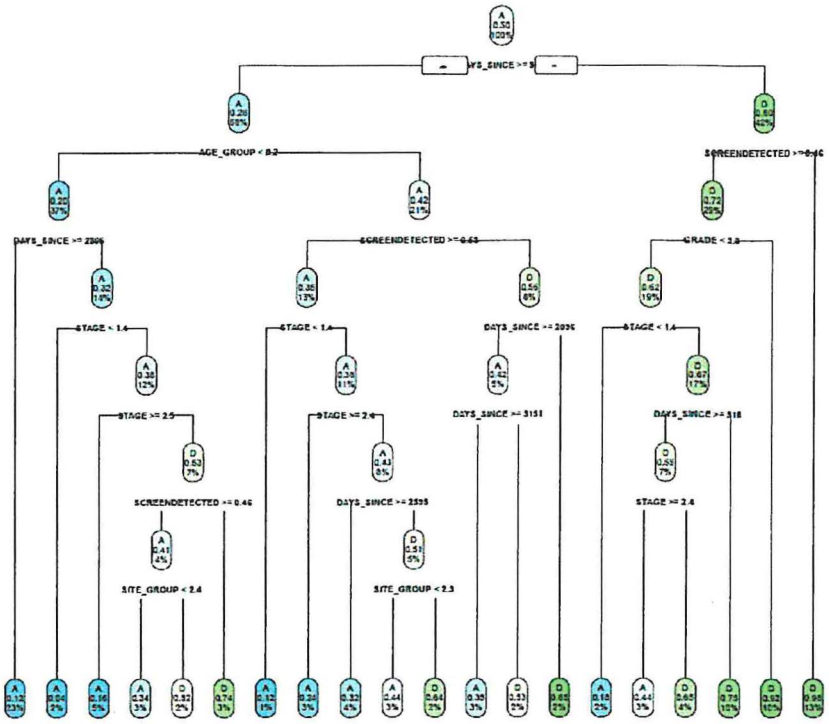


*Figure 4.9: The Decision Tree*

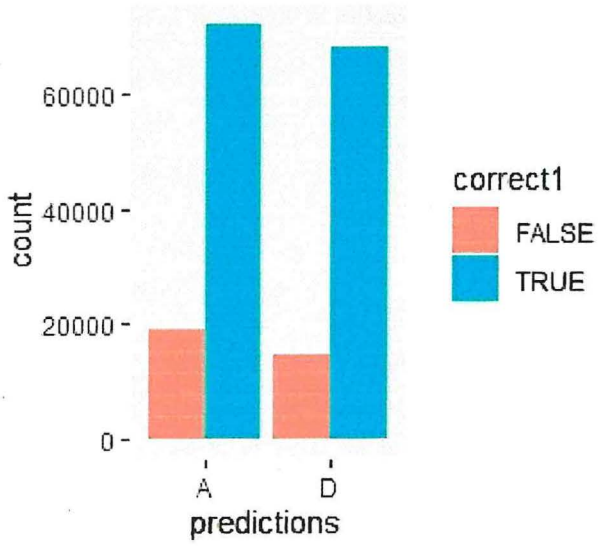
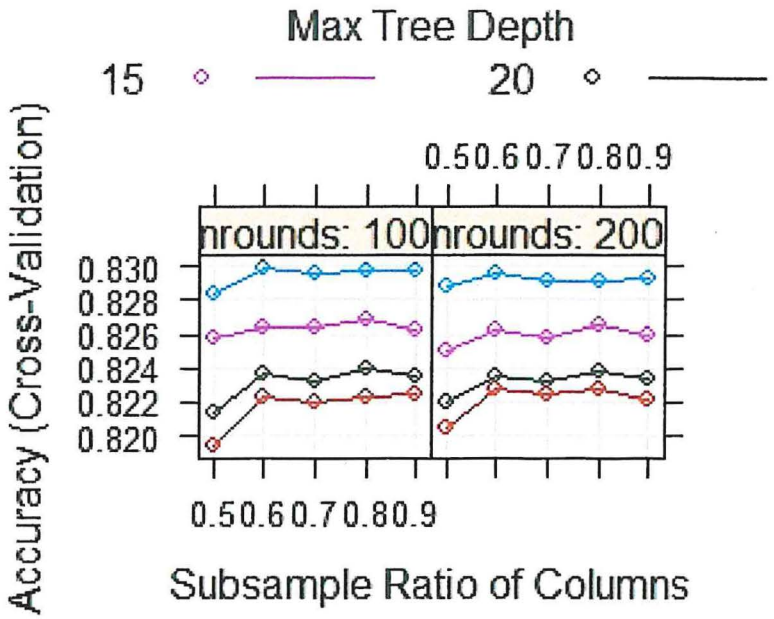



*Figure 4.10: Graph showing correct predictions vs incorrect predictions.*

### 4.4.4 Extreme Gradient Boosting Tree

An extreme gradient boosting tree is trained using the xgbtree method. Although the same training and testing data sets are used, two matrices one from the training and testing data sets are created. How the training is controlled is also specified. 5 cross validations are used. The maximum tree depth is also set to either 10,15, 20 or 25.

As with the others, predictions are made on the testing data. A confusion matrix is called for the tree. The results were - it classified 74331 alive patients correctly and 69658 dead patients correctly. Conversely, it misclassified 12531 alive patients as dead and 17095 dead patients as alive. The accuracy for the model chosen after the 5 cross validations is then 0.829 which is 83%.




4

*Figure 4.11: The Extreme Gradient Boosting Tree*

From figure 4.11, it can be seen that with each subsequent cross validation, the accuracy of the model increases.

# CHAPTER 5: SUMMARY OF FINDINGS, CONCLUSIONS AND RECOMMENDATIONS

## 5.1 Introduction

This chapter presents the summary of the findings of the study, the conclusions, and the relevant recommendations. It also provides some suggestions for further study.

## 5.2 Summary of Findings.

The research set out to build predictive models for colorectal cancer in the UK to establish whether a patient diagnosed today would die or survive. This was an explanatory study based on a causal design that utilized secondary data collected through healthcare data collected by the NHS. The data contained information for the years 2000 to 2017 and contained a host of variables like site of the tumor, grade of differentiation and the stage of the cancer progression that were unique to a certain patient.

In order to achieve the research objective, four models were built, and their performance assessed. The regression model was based on a GLM that had the vital status - alive or dead as the dependent variable. The model had an accuracy of 0.7742583. The coefficients for the variables were mostly positive – all but the screening variable.

The next model built was a decision tree. The decision tree showed possible outcomes and probabilities for each course of action. The decision tree had an accuracy of 81% correct predictions. The decision tree had a number of branches – it mostly used age group, screening, stage, and grade as its parameters for branching.

The next model was an extreme gradient boosting tree that was built for its speed and tuning. The extreme gradient boosting tree was subject to a number of repeat cross validations. All in all, it came up with an accuracy of 83% of correct predictions. Feature importance was also calculated for this model. From the feature importance – the most relevant variables were the stage of the cancer followed by the site group, the age group, the grade of differentiation, days since diagnosis, screening and the least important the sex.

## 5.3 Conclusions.

This paper set out to build predictive models for colorectal cancer in the UK. The study specific objectives of the study were to assess the performance of the models, and thereafter their applicability and whatever insights that can be drawn from them.

First and foremost, the findings showed that the extreme gradient boosting tree performed the best and had better accuracy compared to the decision tree and lastly the regression model. This is consistent with the findings of the study conducted by Uddin & others in 2019.

The findings of this study also indicate that indicate that the most significant features in the prediction were the stage of the cancer followed by the site group, the age group, the grade of differentiation. These insights are drawn from the feature importance. This means that, for the stage, as the cancer keeps on towards the latter stages, the chance of someone dying from the cancer keeps increasing.

Looking at the age, the more advanced a person is in age it is easier to predict that that person would die from the cancer should they be diagnosed. The grade of differentiation also is very significant in predicting survival or death for a patient. Undifferentiated cells which are in grade 1 look like normal cells and are not growing rapidly. As the grade increases to 3 or 4, the cells start looking more abnormal and this is a sign they are growing rapidly. At these stages, it is then easier to predict death for a patient.

Conversely, days since diagnosis, screening and sex of the patient were less important. This is because cancer develops differently for different people – it may be aggressive for one and not the same for another so the days since diagnosis only may not be reliable to predict whether one dies or not. Moreover, although screening presents an opportunity to catch the cancer early and start treatment earlier increasing chances of survival, it is not certain that this treatment will work or not.

Although these are insignificant, they should not be ignored all together since they still have some effect to some extent, on whether they live or die.

This study also ascertained that the model is applicable for use in the UK health sector. Not only is colorectal cancer a current concern in the country, but the ageing population also coupled with growing lifestyle diseases like obesity means that it will continue being a concern for the health care providers. Moreover, the vast amount of data being collected and released by the NHS on a regular basis only shows the great opportunities to use this data for insights and to build such models. Also, with 250 million pounds pledged by the Health Secretary, Matt Hancock, towards AI in healthcare (Brickwood, 2020) such a study becomes even more applicable and relevant.

### 5.4 Limitations.

Some of the models built in this study are computationally heavy to come up with – running for hours to just come up with the base model. These models have not been subjected to rigorous hyperparameter tuning which would make them sounder and improve their accuracy by far. This is a process that for some heavy models would require even weeks and machines with greater computation power.

Moreover, the accuracy while good could be better. A score of 83% means 4 in 5 patients would get a right prediction. However, this does not bode well for that one who would get a wrong prediction.

### 5.5 Recommendations.

The first recommendation would be further study into developing and implementing predictive models in the health sector. Moreover, for this particular study, the models can be advanced to accommodate larger datasets and in the presence of vast computational power, tuned further to improve their accuracy to more supreme models.

Moreover, to the governments and policy-makers, to create and sustain environments in which AI is embraced particularly in the healthcare sector not only for cancer but for other diseases.

The advancements that can be made from these methods would help in help in the formulation of good policies to combat cancer, also help in the budget formulation and aid in government service delivery.

Not only the government, but also insurance companies should embrace these models and use them where possible. This can help in proper contract design, assumption setting, setting of reserves, and pricing of their products wherever applicable.

Pharmaceutical companies also should be encouraged to use these models to forecast and establish the need for their products. This would help in their business strategy decisions – gearing their companies towards the best possible experience.

Finally, the variables used here are not the only ones that can affect whether one dies from cancer or not – there are host of many others. Extensive research and data collection should be done to find other indicators and predictors. Innovation should also be encouraged to build more creative models up until proper, robust, foolproof models are found.

## LIST OF REFERENCES.


Admin, A. (2011, September 6). *Health Care Reform: Duties and Responsibilities of the Stakeholders*. Retrieved from https://sites.sju.edu/icb/health-care-reform-duties-and-responsibilities-of-the-stakeholders/

Atlas, L. et al. (1990, October 1). *A performance comparison of trained multilayer perceptrons and trained classification trees - IEEE Journals & Magazine*. Retrieved from https://ieeexplore.ieee.org/abstract/document/58347

Barcelona Institute for Global Health. (2017, October 26). *Kenya's Struggling Health System - Blog*. Retrieved from https://www.isglobal.org/en/healthisglobal/-/custom-blog-portlet/kenya-s-struggling-health-system/5083982/8601

Brickwood, B. (2020, April 30). *How is AI and machine learning benefiting the healthcare industry?* Retrieved from https://www.healtheuropa.eu/how-is-ai-and-machine-learning-benefiting-the-healthcare-industry/98260/

Brownlee, J. (2019, August 12). *Supervised and Unsupervised Machine Learning Algorithms*. Retrieved from https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

Cambridge Wireless. (2019, December 2). *Top 10 AI in Healthcare Start-Ups in the UK.* Retrieved from https://www.cambridgewireless.co.uk/news/2019/dec/2/top-10-ai-healthcare-start-ups-uk/

Cancer Organization. *How Common Is Breast Cancer? | Breast Cancer Statistics. (n.d.)*. Retrieved from https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html

Cancer Research Institute. *Screening for Cancer*. (2018, April 9). Retrieved from https://www.cancer.gov/about-cancer/screening

Cancer Research UK. (2020, September 30). *Cancer incidence statistics*. Retrieved from https://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence#heading-Zero

Cancer Research UK. (2021, January 29). *Bowel cancer statistics*. Retrieved from https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer#:%7E:text=Bowel%20cancer%20risk,bowel%20cancer%20in%20their%20lifetime

Castañón, J. (2019, May 2). *10 Machine Learning Methods that Every Data Scientist Should Know*. Retrieved from https://towardsdatascience.com/10-machine-learning-methods-that-every-data-scientist-should-know-3cc96e0eeee9

Chitunhu, S. et al. (2015, July 8). *Direct and indirect determinants of childhood malaria morbidity in Malawi: a survey cross-sectional analysis based on malaria indicator survey data for 2012*. Retrieved from https://link.springer.com/article/10.1186/s12936-015-0777-1?error=cookies_not_supported&code=384bdec8-4a13-44c5-9165-87b68914f26b

Davenport, T. (2019, June 1). *The potential for artificial intelligence in healthcare*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181/

Flat World Solutions. *Top 10 Applications of Machine Learning in Healthcare - FWS. (n.d.)*. Retrieved April 19, 2020, from https://www.flatworldsolutions.com/healthcare/articles/top-10-applications-of-machine-learning-in-healthcare.php

Global Cancer Observatory. (2019, July 21). *Cancer today*. Retrieved from https://gco.iarc.fr/today/online-analysis-table?v=2020&mode=cancer&mode_population=continents&population=900&populations=900&key=asr&sex=0&cancer=39&type=0&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_group%5B%5D=17&group_cancer=1&include_nmsc=1&include_nmsc_other=1

Guru, U. (2020, March 17). *Supervised vs Unsupervised Learning: Key Differences*. Retrieved from https://www.guru99.com/supervised-vs-unsupervised-learning.html

Halliday, K. et al. (2014, February). *Impact of Intermittent Screening and Treatment for Malaria among School Children in Kenya: A Cluster Randomized Trial*. Retrieved from https://elibrary.worldbank.org/doi/abs/10.1596/1813-9450-6791

Institute of Economic Affairs. (2018, April 20). *Major Causes of Mortality In Kenya*. Retrieved from https://www.ieakenya.or.ke/number_of_the_week/major-causes-of-mortality-rates#:%7E:text=Major%20Causes%20of%20Mortality%20Deaths&text=According%20to%20the%20Kenya%20National,4%25%20and%203%25%20respectively.

Jamgade & Zade. (2019, May). *Disease Prediction Using Machine Learning*. Retrieved from https://www.irjet.net/archives/V6/i5/IRJET-V6I5977.pdf

Konerman, M. A. (2019, January 4). *Machine learning models to predict disease progression among veterans with hepatitis C virus*. Retrieved from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0208141

Liu, Zhang & Zaravian. (2018, August 15). *Deep EHR: Chronic Disease Prediction Using Medical Notes*. Retrieved from https://arxiv.org/pdf/1808.04928v1.pdf

Makau-Barasa, L. K. (2020, January 7). *A review of Kenya's cancer policies to improve access to cancer testing and treatment in the country*. Retrieved from https://health-policy-systems.biomedcentral.com/articles/10.1186/s12961-019-0506-2#availability-of-data-and-materials

Marr, B. (2019, October 4). *4 Powerful Examples Of How AI Is Used In The NHS.* Retrieved from https://bernardmarr.com/default.asp?contentID=1834#:%7E:text=HeartFlow's%20AI%20technology%20is%20also,flow%20is%20disrupted%20by%20blockages

Mitchell, T. M. (1997). *Machine Learning*. New York, United States: McGraw-Hill Education.

Morde, V. (2019, April 15). *XGBoost Algorithm: Long May She Reign! - Towards Data Science*. Retrieved from https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

Muinga, N. (2020, January 6). *Digital health Systems in Kenyan Public Hospitals: a mixed-methods survey*. Retrieved from https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-1005-7

Murphy, K. P. (2012). *Machine Learning*. Amsterdam, Netherlands: Amsterdam University Press.

Nanyingi, M. (2020, March 30). *Predicting COVID-19: what applying a model in Kenya would look like*. Retrieved from https://theconversation.com/predicting-covid-19-what-applying-a-model-in-kenya-would-look-like-134675

Nevin, L. (2018, November 30). *Advancing the beneficial use of machine learning in health care and medicine: Toward a community understanding*. Retrieved from https://journals.plos.org/plosmedicine/article?id=10.1371%2Fjournal.pmed.1002708

Qianfan et al., (2020, May 17). *Deep Learning Methods for Predicting Disease Status Using Genomic Data*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6530791/

Sangwon, Sungjun, and Donghyun. (2018, August 1). *Predicting Infectious Disease Using Deep Learning and Big Data*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6121625/

Sewe, M. et al. (2017, June 1). *Using remote sensing environmental data to forecast malaria incidence at a rural district hospital in Western Kenya.* Retrieved from https://umu.diva-portal.org/smash/get/diva2:1116951/FULLTEXT01.pdf

Sperrin, M. (2020, October 20). *Prediction models for covid-19 outcomes*. Retrieved from https://www.bmj.com/content/371/bmj.m3777

Team, E. S. (2019, November 11). *What is Machine Learning? A definition.* Retrieved from https://expertsystem.com/machine-learning-definition/

The Medic Portal. (2019, December 13). *Challenges Facing The NHS and Current NHS Issues*. Retrieved from https://www.themedicportal.com/application-guide/the-nhs/challenges-facing-the-nhs/

Uddin et al., (2019, December 21). *Comparing different supervised machine learning algorithms for disease prediction.* Retrieved from https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-1004-8

Waweru, D. (2019, November 13). *Kenya Leads Africa In Artificial Intelligence. Here's Why*. Retrieved from https://gadgets-africa.com/2019/11/13/kenya-leads-africa-in-artificial-intelligence-heres-why/

Wiens, Jenna & Shenoy, Erica. (2017*). Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology*. Boston, United States: an official publication of the Infectious Diseases Society of America.