SOLENT
UNIVERSITY

Applications of Machine Learning in Predicting the Risk of Cancer in Adults: A Quantitative Risk Assessment Model

Dissertation Research Project

For

MSc. Applied AI and Data Science
Faculty of Business Law and Digital Technologies
Solent University

Abiodun Gabriel Ajanaku
Student ID: 15798682

September 9, 2022

Supervisor: Dr. Olufemi Isiaq

SOLENT UNIVERSITY
FACULTY OF BUSINESS LAW AND DIGITAL TECHNOLOGIES

MSc. Applied AI and Data Science
Academic Year 2021-2022

Applications of Machine Learning in Predicting the Risk of Cancer in Adults: A Quantitative Risk Assessment Model

Abiodun Gabriel Ajanaku
Student ID: 15798682

Supervisor: Dr. Olufemi Isiaq

September 2022

This dissertation is submitted in partial fulfilment of the requirements of Solent University for the degree of MSc. Applied Artificial Intelligence and Data Science

TABLE OF CONTENTS

Certification

I hereby certify that the dissertation titled "***Applications of Machine Learning in Predicting the Risk of Cancer in Adults: A Quantitative Risk Assessment Model***", submitted by **Abiodun Gabriel Ajanaku,** with student ID: **15798682** for the award of MSc. Degree in Applied Artificial Intelligence and Data Science, embodies original work conducted for the project by me. All information other than my own contribution will be fully referenced at the rear of the project.

Signature: Abiodun Gabriel Ajanaku

Date: September 9, 2022

Acknowledgement

I am heartily thankful to my supervisor, Dr. Olufemi Isiaq whose continuous supervision, encouragement, guidance, and support from the initial to the final level enabled me to develop an interesting thesis.

Thanks to my wife, Emmanuella Ajanaku and my entire family members for their moral, and material support that have made this work successful.

ABSTRACT

Background

Several studies have shown that everyone has a certain risk of developing cancer. The interaction and combination of genes, lifestyle, chronic diseases (comorbidities), and environment can influence this risk. Hence, early detection of the risk factors through the application of machine learning will enhanc personalized predictive healthcare, reduce the incidence and mortality of cancer.

Objectives

This research study is aimed at: (1) Identifying different risk factors and their effects in cancer risk classification for both symptomatic and asymptomatic adults using machine learning (ML) algorithms, (2) Identify, compare the performance of 6 machine learning models, and select the most suitable model for prediction, (3) synthesise existing knowledge, identify gaps and patterns necessary to train and deploy a cancer risk predictive model.

Methods

A quantitative analysis of the primary dataset of 580 records with 58 features obtained through survey questionnaire from participants (aged 18 and above) was carried out. Clustering analysis was deployed using KMeans and KModes to extract knowledge, patterns and group the datapoints in two class labels: Low and High cancer risk. The dataset was divided into training (n=371), validation (n=93) and testing (n=116) for building, training, and testing the model respectively. These techniques adopt the 80/20 train-test split. The data were used to construct six machine learning models including decision tree (DT), random forest (RF), logistic regression (LR), support vector machine (SVM), naïve bayes (NB) and K-Nearest Neighbour (KNN) to predict and classify the risk of cancer in adults. A total of 10 relevant variables were input into these models. The performance of the models was measured using the area under receiver operating characteristic curve (AUC-ROC) and the confusion metrics.

Results

Two (2) cancer risk class (Low: 1 and High: 0) were predicted. Of the six machine learning models, the Random Forest ensemble learning methods had the best performance in predicting cancer risk (AUC: 0.93), outperformed conventional logistic regression (AUC: 0.83). While the decision tree, support vector machine, naïve bayes and KNN have AUC of 0.89, 0.87, 0.80, 0.80 respectively. With the Random Forest model, the number of people considered as having a high-risk of cancer (that is, RECALL, otherwise known as Sensitivity) was 93% with miss rate (false negative) of 7%. Gender, sugar intake, frequency of exercise, smokes per pack year, level of alcohol intake, comorbid such as frequent cold, exposure to industrial pollution, domestic pollution, sun and co2 emissions were the 10 relevant risk factors used by the RF model.

Conclusion

The results of the study showed that based on baseline information of a person, machine learning can accurately predict and classify the risk level of cancer. Beyond the lifestyle and environmental risk factors, the study included chronic diseases (comorbidities) in a person in building the predictive model which have been overlooked in many previous studies. In the context of this study and the demographic of the participants, risk factors such as gender, exposure to sun, alcohol intake, chronic diseases, pollution resulting from domestic, co2 emissions, and industrial practices contributed significantly in classifying the cancer risk level for Black African, Asian-Indian, Asian-Pakistani, White British and White Irish.

Keywords: Machine learning in healthcare; cancer; cancer risk factors; comorbidities; decision support system; predictive modelling; machine learning; risk assessment; oncology.

CHAPTER ONE

Introduction and Background

1.1. Background

Cancer is a leading cause of deaths globally and co-exist with other chronic diseases such as COVID-19 and tuberculosis. It is a non-communicable disease which has a high potential of morbidity and mortality; caused by a rapid development of abnormal cell that grows beyond their usual boundaries which have the tendencies to invade adjoining part of the body and spread to other organs(World Health Organization 2022). In other words, the disease could be called a malignant tumours or neoplasms.

The Problem: a global perspective

In 2020, cancer accounted for nearly 10 million deaths globally, representing one in six deaths. Around one-third(33.33%) of the global deaths from cancer were due to lifestyle (modifiable) risk factors including: use of tobacco, high body mass index, alcohol consumption, low fruit and vegetables intake, and lack of physical activities (World Health Organization 2022). The combination of lifestyle and other risk factors such as family and past health history have influenced the development of many popular types of cancers such as breast, lung, colon, rectum, and prostate cancers. Furthermore, chronic infections as human papillomavirus (HPV) and hepatitis which are cancer causing infections are responsible for 30% of the cancer cases in low and lower-middle - income countries.

In the UK for example, one in every two people will be told they have cancer at some point in their lifetime (GOV.UK 2022). Based on recent available data, around 366,303 cases of cancer were diagnosed in 2017. Out of the total cases of 366,303, 51% are in men and the remaining 49% were in women.

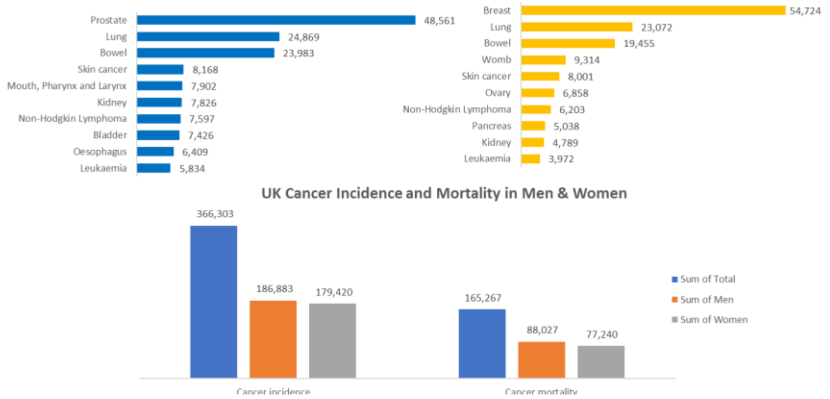


Fig 1: UK Cancer Statistics (Adapted from: World Cancer Research Fund)

In contrast, 1,708,921 new cancer cases were reported and 599,265 people died of cancer in the United States in 2018. In addition, 436 new cancer cases were reported and 149 people died of cancer per 100,000 population in the US (Centers for Disease Control and Prevention 2021). Whereas, In Africa Cancer is an emerging health problem, and it been projected that by 2030 there will be a 70% increase in the incidence and mortality rate due to accelerated population growth, aging, shortage of medical equipment, research resources and epidemiological expertise. Around 57% of all new cancer cases globally occurred in low-income countries, and significant part of this were from Africa (Hamdi et al 2021).

The Root Cause of Cancer

Cancer is directly caused by the interaction between a person's genetic factors and three main external agents including physical, chemical, and biological carcinogens. In other words, It is a genetically driven disease that interacts with other risk factors to determine an individual's risk (Hamdi et al 2021). The carcinogens include ultraviolet and ionizing radiation for the physical agents; asbestos and aflatoxin associated with tobacco smoke and alcohol respectively make up the chemical components. While the biological external agents are made up from infections from certain viruses, bacteria, and parasites. Hence, the development of cancer starts when cells change abnormally, divide in an uncontrollable way, and spread into other tissues to develop from a pre-cancerous lesion to a malignant tumour. Furthermore, the lack of early awareness, lack of preventive strategies, and increased life expectancies have accelerated the incidence of cancer globally.

Cancer Risk Factors

A cancer risk is anything including modifiable behaviours and unmodifiable factors that can potentially increase a person's chance of getting cancer (Ghanizada et al 2020). Although most of the risk factors do not directly cause cancer, however, they can influence the development of cancer (American Society of Clinical Oncology 2022). Modifiable risk factors include but not limited to lifestyle behaviours such as alcohol consumption, use of tobacco, unhealthy diet, physical inactivity. This also includes air pollution, and occupational risk. The unmodifiable risks include age, genetic or other family medical history of a person. Beyond the modifiable and unmodifiable risk factors, some chronic infections associated with geographical locations are risk factors for cancer which are common in low-and-middle income countries and accounted for 13% of the global cancers diagnosed in 2018 (World Health Organization 2022). This includes chronic infections such as Helicobacter pylori, human papillomavirus (HPV), hepatitis B virus, hepatitis C virus, and Epstein-Barr virus.

Related Work and Current State of the Art

As a result of the accelerating increase in the incidence and mortality of cancer, research on the risks of cancer has also increased in the past few years. Many recent studies have focused on the applications of machine learning in predicting the different types of cancers, only a few researchers have taken the effect of risk factors on level of cancer risk into consideration. Hence, this provides the opportunity for further investigation.
In recapping what has been on the research topic, the PubMed database was utilised for the systematic review of literature as detailed in chapter 2 of the research project.

Alfayez et al. 2021, carried out a systematic scoping review on the application of machine learning in predicting the risk of cancer in adults. The authors used a scoping review based on population, context and concepts to identify and understand the various machine learning that have been deployed in predicting the risk of cancer using clinical, demographic, and lifestyle behaviour datasets. The finding of the research showed that wide variety of ML models were used in different studies including Artificial Neural Networks(ANN: 8 out of 10 studies), Logistic Regression (2/10 studies), Gaussian naïve Bayes (1 out of 10 studies), Bayesian network inference (1/10 studies), DTs (1/10 studies) and RFs (2/10 studies), linear discriminant analysis (LDA) (1/10 studies), and SVMs (1/10 studies). The scoping review did not identified a single 'best' method. This is because not all models generalised well to the validation datasets.

Bundazak et al 2019 analysed and predicted the risk of cancer using Decision Tree Classifier using 1,030 datasets obtained from a survey. The research used 10 variables including participants smoking behaviour, dietary, alcohol consumption, exercise, occupational risk, environmental risk, medical, personal history, and other demographic profile of participants. The authors deployed two machine learning models: decision tree classifier and neural network. The results of the research showed that decision tree (DT) model performed better that than the Neural Network model. The DT model produced 85.63%, 16.7%, and 35% for accuracy, mean-squared error, and root mean squared error respectively. Whereas the neural network produced a score 45.3, 0.5013 and 0.5017 for accuracy, mean-squared error, and root mean squared error respectively. Beyond the deployment of the machine learning, there are other risk factors that influence a person's level of cancer risk. According to Hamdi, et al 2021, the diverse ethnicities, and sub-populations in some parts of the globe more particularly associated with low-income countries have also contributed to the increase in the incidence of cancer globally. These factors manifest a number genetically associated cancers that affect different groups over others in a very disproportionate manner. Hence, as the global risk of cancer declines, the genetically associated cancers become very more obvious and spread relatively. The increase in life expectancy resulting from increased age has exacerbated the incidence of cancer significantly (Hamdi, et al 2021). This implies that the incidence and mortality of cancer increases with age; age increase directly increases life expectancy which in turns leads to more cases of cancer globally.

In terms of the state- of-the- art, the major approach in identifying the risk of cancer at the early stage is the population-wide asymptomatic screening, which is a pathological approach that is aimed at identifying individuals who do not show symptoms of cancer and could be at risk. This screening includes but not limited to mammography, cervical, and faecal occult blood testing for breast, cervical, and colorectal cancers respectively. There are numerous nation-wide screening programmes in the US and UK designed to specifically for early detections of the most common cancers including breast, colorectal, bowel and cervical. For instance, the US has the National Breast and Cervical Cancer Early Detection Programme (NBCCEDP). Despite the numerous nation-wide efforts and resources that have mobilized to identify early risk of cancer, majority of the people eligible for the screening programme do not participate due to fear, not creating time for the screening or they do not see any value in the programme. Secondly, both the World Cancer Research Fund and the Cancer Research UK have an online web cancer risk quiz platform that people complete in understanding their risk of cancer. These platforms are merely informational without categorizing a person's risk of cancer accurately into class.

The Applications and Brief Description of Machine Learning

Machine Learning (ML) is a core part of Artificial Intelligence and extension of traditional statistical techniques that uses computational resources to detect underlying patterns in high-dimensional data, and it is increasingly being used in different areas in healthcare and other domains requiring predictions. The goal of machine learning models/algorithms is to facilitate and establish the best way in mapping of features (inputs) to output(labels) through learning from the dataset(Muller and Guido 2016). The five models selected for this research project includes Decision Tree Classifier, Random Forest Classifier, Support Vector Machine, Logistic Regression Classifier, and Deep Learning Neural Network. The meaning, advantages, and disadvantages of each of the selected model are described in detail in Appendix 12 of this report.

1.2. Research questions

The research question this study aimed to answer is to establish the effect of different risk factors in classifying the level of cancer risk using machine learning? The answer to this question will assist in achieving the overall aim and objectives of the project.

1.3. The Statement of Hypotheses

Below are the hypotheses in the study

I. The Null Hypothesis **(H_0)**: Cancer risk factors (predictors) do not have any effect on the target variable [Level of cancer risk]. That is, the coefficients of the predictors (β_1, β_2… β_{10}) are equal to Zero.

II. The Alternate Hypothesis **(H_1)**: Cancer risk factors (predictors) have significant effect on the target variable [Level of cancer risk]. That is, the coefficients of the predictors (β_1, β_2… β_{10}) are significant and not equal to zero.

1.4. Research Aim and Specific Objectives

The aim of the study is to investigate different risk factors in adults and identify relevant variables for machine learning models using 9 risk categories: anthropometric, demographic, environmental, personal history, family history, social and lifestyle, occupational, economic class, and co-existing diseases in a person. The identified relevant features will be used to map and predict the cancer risk level of a person into one of two class labels: low, or high risk.

The research objectives are:

1. To critically review literature on the risk of cancer in adults. This will help to identify gaps in research, the state-of-art, and relevant data.

2. To identify relevant features (cancer risk factors) in adults and establish the relationship and effect of the risk factors in predicting level of cancer risk: low or high.
3. Carry out data analysis and use the Python Scikit-learn libraries to build and evaluate the performance of the models using the confusion matrix and area under the receiver operating curve (AUC-ROC) which emphasis success metrics such as accuracy, recall, precision and F1-score.
4. Create and deploy a web application using Streamlit for cancer risk prediction. The deployment will make the prediction system consumable and available to the beneficiaries and other stakeholders.

1.5. Significance and Justification of the Study

The application of machine learning in healthcare domain is a relatively new technique when compared to pathological approach to detecting the risk of cancer. Therefore, this study is aimed to utilize machine learning techniques for early prediction of the risk of cancer in adults within the scope of general adults (aged 18 years and above). The below highlights the significant of the project:

1. The cancer risk prediction system will help beneficiaries detect the risk of cancer early and gain insights into the relevant cancer risk factors. The beneficiaries of the research project include individuals, health professionals, governments, and the global health community.
2. Existing cancer risk profiling platforms are merely informational and not designed based on any machine learning model to accurately classify the risk of cancers (Cancer Research UK 2022; World Cancer Research Fund 2022). This research provides practical solution to the gap through the applications of machine learning algorithms in predicting the risk of a cancer and makes a far- reaching departure from existing state of art. It combines both accuracy and speed in learning from data, finds patterns, trains itself using the labelled data to predict an outcome: low, medium, or high risk of cancer.

3. Existing studies has shown that there are barriers to cancer screening resulting from fear or concerns about medical procedures, lack of knowledge of risk factors, stigmatization, and difficulty in navigating the healthcare system. These barriers have contributed to people not interested in cancer screening resulting to high incidence and mortality of cancer. The findings from this study will directly benefit individuals and

will serves as the basis for making informed decision about early screening and diagnosis.

4. Healthcare professionals can adopt the prediction system as a preliminary (fact-finding) risk factors and vulnerability assessment to form the basis for referring patients for cancer screening and diagnosis.
5. The findings and insights from the research project will serve as a contribution to the body of knowledge especially on cancer epidemiology and new datasets.

Cancer incidence and mortality can be reduced when different risk factors are spotted and treated early. Failure to detect and manage the risk factors properly can lead to significant loss of lives globally due to cancer, increase in cancer burden and high cost associated with the treatment of the disease.

1.6. Dissertation Structure

Chapter 1: Introduction and Background

This chapter provides background information in relation to the applications of machine learning in predicting the risk of cancers. It summarizes the motivation and significance of the study. In addition, it outlines the research aims, objectives and overview of the research plan.

Chapter 2: Literature Review

This chapter carries out a critical review of related work on the application of machine learning from a global perspective, different context and countries including the United States, United Kingdom, Africa, and the rest of the world. It outlines what was done, the methods deployed, and the results or findings from previous research. This is crucial in shaping and validating the study.

Chapter 3: Research Methodology

This chapter outlines the principles, strategies, methods of data collection, analysis, evaluation, and design of the research project as well as the justification for specific issues highlighted in the literature review. The validity of this section is vitally important as it enhances the reproduction of the research and how conclusions have been established.

Chapter 4: Presentation of Results

This section presents the results of the study in tabular and visuals format ahead of the discussions of findings.

Chapter 5: Discussions and Presentation of Findings

This chapter presents the findings of the data analytics of the primary data collected through survey questions and machine learning activities. Previous research are also presented for comparative analysis to establish whether the study support or challenge empirical findings from reviewed literature.

Chapter 6: Conclusion and Recommendations for Future work

This chapter concludes based on extensive reflection on objectives of the study: analysis what worked, what did not work, limitations and the need for future work where necessary.

CHAPTER TWO

Literature Review

This section provides a detailed and critical review of related work on the study including books, journal articles and other resources. This is more important to investigate existing knowledge and the methodological approaches.

The accelerating increase in the global cancer burden and mortality has resulted into more research on the risks of cancer in the past few years. The recent demands for machine learning in healthcare and the growing trends towards personalised predictive medicine have facilitated the research process. Many recent studies have focused on the applications of machine learning in predicting the different types of cancer with emphasis on lifestyle and environmental factors, only a few research has taken cancer risk factors combined with comorbidities into consideration. Hence, this study provides the opportunity for further investigation.

In recapping related works on the research topic, the PubMed database was searched from inception to September 1 using the search string: ***(("Supervised Machine Learning"[Mesh] OR "machine learning in healthcare" OR "prediction of cancer" OR "ML") AND ("Neoplasms"[Mesh] OR "cancer" OR "Cancers" OR "Oncology" OR "prognosis")) AND ("Risk Assessment"[Mesh] OR "risk factors" OR "comorbidities" OR "symptoms").***

The search returned 594 unique articles of which 550 articles were excluded because they did not meet the inclusion criteria, were duplicates or were related to other studies. Full-text reviews were conducted for 44 articles and a final set of 10 articles were included for literature review because they deployed ML to predict cancer risk or cancer related diseases in individuals. Nevertheless, the search was supported with manual search of previously published related works. It is worth to mention that most of the previous research reported area under the receiver operating characteristics curve (AUC-ROC) as the performance evaluation metric, but just a few of the related work reported the model calibration, the number of runs and hyperparameters tuning.

Leug et al.(2021) applied machine learning models in predicting gastric cancer risk in patients after helicobacter pylori eradication. The research utilised a total dataset (n=89,568) of H. pylori-infected patients who had received clarithromycin-based triple therapy between 2003 and 2014 in Hong Kong with a training and validation split of 70% and 30% respectively. The dataset was deployed in constructing seven machine learning models for predicting gastric cancer over a period of 5 years H. pylori-infection treatment. A total of 26 relevant variables were inputs for the model: age, presence of intestinal metaplasia, and gastric ulcer were the heavily weighted risk factors used for the models. The performance evaluation of the models was carried out using area under receiver operating characteristic curve (AUC) analysis. The results of the study showed that out of the seven machine learning models, extreme gradient boosting (XGBoost) had the best performance in predicting cancer development (AUC 0.97, 95%CI 0.96-0.98). The model performed better than logistic regression (AUC 0.90, 95% CI 0.84-0.92). The work of Moncada-Torres et al.(2021) shows *that explainable machine learning models can outperform cox regression predictions (including other standard for survival analysis in oncology) and provide insights in breast cancer survival if the process of the decision making are transparent and explainable*. These are important factors to be considered for the adoption of the ML models in clinical settings. The authors utilised dataset (n=36,658) of non-metastatic breast cancer patients obtained from the Netherlands Cancer Registry for comparison of the performance of the standard Cox Proportional Hazards (CPH) analysis and machine learning models (ML) including Random Forest(RF), support vector machine(SVM), and extreme gradient boosting(XGB) in predicting the predicting survival of breast cancer patients. The study demonstrated that machine learning models such as RF, SVM and XGB outperformed the CPH due to the ability of the ML classification models to capture non-linear relationships and complex interactions of the multiple variables. In addition, the authors concluded that the ML classification models are better algorithms for predicting survival because they depict concise knowledge of the decision-making process and how the predictions were established, this is the critical success factors in accelerating the trust and adoption of innovative ML techniques in oncology and healthcare in general.

Ye et al. (2019) deployed the ensemble feature learning in identifying risk factors for predicting secondary cancer with due consideration for class imbalance and patients' heterogeneity. The authors utilised spectral clustering in dividing the patients into heterogeneous groups and oversampling was applied to each group to balance the sample in each group for the training of the model. Three ensemble models: decision tree (DT), support vector machine (SVM) and k-nearest neighbour (KNN) were used for the classification problem. Amongst the three classifiers, DT produced the best result for predicting secondary cancer with 0.72 and 0.38 in terms of AUC and F1-score when the patients were divided into 15 and 20 groups respectively. 20 variables were used for the predictions and the performance of the three classifiers improved when selected important features were used as predictors for the model. Achilonu et al.(2021) have argued that accurate prediction of patients at risk of cancer can enhance clinical expectations and decisions. The authors applied supervised machine learning approach in predicting the colorectal cancer recurrence and patients' survival: a South-African population-based approach. For higher predictive performance and interpretability, six supervised machine learning models were evaluated. The models included logistic regression(LR), naïve bayes(NB), decision tree C5.0, random forest (RF), support vector machine (SVM) and artificial neural network(ANN). Although the six algorithms produced high accuracy in terms of AUC-ROC and without much significant difference, however, ANN outperformed the other models with highest AUC-ROC for recurrence (87.0%) and survival (82.0%). The variables used as inputs for the modelling includes patients' age, histology, radiology stage which are relevant features for both cancer recurrence and survival. Stark et al.(2019) classified and predicted breast cancer risk using personal health data and six machine learning models including logistic regression, Gaussian Naive Bayes, decision tree, linear discriminant analysis, support vector machine, and feed-forward artificial neural network. The authors utilised the prostate, lung, colorectal and ovarian cancers (PLCO) data (n=78,215 for women ages 50-78) obtained from the National Cancer Institute sponsored screening trial for PLCO. The dataset was generated from randomized, controlled, prospective study that sought to determine the effectiveness of different cancers screening. The six models were trained based on 13 important features including age, age at menarche, age at menopause, age at first live birth, number of first-degree relatives who have had breast cancer, race / ethnicity, BMI, packed year of cigarettes smoked, an indicator of current hormone usage, number of years of hormone usage, BMI, years of birth control usage, number of live births, and an indicator of personal prior history of cancer. The performance of the models was evaluated based on the area under the curve (AUC) as well as sensitivity (otherwise known as recall), specificity, and precision. The logistic regression and linear discriminant analysis models have the highest AUC (0.613) and outperformed the other models. The logistic regression, linear discriminant analysis, and neural network has sensitivity score of 0.476, 0.688 and 0.599 respectively. In terms of specificity, logistic regression, linear discriminant analysis, and neural network has a score of 0.691, 0.562, and 0.467 respectively. Whereas all three models have a low precision score of 0.0323, 0.0272, and 0.0287 respectively. Duan et al.(2020) carried out research on the development of machine learning-based multi-mode diagnosis system for lung cancer. The authors deployed three machine learning algorithms including decision tree C5.0, artificial neural networks(ANN) and support vector machine (SVM) on 14 epidemiological data and clinical symptoms. The performance of the models were evaluated based on the area under the curve (AUC). The model produced results of 0.676, 0.736 and 0.640 in the first phase for C5.0, ANN and SVM respectively. In the second phase, the AUC score was 0.804, 0.889 and 0.825 for C5.0, ANN and SVM respectively. Furthermore, in the third layer better sensitivity score of 94.12% was found for the C5.0 supported by AUCs of 0.908, 0.910 and 0.849 for C5.0, ANN and SVM respectively. The work of Ali et al (2021) deployed machine learning-based statistical analysis for early-stage detection of cervical cancer: a cancer that is very common in women, particularly in less developed countries. The authors utilised the cervical cancers patient's dataset from the Kaggle repository including four main class of attributes: biopsy, cytology, Hinselmann, and Schiller. The study used three feature scaling techniques including log, sine, and z-score to transform the dataset and get them ready for machine learning. Several supervised machine learning models were evaluated based on accuracy and their corresponding performance in classification . The Random Forest (RF), and Instance-Based K-nearest neighbour (IBK) outperformed other models in the classification of Hinselmann and Schiller with accuracy score of 99.16% and 98.56% respectively. Choudhury (2021) carried out research on predicting cancer using supervised machine learning: mesothelioma. The author used patients' clinical data collected from by Dicle University, Turkey and applied eight machine learning models including: multilayered perceptron (MLP), voted perceptron (VP), Clojure classifier (CC), kernel logistic regression (KLR), stochastic gradient decent (SGD), adaptive boosting (AdaBoost), Hoeffding tree (VFDT), and primal estimated sub-gradient solver for support vector machine (s-Pegasos). The models were evaluated using the confusion matrix including evaluation metrics such as accuracy, precision, recall, f-statistics, root mean squared error and receivers' characteristic curve (ROC). The modelling was carried out in two phases, in phase 1: the highest and optimal performance was obtained for SGD, AdaBoost.M1, KLR, MLP, VFDT. Whereas, in phase 2, the best model was AdaBoost and outperformed other algorithms with classification accuracy of 71.29%. The relevant features which served as inputs for the model includes C-reactive protein, platelet count, duration of symptoms, gender, and pleural protein were found to be the most relevant predictors that can prognosticate

Mesothelioma. Cruz and Wishart (2006) carried out a detailed systematic review of the different machine learning techniques, the data types being used, and the performance of the various models in predicting cancers and prognosis. The main findings from the review were a bias towards the application of older technologies such as artificial neural networks (ANNs) in place of more interpretable and recent machine learning models such as logistic regression, naïve bayes, decision tree and other ensemble methods. In addition, the authors stated there were absence and lack of appropriate level of testing and validation of the models in several published studies. The authors concluded that the well designed and validated studies show a substantial improvement of around 15-25% accuracy in predicting the cancer susceptibility, recurrence, and mortality. In predicting the future cancer burden in the United States, Piva et al (2021) deployed artificial neural networks to capture the complex relationship between risk factors and cancer burden in the US. The authors utilised data from National Cancer Institute online datasets on the four most common tumors including breast, colorectal, lung and prostate for the period 1992 to 2006. The research deployed two artificial neural network (ANN) models: a multilayer feed-forward network (MLFFNN) and a nonlinear autoregressive network with eXogenous inputs (NARX). The ANN performed better in predicting the incidence of prostate cancer to decline from 2010 to 2019, and to reach a plateau by 2050. In addition, the study predicted colorectal and lung cancer incidence to reach a minimum value of 35 per 100,000 and per 100,000 (2017) from 50 to 31 per 100,000 in 2030 respectively.

As evidenced in the all the literature review above, the current healthcare is reactive, which in turns impede early diagnosis of diseases, and increase the risk of undetected illness. Hence, the application of machine learning in personalised preventive medicines using simple baseline information of a person is needed to accelerate early detection of cancer risk and enhance the ability of health professionals to deliver personalised and proactive treatment of a wide range diseases such as cancer and many more.

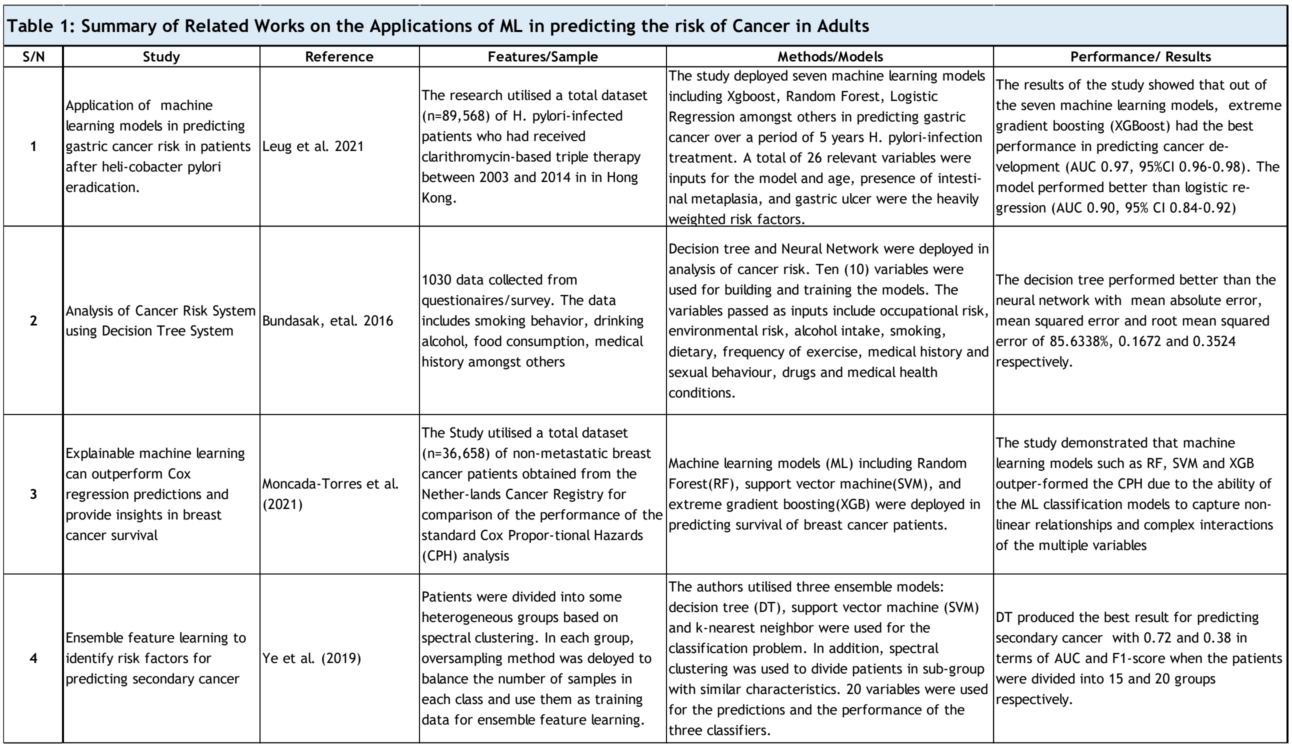


Table 1: Summary of Related Works on the Applications of ML in predicting the risk of Cancer in Adults

S/N	Study	Reference	Features/Sample	Methods/Models	Performance/ Results
1	Application of machine learning models in predicting gastric cancer risk in patients after heli-cobacter pylori eradication.	Leug et al. 2021	The research utilised a total dataset (n=89,568) of H. pylori-infected patients who had received clarithromycin-based triple therapy between 2003 and 2014 in in Hong Kong.	The study deployed seven machine learning models including Xgboost, Random Forest, Logistic Regression amongst others in predicting gastric cancer over a period of 5 years H. pylori-infection treatment. A total of 26 relevant variables were inputs for the model and age, presence of intesti-nal metaplasia, and gastric ulcer were the heavily weighted risk factors.	The results of the study showed that out of the seven machine learning models, extreme gradient boosting (XGBoost) had the best performance in predicting cancer de-velopment (AUC 0.97, 95%CI 0.96-0.98). The model performed better than logistic re-gression (AUC 0.90, 95% CI 0.84-0.92)
2	Analysis of Cancer Risk System using Decision Tree System	Bundasak, etal. 2016	1030 data collected from questionaires/survey. The data includes smoking behavior, drinking alcohol, food consumption, medical history amongst others	Decision tree and Neural Network were deployed in analysis of cancer risk. Ten (10) variables were used for building and training the models. The variables passed as inputs include occupational risk, environmental risk, alcohol intake, smoking, dietary, frequency of exercise, medical history and sexual behaviour, drugs and medical health conditions.	The decision tree performed better than the neural network with mean absolute error, mean squared error and root mean squared error of 85.6338%, 0.1672 and 0.3524 respectively.
3	Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival	Moncada-Torres et al. (2021)	The Study utilised a total dataset (n=36,658) of non-metastatic breast cancer patients obtained from the Nether-lands Cancer Registry for comparison of the performance of the standard Cox Propor-tional Hazards (CPH) analysis	Machine learning models (ML) including Random Forest(RF), support vector machine(SVM), and extreme gradient boosting(XGB) were deployed in predicting survival of breast cancer patients.	The study demonstrated that machine learning models such as RF, SVM and XGB outper-formed the CPH due to the ability of the ML classification models to capture non-linear relationships and complex interactions of the multiple variables
4	Ensemble feature learning to identify risk factors for predicting secondary cancer	Ye et al. (2019)	Patients were divided into some heterogeneous groups based on spectral clustering. In each group, oversampling method was deloyed to balance the number of samples in each class and use them as training data for ensemble feature learning.	The authors utilised three ensemble models: decision tree (DT), support vector machine (SVM) and k-nearest neighbor were used for the classification problem. In addition, spectral clustering was used to divide patients in sub-group with similar characteristics. 20 variables were used for the predictions and the performance of the three classifiers.	DT produced the best result for predicting secondary cancer with 0.72 and 0.38 in terms of AUC and F1-score when the patients were divided into 15 and 20 groups respectively.

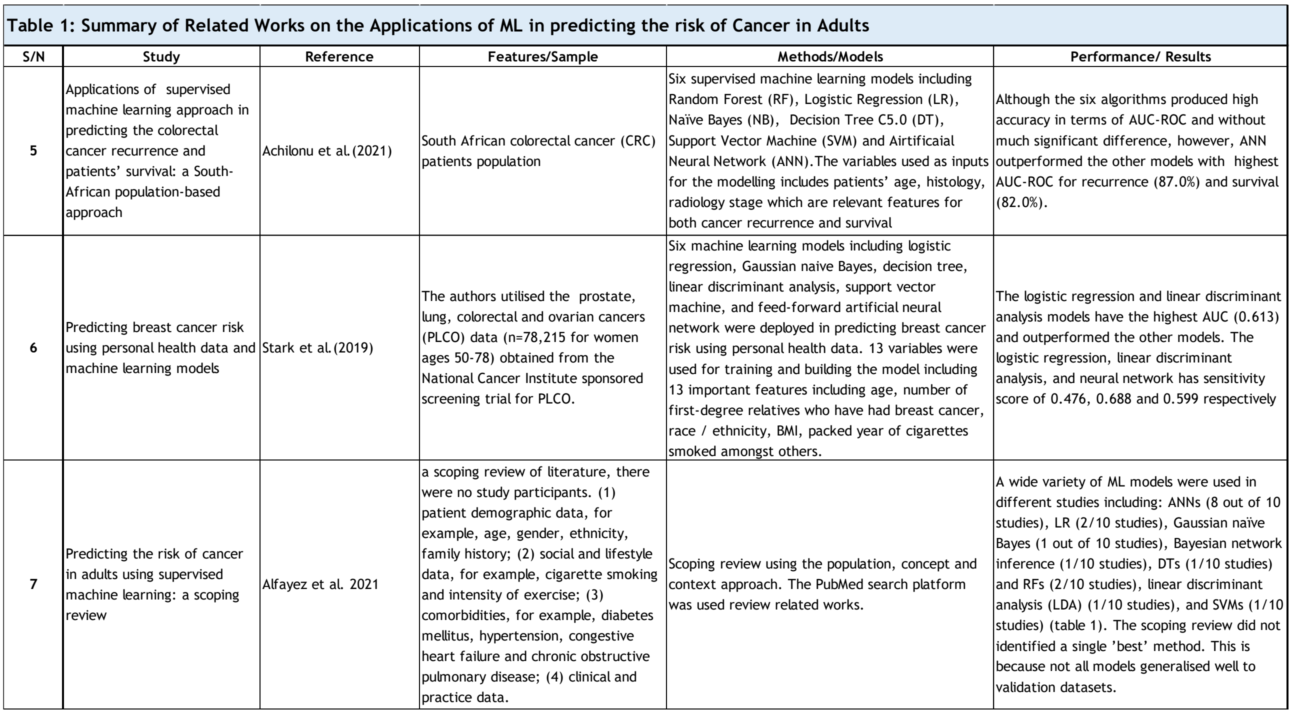


Table 1: Summary of Related Works on the Applications of ML in predicting the risk of Cancer in Adults

S/N	Study	Reference	Features/Sample	Methods/Models	Performance/ Results
5	Applications of supervised machine learning approach in predicting the colorectal cancer recurrence and patients' survival: a South-African population-based approach	Achilonu et al.(2021)	South African colorectal cancer (CRC) patients population	Six supervised machine learning models including Random Forest (RF), Logistic Regression (LR), Naïve Bayes (NB), Decision Tree C5.0 (DT), Support Vector Machine (SVM) and Airtificaial Neural Network (ANN).The variables used as inputs for the modelling includes patients' age, histology, radiology stage which are relevant features for both cancer recurrence and survival	Although the six algorithms produced high accuracy in terms of AUC-ROC and without much significant difference, however, ANN outperformed the other models with highest AUC-ROC for recurrence (87.0%) and survival (82.0%).
6	Predicting breast cancer risk using personal health data and machine learning models	Stark et al.(2019)	The authors utilised the prostate, lung, colorectal and ovarian cancers (PLCO) data (n=78,215 for women ages 50-78) obtained from the National Cancer Institute sponsored screening trial for PLCO.	Six machine learning models including logistic regression, Gaussian naive Bayes, decision tree, linear discriminant analysis, support vector machine, and feed-forward artificial neural network were deployed in predicting breast cancer risk using personal health data. 13 variables were used for training and building the model including 13 important features including age, number of first-degree relatives who have had breast cancer, race / ethnicity, BMI, packed year of cigarettes smoked amongst others.	The logistic regression and linear discriminant analysis models have the highest AUC (0.613) and outperformed the other models. The logistic regression, linear discriminant analysis, and neural network has sensitivity score of 0.476, 0.688 and 0.599 respectively
7	Predicting the risk of cancer in adults using supervised machine learning: a scoping review	Alfayez et al. 2021	a scoping review of literature, there were no study participants. (1) patient demographic data, for example, age, gender, ethnicity, family history; (2) social and lifestyle data, for example, cigarette smoking and intensity of exercise; (3) comorbidities, for example, diabetes mellitus, hypertension, congestive heart failure and chronic obstructive pulmonary disease; (4) clinical and practice data.	Scoping review using the population, concept and context approach. The PubMed search platform was used review related works.	A wide variety of ML models were used in different studies including: ANNs (8 out of 10 studies), LR (2/10 studies), Gaussian naïve Bayes (1 out of 10 studies), Bayesian network inference (1/10 studies), DTs (1/10 studies) and RFs (2/10 studies), linear discriminant analysis (LDA) (1/10 studies), and SVMs (1/10 studies) (table 1). The scoping review did not identified a single 'best' method. This is because not all models generalised well to validation datasets.

Table 1: Summary of Related Works on the Applications of ML in predicting the risk of Cancer in Adults

S/N	Study	Reference	Features/Sample	Methods/Models	Performance/ Results
8	A Support Vector Machine Model Predicting the Risk of Duodenal Cancer in Patients with Familial Adenomatous Polyposis at the Transcript Levels	Liu et al 2020	A total of 196 differentially expressed genes. Genes were identified by FAP vs. normal samples and FAP and duodenal cancer vs. normal samples. Microarray datasets related with FAP were retrieved from the Gene Expression Omnibus (GEO) database	The support vector machine (SVM) was utilised to train and validate cancer risk prediction model	After validation, the SVM model accurately distinguish FAP patients with high risk from those with low risk for duodenal cancer.
9	Predicting cancer using supervised machine learning: Mesothelioma	Choudhury 2021	C-reactive protein, platelet count, duration of symptoms, gender, and pleural protein were found to be the most relevant predictors that can prognosticate Mesothelioma. Patients' clinical data collected by Dicle University, Turkey	Multilayered perceptron (MLP), voted perceptron (VP), Clojure classifier (CC), kernel logistic regression (KLR), stochastic gradient decent (SGD), adaptive boosting (AdaBoost), Hoeffding tree (VFDT), and primal estimated sub-gradient solver for support vector machine (s-Pegasos)	In phase 1, SGD, AdaBoost.M1, KLR, MLP, VFDT generate optimal results with the highest possible performance measures. In phase 2, AdaBoost, with a classification accuracy of 71.29%, outperformed all other algorithms
10	Machine learning-based statistical analysis for early stage detection of cervical cancer	Ali etal. 2021	Clinical data for cervical cancer patients. Biopsy, cytology, Hinselmann, and Schiller.	Random Forest and Instance-Based K-nearest neighbor (IBk) were deployed for the classification problem.	A Random Tree (RT) algorithm provided the best classification accuracy for the biopsy (98.33%) and cytology (98.65%) data, whereas Random Forest (RF) and Instance-Based K-nearest neighbor (IBk) provided the best performance for Hinselmann (99.16%),
11	Machine Learning and Deep Learning Approaches in Breast Cancer Survival Prediction Using Clinical Data	Kalafi et al. 2019	4,902 patient records from the University of Malaya Medical Centre Breast Cancer Registry.	Multilayer perceptron (MLP), random forest (RF), decision tree (DT) classifiers, and Support vector machine (SVM) were deployed in the prediction and classification problem.	The results indicated that the MLP, RF DT classifiers could predict survivorship, respectively, with 88.2 %, 83.3 % and 82.5 % accuracy in the tested samples. SVM has the lowest score of 80.5 %.

Cancer Risk Factors: Ranking and Mapping to Cancer Types

Cancer risk factors are anything that could increase the chance of a person developing cancer (Macmillan Cancer Support, 2018). The interaction of a person's of gene, lifestyle and environment can influence the probability of developing cancer. American Cancer Society (2017) states that cancers are caused by a range of factors – some of these factors are modifiable and some are non-modifiable. The modifiable risks factors include but not limited to lifestyle risk factors. On the other hand, the non-modifiable risk factors include a person's age or genes. However, between 30% and 50% of cancers can be prevented through strategies to reduce behavioural and dietary risk factors. According to the Cancer Research UK (2022), smoking has been identified as the largest cause of cancer in the UK; this is followed by overweight and obesity. However, around 4 out of 10 cancer cases (around 135,000 every year) can be prevented with small change in daily routine and behaviours.

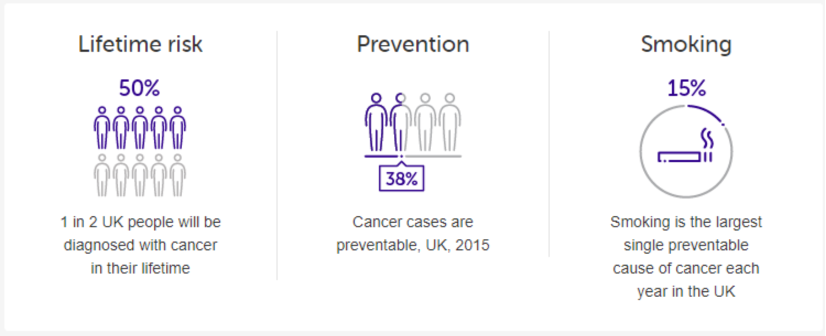


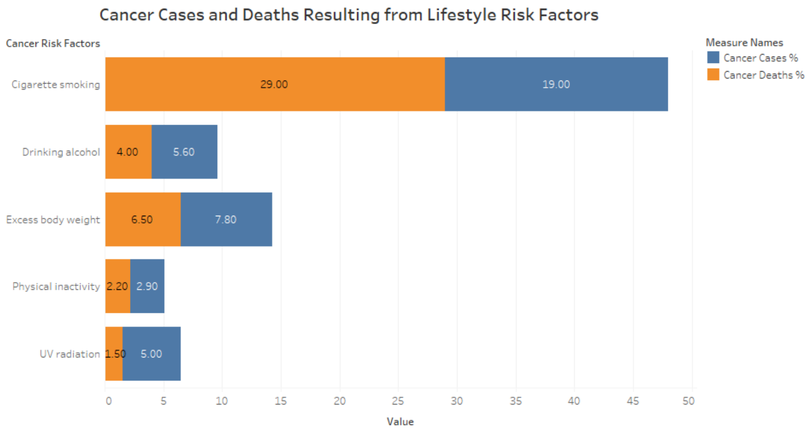
Fig 2: Cancer risk statistics in the UK. Source: Cancer Research UK 2022

Lifestyle risk factors are attributed to over 26 different types of cancers among adults aged 30 and older; and the risk factors include, but not limited to: use of tobacco, second-hand smoke, excessive body weights, drinking alcohol, eating red and processed meat, low in fruits and vegetables, dietary fiber, and dietary calcium, physical inactivity. While environmental risk factors include ultraviolet (UV) radiation from the sun, air pollution resulting from Co2 emission from automobiles, industrial, domestic activities and water pollution due to oil spillage, sewage, or chemical end products. The list of risk factors also include non-communicable chronic infection resulting from Helicobacter pylori, hepatitis B virus (HBV), hepatitis C virus (HPC), human herpes virus type 8 (HHV8), human immuno-deficiency virus (HIV), and human papillomavirus (HPV).

Cancer Risks Ranking

Cancer is commonly believed to be prevented and around 1 in 3 cases of the most common cancers (about 33%) could be prevented by eating a healthy diet, keeping to a healthy weight and being more active (Macmillan Cancer, 2022). In the work of Anand et al. 2018, 90-95% of cancer cases can be attributed to environmental and lifestyle risk factors and the remaining 5-10% are associated with age or genetic defects. The authors further explained that around 30-35% of all cancer-related deaths originated from dietary habits, 25-30% are associated with the use of tobacco and harmful substance, 15% can be attributed to infections and the remaining 10-20% resulting from other factors like radiation, stress, physical activity, environmental pollutants amongst others.

Cancers are caused by a range of factors, however, between 30% and 50% of cancers can be prevented through early detection of cancer risk and healthier lifestyle. The study from American Cancer Society (2017) estimated that 42% of cancer cases and 45% of cancer deaths in the United States are linked to preventable (modifiable) risk factors as detailed in fig 3 below.



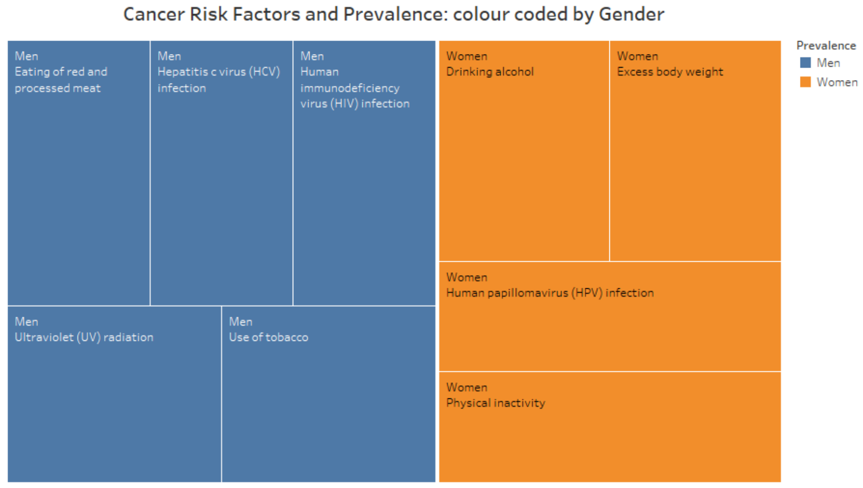
Cancer Cases % and Cancer Deaths % for each Cancer Risk Factors. Color shows details about Cancer Cases % and Cancer Deaths %. The marks are labeled by Cancer Cases % and Cancer Deaths %. Details are shown for Cancer Cases % and Cancer Deaths %.

Fig 3: Cancer cases and deaths resulting from lifestyle risk factors

Adapted from American Cancer Society (2017)

The above figure shows the total cancer cases and deaths in the United States based on available data for 2014. Out of the 42% cancer cases (659,640 cancer cases (out of 1,570,975)) attributable to lifestyle risk factors; cigarette smoking, excess body weight, drinking alcohol, UV radiation and physical inactivity accounted for 19%, 7.8%, 5.6%, 5%, 2.95 respectively. Whereas, cigarette smoking, excess body weight, drinking alcohol, UV radiation and physical inactivity accounted for 29%, 6.5%, 4%, 1.5%, 2.2% respectively from the total cancers deaths (265,150 cancer deaths (out of 587,521)) for the period.

Cancer Risks Factors: Prevalence in Women and Men



Prevalence and Cancer Risk factors. Color shows details about Prevalence. Size shows count of Prevalence (2). The marks are labeled by Prevalence and Cancer Risk factors.

Fig 4: Cancer risk factors: prevalence by gender. Adapted from American Cancer Society (2017)

- Historically, the cancer risk factors depicted in the figure above are more prevalent among men than women due to psychological, cultural and behavioural factors .

- Similarly, risk factors depicted for women were higher than men largely due to the high burden of breast, endometrial, and cervical cancers traceable to the risk factors.

Cancer risk associated with comorbidities

Beyond the cancer risk factors, co-existing chronic health conditions (comorbidities) in a person jointly interact with the risk factors to influence the development of cancer. Many studies have overlooked the influence of comorbidities as risk factors for cancer. Chronic diseases such as diabetes, lung problems, kidney, heart, respiratory diseases accounted for 71% of the global deaths in 2015, and attributable to cancer amongst other diseases (Yu et al. 2018). Furthermore, many recent studies and current cancer prevention strategies focus on lifestyle risk factors and overlooked chronic diseases as important risk factors partly because of the modest relationship between these diseases and cancer risk factors. Chronic diseases share common risk factors with different types of cancers and may jointly influence the development of cancer regardless of the interaction with the risk factors. The figure below summarises the proportion of different risk factors contribution to cancer cases. The figure shows that the combination of lifestyle and environmental risk factors contributed to around 90% of the cancer cases. Whereas age and genetic risk contributed 5%, and chronic diseases (comorbidities) accounted for the remaining 5% of cancer cases in America.

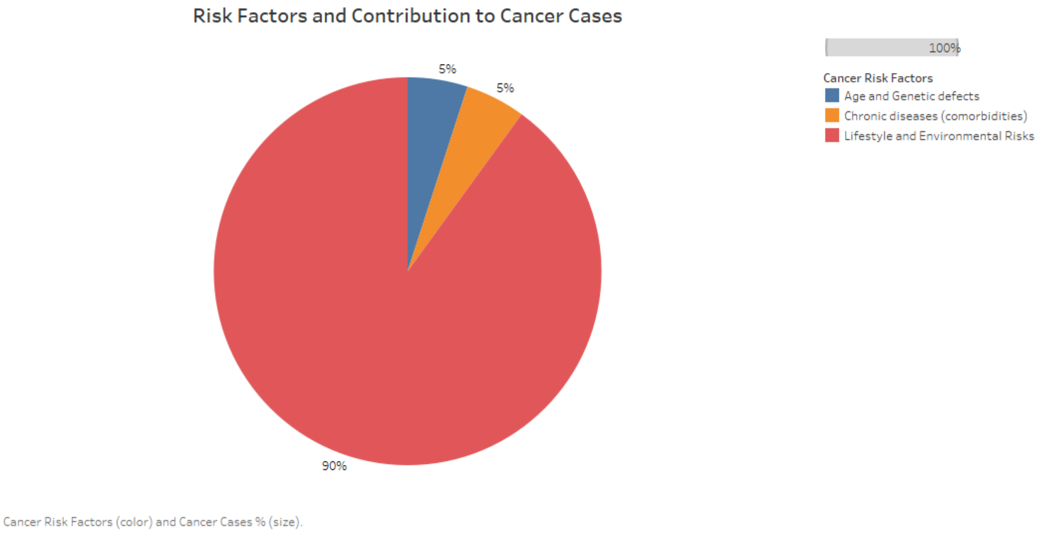
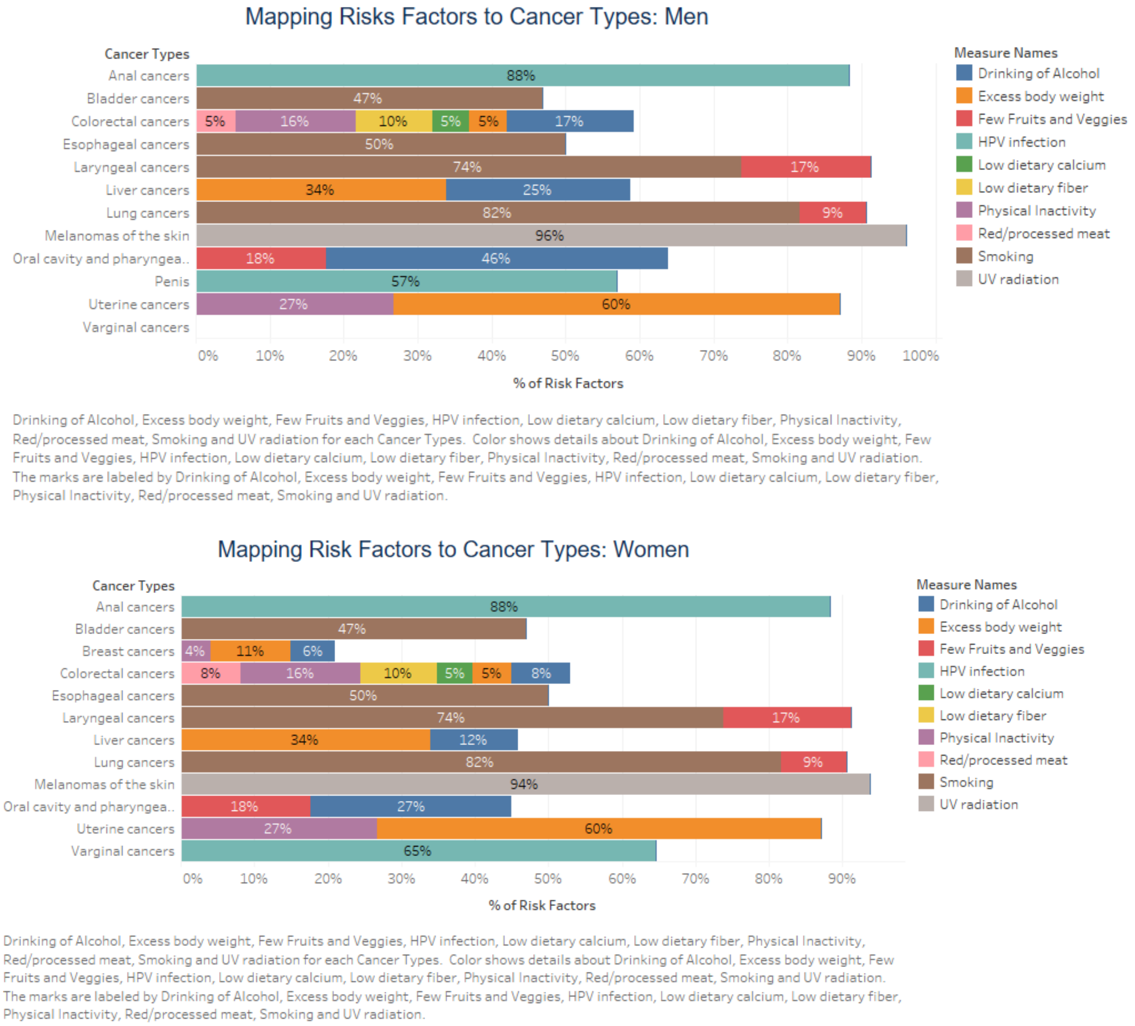


Fig 5: Proportion of different risk factors to cancer cases.

Adapted from American Cancer Society (2017) and Hoang, Lee, and Kim (2020)

Mapping Cancer Risks to Cancer Types and Prevalence in Gender

The figure below summarises 13 different cancers individually and their association to lifestyle risk factors. The modifiable lifestyle risk factors were greater than 50% for 6 of the 13 cancers with proportion ranging from as high as 96% to as low as 5%.



Mapping Risks Factors to Cancer Types: Men

Drinking of Alcohol, Excess body weight, Few Fruits and Veggies, HPV infection, Low dietary calcium, Low dietary fiber, Physical Inactivity, Red/processed meat, Smoking and UV radiation for each Cancer Types. Color shows details about Drinking of Alcohol, Excess body weight, Few Fruits and Veggies, HPV infection, Low dietary calcium, Low dietary fiber, Physical Inactivity, Red/processed meat, Smoking and UV radiation. The marks are labeled by Drinking of Alcohol, Excess body weight, Few Fruits and Veggies, HPV infection, Low dietary calcium, Low dietary fiber, Physical Inactivity, Red/processed meat, Smoking and UV radiation.

Mapping Risk Factors to Cancer Types: Women

Drinking of Alcohol, Excess body weight, Few Fruits and Veggies, HPV infection, Low dietary calcium, Low dietary fiber, Physical Inactivity, Red/processed meat, Smoking and UV radiation for each Cancer Types. Color shows details about Drinking of Alcohol, Excess body weight, Few Fruits and Veggies, HPV infection, Low dietary calcium, Low dietary fiber, Physical Inactivity, Red/processed meat, Smoking and UV radiation. The marks are labeled by Drinking of Alcohol, Excess body weight, Few Fruits and Veggies, HPV infection, Low dietary calcium, Low dietary fiber, Physical Inactivity, Red/processed meat, Smoking and UV radiation.

Fig 6: Mapping of Cancer Risks to Cancer Types and Gender: Adapted from American Cancer Society (2017)

The Application of Machine Learning in Cancer Risk Detection and Diagnosis: Oncology

Machine learning applications in healthcare spanned over 20 years considering the use of artificial neural networks (ANN) and decision trees (DT) in cancer detection and diagnosis (Simes 1985; Maclin et al.1991; Ciccheti 1992 cited in Cruz and Wishart 2006). As a core branch of artificial intelligence, machine learning (ML) utilises statistical, probabilistic, computational and optimisation techniques to enhance the capability of computers to learn patterns from noisy, complex, and big datasets. In recent times, ML has been applied to wide range applications from detection to classification of cancer risks, tumors amongst others. The capability of computers in discerning complex patterns from data is compatible with the applications in healthcare considering the heavy reliance proteomic and genomic measurements. Hence, machine learning usability in detection, screening and diagnosis of different cancers is increasing at an accelerating rate. In addition, the increasing use of machine learning in oncology has sparked the demand for personalised and predictive medicine.

In the latest report of PubMed statistics, over 1500 papers have been published on machine learning and cancers (PubMed.gov 2022). However, majority of the studies focused on using machine learning techniques to identify, classify, detect, or distinguish tumors and other malignancies. In addition, most of the studies on predicting the risk of cancer utilized the lifestyle risk factors and completely ignored chronic diseases (comorbidities). Hence, the body of literature in the field of machine learning and cancer risk predictions using different risk factors including comorbidities is relatively small (less than 50 papers).

Machine Learning Methods

Machine learning is divided into three major algorithms including (1) supervised machine learning, (2) unsupervised machine learning and (3) reinforcement learning. The classification of the machine learning methods is based on the desired outcomes (Mitchell, 1997; Duda et al. 2001 cited in Cruz and Wishart 2006). In the supervised machine learning, a labelled training data set is provided to the algorithm. The labelled training set are what the computers can learn about to map the input to the desired output.

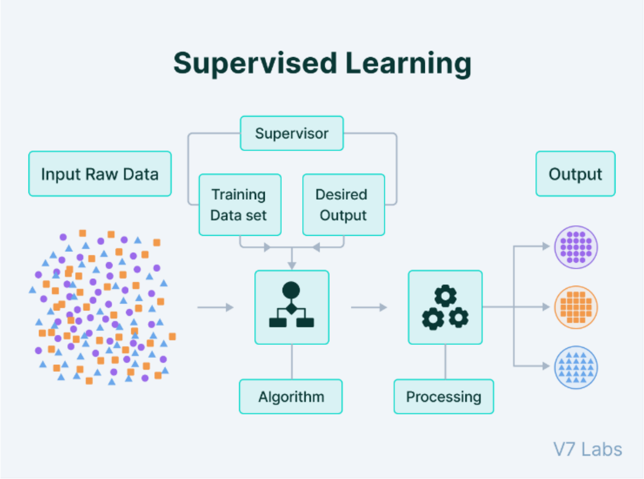


Fig 7: Supervised learning process. Source: V7 Labs 2022

In the above figure, the dataset has labels corresponding to the input data. The machine is able to learn, detect patterns, classify, and predict the output label. The supervised machine learning can be divided into classification and regression. Classification problem involves mapping inputs to output to predict and classify a discrete output or label. The output of a classification problem usually consists of classes or categories. For instance, predicting a person's risk of cancer as either high, medium, or low. On the other hand, regression problem involves mapping inputs to output to predict a continuous output. A good example is predicting prices of house, stock, cryptocurrencies amongst others.

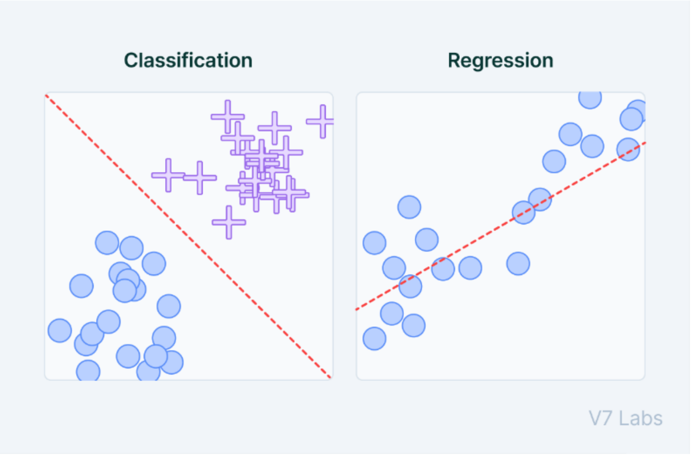


Fig 8: Classification and Regression Problem. Source: V7 Labs 2022

In unsupervised machine learning, no class labels are provided, it involves a self-learning process without any supervision from the machine (algorithm) to find hidden patterns, group datapoints with similar attributes into sub-groups and cluster them into different class. Example of unsupervised machine learning methods include clustering, dimensionality reduction, association amongst others. The goal with clustering is to find hidden patterns in unlabelled dataset and group data into sub-groups called clusters based on similarities or differences. Examples of clustering methods includes, K-Mean clustering, DBSCAN, hierarchical clustering amongst others. Association in unsupervised machine learning is finding the relationship amongst variables or datapoints. A good example is the correlation amongst different cancer risk factors, consumer buying choice and the demand for a particular product. Whereas, dimensionality reduction in unsupervised machine learning helps to extract the most important features and reduce noise and unnecessary features from dataset for machine learning activities.

In reinforcement machine learning, the machine learns through trial and error using feedback from its actions and take suitable actions to maximize rewards (positives) or a minimize risk (negatives) in a given situation. Reinforcement machine learning is different from supervised machine learning. In supervised machine learning a label is given and a model is trained based on the label, whereas in reinforcement learning no label or answer is given, but the reinforcement agent decides what to do to perform the given task.

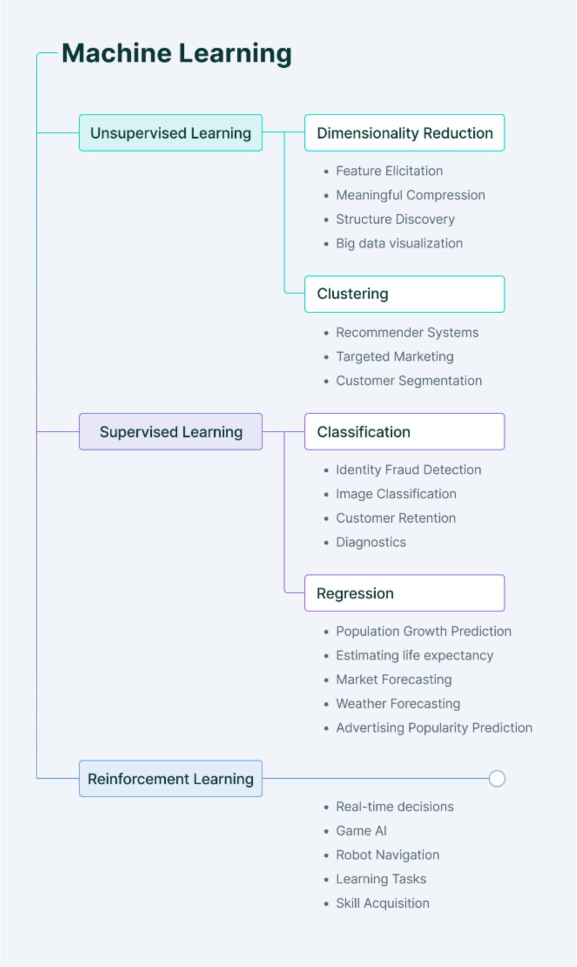


Fig 9: Types of learning in machine learning. Source: V7 Labs 2022

In the context of healthcare, personalised and predictive medicine, supervised machine learning algorithms are used for cancer prediction and diagnosis because the models employ applicable category of classifiers and perform classification of cancer risk based on conditional probabilities or decisions.

Machine learning models in cancer risk prediction

Machine learning models are programs with capability to find patterns or make decisions from previously unseen datasets (Muller and Guido 2016). The goal with the application of supervised machine learning models in cancer predictions is to achieve best approximation (or function) that maps the existing between features (cancer risk factors) and labels (classes of risk). The function to map this relationship is unknown. Hence, a model helps to achieve the representation or best approximation of mapping of features and labels using specific rules and data structure. Machine learning algorithms helps to express the mapping of relationship between inputs and labels in a mathematical form to find patterns in a given dataset. There are many machine learning models, and these models are based on specific machine learning algorithms. The commonly used supervised machine learning models in cancer risk prediction includes logistic regression (LR), support vector machine (SVM), Naïve Bayes (NB), decision tree (DT), k Nearest Neighbors (KNN), random forest, extreme gradient boosting (XGB) amongst others. The logistic regression (LR) helps the process of modeling the probability of a discrete outcome given an input variable. It also helps to establish relationship between a dependent variable and one or several features (inputs) to make a prediction about a categorical variable (Ali et al. 2021; Cruz and Wishart 2006). Examples of categorical output includes have cancer or no cancer; yes or no; high, medium, or low risk of cancer amongst others. Some of the advantage of the LR model are that is very easy to implement and interpret. However, the major drawback is that the model tends to overfit and always assume a linear boundary or relationship. The support vector machine (SVM) creates coordinates for each object in an n-dimensional space and uses a hyperplane to group objects by common features (Muller and Guido 2016). The advantage attributed to the SVM model over logistic regression is that it can model non-linear boundaries, overfitting is relatively low and essentially easy to interpret. The limitation with the model is that it requires high training time and computational resources compared to the decision tree and naïve bayes. The naïve bayes (NB) algorithms is based on the assumption that there is independence amongst the variables and deploys probability techniques to classify a label based on features (Nafizatus and Rustam 2019). The biggest advantage of the model is that it has a wide range application in healthcare and other domains, easy to understand, and train. The limitation of the model includes heavy assumptions that the dataset is normally distributed, attributes are statistically independent, and classes must be mutually exclusive. The decision trees (DT) are classifiers deploy to establish what category an input falls into by traversing the leaf's and nodes of a tree(Bundasak 2016). The main strengths of the DT are that they are widely used in the medical settings because of easy of interpretability and transparency in the decision-making process of the classifier. In addition, the algorithm is not easily affected by outliers. The drawback with the model is that it tends to overfit, and assumes attributes are independent, and classes are mutually exclusive. The k Nearest Neighbors (KNN) is the process of grouping the closest objects in a dataset together and finding the average representation (mean), most frequent (mode) attributes among the objects. KNN has the advantages of very efficient for non-linear classification problems, easy and fast to train, and can be used for both classification and regression problems (Ali et al. 2021; Cruz and Wishart 2006). The drawback with the model is assumption that the features are equally relevant, and the model tends to become computationally complex as the number of features increase. The random forest (RF) has the highest usage in cancer risk prediction. RF model is based on the bagging algorithm and uses Ensemble Learning technique, leverages the power of collection of many decision trees from random subsets of the data (Alfayez et al. 2021; Cruz and Wishart 2006). The reliance and combination of trees makes the decision of the RF model more accurate in prediction than a single decision tree. The model can be used to solve both classification and regression problems, it is efficient for both categorical and continuous variables, can handle outliers and missing values efficiently, can handle and capture non-linear relationship in dataset and it does not require any feature scaling on the input variables. The drawback with the model is that it requires much computational power and resources because it relies on collection of random trees to make its decision. In addition, the RF model requires long time to train compared to a single decision tree. The extreme gradient boosting (XGB) deploys ensemble learning techniques. XGB captures and combines the predictions from multiple algorithms (such as decision trees) and take into account the error from the previous algorithm(Kabiraj et al 2020). The advantages of the XGB includes efficient for small and medium datasets, it is flexible and very fast to train, it can handle missing values efficiently, it supports regularisation and the leverages the power of parallel processing. The main drawback with the model is that is hardly scalable, not efficient for sparse and unstructured dataset and very sensitive to outliers.

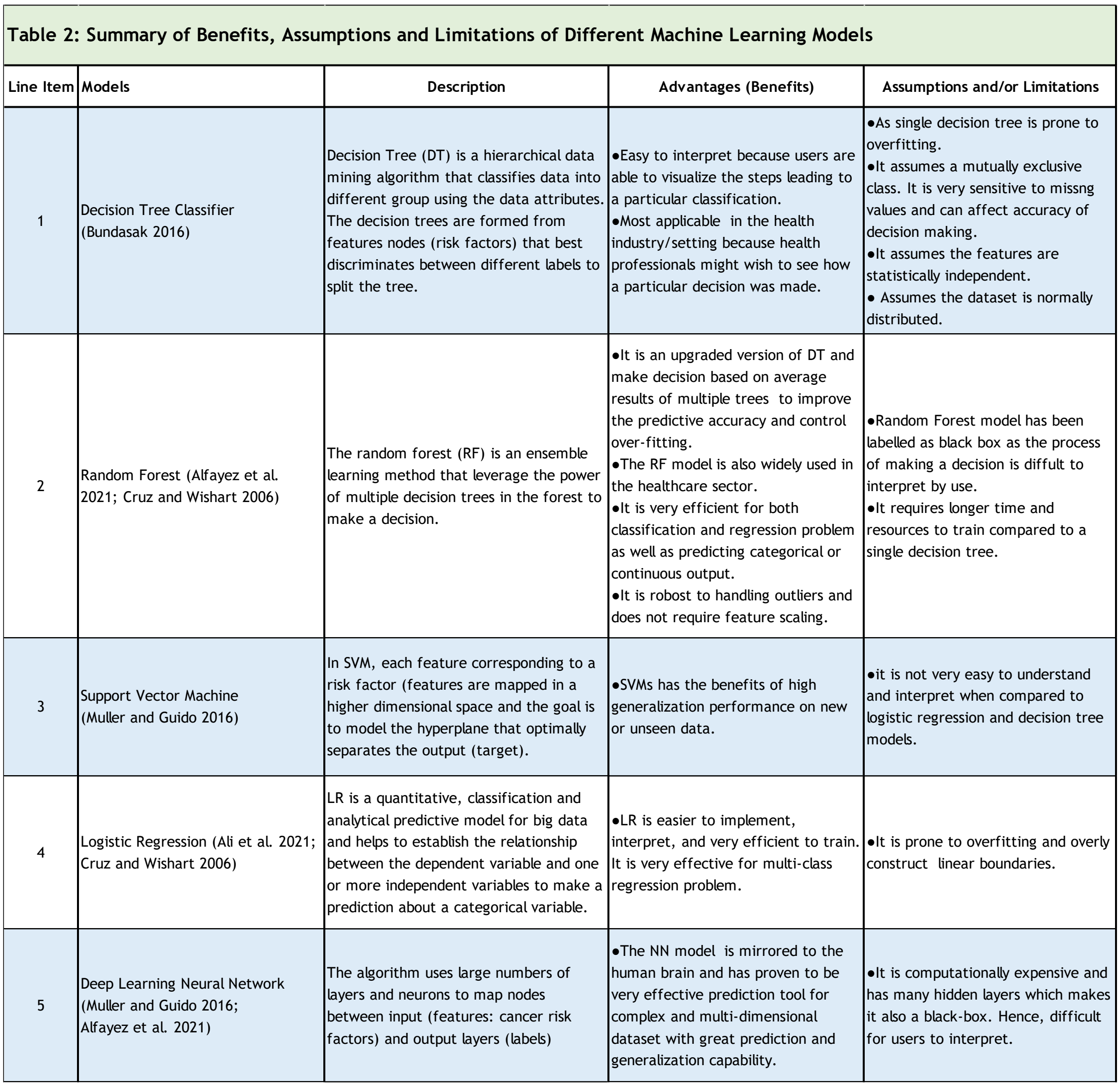


Table 2: Summary of Benefits, Assumptions and Limitations of Different Machine Learning Models

Line Item	Models	Description	Advantages (Benefits)	Assumptions and/or Limitations
1	Decision Tree Classifier (Bundasak 2016)	Decision Tree (DT) is a hierarchical data mining algorithm that classifies data into different group using the data attributes. The decision trees are formed from features nodes (risk factors) that best discriminates between different labels to split the tree.	●Easy to interpret because users are able to visualize the steps leading to a particular classification. ●Most applicable in the health industry/setting because health professionals might wish to see how a particular decision was made.	●As single decision tree is prone to overfitting. ●It assumes a mutually exclusive class. It is very sensitive to missng values and can affect accuracy of decision making. ●It assumes the features are statistically independent. ● Assumes the dataset is normally distributed.
2	Random Forest (Alfayez et al. 2021; Cruz and Wishart 2006)	The random forest (RF) is an ensemble learning method that leverage the power of multiple decision trees in the forest to make a decision.	●It is an upgraded version of DT and make decision based on average results of multiple trees to improve the predictive accuracy and control over-fitting. ●The RF model is also widely used in the healthcare sector. ●It is very efficient for both classification and regression problem as well as predicting categorical or continuous output. ●It is robost to handling outliers and does not require feature scaling.	●Random Forest model has been labelled as black box as the process of making a decision is diffult to interpret by use. ●It requires longer time and resources to train compared to a single decision tree.
3	Support Vector Machine (Muller and Guido 2016)	In SVM, each feature corresponding to a risk factor (features are mapped in a higher dimensional space and the goal is to model the hyperplane that optimally separates the output (target).	●SVMs has the benefits of high generalization performance on new or unseen data.	●it is not very easy to understand and interpret when compared to logistic regression and decision tree models.
4	Logistic Regression (Ali et al. 2021; Cruz and Wishart 2006)	LR is a quantitative, classification and analytical predictive model for big data and helps to establish the relationship between the dependent variable and one or more independent variables to make a prediction about a categorical variable.	●LR is easier to implement, interpret, and very efficient to train. It is very effective for multi-class regression problem.	●It is prone to overfitting and overly construct linear boundaries.
5	Deep Learning Neural Network (Muller and Guido 2016; Alfayez et al. 2021)	The algorithm uses large numbers of layers and neurons to map nodes between input (features: cancer risk factors) and output layers (labels)	●The NN model is mirrored to the human brain and has proven to be very effective prediction tool for complex and multi-dimensional dataset with great prediction and generalization capability.	●It is computationally expensive and has many hidden layers which makes it also a black-box. Hence, difficult for users to interpret.

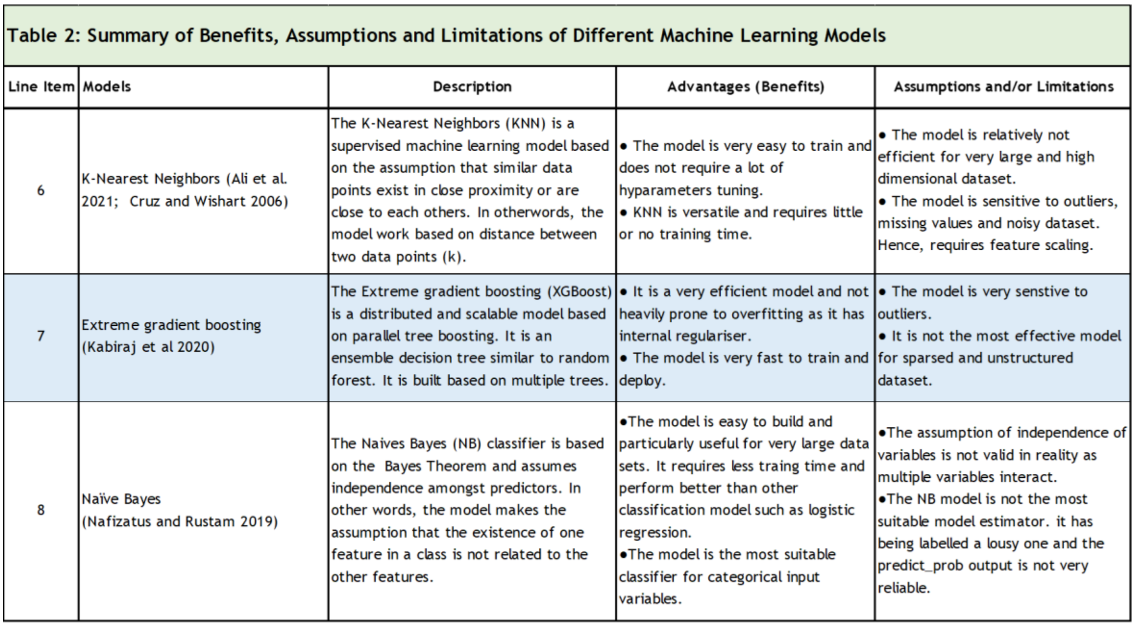


Table 2: Summary of Benefits, Assumptions and Limitations of Different Machine Learning Models

Line Item	Models	Description	Advantages (Benefits)	Assumptions and/or Limitations
6	K-Nearest Neighbors (Ali et al. 2021; Cruz and Wishart 2006)	The K-Nearest Neighbors (KNN) is a supervised machine learning model based on the assumption that similar data points exist in close proximity or are close to each others. In otherwords, the model work based on distance between two data points (k).	● The model is very easy to train and does not require a lot of hyparameters tuning. ● KNN is versatile and requires little or no training time.	● The model is relatively not efficient for very large and high dimensional dataset. ● The model is sensitive to outliers, missing values and noisy dataset. Hence, requires feature scaling.
7	Extreme gradient boosting (Kabiraj et al 2020)	The Extreme gradient boosting (XGBoost) is a distributed and scalable model based on parallel tree boosting. It is an ensemble decision tree similar to random forest. It is built based on multiple trees.	● It is a very efficient model and not heavily prone to overfitting as it has internal regulariser. ● The model is very fast to train and deploy.	● The model is very senstive to outliers. ● It is not the most effective model for sparsed and unstructured dataset.
8	Naïve Bayes (Nafizatus and Rustam 2019)	The Naives Bayes (NB) classifier is based on the Bayes Theorem and assumes independence amongst predictors. In other words, the model makes the assumption that the existence of one feature in a class is not related to the other features.	●The model is easy to build and particularly useful for very large data sets. It requires less traing time and perform better than other classification model such as logistic regression. ●The model is the most suitable classifier for categorical input variables.	●The assumption of independence of variables is not valid in reality as multiple variables interact. ●The NB model is not the most suitable model estimator. it has being labelled a lousy one and the predict_prob output is not very reliable.

CHAPTER THREE

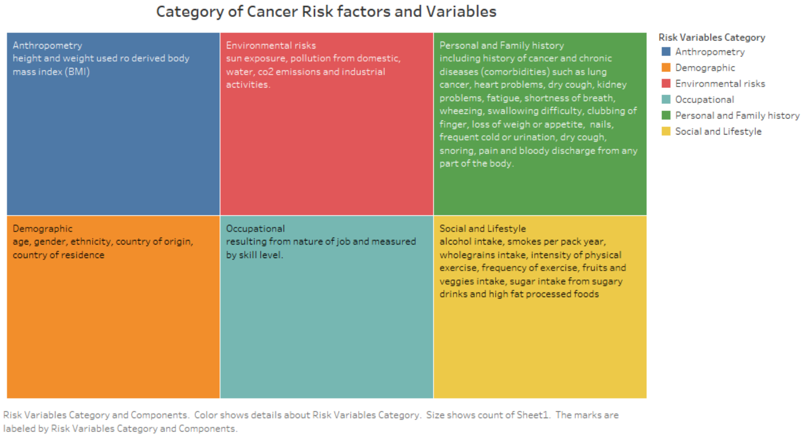
Research Design/Methodology

This section lay down the foundation of the research project and covers both the techniques (methods) used in data collection, the overall systematic approach, and set of principles that guide how the research project was carried out.

3.1. Study Design

The study is quantitative research aimed at creating a cancer risk prediction system using evidence-based data obtained through a primary data collection method. The quantitative methodology was selected in line with the aim of the study and to produce generalisable knowledge about the different risk factors for cancer, enhance personalised, predictive healthcare, and achieve higher interpretability in findings to accelerate adoption in the medical settings.

Based on the study objectives, a range of different risk factors were included in the study to build and train the ML model. These risk factors(variables) were grouped into the following 9 categories: **(1) demographic** including: age, gender, ethnicity, country of origin, country of residence; **(2) social and lifestyle** including: alcohol intake, smokes per pack year, intensity of physical exercise, frequency of exercise, fruits and veggies intake, sugar intake from sugary drinks and high fat processed foods; **(3) anthropometry** including: height, weight; **(4) personal history including: history of cancer and existing medical conditions, 5) family history** including history of cancer **and existing medical conditions** (6) Comorbidities including chronic diseases such as lung cancer, heart problems, dry cough, fatigue, shortness of breath, wheezing, swallowing difficulty, clubbing of finger nails, frequent cold or urination, dry cough, snoring, pain and bloody discharge from any part of the body; **(7) Environmental risks including**: sun exposure, pollution from domestic, water, co2 emissions and industrial activities; **(8) Occupational** resulting from nature of job measured by skill level and (9) economic class.



Risk Variables Category and Components. Color shows details about Risk Variables Category. Size shows count of Sheet1. The marks are labeled by Risk Variables Category and Components.

Fig 10: Category of cancer risk factors and variables

Ethical Considerations

In designing the study, several principles were considered to protect the rights of the participants, enhance the research validity, and maintain scientific integrity. The following are the main ethical considerations reflected in the study design and their corresponding meaning:

Table 3: Ethical considerations for the research study: Adapted from Scribbr 2022

Ethical Considerations	Meaning and Significance
Anonymity	Data was collected from participants without any knowledge of who they are and cannot be linked to anyone. Hence, no personally identifying information such as name, phone number, email addresses were collected from participants.
Informed consent	The introductory page of the research questionnaire contained a statement informing potential participants of the study's benefits, and institutional approval. Hence, this provides the participants with prior information and clear understanding of the context of the study before making the decision to participate in the survey.
Voluntary participation	A section of the cover page of the questionnaire distributed informed all potential participants that they are free to choose whether they want to participant in the study and can opt out at any time without any negative consequences.
Confidentiality	No identifying information for known participants including health professionals and domain experts who helped in circulating the survey in their respective networks have been included in the study or the required report.
Potential for harm	Various possibility of harm were considered when designing and distributing the questionnaire to ensure that sensitive information or questions were not included that might trigger anxiety, shame, stigmatization, pain amongst others. In addition, due care was taken not to reveal sensitive data that will lead to legal action or breach of privacy.
Results communication	Due care has been taken to avoid misrepresentation of results, plagiarism and any research misconduct including falsing data, manipulating data analysis amongs others.

Considering the quantitative methodology and domain of the project, the cross Industry standard process for data mining (CRISP-DM) model was deployed for a scientific and systematic approach to data exploration/understanding, data preparation, modelling, evaluation, and deployment. With the CRISP model, learning is continuous and does not end when the model is created and deployed as depicted in the outer layer of the

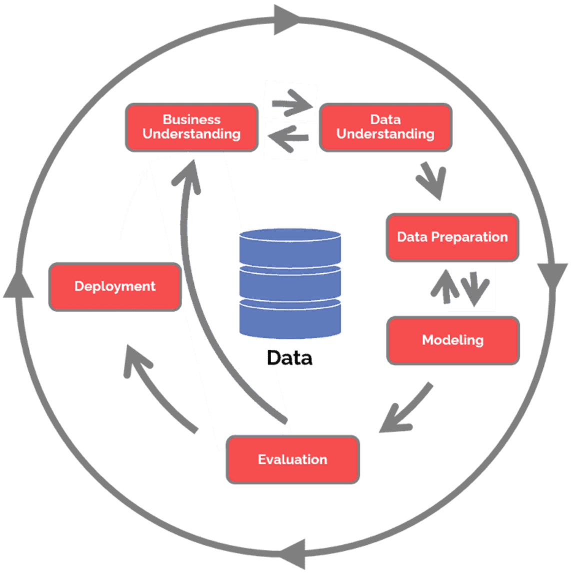


Figure 11: The CRISP-DM Model (Data science process alliance, 2021)

3.2. Data Source

The research project utilised primary data collected through survey questionnaire. The questionnaire was designed using survey the Zoho survey platform and was distributed online through social media platforms including Facebook, Twitter, Instagram, LinkedIn, WhatsApp group amongst others. In addition, participants who completed the questionnaire also referred others to complete the survey. Hence, word of mouth/referrals had the highest reach of 46.01% followed by WhatsApp group with 36.64% as shown in figure 14 below.

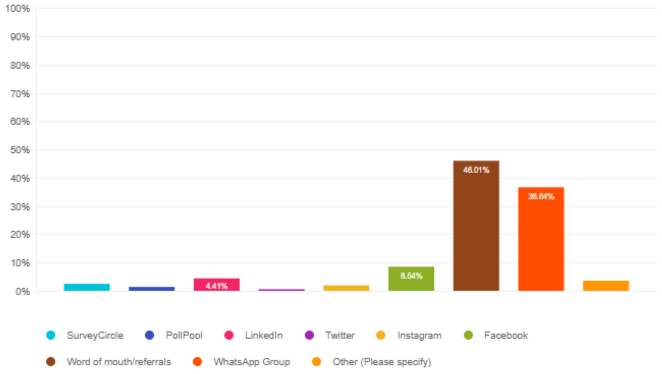
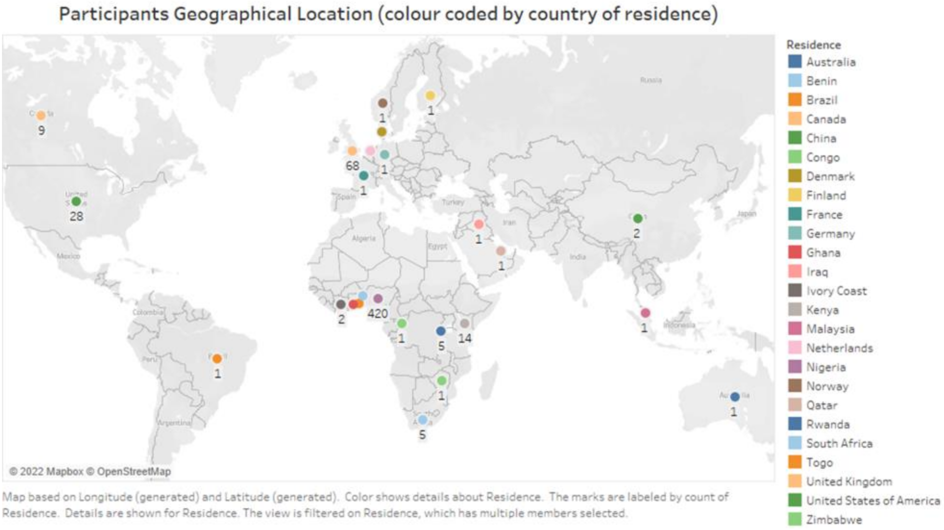


Figure 12: Survey reach and how questionnaire was distributed: Adapted from Zoho Survey 2022

Survey was selected as the preferred method of data collection to reach a wider target population (Adults aged 18 and above). Furthermore, the data collection method provided the flexibility in collecting data that reflects the characteristics and demographic of the target population, which is important for quantitative analysis and generalisation of the research results. The method has proven to be very effective for collecting data from people about possible exposure to risk factors, symptoms, and co-existing medical conditions. Hence, it is widely used for epidemiology research in the healthcare and amongst others (Scribbr 2022). The dataset collected was anonymised evidence-based data from 580 participants aged 18 and above across 25 countries. Nigeria has the highest participants of 420, United Kingdom: 68, United States of America: 28 and the remaining participants reside in other part of the world. The original dataset has 580 instances and 58 features, and the features were reduced from 58 to 38 through data cleaning and preparation.



Participants Geographical Location (colour coded by country of residence)

© 2022 Mapbox © OpenStreetMap

Map based on Longitude (generated) and Latitude (generated). Color shows details about Residence. The marks are labeled by count of Residence. Details are shown for Residence. The view is filtered on Residence, which has multiple members selected.

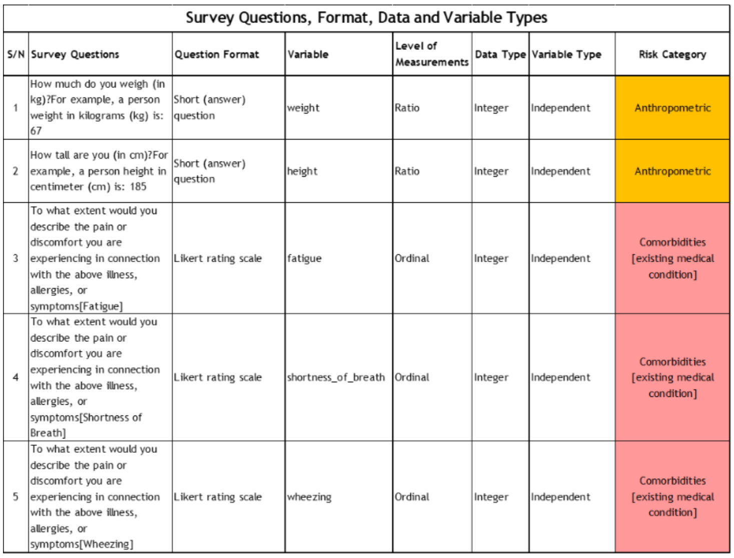
Figure 13: Geographical distribution of participants: Developed by Author using Tableau

The survey questionnaire was designed using a combination of multiple choice, Likert-scale, matrix choice, slider scale and short answers questions. There were 27 main questions (including 31 sub-questions) as detailed in table 2 and 3 below:

Table 4: Survey questionnaire design

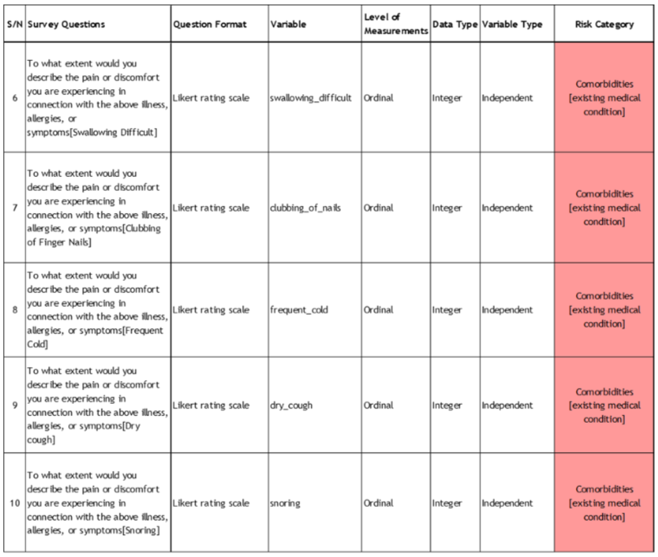
Type of Questions	Number/Frequency of Questions
Short (answer) question	5
Closed-ended multiple choice questions	9
Matrix Choice (one answer)	1
Slider Scale	1
Likert scale (slider, matrix, and rating)	11

Table 5: Survey questions, data, and variable types

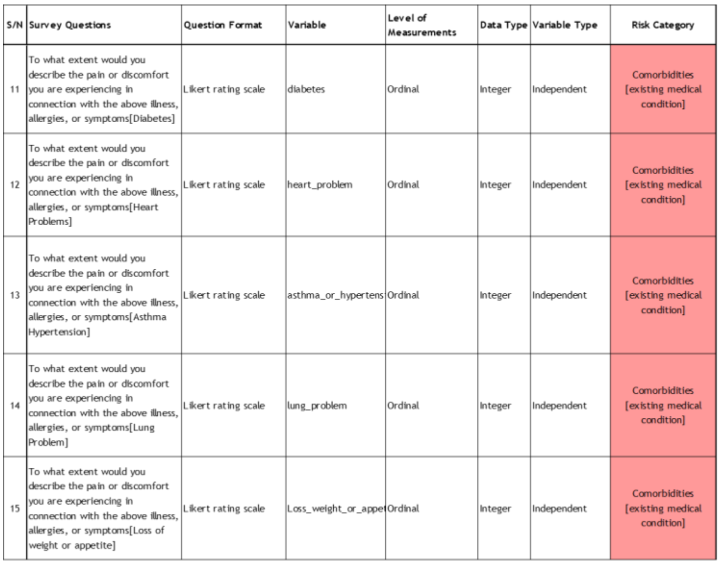


Survey Questions, Format, Data and Variable Types

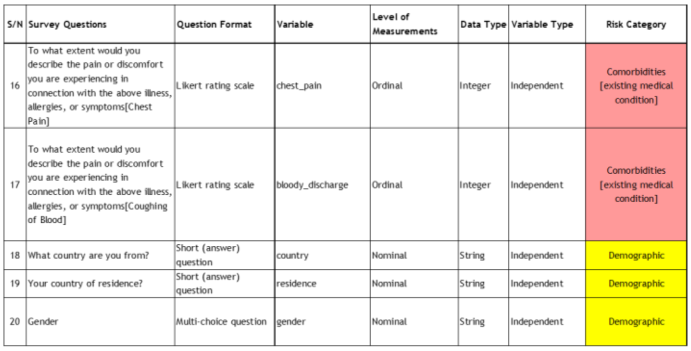
S/N	Survey Questions	Question Format	Variable	Level of Measurements	Data Type	Variable Type	Risk Category
1	How much do you weigh (in kg)?For example, a person weight in kilograms (kg) is: 67	Short (answer) question	weight	Ratio	Integer	Independent	Anthropometric
2	How tall are you (in cm)?For example, a person height in centimeter (cm) is: 185	Short (answer) question	height	Ratio	Integer	Independent	Anthropometric
3	To what extent would you describe the pain or discomfort you are experiencing in connection with the above illness, allergies, or symptoms[Fatigue]	Likert rating scale	fatigue	Ordinal	Integer	Independent	Comorbidities [existing medical condition]
4	To what extent would you describe the pain or discomfort you are experiencing in connection with the above illness, allergies, or symptoms[Shortness of Breath]	Likert rating scale	shortness_of_breath	Ordinal	Integer	Independent	Comorbidities [existing medical condition]
5	To what extent would you describe the pain or discomfort you are experiencing in connection with the above illness, allergies, or symptoms[Wheezing]	Likert rating scale	wheezing	Ordinal	Integer	Independent	Comorbidities [existing medical condition]



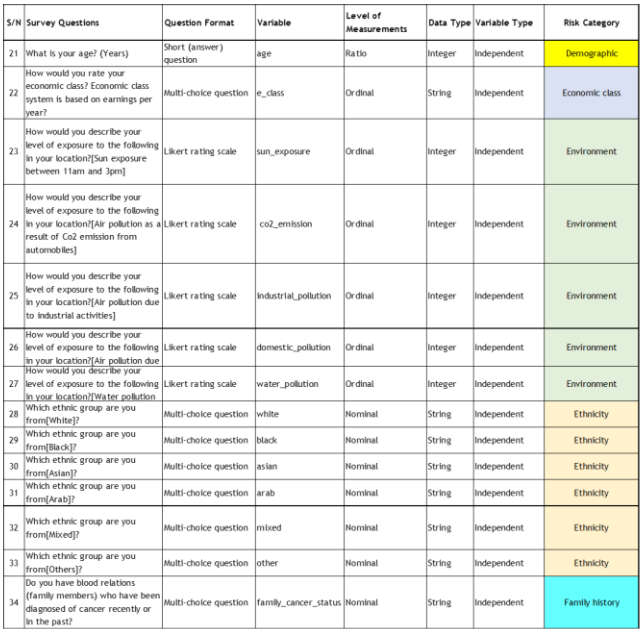
S/N	Survey Questions	Question Format	Variable	Level of Measurements	Data Type	Variable Type	Risk Category
6	To what extent would you describe the pain or discomfort you are experiencing in connection with the above illness, allergies, or symptoms[Swallowing Difficult]	Likert rating scale	swallowing_difficult	Ordinal	Integer	Independent	Comorbidities [existing medical condition]
7	To what extent would you describe the pain or discomfort you are experiencing in connection with the above illness, allergies, or symptoms[Clubbing of Finger Nails]	Likert rating scale	clubbing_of_nails	Ordinal	Integer	Independent	Comorbidities [existing medical condition]
8	To what extent would you describe the pain or discomfort you are experiencing in connection with the above illness, allergies, or symptoms[Frequent Cold]	Likert rating scale	frequent_cold	Ordinal	Integer	Independent	Comorbidities [existing medical condition]
9	To what extent would you describe the pain or discomfort you are experiencing in connection with the above illness, allergies, or symptoms[Dry cough]	Likert rating scale	dry_cough	Ordinal	Integer	Independent	Comorbidities [existing medical condition]
10	To what extent would you describe the pain or discomfort you are experiencing in connection with the above illness, allergies, or symptoms[Snoring]	Likert rating scale	snoring	Ordinal	Integer	Independent	Comorbidities [existing medical condition]



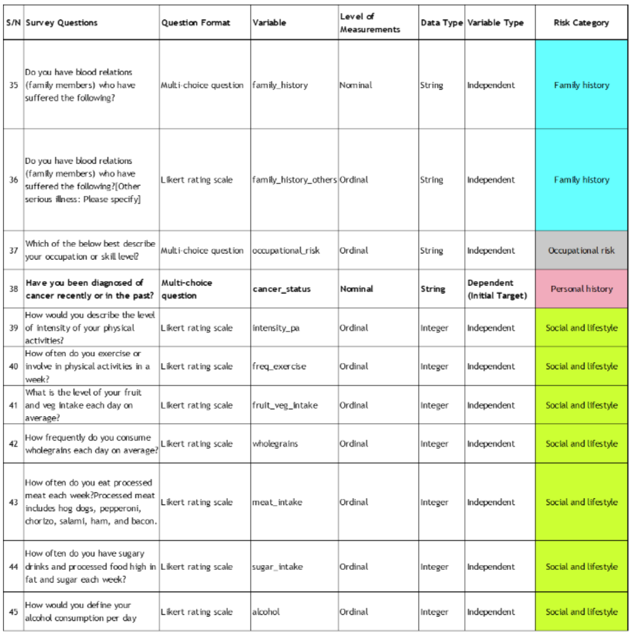
S/N	Survey Questions	Question Format	Variable	Level of Measurements	Data Type	Variable Type	Risk Category
11	To what extent would you describe the pain or discomfort you are experiencing in connection with the above illness, allergies, or symptoms[Diabetes]	Likert rating scale	diabetes	Ordinal	Integer	Independent	Comorbidities [existing medical condition]
12	To what extent would you describe the pain or discomfort you are experiencing in connection with the above illness, allergies, or symptoms[Heart Problems]	Likert rating scale	heart_problem	Ordinal	Integer	Independent	Comorbidities [existing medical condition]
13	To what extent would you describe the pain or discomfort you are experiencing in connection with the above illness, allergies, or symptoms[Asthma Hypertension]	Likert rating scale	asthma_or_hypertens	Ordinal	Integer	Independent	Comorbidities [existing medical condition]
14	To what extent would you describe the pain or discomfort you are experiencing in connection with the above illness, allergies, or symptoms[Lung Problem]	Likert rating scale	lung_problem	Ordinal	Integer	Independent	Comorbidities [existing medical condition]
15	To what extent would you describe the pain or discomfort you are experiencing in connection with the above illness, allergies, or symptoms[Loss of weight or appetite]	Likert rating scale	Loss_weight_or_appet	Ordinal	Integer	Independent	Comorbidities [existing medical condition]



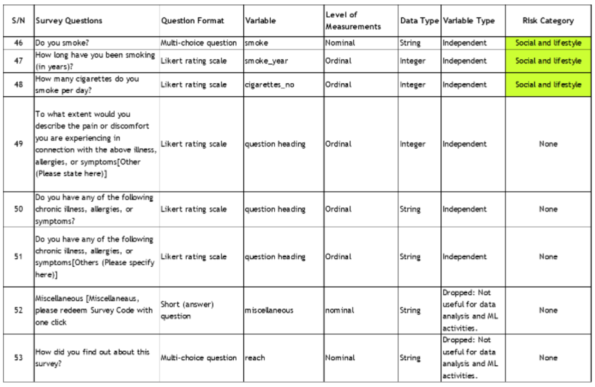
S/N	Survey Questions	Question Format	Variable	Level of Measurements	Data Type	Variable Type	Risk Category
16	To what extent would you describe the pain or discomfort you are experiencing in connection with the above illness, allergies, or symptoms[Chest Pain]	Likert rating scale	chest_pain	Ordinal	Integer	Independent	Comorbidities [existing medical condition]
17	To what extent would you describe the pain or discomfort you are experiencing in connection with the above illness, allergies, or symptoms[Coughing of Blood]	Likert rating scale	bloody_discharge	Ordinal	Integer	Independent	Comorbidities [existing medical condition]
18	What country are you from?	Short (answer) question	country	Nominal	String	Independent	Demographic
19	Your country of residence?	Short (answer) question	residence	Nominal	String	Independent	Demographic
20	Gender	Multi-choice question	gender	Nominal	String	Independent	Demographic



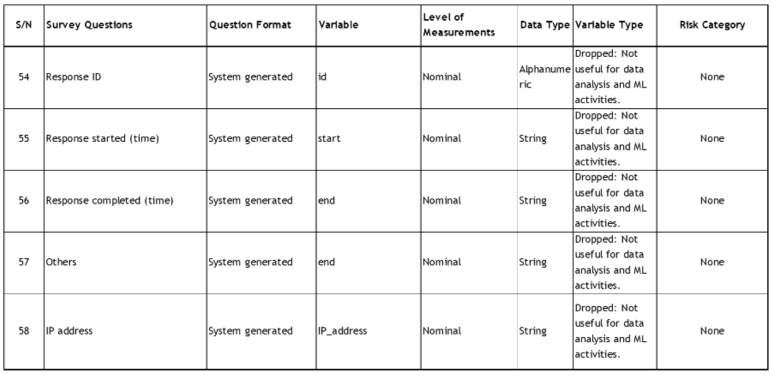
S/N	Survey Questions	Question Format	Variable	Level of Measurements	Data Type	Variable Type	Risk Category
21	What is your age? (Years)	Short (answer) question	age	Ratio	Integer	Independent	Demographic
22	How would you rate your economic class? Economic class system is based on earnings per year?	Multi-choice question	e_class	Ordinal	String	Independent	Economic class
23	How would you describe your level of exposure to the following in your location?[Sun exposure between 11am and 3pm]	Likert rating scale	sun_exposure	Ordinal	Integer	Independent	Environment
24	How would you describe your level of exposure to the following in your location?[Air pollution as a result of Co2 emission from automobiles]	Likert rating scale	co2_emission	Ordinal	Integer	Independent	Environment
25	How would you describe your level of exposure to the following in your location?[Air pollution due to industrial activities]	Likert rating scale	industrial_pollution	Ordinal	Integer	Independent	Environment
26	How would you describe your level of exposure to the following in your location?[Air pollution due	Likert rating scale	domestic_pollution	Ordinal	Integer	Independent	Environment
27	How would you describe your level of exposure to the following in your location?[Water pollution	Likert rating scale	water_pollution	Ordinal	Integer	Independent	Environment
28	Which ethnic group are you from[White]?	Multi-choice question	white	Nominal	String	Independent	Ethnicity
29	Which ethnic group are you from[Black]?	Multi-choice question	black	Nominal	String	Independent	Ethnicity
30	Which ethnic group are you from[Asian]?	Multi-choice question	asian	Nominal	String	Independent	Ethnicity
31	Which ethnic group are you from[Arab]?	Multi-choice question	arab	Nominal	String	Independent	Ethnicity
32	Which ethnic group are you from[Mixed]?	Multi-choice question	mixed	Nominal	String	Independent	Ethnicity
33	Which ethnic group are you from[Others]?	Multi-choice question	other	Nominal	String	Independent	Ethnicity
34	Do you have blood relations (family members) who have been diagnosed of cancer recently or in the past?	Multi-choice question	family_cancer_status	Nominal	String	Independent	Family history



S/N	Survey Questions	Question Format	Variable	Level of Measurements	Data Type	Variable Type	Risk Category
35	Do you have blood relations (family members) who have suffered the following?	Multi-choice question	family_history	Nominal	String	Independent	Family history
36	Do you have blood relations (family members) who have suffered the following?[Other serious illness: Please specify]	Likert rating scale	family_history_others	Ordinal	String	Independent	Family history
37	Which of the below best describe your occupation or skill level?	Multi-choice question	occupational_risk	Ordinal	String	Independent	Occupational risk
38	**Have you been diagnosed of cancer recently or in the past?**	**Multi-choice question**	**cancer_status**	**Nominal**	**String**	**Dependent (Initial Target)**	Personal history
39	How would you describe the level of intensity of your physical activities?	Likert rating scale	intensity_pa	Ordinal	Integer	Independent	Social and lifestyle
40	How often do you exercise or involve in physical activities in a week?	Likert rating scale	freq_exercise	Ordinal	Integer	Independent	Social and lifestyle
41	What is the level of your fruit and veg intake each day on average?	Likert rating scale	fruit_veg_intake	Ordinal	Integer	Independent	Social and lifestyle
42	How frequently do you consume wholegrains each day on average?	Likert rating scale	wholegrains	Ordinal	Integer	Independent	Social and lifestyle
43	How often do you eat processed meat each week?Processed meat includes hog dogs, pepperoni, chorizo, salami, ham, and bacon.	Likert rating scale	meat_intake	Ordinal	Integer	Independent	Social and lifestyle
44	How often do you have sugary drinks and processed food high in fat and sugar each week?	Likert rating scale	sugar_intake	Ordinal	Integer	Independent	Social and lifestyle
45	How would you define your alcohol consumption per day	Likert rating scale	alcohol	Ordinal	Integer	Independent	Social and lifestyle



S/N	Survey Questions	Question Format	Variable	Level of Measurements	Data Type	Variable Type	Risk Category
46	Do you smoke?	Multi-choice question	smoke	Nominal	String	Independent	Social and lifestyle
47	How long have you been smoking (in years)?	Likert rating scale	smoke_year	Ordinal	Integer	Independent	Social and lifestyle
48	How many cigarettes do you smoke per day?	Likert rating scale	cigarettes_no	Ordinal	Integer	Independent	Social and lifestyle
49	To what extent would you describe the pain or discomfort you are experiencing in connection with the above illness, allergies, or symptoms[Other (Please state here)]	Likert rating scale	question heading	Ordinal	Integer	Independent	None
50	Do you have any of the following chronic illness, allergies, or symptoms?	Likert rating scale	question heading	Ordinal	String	Independent	None
51	Do you have any of the following chronic illness, allergies, or symptoms[Others (Please specify here)]	Likert rating scale	question heading	Ordinal	String	Independent	None
52	Miscellaneous [Miscellaneaus, please redeem Survey Code with one click	Short (answer) question	miscellaneous	nominal	String	Dropped: Not useful for data analysis and ML activities.	None
53	How did you find out about this survey?	Multi-choice question	reach	Nominal	String	Dropped: Not useful for data analysis and ML activities.	None

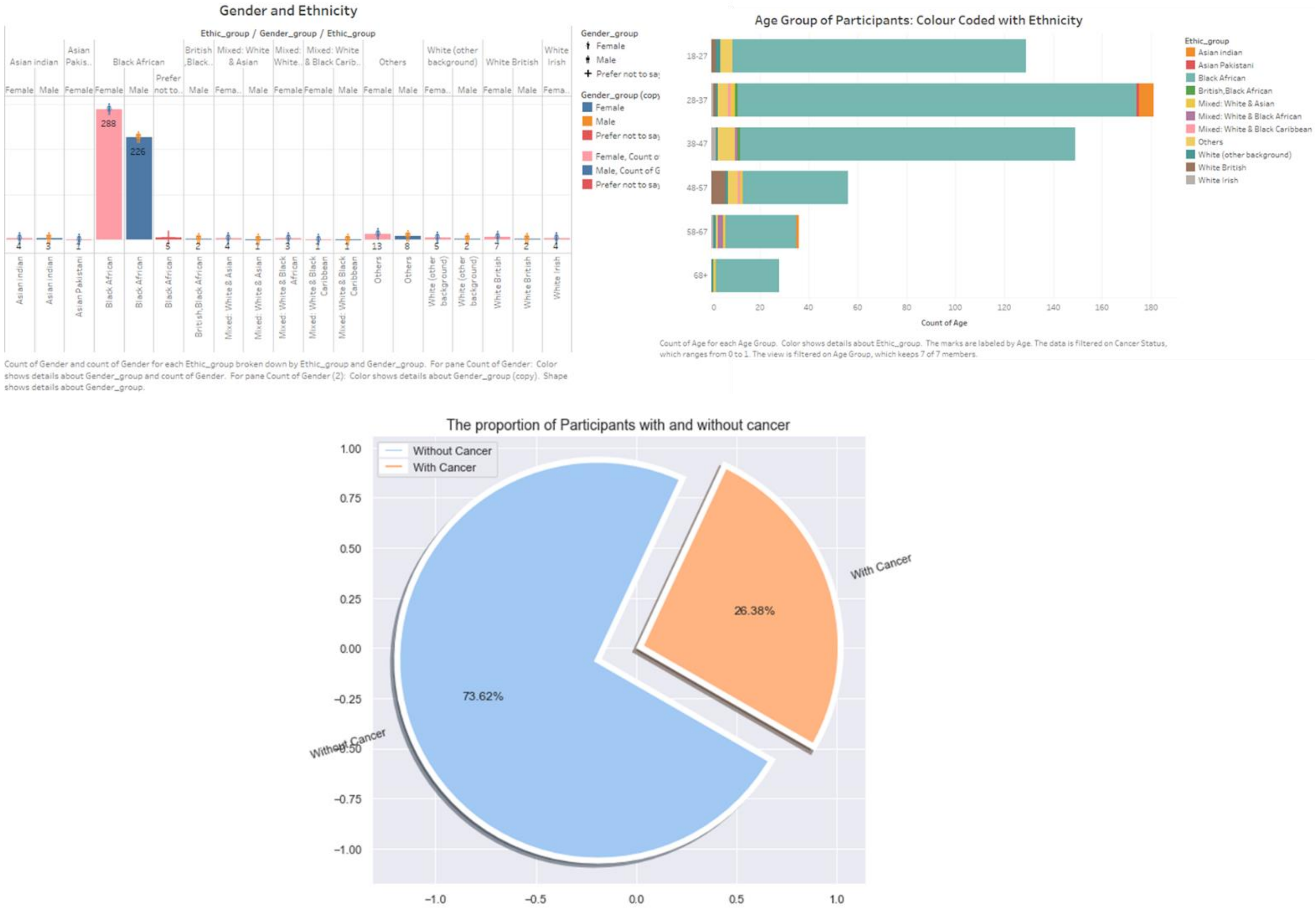


S/N	Survey Questions	Question Format	Variable	Level of Measurements	Data Type	Variable Type	Risk Category
54	Response ID	System generated	id	Nominal	Alphanumeric	Dropped: Not useful for data analysis and ML activities.	None
55	Response started (time)	System generated	start	Nominal	String	Dropped: Not useful for data analysis and ML activities.	None
56	Response completed (time)	System generated	end	Nominal	String	Dropped: Not useful for data analysis and ML activities.	None
57	Others	System generated	end	Nominal	String	Dropped: Not useful for data analysis and ML activities.	None
58	IP address	System generated	IP_address	Nominal	String	Dropped: Not useful for data analysis and ML activities.	None

Table 4 and 5 above show how the questionnaire was designed, the corresponding question, data, and variable types. The multiple-choice questions were used for the nominal variables because they lack numerical significance and can be measured using the frequency distribution table to depict each option and the frequency at which these options were selected. Whereas the Likert-scale questions were used for the ordinal variables because they depict the order of value. There are quantitative values because one rank is higher than the other. The Likert-scale questions were measured using Mode, Median and cross tabulation analysis. In similar vein, the short (answer) questions were used for the ratio variables. The ratio variables depict the order and differences between values and can take true zero point. It is quantitative and the absence of a value can still provide information. These variables were measured using mode, median and mean. In addition, the scale can be analysed using t-tests, ANOVA and correlation analyses. ANOVA tests the significance of the survey results. While t-tests and correlation establish variables relationship. **Table 5** contains 55 categorical variables (20 nominal and 35 ordinal) and 3 numerical variables (ratio). This implies that the data table contains 57 independent variables and one initial (target) variable called cancer_status.

3.3. Study Population and Scope

The target population for the study includes adults (aged 18 and above) both asymptomatic and symptomatic for ease of generalization. The scope of the research project is to predict and classify cancer risk level based on selected importance features which form part of the simple baseline information of a person. Probability sampling method was used to select the target population that is representative of the of the entire adult population. The choice of the sampling method is to produce results that are representative of the whole population. Clustering (sampling) techniques in dividing the population into subgroups (clusters) with similar characteristics. The resulting clusters enhanced the risk assessment and classification of cancer risk level as either low or high. The below figures summarize the target population.



Count of Gender and count of Gender for each Ethic_group broken down by Ethic_group and Gender_group. For pane Count of Gender: Color shows details about Gender_group and count of Gender. For pane Count of Gender (2): Color shows details about Gender_group (copy). Shape shows details about Gender_group.

Count of Age for each Age Group. Color shows details about Ethic_group. The marks are labeled by Age. The data is filtered on Cancer Status, which ranges from 0 to 1. The view is filtered on Age Group, which keeps 7 of 7 members.

Figure 14: Ethnicity, age distribution, and cancer status of participants: Developed by Author using Tableau and Python

In the above figure, participants who completed the questionnaire where from 10 different ethnic groups including ***Black African***, ***White British***, ***White & Black African***, ***Asian Indian***, ***White Irish***, ***Mixed***, ***White & Asian***, ***White (any other white background)***, ***Mixed: White & Black African***, ***British: Black African***, ***Mixed: White & Black Caribbean***, ***Asian Pakistani and others***. Black African has the largest representation of 517. Out of the 580 participants who took part in the survey, the highest age group was 28-37 years. In addition, participants who has been diagnosed of cancer out of the total participants was 153. Whereas the total participants without cancer was 427.

3.4. Measures and Methods of Analysis

Considering the quantitative nature of the research project, before the exploratory data analysis (EDA), the collected data was prepared and cleaned. The original dataset cannot be used immediately because it is complex, inconsistent, noisy, and not in the right shape due to the heterogeneous origin. The headings in the survey questionnaire were mapped and renamed to appropriate variables, the dataset was checked for missing values, outliers, and categorical data were transformed to numerical value using the label encoding. The Python-Sklearn libraries including but not limited to numpy, pandas, seaborn, matplotlib, scikit learn were used to carry out mathematical and statistical analysis, loading and manipulation of data frames, visualisation, scikit learn machine learning activities respectively. The data pre-processing is necessary to convert the raw data into a clean data set and improve the overall data quality. It includes the following steps:

- **Data Cleaning**: the aim is to extract the answers from the questionnaire into a variable, handling missing data, detection, and removal of outliers, minimizing duplication and computed biases within the data.
- **Feature Encoding and Transformation:** This entails using data transformation strategies such as Min-Max scaler amongst others to consolidate data or numerical attributes into alternate forms or structures by scaling up or down within a specified range.
- **Prediction of the Class Label:** The raw data collected during the survey do not have a class label. Hence, the unsupervised machine learning techniques such as clustering was developed in dividing the dataset into sub-group with similar characteristics to establish the appropriate clusters (class label or target) to use for machine learning activities.
- **Feature Selection:** This is process of selecting features that are representations or attributes that describe the dataset and enhance the machine learning model in predicting the classes accurately. In other words, feature selection is a reduced representation of the data; smaller in volume but produce quality results. Feature selection techniques such as Random Forest classifier, Chi-square, mutual information classification, forward feature selection and backward feature elimination were deployed to execute the tasks.

Data Cleaning

The answers from the survey questionnaire in table 3 above were extracted, stored in appropriate variables and risk category as part of the data cleaning process. For instance, the question:

- How much do you weigh (in kg)? For example, a person weight in kilograms (kg) is: 67? The answer to this question was mapped using user defined function in python and stored in the variable **weight** as shown in table 3 above.
- The same is applicable to the question: How tall are you (in cm)?For example, a person height in centimeter (cm) is: 185? The response to this was mapped and stored in the variable name **height.** This process was also applied to the remaining 56 questions of the survey.

In the table below, 10 out of the initial 58 features were dropped because they contained question headings and data that were neither useful for exploratory data analysis nor machine learning activities. The features(columns) dropped includes the below:

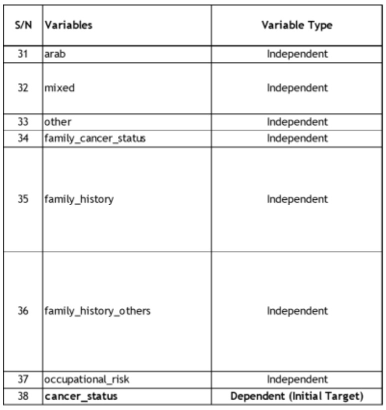
Table 6: Dropped features from the original data collection shown in table 3 above

S/N	Survey Questions	Variable	Variable Type	Risk Category
49	To what extent would you describe the pain or discomfort you are experiencing in connection with the above illness, allergies, or symptoms[Other (Please state here)]	question heading	Independent	None
50	Do you have any of the following chronic illness, allergies, or symptoms?	question heading	Independent	None
51	Do you have any of the following chronic illness, allergies, or symptoms[Others (Please specify here)]	question heading	Independent	None
52	Miscellaneous [Miscellaneaus, please redeem Survey Code with one click	miscellaneous	Dropped: Not useful for data analysis and ML activities.	None
53	How did you find out about this survey?	reach	Dropped: Not useful for data analysis and ML activities.	None
54	Response ID	id	Dropped: Not useful for data analysis and ML activities.	None
55	Response started (time)	start	Dropped: Not useful for data analysis and ML activities.	None
56	Response completed (time)	end	Dropped: Not useful for data analysis and ML activities.	None
57	Others	end	Dropped: Not useful for data analysis and ML activities.	None
58	IP address	IP_address	Dropped: Not useful for data analysis and ML activities.	None

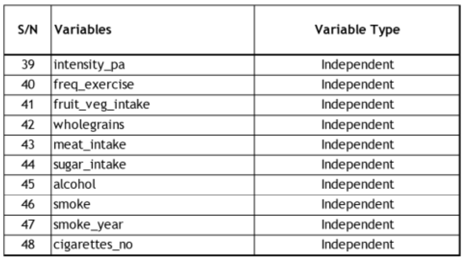
After dropping the above features due to redundancy, the below shows the features that were retained and their corresponding variable names. The reduced and new dataset has 580 instances and 48 features:

Table 7: Missing values in the data collected

S/N	Variables	Variable Type
1	weight	Independent
2	height	Independent
3	fatigue	Independent
4	shortness_of_breath	Independent
5	wheezing	Independent
6	swallowing_difficult	Independent
7	clubbing_of_nails	Independent
8	frequent_cold	Independent
9	dry_cough	Independent
10	snoring	Independent
11	diabetes	Independent
12	heart_problem	Independent
13	asthma_or_hypertension	Independent
14	lung_problem	Independent
15	Loss_weight_or_appetite	Independent
16	chest_pain	Independent
17	bloody_discharge	Independent
18	country	Independent
19	residence	Independent
20	gender	Independent
21	age	Independent
22	e_class	Independent
23	sun_exposure	Independent
24	co2_emission	Independent
25	industrial_pollution	Independent
26	domestic_pollution	Independent
27	water_pollution	Independent
28	white	Independent
29	black	Independent
30	asian	Independent



S/N	Variables	Variable Type
31	arab	Independent
32	mixed	Independent
33	other	Independent
34	family_cancer_status	Independent
35	family_history	Independent
36	family_history_others	Independent
37	occupational_risk	Independent
38	**cancer_status**	**Dependent (Initial Target)**



S/N	Variables	Variable Type
39	intensity_pa	Independent
40	freq_exercise	Independent
41	fruit_veg_intake	Independent
42	wholegrains	Independent
43	meat_intake	Independent
44	sugar_intake	Independent
45	alcohol	Independent
46	smoke	Independent
47	smoke_year	Independent
48	cigarettes_no	Independent

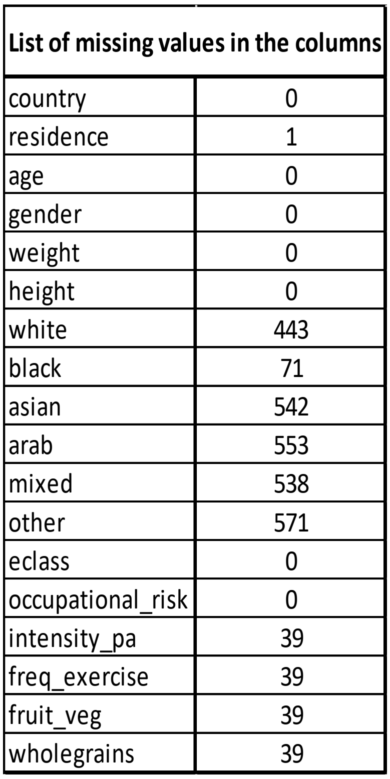
Checking Missing Values

As part of the data cleaning process, missing values were checked across the data frame rows and columns using the df.isnull().any(). The major cause of the missing values in the dataset were due to incomplete or partial responses from participants, and the selection of not applicable (N/A) in some of the multichoice and Likert scale questions. For instance, the questions on:

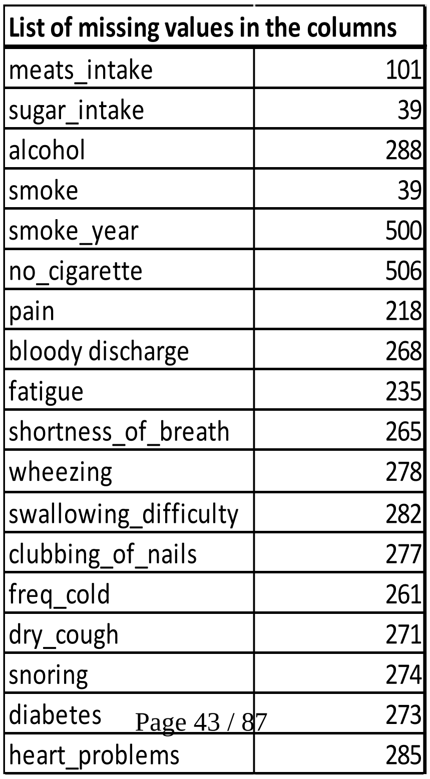
- **How would you define your alcohol consumption per day?** The responses contained 249 participants who do not take alcohol and selected N/A. 39 participants partially completed the questionnaire resulting into blank cells. Hence, the total missing value of 288 for the column.
- **Do you smoke?** There were 39 participants who partially completed the questionnaire resulting into the missing values.
- **How long have you been smoking (in years)?** The 461 participants who do not smoke selected N/A and 39 participants partially completed the questionnaire resulting in a total of 500 missing values for the column.
- **Which ethnic group are you from[Arab]?** This question was designed using the matrix choice (one answer) format. This implies if a field is not related or selected for the column, it remains blank. Hence, there was 553 blank fields in the column not related to the Arab. The same applies to the remaining ethnic groups with missing values.
- **To what extent would you describe the pain or discomfort you are experiencing in connection with the above illness, allergies, or symptoms[Bloody discharge (cough, stool, nose, private parts etc)]?** 266 fields in this column were blank because the question was partially or not completed. The same applies to the remaining columns with questions relating to comorbidities or history of a person's medical condition.
- **How would you describe your level of exposure to the following in your location?[Sun exposure between 11am and 3pm]?** 60 fields in this column were blanks resulting from questions in the survey that were partially completed. The same is applicable to the remaining questions regarding pollution.

The below gives a tabular representation of the missing values.

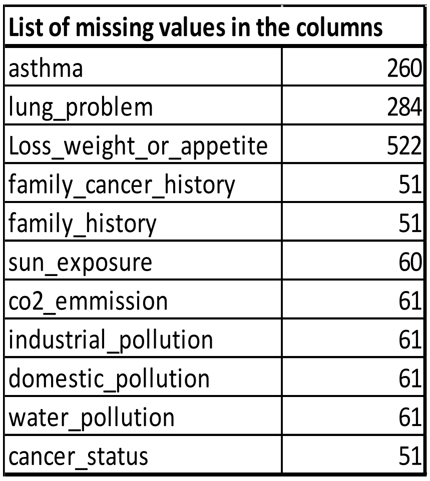
Table 6: Tabular representation of missing values



List of missing values in the columns	
country	0
residence	1
age	0
gender	0
weight	0
height	0
white	443
black	71
asian	542
arab	553
mixed	538
other	571
eclass	0
occupational_risk	0
intensity_pa	39
freq_exercise	39
fruit_veg	39
wholegrains	39



List of missing values in the columns	
meats_intake	101
sugar_intake	39
alcohol	288
smoke	39
smoke_year	500
no_cigarette	506
pain	218
bloody discharge	268
fatigue	235
shortness_of_breath	265
wheezing	278
swallowing_difficulty	282
clubbing_of_nails	277
freq_cold	261
dry_cough	271
snoring	274
diabetes	273
heart_problems	285



List of missing values in the columns	
asthma	260
lung_problem	284
Loss_weight_or_appetite	522
family_cancer_history	51
family_history	51
sun_exposure	60
co2_emmission	61
industrial_pollution	61
domestic_pollution	61
water_pollution	61
cancer_status	51

Handling of missing values

Missing values in dataset is very common phenomenon for primary data collection through survey because of the heterogeneous origin (Analytics Vidya 2022; Scribbr 2022). The correction of missing is essential to reduce bias and produce a very efficient and suitable machine learning model. Considering the context of the problem and the fact that majority of the features are categorical variables, the missing values were replaced using the most frequent value: imputed based on mode in the case of categorical variables, and constant such as 0 was as direct replacement for some of the features as well. The justification for using the most frequent value as a replacement technique was because:

- It is the widely used missing value strategy for categorical variables in the machine learning community and medical settings (Analytics Vidya 2022; Scribbr 2022).
- The imputation method based on mode is the most appropriate measurement technique for nominal and ordinal variables (Analytics Vidya 2022; Scribbr 2022).
- Considering the size of the dataset and the percentage of availability of data for all the features, the replacement used the most frequent is the most suitable option for missing values.

The below mapping and consolidation of features were done before the replacement of missing values was carried out.

- The columns containing white, black, asian, arab, mixed [6:11] were combined into one new single column called ethnicity using the below code:

```
df_survey['ethnicity']= df_survey[df_survey.columns[6:11]].apply(lambda x:
','.join(x.dropna().astype(str)), axis=1)
```

- As a result of the above, the columns white, black, asian, arab, and mixed were dropped from the data frame.

```
df_survey = df_survey.drop(columns = ['white', 'black', 'asian', 'arab',
'mixed'], axis=1)
```

- The missing values for number of smoking year and cigarettes taken on a daily basis for the participants who do not smoke were replaced by constant 0.

```
df_survey['smoke_year'].fillna(value = 0, inplace = True)
df_survey['no_cigarette'].fillna(value = 0, inplace = True)
```

After step 1-3 of the above were carried out, the remaining missing values were replaced using the most frequent values imputed using the mode as detailed in the below code snippet.

```
df = df_survey.fillna(df_survey.mode().iloc[0])
```

Handling of Outliers and Creation of Derived Features

There was presence of extreme values (outliers) in the features: height, weight, and age due to inconsistent or erroneous values input by the survey participants. Hence, this may pose problem for exploratory data analysis and machine learning model if not handled properly. The below are the outliers noticed in the data collected based on preliminary observations:

- The height column (in cm) had a very extreme value of 9999, 80 and 205.

Considering height is one of the three numerical (ratio) variables in the dataset, the outlier was handled by replacement using the median (average) height. The code snippet below was used to set up a logical operation to clean up outlier in the height variable by constraining it to the median value.

```
df['height'] = np.where(df['height'] > 9999, 165, df['height'])
df['height'] = np.where(df['height'] <= 80, 165, df['height'])
df['height'] = np.where(df['height'] > 205, 165, df['height'])
```

In addition to the above, calculated fields were created for Body Mass Index (BMI) using the ratio of weight and height (in meter)^2.

- The BMI was derived using the below code snippet.

```
#create a calculated field for BMI
df['height_m'] = (df['height'] / 100)
df['bmi'] = (df['weight'] / df['height_m']**2)
```

- The smokes per person year (sppy) was also derived from the combination of number of cigarettes taken per day and the number of smoke year as detailed in the code below:

The below code snippet shows the dataframe after replacement of the missing value

```
# double checking of the dataframe after replacement of missing values
print("\n count of all NaN in the Dataframe (columns and rows)after replacement\n", df.isnull().sum().sum())
```

Output:

count of all NaN in the Dataframe (columns and rows)after replacement
 0

Other cleaning tasks performed on the data

There was inconsistent value input from participants relating to their country of origin and country of residence. A good example, Nigeria as a country was entered by some participants as Nigeris, nigeria and in some cases Lagos, Ogun. Another example was participants entered United States of America as usa, US, USA etc. These input from users were inconsistent and had to be renamed correctly before storing in the appropriate variable. The process carried out for correction of this inconsistent responses are detailed in the appendix section of the report.

Feature Encoding and Transformation

Upon completion of the data cleaning process, the Python-Sklearn pre-processing library: label encoding was utilised in converting the categorical variables to numerical variables. This categorical encoding was necessary to prepare the dataset in the right form, structure and format that is readable and processed for machine learning activities. Although the label encoding uses alphabetical ordering to assign values to features when there is no specific order or rank, it has been selected as the most suitable encoding technique over one-hot encoding because the list of features to encode is relatively large(n=7) with sub-categories. Hence, it is not efficient to use one hot encoding as it will lead to high memory consumption and problems of dummy variables. In addition, 3 out of the 7 features to be encoded are ordinal (categorical) variables as shown in the code snippet below:

```
objlist = ['gender', 'eclass', 'occupational_risk', 'smoke', 'family_cancer_history', 'ethnicity', 'cancer_status']
#converting  the above objList features into numeric type using forloop as given below:
le = LabelEncoder()
df[objlist] = df[objlist].apply(le.fit_transform)
```

The below were the output of the label encoding exercise and the corresponding assigned values:

- gender: {Female: **0**, Male: **1**, Prefer not to say: **2**}
- economic class (eclass): {Lower class: **0**, Lower-middle class:**1**, Middle-class: 2 , Upper-middle class: **4**, Upper-higher class: **5** }
- occupational_risk: {Skill 1: **3**, Skill 2: **0**, Skill 3: 1, Skill 4: **2**}
- smoke: {No: 0, Yes: 1}
- family_cancer_history: {No history of cancer: 0, history of cancer: 1}
- cancer_status: {No Cancer: 0, Diagnosed of cancer: 1}
- ethnicity:

Ethnic group	**Encoded as:**
Black African	2
White British	9

Black (any other black background)	7
Asian indian	1
White Irish	10
Mixed: White & Asian	4
White (any other white background)	8
Mixed: White & Black African	5
British: Black African	3
Mixed: White & Black Caribbean	6
Asian Pakistani	0

Applications of unsupervised machine learning techniques to determine the optimal number of clusters in the dataset

The primary data collected during the survey does not have a class label (target) variable resulting into the use of the KMEANS and the KMODE clustering methods in establishing the appropriate class label. Both the KMEANS and KMODES clustering were utilised to divide the entire dataset into sub-group with similar characteristics. The KMODES carried out the clustering based on dissimilarity or similarity amongst the sub-group to predict the class label. On the other hand, the KMEAN clusters data points based on average distance amongst the sub-group to establish the class label. The KMODES was selected as the most suitable clustering techniques because it is the valid choice for categorical variables.

The clustering process starts with finding the optimal number of clusters. The elbow method was utilised in find the optimal number of clusters as shown in the figure 16 below:

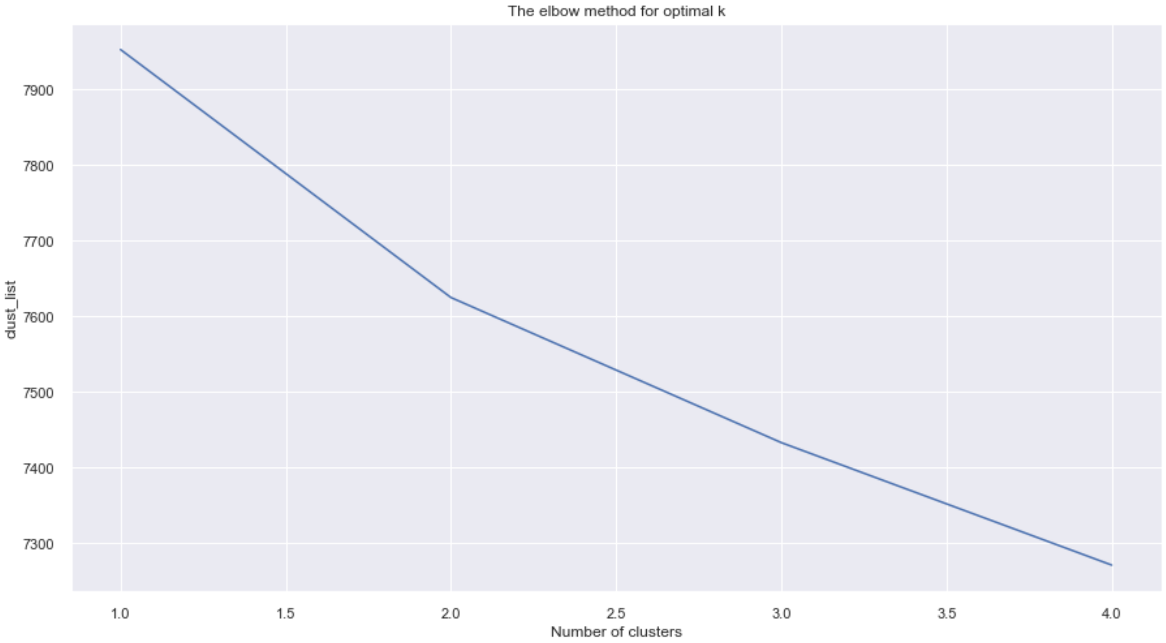


Figure 15a: The optimal number of clusters (k) using the KMODES elbow method

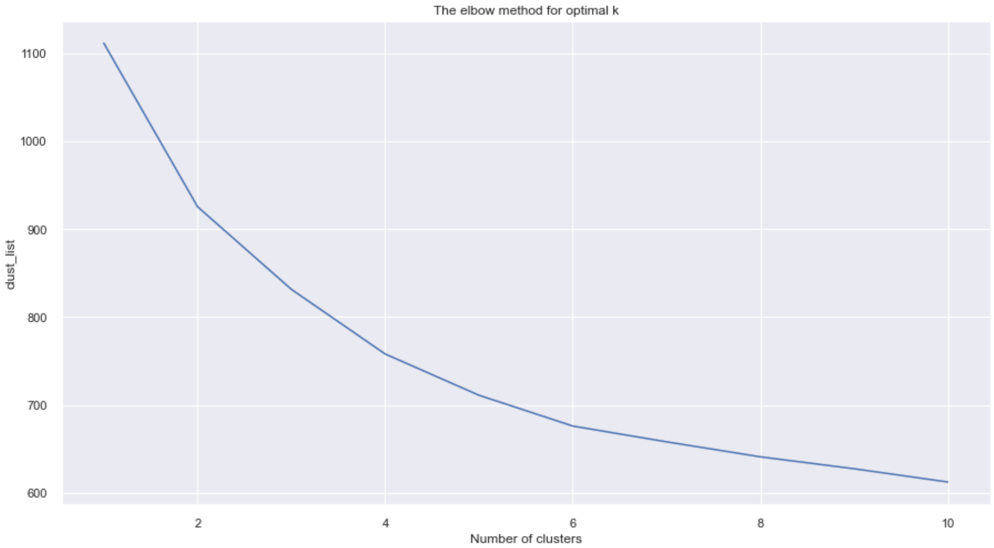


Figure 15b: The optimal number of clusters (k) using the KMEANS elbow method

Considering the elbow method does not always depict a succinct graphical representation of the optimal cluster, the Silhouette method was deployed to validate the consistency within the clusters and avoid any ambiguity regarding the optimal number of clusters. a high Silhouette Score is desirable. The Silhouette Score reaches its global maximum at the optimal k which was determined using the KMODES as shown in the figure below:

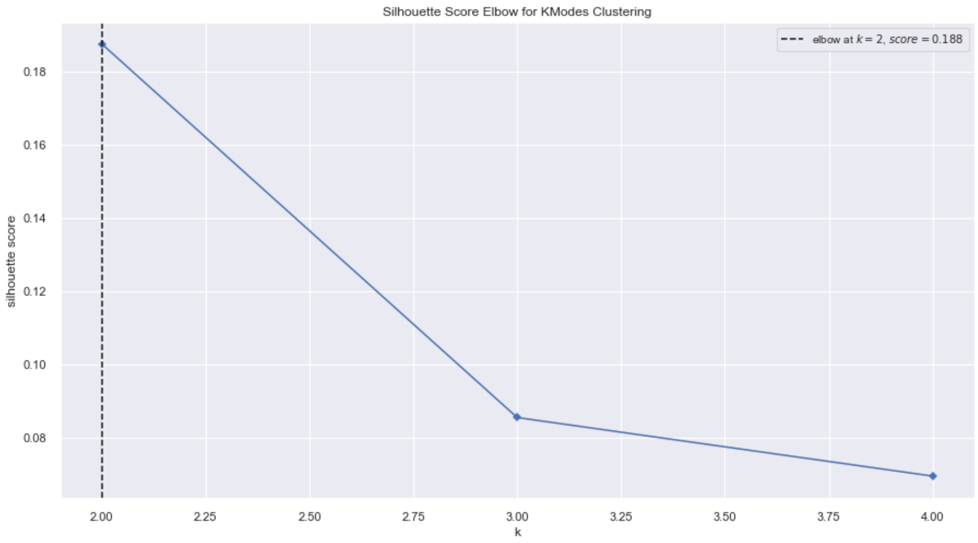


Figure 16a: Silhouette Score using the KMODES clustering method

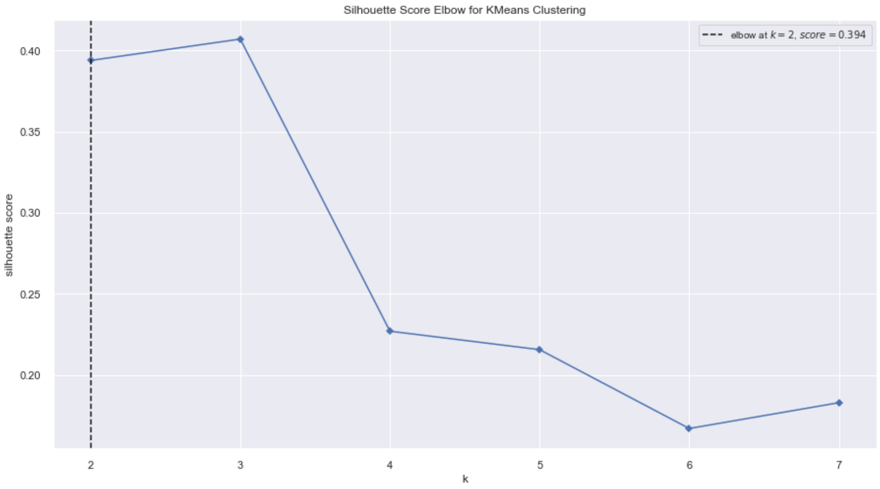


Figure 16b: Silhouette Score using the KMEANS clustering method

After finding the optimal number of clusters, the KMODES clustering model is fit to the dataset to predict clusters for each of the data elements as shown below. The new class label from the clustering analysis was **risk_level**.

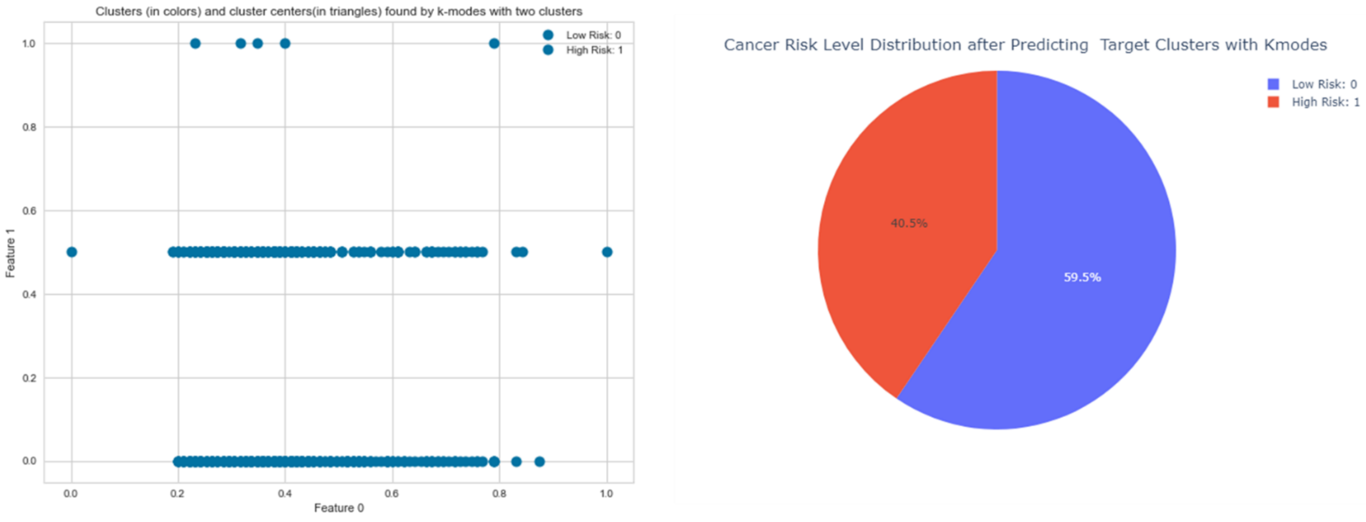


Figure 17: Cancer risk level distribution after predicting the target label with KMODES

Feature Selection

Before developing the machine learning model, it is important to select the features (independent variables) that give an approximate representation of the entire dataset and enhance the model capability in predicting cancer risk classes accurately. The features in the dataset are referred as the attributes, variables, or columns in the data frame(in this case the csv file) that are measurable data for exploratory analysis and machine learning activities. The process of the selecting important feature was carried out using five different techniques including:

1. **Chi-Square**: This is one of the filter methods that uses statistical techniques to select feature based on their scores in statistical test and corresponding correlation in relation to the target (dependent) variable.
2. **Forward Feature Selection**: This technique is one of the wrapper methods that select features based iterative process of starting without any features and then progressively introduces each of the feature that improves the model until the additional feature does not improve the model performance anymore.
3. **Back Feature Elimination method**: This is another wrapper method which is the reverse of the forward feature. This process iteratively starts with all the features, and gradually remove the least significant features until optimal features that give the best performance for the model is reached.

4. **The Random Forest Classifier**: It is an ensemble or embedded method that combines the quality of the filter and wrapper techniques in selecting features that give the best possible model performance. The random forest leverages on the power of multiple trees to make the selection decision.
5. **The mutual information (MI)** classification: The mutual information technique selects features based on the combination that gives or leaks information about the target variable. The MI score ranges from 0 to infinity. The higher the MI score, the closer the connection between the feature and the target. Whereas low score suggests a weak connection between the target and the feature.

The above techniques have been selected to experiment with the selection of the most appropriate features because they are the most suitable methods for selection of features for categorical variables. Machine learning models performance are enhanced when they are trained with the appropriate set of features, below detailed the importance of the feature selection:

- To reduce the training and computational time
- To reduce overfitting of the model
- To enhance the interpretability of the model and reduce any complexity
- To improve the predictive performance and accuracy of the model

Table 7: Summary of the 10 most important Feature Selection out of 37: using the the Chi-square, Random Forest Classifier, Mutual Information, Forward Features Selection and Backward Features Elimination

Chi-square	Random Forest Classifier	Mutual Information	Forward Features Selection	Backward Features Elimination
gender	gender	occupational_risk	age	gender
freq_cold	industrial_pollution	freq_exercise	eclass	occupational_risk
age	domestic_pollution	fruit_veg	occupational_risk	freq_exercise
bmi	sugar_intake	wholegrains	intensity_pa	meat_intake
sugar_intake	smoke_ppy	domestic_pollution	freq_exercise	pain
dry_cough	freq_cold	water_pollution	smoke_ppy	shortness_of_breath

snoring	alcohol	co2_emission	fatigue	wheezing
pain	co2_emission	industrial_pollution	wheezing	dry_cough
smoke_ppy	freq_exercise	meat_intake	clubbing_of_nails	domestic_pollution
alcohol	sun_exposure	sugar_intake	snoring	bmi

3.5. Design Specification

This research project has used the **Python programming together with the Sklearn (Scikit learn) libraries** for the data analysis, training and building of the machine learning model. **Tableau** was used as a supplementary tool for the data analytics and visualisation. The codes are written in Python, which is widely used programming language for machine learning project. All codes relating to this project can be found in Appendix 3 of the report.

3.6. Statistical Test for Hypothesis

The Kruskal-Wallis H Test, a non-parametric hypothesis testing was used to test whether or not the independent variables simultaneously explain a statistically significant amount of variance in the dependent(target) variable. The non-parametric statistical testing was chosen because the variables do not follow a normal distribution. Hence, not compatible to two of the three assumptions of the parametric testing.

3.7. Model Modelling, Evaluation, and Deployment

The modelling was carried out using supervised machine learning as the aim of the study was to predict and classify the risk level of cancer. Several categorical predictor variables were mapped to the target output to classify cancer risk level into two categories: Low or High risk of cancer. For training and testing of the model, the dataset was divided into three distinct parts: training (n=371), validation (n=93) and testing (n=116) for building, training, and testing the model

```
train_features, X_test, train_target, y_test = train_test_split(X, y, test_size
= 0.20, random_state=SEED)
# splitting the training set into training and validation for model building and
training
X_train, X_val, y_train, y_val = train_test_split(train_features, train_target,
test_size = 0.20, random_state=SEED)
```

The model performance evaluation was done using the success metrics: **"Area Under the Curve" (AUC) of "Receiver Characteristic Operator" (ROC).** The ROC curve is a probability curve that depicts the true positive rate (TP) against the false negative rate (FN) at various threshold. Whereas the AUC evaluates the ability of the model (or classifier) to separate the positives and negatives classes. The higher the AUC score, the better the separability capacity of the classifier. The AUC-ROC curve is based on the principles of the confusion matrix such as Accuracy, Precision, and Recall. In the context of the problem (in this case, healthcare), the Confusion Matrix will be used for the model evaluation. RECALL will be used to measure how well the algorithm fits the problem because False Negative**(FN)**: wrong classification as low risk when it is high is of higher concern than False Positive **(FP)**: wrong classification as High cancer risk. True Positive(TP): high cancer risk) should be detected with a higher accuracy and should not go undetected. The model will be deployed using the STREAMLIT web application for ease of usability and consumption by the users.

3.8. Justification for Methodology Choice

In the context of the research project aims and objectives, a quantitative methodology has been selected as the most appropriate because it provides:

- flexibility to make generalization and generate reproductive knowledge.
- expressing the data in numbers and carrying out the exploratory data analysis using statistical and visualisation techniques make the results more interpretable and usable by the target audience.

The data collection for the research implementation was done using the primary data collection method through survey questionnaire. The questionnaires were distributed online via social media platforms, in-person and were completed by the participants by themselves.

The justifications for this choice of data collection are as detailed below:

- to get first-hand information and gain an in-depth understanding of the demographic (general characteristics) of the target population such as country of origin, country of residence, age, gender, ethnicity, economic class amongst others.
- It is widely used for health-related research such as collecting data from people about symptoms, risk factors and treatments (Scribbr 2022; Institute for work and health 2015).

- It gives the flexibility to tailor the questionnaires to elicit the data that will help provides answer to the research problem or questions to addressed.
- It provides significant control over sampling and measurements.

Although the secondary data collection was utilised during the pilot study to synthesise existing knowledge and identify patterns on a large scale. In addition, it was easier and faster to collect. Despite the numerous advantages of the primary data collection, it is more expensive and time-consuming process.

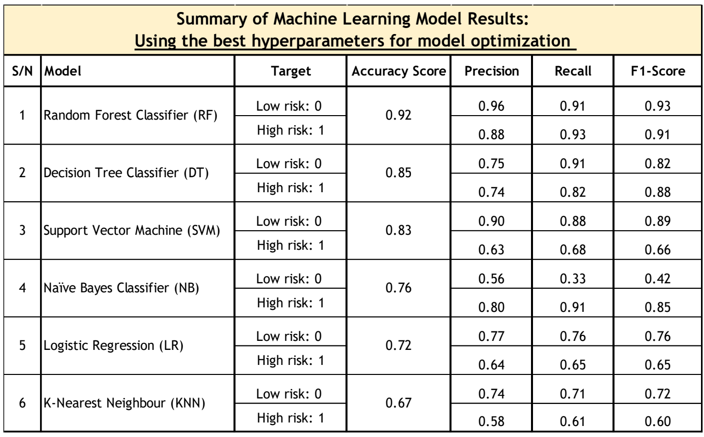
CHAPTER FOUR

Results

This section presents the findings of the research project in a concise, logical, and objective manner in line with the study aim, objectives and research questions.

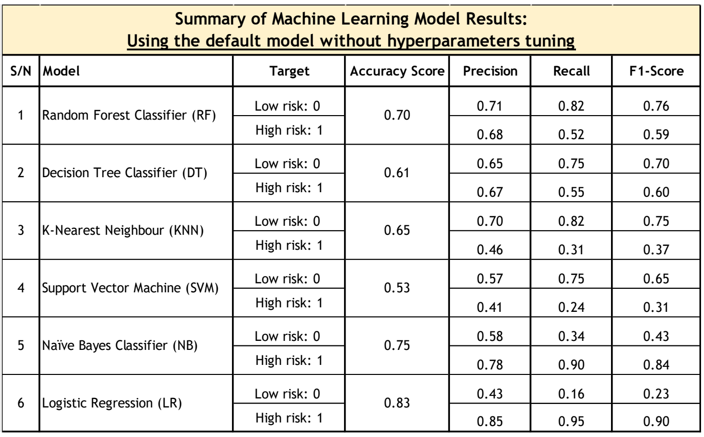
The Kruskal-Wallis hypothesis testing (**shown in fig. 18i**) was used to test the hypothesis that different risk factors have effect on the level of cancer risk. Apart from age, height, and weight; other risk factors were recorded on a Likert-rating scale of 1 to 10 as the predictors; and the level of cancer risk was classified into two categories (Low risk: 0 and High Risk: 1) as the target variable. The Kruskal-Wallis hypothesis showed a t-statistics (95% confidence level) of 54.9 with p-value of 0.000 which is less than the critical value of 0.05. This contradicts the initial hypothesis(H0) that the different risk factors are not statistically significant and do not have any effect on the level of cancer.

Both the random forest classifier and the Chi-square test of association (otherwise known as test of independence) in **fig. 18c and 18j** were deployed to find the relevant (significant) features with strong correlation and information about the target variable. Similarly, the machine learning algorithms were tuned using the grid search and random search to select the best hyperparameters to optimise the model performance as also customise it **(fig. 18a, b, d).** In order to establish the average generalisation performance of the model, the 10 folds cross validation was deployed to validate how well the model generalises its learning **(fig. d). Fig. g** summarizes the K-MEANS cluster analysis carried out using Tableau to show number of clusters and patterns in each cluster. The AUC-ROC (**in fig. 18f**) shows the separability power of the model in classifying positives and negatives risk of cancer. Exploratory data analysis was carried out using visualisation techniques. The density plot, pair plot and correlation heatmap were used to establish univariate, bivariate, and multi-variate relationship respectively (**see fig. 18j and k**). Finally, the logistic regression summary results show the relationship and magnitude of the association between each of the significant features and the target variables as shown in **fig.l** below.



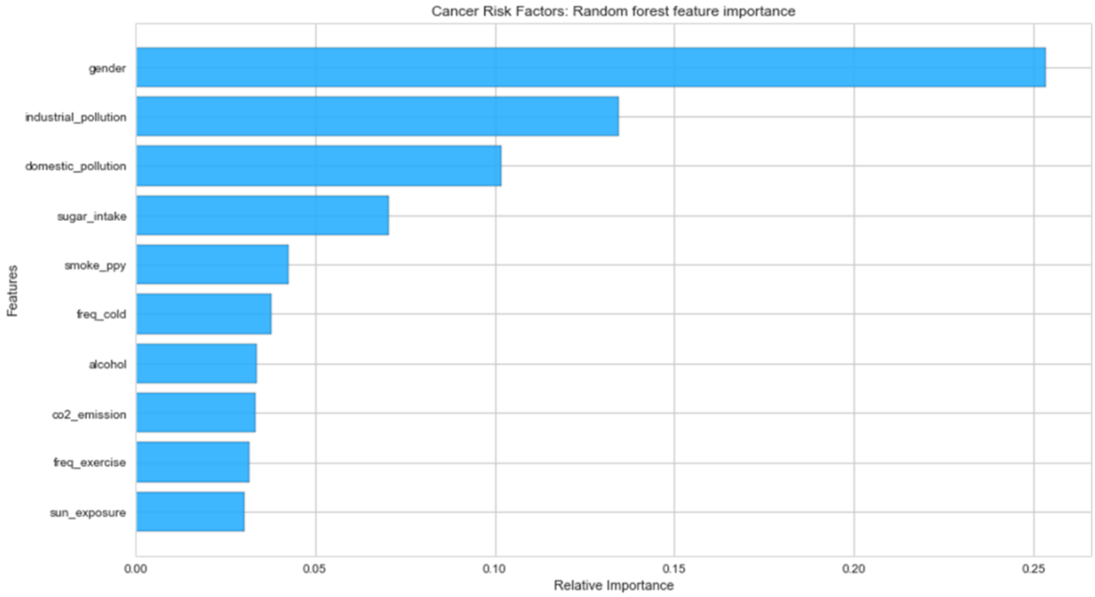
Summary of Machine Learning Model Results:
Using the best hyperparameters for model optimization

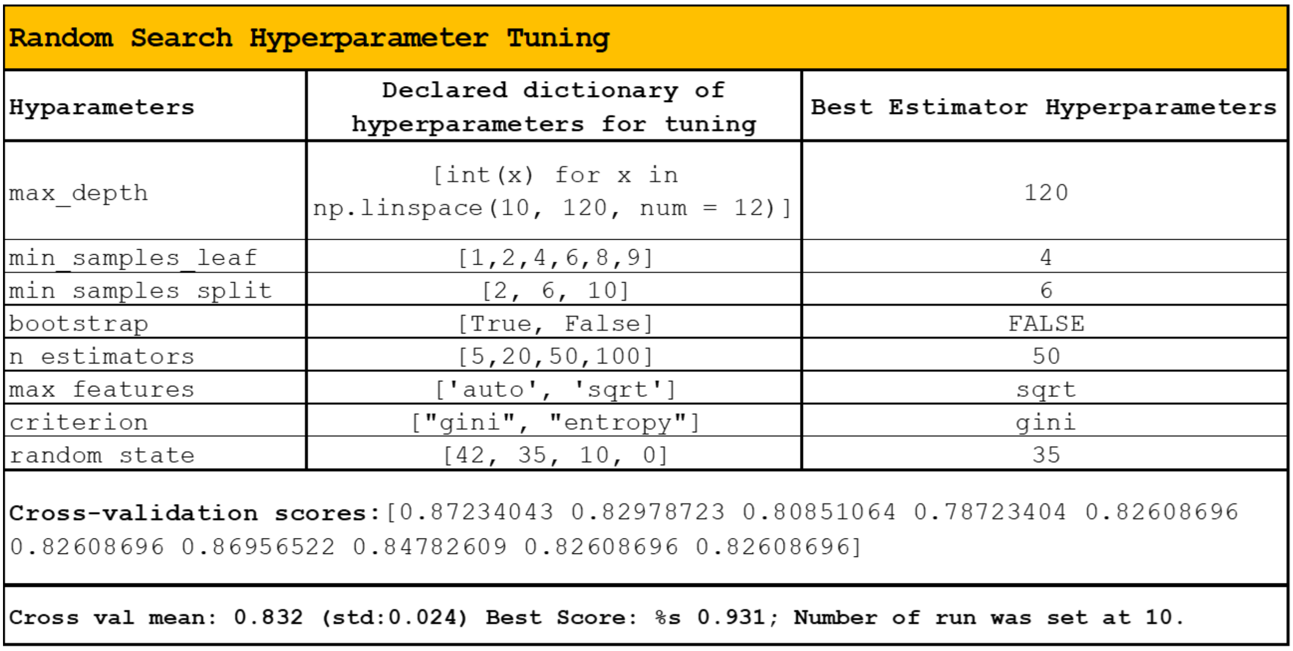
S/N	Model	Target	Accuracy Score	Precision	Recall	F1-Score
1	Random Forest Classifier (RF)	Low risk: 0	0.92	0.96	0.91	0.93
		High risk: 1		0.88	0.93	0.91
2	Decision Tree Classifier (DT)	Low risk: 0	0.85	0.75	0.91	0.82
		High risk: 1		0.74	0.82	0.88
3	Support Vector Machine (SVM)	Low risk: 0	0.83	0.90	0.88	0.89
		High risk: 1		0.63	0.68	0.66
4	Naïve Bayes Classifier (NB)	Low risk: 0	0.76	0.56	0.33	0.42
		High risk: 1		0.80	0.91	0.85
5	Logistic Regression (LR)	Low risk: 0	0.72	0.77	0.76	0.76
		High risk: 1		0.64	0.65	0.65
6	K-Nearest Neighbour (KNN)	Low risk: 0	0.67	0.74	0.71	0.72
		High risk: 1		0.58	0.61	0.60



Summary of Machine Learning Model Results:
Using the default model without hyperparameters tuning

S/N	Model	Target	Accuracy Score	Precision	Recall	F1-Score
1	Random Forest Classifier (RF)	Low risk: 0	0.70	0.71	0.82	0.76
		High risk: 1		0.68	0.52	0.59
2	Decision Tree Classifier (DT)	Low risk: 0	0.61	0.65	0.75	0.70
		High risk: 1		0.67	0.55	0.60
3	K-Nearest Neighbour (KNN)	Low risk: 0	0.65	0.70	0.82	0.75
		High risk: 1		0.46	0.31	0.37
4	Support Vector Machine (SVM)	Low risk: 0	0.53	0.57	0.75	0.65
		High risk: 1		0.41	0.24	0.31
5	Naïve Bayes Classifier (NB)	Low risk: 0	0.75	0.58	0.34	0.43
		High risk: 1		0.78	0.90	0.84
6	Logistic Regression (LR)	Low risk: 0	0.83	0.43	0.16	0.23
		High risk: 1		0.85	0.95	0.90



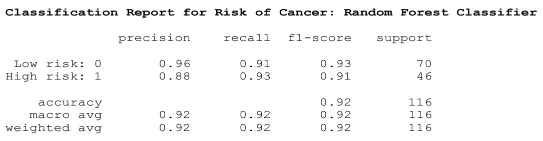


Random Search Hyperparameter Tuning

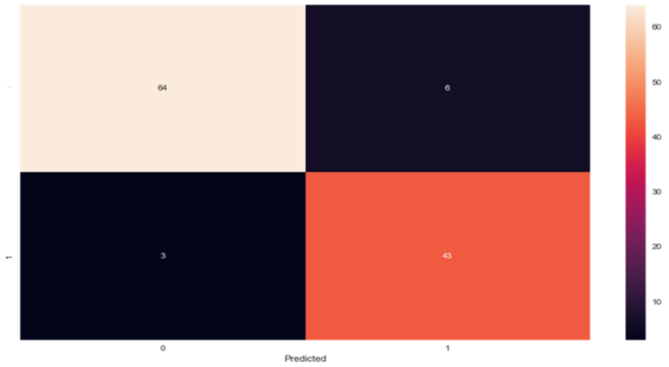
Hyparameters	**Declared dictionary of hyperparameters for tuning**	**Best Estimator Hyperparameters**
max_depth	[int(x) for x in np.linspace(10, 120, num = 12)]	120
min_samples_leaf	[1,2,4,6,8,9]	4
min samples split	[2, 6, 10]	6
bootstrap	[True, False]	FALSE
n estimators	[5,20,50,100]	50
max features	['auto', 'sqrt']	sqrt
criterion	["gini", "entropy"]	gini
random state	[42, 35, 10, 0]	35

Cross-validation scores:[0.87234043 0.82978723 0.80851064 0.78723404 0.82608696 0.82608696 0.86956522 0.84782609 0.82608696 0.82608696]

Cross val mean: 0.832 (std:0.024) Best Score: %s 0.931; Number of run was set at 10.



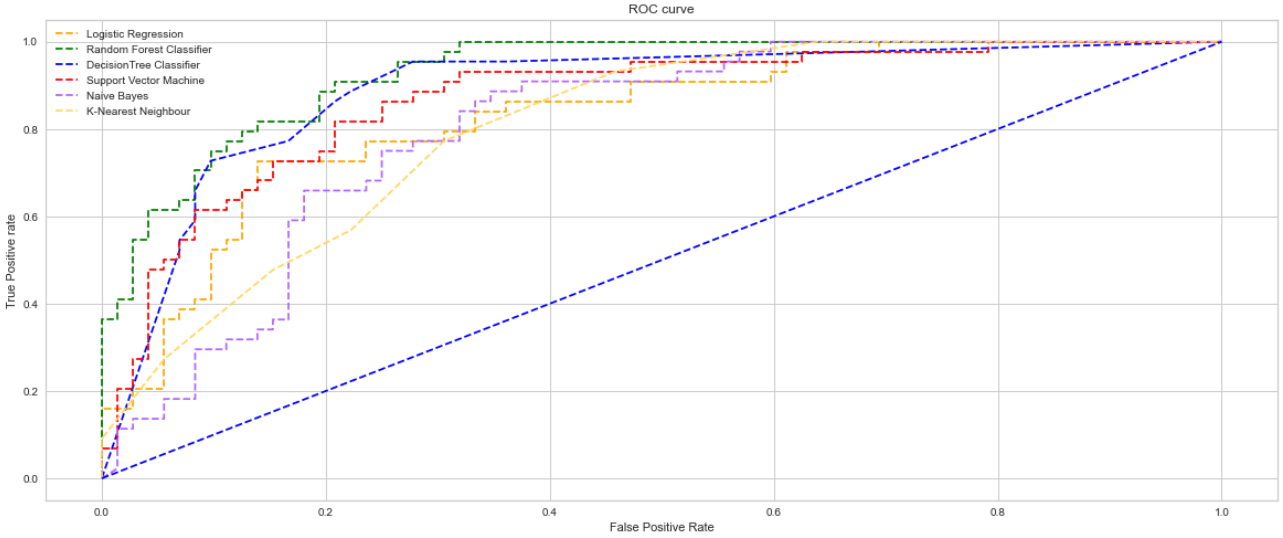
```
Classification Report for Risk of Cancer: Random Forest Classifier

              precision    recall  f1-score   support

  Low risk: 0       0.96      0.91      0.93        70
 High risk: 1       0.88      0.93      0.91        46

     accuracy                           0.92       116
    macro avg       0.92      0.92      0.92       116
 weighted avg       0.92      0.92      0.92       116
```



The AUC and ROC curve for the six models

The below summarises the AUC score for logistic regression, random forest classifier, decision tree classifier, support vector machine, naïve bayes and K-nearest neighbour respectively.

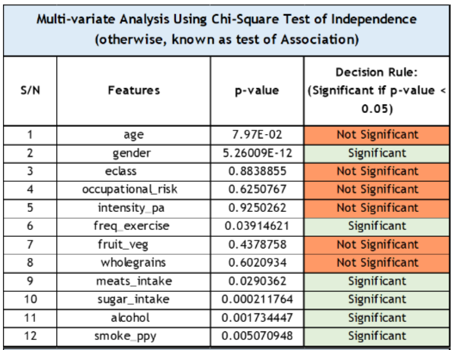
```
0.8295454545454546 0.9283459595959596 0.8890467171717171 0.8686868686868686
0.8012941919191918 0.8044507575757576
```



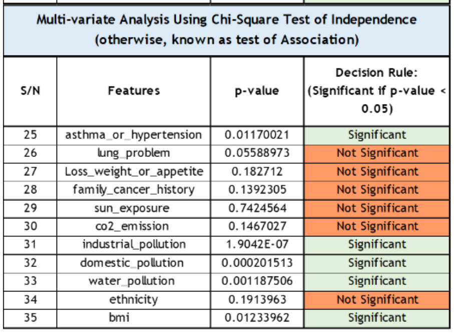
Cancer Risk Level: Clustering(C0-1) Matrix		
Features (Risk Factors)	C0- Low	C1-High
gender	🔵	🔴
industrial_pollution	🔵	🔴
domestic_pollution	🟡	🟡
sugar_intake	🟡	🟡
smoke_ppy	🔴	🔵
freq_cold	🔵	🔴
alcohol	🔵	🔴
co2_emission	🔵	🔴
freq_exercise	🔵	🟡
sun_exposure	🔵	🔴

Ordinal Scaling	Risk Class	Interpretation
0 - 4.5	Low	Many of the risk factors and comorbidities are low
4.6 - 10	High	Many of the risk factors and commobidities are very high

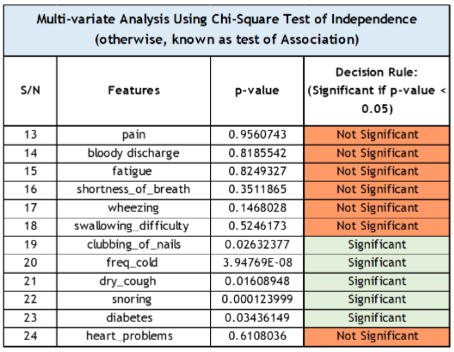
Colour	Legend
🔵	Low
🟡	Medium (Mix)
🔴	High



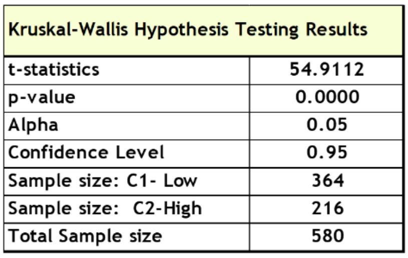
Multi-variate Analysis Using Chi-Square Test of Independence (otherwise, known as test of Association)			
S/N	Features	p-value	Decision Rule: (Significant if p-value < 0.05)
1	age	7.97E-02	Not Significant
2	gender	5.26009E-12	Significant
3	eclass	0.8838855	Not Significant
4	occupational_risk	0.6250767	Not Significant
5	intensity_pa	0.9250262	Not Significant
6	freq_exercise	0.03914621	Significant
7	fruit_veg	0.4378758	Not Significant
8	wholegrains	0.6020934	Not Significant
9	meats_intake	0.0290362	Significant
10	sugar_intake	0.000211764	Significant
11	alcohol	0.001734447	Significant
12	smoke_ppy	0.005070948	Significant



Multi-variate Analysis Using Chi-Square Test of Independence (otherwise, known as test of Association)			
S/N	Features	p-value	Decision Rule: (Significant if p-value < 0.05)
25	asthma_or_hypertension	0.01170021	Significant
26	lung_problem	0.05588973	Not Significant
27	Loss_weight_or_appetite	0.182712	Not Significant
28	family_cancer_history	0.1392305	Not Significant
29	sun_exposure	0.7424564	Not Significant
30	co2_emission	0.1467027	Not Significant
31	industrial_pollution	1.9042E-07	Significant
32	domestic_pollution	0.000201513	Significant
33	water_pollution	0.001187506	Significant
34	ethnicity	0.1913963	Not Significant
35	bmi	0.01233962	Significant

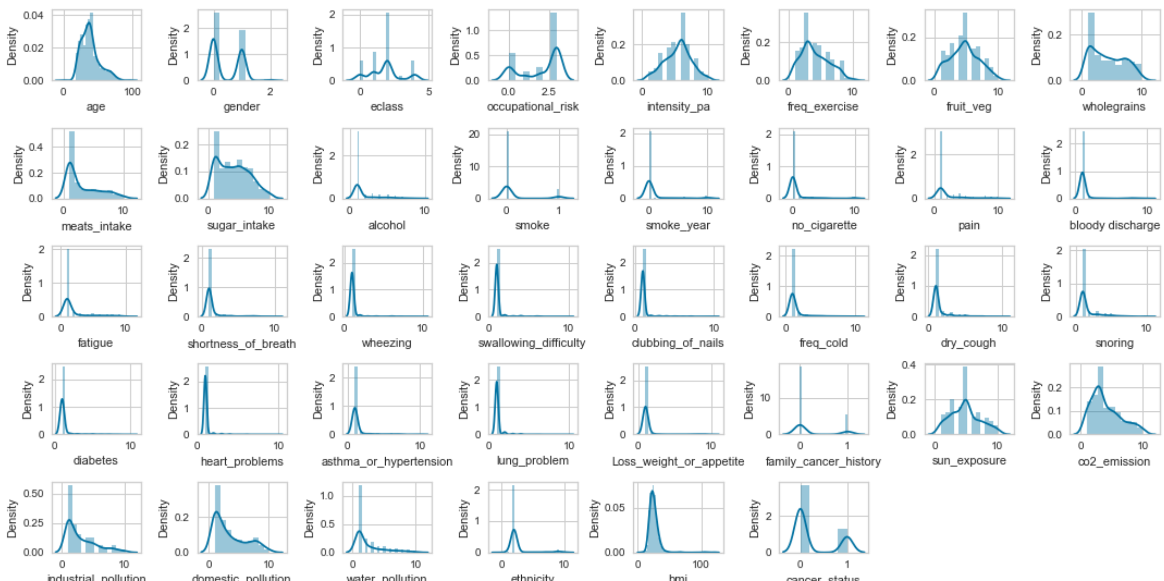


Multi-variate Analysis Using Chi-Square Test of Independence (otherwise, known as test of Association)			
S/N	Features	p-value	Decision Rule: (Significant if p-value < 0.05)
13	pain	0.9560743	Not Significant
14	bloody discharge	0.8185542	Not Significant
15	fatigue	0.8249327	Not Significant
16	shortness_of_breath	0.3511865	Not Significant
17	wheezing	0.1468028	Not Significant
18	swallowing_difficulty	0.5246173	Not Significant
19	clubbing_of_nails	0.02632377	Significant
20	freq_cold	3.94769E-08	Significant
21	dry_cough	0.01608948	Significant
22	snoring	0.000123999	Significant
23	diabetes	0.03436149	Significant
24	heart_problems	0.6108036	Not Significant

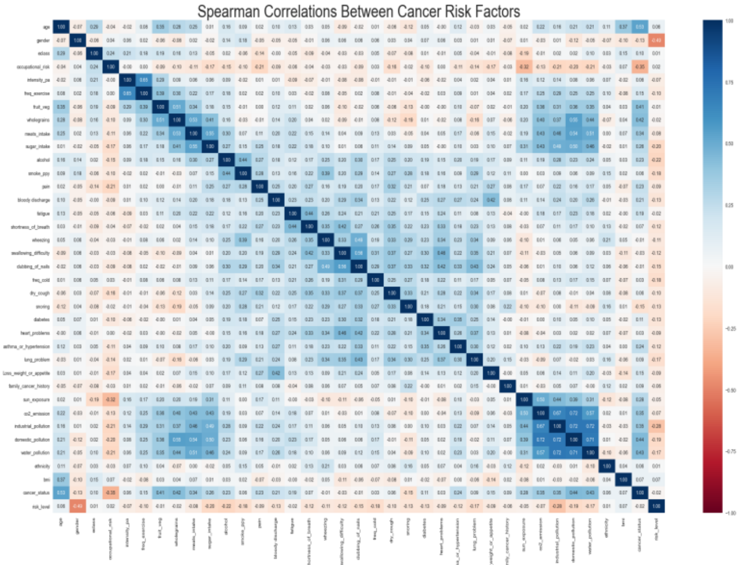


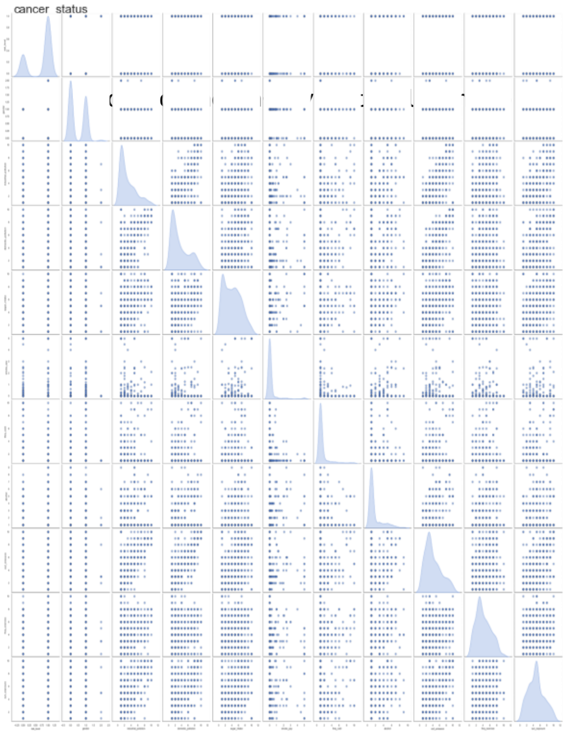
Kruskal-Wallis Hypothesis Testing Results	
t-statistics	54.9112
p-value	0.0000
Alpha	0.05
Confidence Level	0.95
Sample size: C1- Low	364
Sample size: C2-High	216
Total Sample size	580

Test of Normality: Cancer Risk Factors Distribution



Spearman Correlations Between Cancer Risk Factors





```
Optimization terminated successfully.
         Current function value: 0.497005
         Iterations 6
                           Results: Logit
====================================================================
Model:               Logit              Pseudo R-squared:  0.277
Dependent Variable:  risk_level         AIC:               598.5256
Date:                2022-09-09 05:06   BIC:               646.5190
No. Observations:    580                Log-Likelihood:    -288.26
Df Model:            10                 LL-Null:           -398.92
Df Residuals:        569                LLR p-value:       5.6921e-42
Converged:           1.0000             Scale:             1.0000
No. Iterations:      6.0000
--------------------------------------------------------------------
                      Coef.  Std.Err.    z    P>|z|   [0.025  0.975]
--------------------------------------------------------------------
const                 3.6259   0.3998  9.0693 0.0000  2.8423  4.4095
gender               -2.3209   0.2348 -9.8841 0.0000 -2.7811 -1.8607
freq_exercise        -0.1095   0.0499 -2.1928 0.0283 -0.2074 -0.0116
sugar_intake         -0.1398   0.0484 -2.8864 0.0039 -0.2348 -0.0449
alcohol              -0.0914   0.0690 -1.3247 0.1853 -0.2267  0.0438
smoke_ppy             0.0568   0.1432  0.3968 0.6915 -0.2238  0.3375
freq_cold            -0.3477   0.0912 -3.8134 0.0001 -0.5264 -0.1690
sun_exposure         -0.0050   0.0531 -0.0937 0.9254 -0.1090  0.0991
co2_emission          0.1755   0.0765  2.2926 0.0219  0.0255  0.3255
industrial_pollution -0.1804   0.0735 -2.4558 0.0141 -0.3244 -0.0364
domestic_pollution   -0.2247   0.0702 -3.1999 0.0014 -0.3623 -0.0871
====================================================================
```

Summary of Top Ten (10) Variable Interaction using Linear regression and corresponding R^2

```
Baseline R2: 0.187
Top 10 interactions: [('industrial_pollution', 'domestic_pollution', 0.234),
 ('sugar_intake', 'gender', 0.22), ('sugar_intake', 'smoke_ppy', 0.22), ('ge
nder', 'domestic_pollution', 0.216), ('sugar_intake', 'interaction', 0.212),
 ('industrial_pollution', 'gender', 0.209), ('sugar_intake', 'freq_exercise'
, 0.208), ('sugar_intake', 'domestic_pollution', 0.206), ('sun_exposure', 'f
req_exercise', 0.205), ('gender', 'co2_emission', 0.204)]
```

Fig 18a: Summary of the machine learning model results: Using the best hyperparameters.

Fig 18b: Summary of the machine learning model results: Without hyperparameters tuning.

Fig 18c: Cancer risk factors: Random Forest classifier features selection.

Fig 18d: Random search hyperparameter tuning.

Fig 18e: Confusion matrix classification report for Random Forest: Visuals and tabular representation.

Fig 18f: The AUC-ROC curve for the six machine learning models.

Fig 18g: Summary report for cluster analysis with K-Means carried out using Tableau to show number of clusters and patterns in each cluster.

Fig 18h: Multi-variate Analysis using the Chi-square test of independence and association.

Fig 18i: Kruskal-Wallis test of hypothesis.

Fig 18j: Univariate Analysis using density plot to test the normality of the dataset.

Fig 18k: Bivariate and Multi-variate Analysis using Pair-plot and Correlation Heatmap to show the normality/ pattern of each features, correlation between variables and the target variables respectively.

Fig 18l: The Logistic Regression Model Summary Report and Variable Interaction

CHAPTER FIVE

DISCUSSIONS

This section details the meanings, importance, and relevance of the study results.

The overall aim of this research project is to understand and establish how different risk factors impact the classification of the cancer risk level using machine learning. To achieve the objectives, six machine learning models were selected, this includes Random Forest Classifier(RF), Decision Tree Classifier(DT), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Naïve Bayes (NB) Classifier and Logistic Regression (LR). The result indicates that Random Forest model is the most suitable machine learning model with an overall accuracy and AUC score of 92% and 93% respectively. The RF model outperformed the other models because it leverages the power of multiple trees for making decisions. In addition, the ensemble tree-based model considers all the variables sequentially and take all interactions amongst the variables in making decision. Hence, both the RF and DT models proved to be the two most effective models for cancer risk prediction and classification. *The results support existing studies that tree-based models are most widely used machine learning models for decision support system in healthcare/medical setting* (Cruz 2022). On the other hand, the DT, SVM, NB, LR and KNN models show accuracy [85%, 83.%, 76%, 72% and 67%] and AUC scores of [89%, 87%, 80%, 83%, 80%] respectively. However, they are not effective to match the predictive performance of the RF classifier. The KNN and logistic regression models produced the least accuracy and AUC score respectively. These are obviously not the most effective model considering for the classification problem classification of cancer risk level.

5.1. Implementation of the Models

The variables selection to use for the models was carried out using five methods including: Random Forest Classifier, Chi-square test of independence, mutual information classifier, forward feature, and back feature elimination method. The top 10 feature selection from the Random Forest Classifier was selected as the most appropriate for the ML models because they combined to account for 77% of the variation in the target variable and coincidentally are included in the top 10 features from the Chi-square selection of the most significant features based on the p-value (fig. 18h). As shown in fig 18b, all six models were first executed with the default configuration which serve as a baseline before hyperparameters tuning. The essence of this is to check for any over or under fitting tendency in the model before applying the necessary tuning. The RF, LR and NB model produced a decent accuracy score for the test set of 0.70, 0.75 and 0.83 respectively. Whereas, DT, KNN and SVM showed a slight under-fitting tendency with score of 0.61, 0.55 and 0.63 respectively. Cross validation was carried out using the 10-Fold split to check the generalisation performance of the model as shown in fig 18a. The average score was 0.832 (std:0.024). This mean on the average the model is 83% accurate with a standard deviation of 0.024.

```
Kfold_split = KFold(n_splits=10, shuffle =False)
CV_score = -cross_val_score(dt_model, X_train, y_train, cv=Kfold_split)
```

Using RandomSearch hyperparameter tuning technique as shown in fig 18a, the best hyperparmeters combination from the declared values in the hyperparameter dictionary (max_depth = 120, min_samples_split = 6, min_samples_leaf =4, random_state =35, bootstrap = False, n_estimators =50, max_features = sqrt and criterion =gini). The RandomSearch hyperparameters retuned cross val mean: 0.832 (std:0.024) and suggested Best Score: %s 0.931. The RandomSearch hyperparameters tuning was utilised considering the computational cost attributed with the Grid Search. The best hyperparameter was then used in refitting the model and the Random Forest returned the best performance amongst the six models. The number of runs was set at 10 as shown in fig. 18a and the best score was selected.

5.2. Evaluation of the Models

The performance measurement(success metric) of the model was based on the AUC-ROC and RECALL (as shown in fig. 18a). The metrics give a vivid visualisation of how well the models performed on the basis of separating low(negative) and high (positive) risk of cancer. The higher the AUC-ROC score, the better the performance of the model in separating between positive and negative class. The Random Forest classifier produced the highest AUC score of 93%. RECALL depicts the sensitivity of the model in how many of the actual positive cases (Low and High Class) the models can predict correctly. Random Forest Classifier produced the highest recall score of 91% and 93% for the low and high risk respectively. The random forest model performance was evaluated based on the confusion matrix classification report in fig 18e. The classification results show a summary of the predictive results of the risk classifier as will now be discussed below:

- **Low Risk**: True positive(TP) was 64, False Negative(FN) was 6, False Positive(FP) was 3 and True Negative (TN) was 43. This implies a 96%, 91% and 93% for precision, recall and F1-score respectively.
- **High Risk**: True positive(TP) was 43, False Negative(FN) was 3, False Positive(FP) was 6 and True Negative (TN) was 64. This implies a 88%, 93% and 91% for precision, recall and F1-score respectively.

In the context of the research problem, Recall(Sensitivity of the model) is the most important success/evaluation metric. Recall are the actual positive cases the model can predict correctly which is 91% and 93% for the Low risk and High risk respectively. This is because False Negative(FN) - Type II error is of higher concern than False Positive(FP) - that is Type 1 error. Hence, True Positive(TP) should not go undetected for the two level of cancer risk. Despite the many advantages of the Random Forest classifier, the model has been labelled as a "Black box" because it there is no specific ways of interpreting the results of the model. This makes the model complex.

5.3. Validation of Hypothesis, Research Questions and Variable Interactions

In line with the hypothesis, the different risk factors are statistically significant in establishing and classifying a person's level of cancer risk. The random forest features importance in fig 18b validates that the top 10 important features combined explained approximately 77% variation in the target variable. In addition, the overall LRR p-value was 5.6921e-42 much less than the 0.05 (alpha: critical value). Therefore, we can reject the null hypothesis (H0) and accept the alternate hypothesis (H1) that the predictors (risk factors) are statistically significant. In addition, the linear regression was used to evaluate the model interaction show that the below interaction.

Top 10 interactions: [('industrial_pollution', 'domestic_pollution', 0.234), ('sugar_intake', 'gender', 0.22), ('sugar_intake', 'smoke_ppy', 0.22), ('gender', 'domestic_pollution', 0.216), ('sugar_intake', 'interaction', 0.212), ('industrial_pollution', 'gender', 0.209), ('sugar_intake', 'freq_exercise', 0.208), ('sugar_intake', 'domestic_pollution', 0.206), ('sun_exposure', 'freq_exercise', 0.205), ('gender', 'co2_emission', 0.204)]. There is a strong interaction between industrial-pollution and domestic_pollution; sugar_intake and gender; sugar_intake and smoke_ppy). These first three interactions resulted in slight improvement in the R2 of 23%, 22% and 22% respectively.

5.4. Main Findings

The Relationship between Chronic Medical Condition (Comorbidities) and Risk Factors

The cluster analysis results and visualisation dashboard in fig 18g provide new insights and contribution of the project in establishing the relationship between different risk factors, comorbidities, ethnicity, and level of cancer risk as detailed below:

- **Cluster C0**: Low risk – individuals in this cluster have many of the risk factors and comorbidities as relatively low.
- **Cluster C1**: High Risk – Individuals in this cluster have many of the risk factors and comorbidities as very high.

Previous research has overlooked the impact of existing medical conditions in people when predicting and classifying risk cancer. Most of these studies have applied machine learning on cancer risk prediction based on lifestyle and environmental. Although, this study shows that commodities have a modest relationship with other risk factors, however, long-term chronic/long-term medical conditions such as high blood pressure (hypertension), obesity, mental health problems, and kidney disease are ***common(shared) risk factors*** that affect people without cancer as well as people with cancer; lifestyle factors such as lack of exercise, and poor diet can aggravate the comorbidities. Hypertension in Men may be associated with the increased risk of developing prostate cancer. Similarly, hypertension in women can be linked to endometrial and breast cancer as well as renal cancer (Mohammed et al 2021).

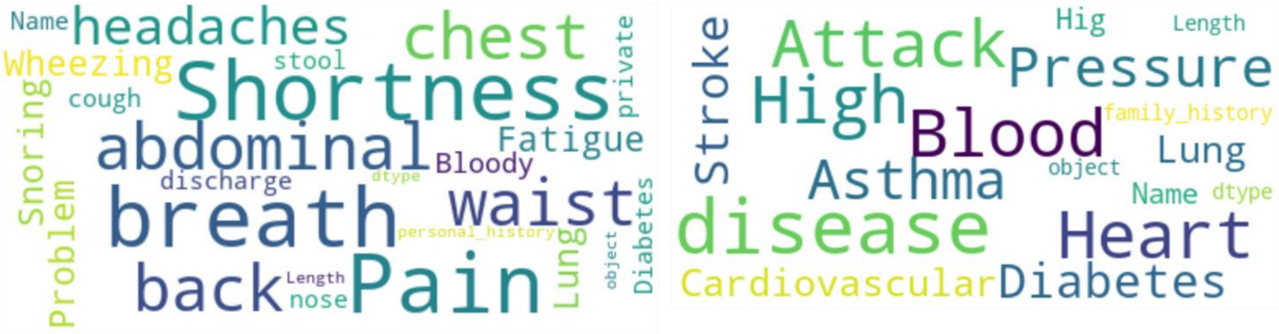


Fig 19: Word Cloud Visualisation of Cancer Risk Factors with Connection to Personal and Family History of Chronic Medical Conditions (Comorbidities)

The word cloud shown above, depicts the most frequent terms used by participants to describe their personal and family medical history. The most frequent word includes shortness of breath, pain in part of the body (back, abdominal, waist, chest, head), heart attack, high blood pressure, diabetes amongst others. The above insights validate underlying medical conditions of a person as key risk factors to consider as well as the interaction of other factors (lifestyle and environmental) when applying machine learning in predicting cancer risk level.

Insights from Data Analytics and Visualisation

The analytics and visualisation below show the prevalence of different risk factors based on gender, and ethnicity. The first figure shows that female has high outliers with regards to intake of alcohol, and the consumption of alcohol tends to decrease further down the gender class. However, the cancer risk level is higher in female than male with a count of 350 and 200 respectively. The remainder are attributed to participants who did not indicate their gender. The age bracket of 25-37 has the highest distribution of cancer risk. Interestingly, the ethnicity: Asian-Indian(0) has the highest class of high cancer risk attributed to exposure to co2 emission followed by Mixed: White & Black African and Mixed: White & Black Caribbean. Whereas, Asian- Pakistani[0], the Asian-Indian[1], Black African [2], White British [9] and White Irish[10] ethnicity group have relatively high risk class of cancer attributable to sun exposure.

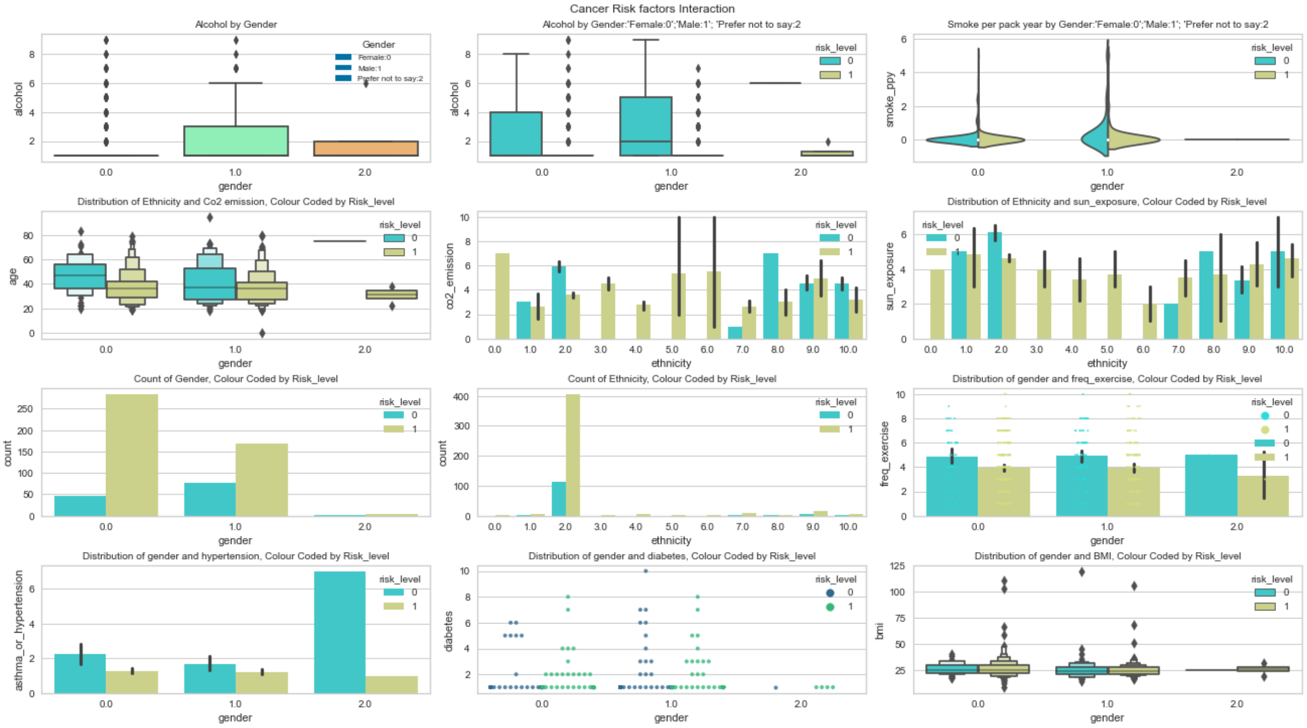


Fig 20: Analytics, Interaction and Visualisation of Cancer Risk factors: Developed by Author using Python

The research project support existing literature that the incidence and mortality of cancer can be reduced when the risk factors are better understood, detected, and treated on early; making early detection of risk vitally important. The insights and results from the research project provide practical solution to existing barriers to cancer screening resulting from fear or concerns about medical procedures, lack of knowledge of risk factors. Individuals, health professionals and the global health community can use the predictive model as a data-driven decision support and fact-finding system to recommend or take decision for cancer screening and diagnosis.

CHAPTER SIX

Conclusion and Recommendations for Future Work

This section summarizes the insights of the research, main findings, limitations, and recommendations for future work.

5.1 Conclusion

Admittedly, machine learning models have been confirmed by previous studies to have significant impact in solving real world problems especially in the healthcare domain. This research aimed to establish the effect of different risk factors on the level of cancer risk. Based on quantitative analysis, it can be concluded that the different risk factors when combined with co-existing diseases in a person are statistically significant in prediction and classification of cancer risk level. Random Forest Classifier is the most suitable decision support system model based on its capabilities to leverage the power of multiple decision trees and classify the individuals with high risk of cancer with the best accuracy, precision and recall score without leaving out any false negative undetected. The results validate that people with high risk of cancer have many risk factors, symptoms, and co-existing diseases all HIGH simultaneously and the model was sensitive significantly in detecting them.

Based on these conclusions, individuals, health professionals and the global health community can enhance capability in spotting the risk of cancer early by adopting the prediction system and deploy it at the comfort of their homes and health facilities as a preliminary fact-finding system for cancer risk assessment and vulnerability. Hence, data-driven decisions can be made for cancer screening and diagnosis for different types of cancer based on the results from the prediction system. Early detection can save lives, reduce cancer incidence and mortality, and huge associated cost with the treatment of the disease.

5.2 . Future Work

This research clearly demonstrates the effect and significance of the different risk factors on level of cancer risk and corresponding general characteristics of the target population such as ethnicities, country of origin, country of residence amongst others. This provided control on the sampling methods, measurements and how the data was generated which in turns enhance the systematic description of the population and ability to generate reproducible knowledge. However, it only focused on binary classification without much information on the threshold of the risk class. This limits the work in certain aspects.

Furthermore, this research is focused on predicting and classifying the level of cancer risk and does not predict the actual type of cancer. Although the research made significant contributions in the identification of different risk factors which are the major influencers of the many types of cancer including lungs, breast, breast cancers and many more. Improving on this aspect will provide for future work in predicting cancer types. In terms of the sample size, the target was to obtain 1000 dataset for the research, however, only 580 dataset was gathered which is another limitation of the study. Hence, further research can be done with primary data source with relatively larger dataset than 580 and better threshold for each of the risk class.

Data Availability Statement

The primary data used for this study was obtained through a survey questionnaire. Participants responses were anonymized, and personal identifier information was collected.

Ethical Statement

Participants gave their consent to participate in the survey and were given the opportunity to withdraw from the survey at any time, without reason.

Ethics Approval

Ethical approval was applied for and granted for this research. However, the study did not engage less privileged, physically challenged individuals nor animal subjects.

REFERENCES

References

ACHILONU O.J, et al., 2021. Predicting colorectal cancer recurrence and patient survival using supervised machine learning approach: a South African population-based study. Front Public Health, 9:694306. doi: 10.3389/fpubh.2021.694306 [viewed 22 August 2022]. Available from: https://pubmed.ncbi.nlm.nih.gov/34307286/

ANAND et al.,2008. Cancer is a preventable disease that requires major lifestyle changes. *Pharmaceutical research. 25(9): 2097–2116* [viewed 20 August 2022]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2515569/

ALFAYEZ, A. et al., 2021. Predicting the risk of cancer in adults using supervised machine learning: A scoping review. *BMJ Open, 11(9).* BMJ Publishing Group [viewed 29 May 2022]. Available from: https://doi.org/10.1136/bmjopen-2020-047755

ALI M.M, et al., 2021. Machine learning-based statistical analysis for early-stage detection of cervical cancer. *Comput Biol Med. 139:104985. doi: 10.1016/j.compbiomed.2021.104985*[viewed 10 August 2022]. Available from: https://pubmed.ncbi.nlm.nih.gov/34735942/

AMERICAN SOCIETY OF CLINICAL ONCOLOGY (ASCO); 2022. *Understanding Cancer Risk*. [Viewed 29 May 2022]. Available from: https://www.Cancer.Net/Navigating-Cancer-Care/Prevention-and-Healthy-Living/Understanding-Cancer-Risk. https://www.cancer.net/navigating-cancer-care/prevention-and-healthy-living/understanding-cancer-risk

AVISHEK, C., 2021. *Predicting cancer using supervised machine learning: Mesothelioma* [viewed 29 May 2022]. Available from: *https://Pubmed.Ncbi.Nlm.Nih.Gov/32568137/*, *29*(1), 45–48.

BOSSERT et al., 2021. Lung cancer patients' comorbidities and attendance of German ambulatory physicians in a 5-year cross-sectional study [Viewed 13 June 2022]. NPJ Prim Care Respir Med, 31(2). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7844218/

BUNDASAK, S; 2016. Analysis of Cancer Risk System Using Decision Tree. Conference: *International Conference on Business and Industrial Research 2016*[viewed 4 June 2022].

Available from: https://www.researchgate.net/publication/342159102_Analysis_of_Cancer_Risk_System_Using_Decision_Tree.

CHOUDHURY A, 2021. Predicting cancer using supervised machine learning: *Mesothelioma. Technol Health Care.29(1):45-58. doi: 10.3233/THC-202237. PMID: 32568137.* [viewed 15 August 2022]. Available from: https://pubmed.ncbi.nlm.nih.gov/32568137/

CRUZ, A.J; AND D.S. WISHART, 2006. Applications of machine learning in cancer prediction and prognosis. *Cancer informatics: sage journal,1(2)* [viewed 4 August 2022]. Available from: https://journals.sagepub.com/doi/10.1177/117693510600200030.

CENTERS FOR DISEASE CONTROL AND PREVENTION; 2021. *Cancer Data and Statistics* [Viewed 3 June 2022]. Available from: https://www.Cdc.Gov/Cancer/Dcpc/Data/Index.Htm.

DATA SCIENCE PROCESS ALLIANCE, 2022. What is CRISP DM? [viewed 10 April 2022]. Available from: https://www.datascience-pm.com/crisp-dm-2/

DUAN S, et al., 2020. Development of a machine learning-based multimode diagnosis system for lung cancer. Aging (Albany NY). 23;12(10):9840-9854. doi: 10.18632/aging.103249 [viewed 15 August 2022]. Available from: https://pubmed.ncbi.nlm.nih.gov/32445550/

GHANIZADA et al., 2020. The impact of comorbidities on survival in oral cancer patients: a population-based, case-control study. *Acta Oncologica, 60(2), 173–179.* [viewed 13 June 2022] Available from: https://www.tandfonline.com/doi/full/10.1080/0284186X.2020.1836393.

GOV.UK., 2022. *Closed consultation: 10-Year Cancer Plan: Call for Evidence* [Viewed on 29 May 2022]. Available from: https://www.Gov.Uk/Government/Consultations/10-Year-Cancer-Plan-Call-for-Evidence.

HAMDI Y; 2021. *Cancer in Africa: The Untold Story. Frontiers in Oncology: Cancer Epidemiology and Prevention* [Viewed 3 June 2022].Available from: https://www.frontiersin.org/articles/10.3389/fonc.2021.650117/full.

INSTITUTE FOR WORK AND HEALTH 2015. Primary data and secondary data[Viewed 3 September 2022). Available from: https://www.iwh.on.ca/what-researchers-mean-by/primary-data-and-secondary-data

KABIRAJ, S. *et al*., (2020). Breast cancer risk prediction using XGBoost and Random Forest Algorithm. *11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1-4, doi: 10.1109/ICCCNT49239.2020.9225451[viewed 6 August 2022]. Available from: https://ieeexplore.ieee.org/document/9225451

LEUNG, W.K et al., 2021. Applications of machine learning models in the prediction of gastric cancer risk in patients after helicobacter pylori eradication. *Alimentary pharmacology and therapeutics*, 53(8), 864-872 [viewed 15 August 2022]. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/apt.16272

Mohammed K, et al., 2021. The risk of chronic diseases in Individuals responding to a measure for the initial screening of depression and reported feelings of being down, depressed, or opeless. *Cureus 13(9):e17634. doi: 10.7759/cureus.17634. PMID: 34646682; PMCID: PMC8486358.* [viewed 14 September 2022] Available from: https://pubmed.ncbi.nlm.nih.gov/34646682/

MACMILLAN CANCER SUPPORT, 2018. Causes and risk factors of cancer [viewed on 3 August 2022). Available from: https://www.macmillan.org.uk/cancer-information-and-support/worried-about-cancer/causes-and-risk-factors

MONCADA-TORRES A, et al., 2021. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Sci Rep. 26;11(1):6968. doi: 10.1038/s41598-021-86327-7* [viewed on 29 July 2022]. Available from: https://pubmed.ncbi.nlm.nih.gov/33772109/

MULLER, A.C., and S. GUIDO, 2016. *Introduction to machine learning with python.*1st ed. Sebastopol: O'Reilly Media, Inc.

NAFIZATUS, S., AND Z. RUSTAM, 2019. Naïve Bayes model for predicting the colon cancer. IOP Conf. Ser. Mater. Sci. Eng. *546 052068* [Viewed on 24 August 2022]. Available from: https://iopscience.iop.org/article/10.1088/1757-899X/546/5/052068.

OPEN DATA SCIENCE CONFERENCE(2019): 15 open datasets for healthcare[viewed 15 June 2022]. Available from: https://odsc.medium.com/15-open-datasets-for-healthcare-830b19980d9.

PUBMED 2022. Review of machine learning models [Viewed 4 September 2022]. Available from: https://pubmed.ncbi.nlm.nih.gov/

SCRIBBR 2022. Research methods: definitions, types and examples [viewed 15 September 2022]. Available from: https://www.scribbr.co.uk/category/research-methods/

STARK, G.F et al., 2019. Predicting breast cancer risk using personal health data and machine learning models. *PLoS ONE 14(12): e0226765. https://doi.org/10.1371/journal.pone.0226765*[viewed 29 July 2022]. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0226765#pone-0226765-t004.

TU et al., 2018. Cancer risk associated with chronic diseases and disease markers: prospective cohort study. BMJ 2018; 360: k134. [viewed 3 August 2022]. Available from: https://www.bmj.com/content/360/bmj.k134.

WORLD HEALTH ORGANIZATION., 2022. Cancer [viewed 3 February 2022]. Available from: https://www.Who.Int/News-Room/Fact-Sheets/Detail/Cancer.

PIVA F, et al., 2021. Predicting future cancer burden in the United States by artificial neural networks. Future Oncol;17(2):159-168. doi: 10.2217/fon-2020-0359 [viewed 12 July 2022]. Available from: https://pubmed.ncbi.nlm.nih.gov/33305617/.

V7LABS, 2022. Supervised and unsupervised learning: differences and examples [viewed 16 August 2022]. Available from: https://www.v7labs.com/

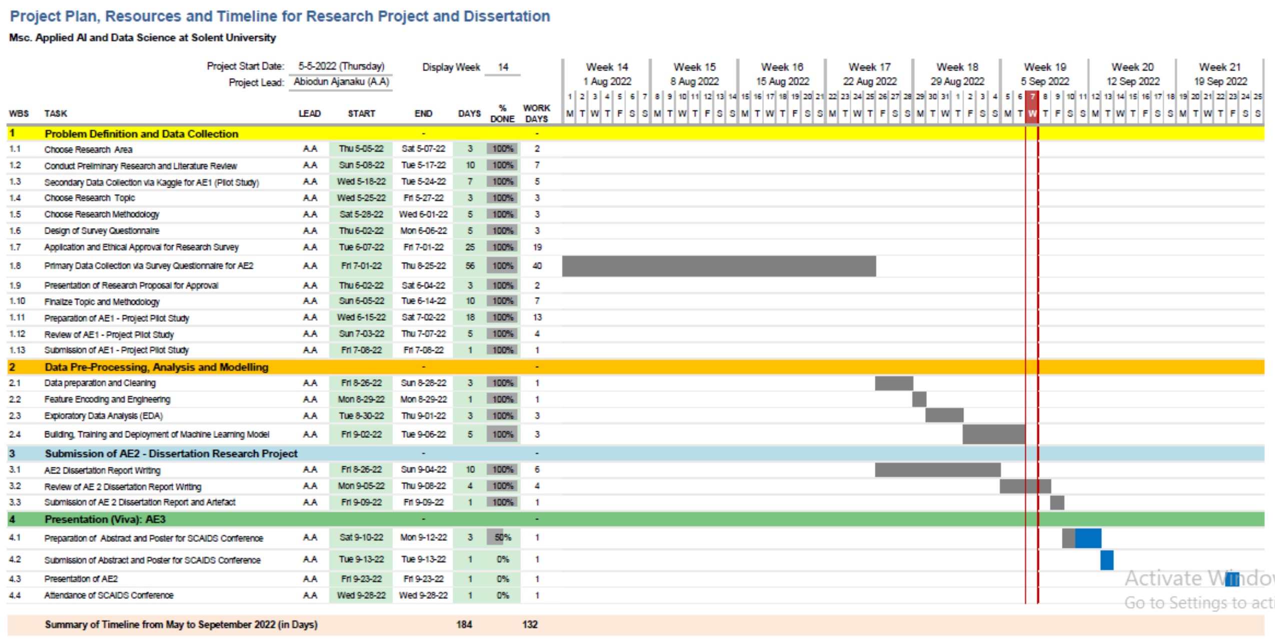
WORLD HEALTH ORGANIZATION REGIONAL OFFICE FOR THE WESTERN PACIFIC REGION, 2009. Pacific Physical Activity Guidelines for Adults: Framework for Accelerating the Communication of Physical Activity Guidelines [Viewed 2 June 2022]. Available from: https://www.change4health.gov.hk/en/physical_activity/facts/classification/index.html

WORLD CANCER RESEARCH FUND., 2022. UK cancer statistics [viewed 29 May 2022]. Available from: https://www.Wcrf-Uk.Org/Preventing-Cancer/Uk-Cancer-Statistics/?Gclid=CjwKCAjws8yUBhA1EiwAi_tpEQ4UhrBBhO-FVdmHxPohNxcrO5NL9dcddL8EtfoGdZkMUxC9n3l-XhoCXe8QAvD_BwE. https://www.wcrf-uk.org/preventing-cancer/uk-cancer-statistics/?gclid=CjwKCAjws8yUBhA1EiwAi_tpEQ4UhrBBhO-FVdmHxPohNxcrO5NL9dcddL8EtfoGdZkMUxC9n3l-XhoCXe8QAvD_BwE

Ye X, et al., 2019. Ensemble feature learning to identify risk factors for predicting secondary cancer. *Int J Med Sci. 16(7):949-959. doi: 10.7150/ijms.33820.* [viewed 18 August 2022]. Available from: https://pubmed.ncbi.nlm.nih.gov/31341408/

APPENDIX

Appendix 1: Project Plan and Resources: Research Project & Dissertation Timeline



Project Plan, Resources and Timeline for Research Project and Dissertation

Msc. Applied AI and Data Science at Solent University

Project Start Date: 5-5-2022 (Thursday)   Display Week 14

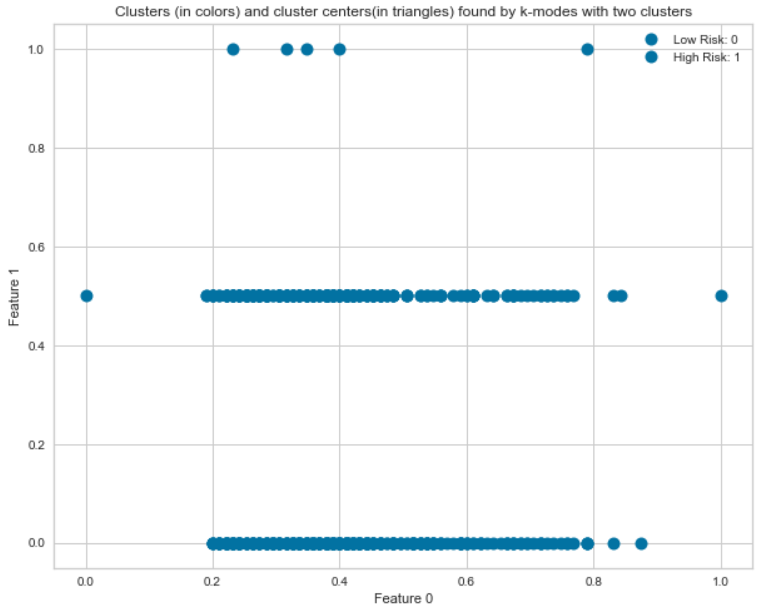
Project Lead: Abiodun Ajanaku (A.A)

Week 14 1 Aug 2022 | Week 15 8 Aug 2022 | Week 16 15 Aug 2022 | Week 17 22 Aug 2022 | Week 18 29 Aug 2022 | Week 19 5 Sep 2022 | Week 20 12 Sep 2022 | Week 21 19 Sep 2022

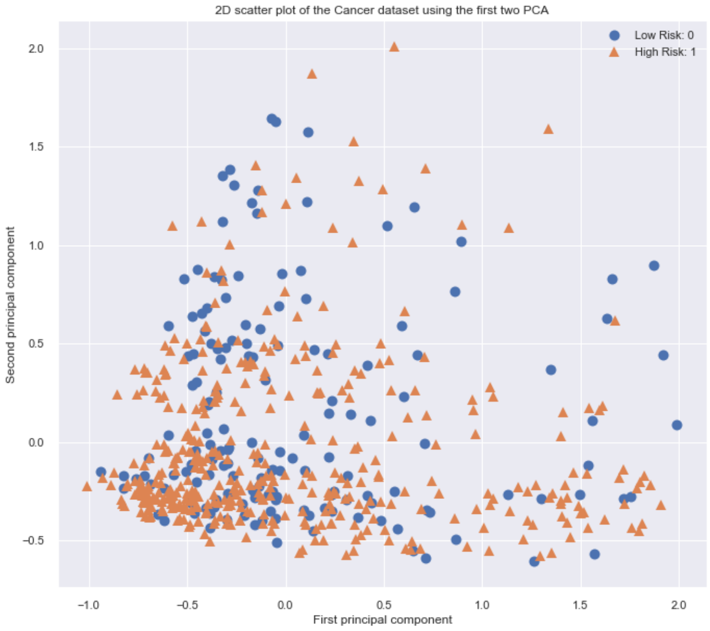
WBS	TASK	LEAD	START	END	DAYS	% DONE	WORK DAYS
1	**Problem Definition and Data Collection**				-		-
1.1	Choose Research Area	A.A	Thu 5-05-22	Sat 5-07-22	3	100%	2
1.2	Conduct Preliminary Research and Literature Review	A.A	Sun 5-08-22	Tue 5-17-22	10	100%	7
1.3	Secondary Data Collection via Kaggle for AE1 (Pilot Study)	A.A	Wed 5-18-22	Tue 5-24-22	7	100%	5
1.4	Choose Research Topic	A.A	Wed 5-25-22	Fri 5-27-22	3	100%	3
1.5	Choose Research Methodology	A.A	Sat 5-28-22	Wed 6-01-22	5	100%	3
1.6	Design of Survey Questionnaire	A.A	Thu 6-02-22	Mon 6-06-22	5	100%	3
1.7	Application and Ethical Approval for Research Survey	A.A	Tue 6-07-22	Fri 7-01-22	25	100%	19
1.8	Primary Data Collection via Survey Questionnaire for AE2	A.A	Fri 7-01-22	Thu 8-25-22	56	100%	40
1.9	Presentation of Research Proposal for Approval	A.A	Thu 6-02-22	Sat 6-04-22	3	100%	2
1.10	Finalize Topic and Methodology	A.A	Sun 6-05-22	Tue 6-14-22	10	100%	7
1.11	Preparation of AE1 - Project Pilot Study	A.A	Wed 6-15-22	Sat 7-02-22	18	100%	13
1.12	Review of AE1 - Project Pilot Study	A.A	Sun 7-03-22	Thu 7-07-22	5	100%	4
1.13	Submission of AE1 - Project Pilot Study	A.A	Fri 7-08-22	Fri 7-08-22	1	100%	1
2	**Data Pre-Processing, Analysis and Modelling**				-		-
2.1	Data preparation and Cleaning	A.A	Fri 8-26-22	Sun 8-28-22	3	100%	1
2.2	Feature Encoding and Engineering	A.A	Mon 8-29-22	Mon 8-29-22	1	100%	1
2.3	Exploratory Data Analysis (EDA)	A.A	Tue 8-30-22	Thu 9-01-22	3	100%	3
2.4	Building, Training and Deployment of Machine Learning Model	A.A	Fri 9-02-22	Tue 9-06-22	5	100%	3
3	**Submission of AE2 - Dissertation Research Project**				-		-
3.1	AE2 Dissertation Report Writing	A.A	Fri 8-26-22	Sun 9-04-22	10	100%	6
3.2	Review of AE 2 Dissertation Report Writing	A.A	Mon 9-05-22	Thu 9-08-22	4	100%	4
3.3	Submission of AE 2 Dissertation Report and Artefact	A.A	Fri 9-09-22	Fri 9-09-22	1	100%	1
4	**Presentation (Viva): AE3**				-		-
4.1	Preparation of Abstract and Poster for SCAIDS Conference	A.A	Sat 9-10-22	Mon 9-12-22	3	50%	1
4.2	Submission of Abstract and Poster for SCAIDS Conference	A.A	Tue 9-13-22	Tue 9-13-22	1	0%	1
4.3	Presentation of AE2	A.A	Fri 9-23-22	Fri 9-23-22	1	0%	1
4.4	Attendance of SCAIDS Conference	A.A	Wed 9-28-22	Wed 9-28-22	1	0%	1
	Summary of Timeline from May to Sepetember 2022 (in Days)				184		132

Activate Windows
Go to Settings to activ

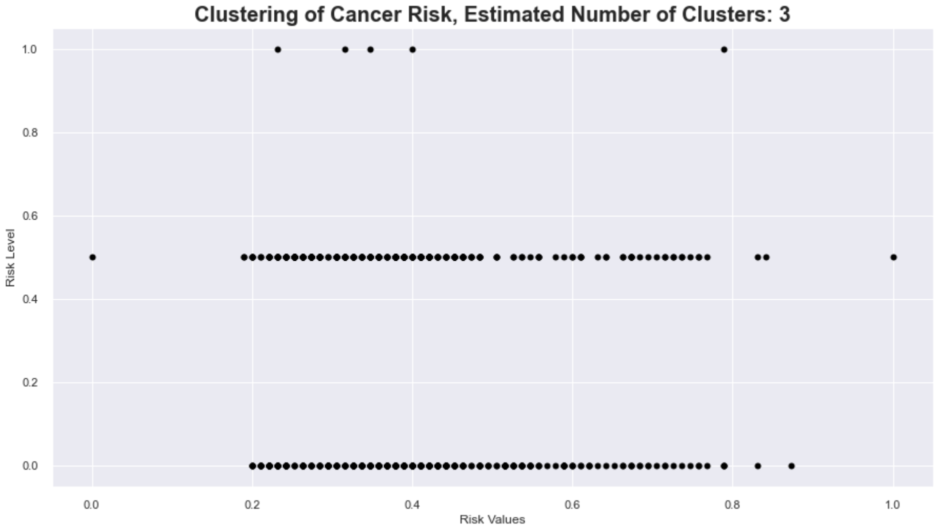
Appendix 2: Clustering Analysis using KMEANS and KMODES



Appendix 3: Principal Component Analysis (PCA)



Appendix 4: Clustering Analysis using DBSCAN



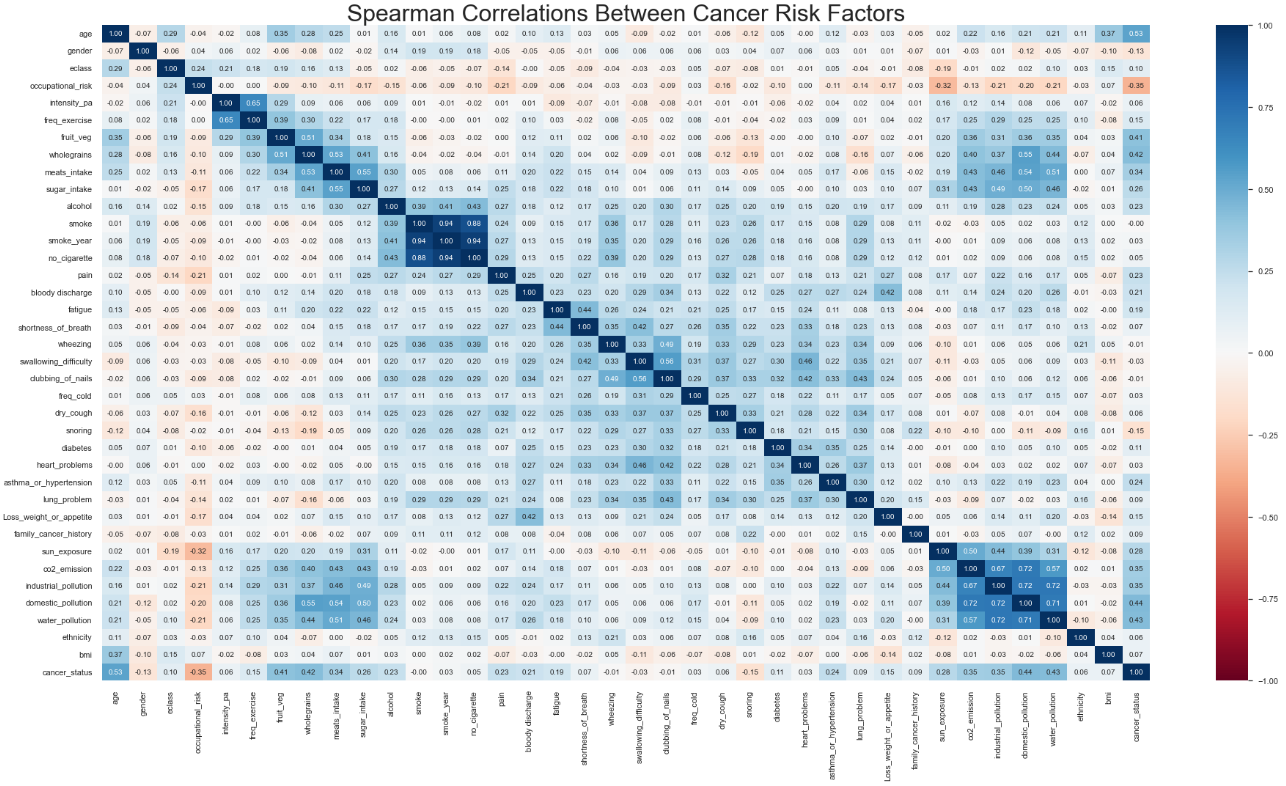
Appendix 5 : Code Snippet for Random Forest (RF), Support Vector Machine(SVM), DecisionTree(DT), Logistic Regression (LR), Decision Tree Classifier (DT), K-Nearest Neighbour (KNN), Naïve Bayes (NB),

```
 #Declaring a seed variable and set to 1 to ensure reproductibility
  SEED =35
```

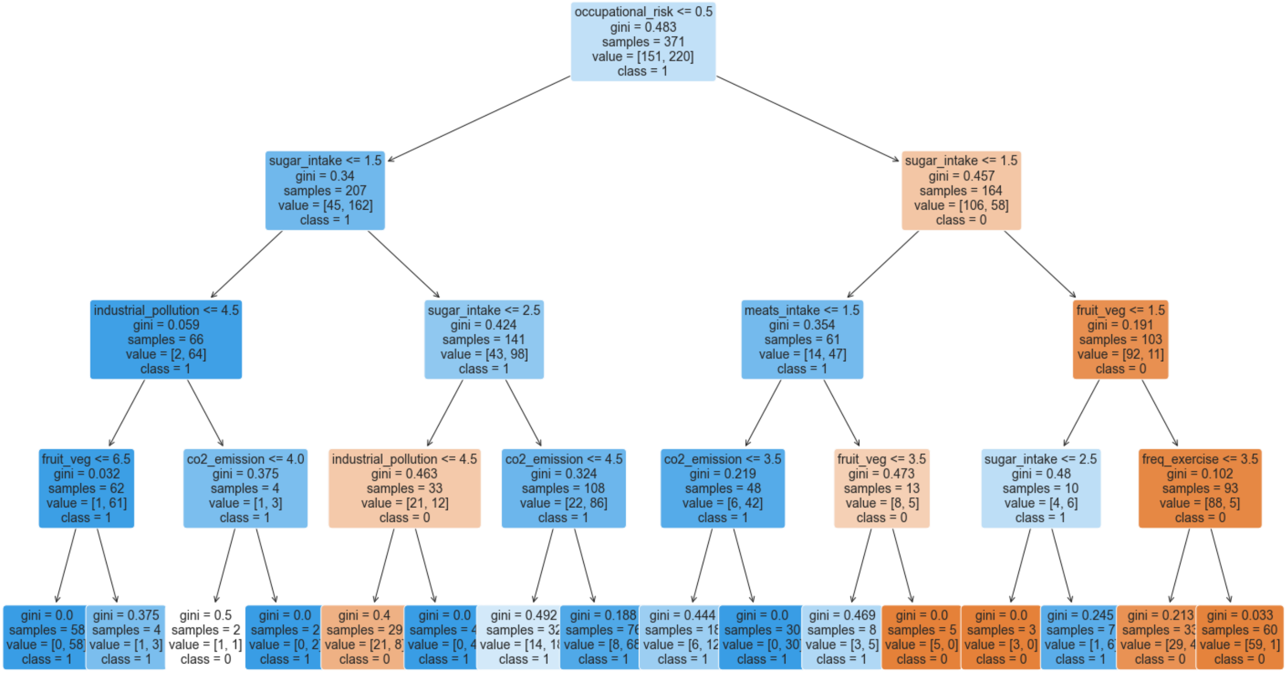
#splitting the dataset into training and test split of 80:20. The testing set will be used for testing and generalisation performance of
#the model. While the training set will be split further into train-validation set to build and train the model

```
train_features, X_test, train_target, y_test = train_test_split(X, y, test_size = 0.20,
random_state=SEED)
# splitting the training set into training and validation for model building and training
X_train, X_val, y_train, y_val = train_test_split(train_features, train_target, test_size = 0.20,
random_state=SEED)

#Instantiate the models
rf_model = RandomForestClassifier(max_depth =120, min_samples_leaf=4, min_samples_split =
6,
                                   random_state =SEED, criterion = 'gini', max_features='sqrt',
                                  n_estimators=50, bootstrap=False)
dt_model = DecisionTreeClassifier(max_depth =120, min_samples_leaf=4, min_samples_split = 6,
                                   random_state =SEED, criterion = 'gini', max_features='sqrt',
                                  )
svm_model =svm.SVC(C= 10, gamma= 'scale', kernel= 'rbf')
lr_model= LogisticRegression()
nb_model = GaussianNB()
knn_model = KNeighborsClassifier(n_neighbors=7)
```

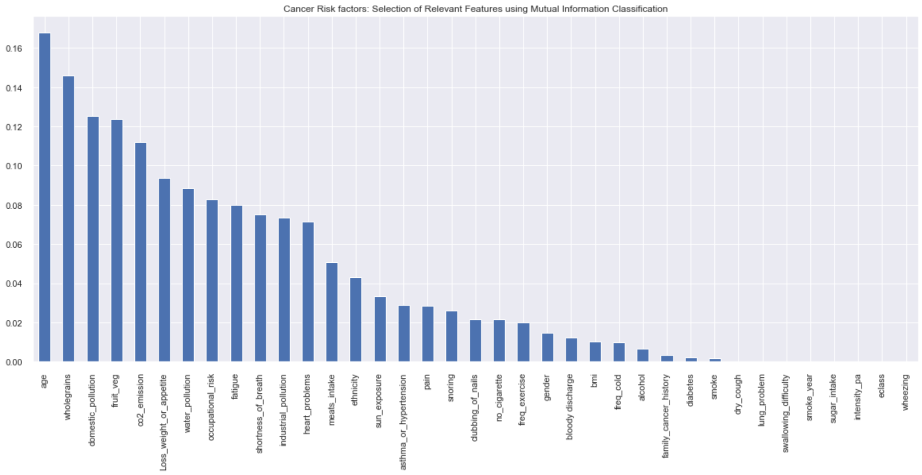
Appendix 6: Clearer Version of the Spearman Correlations Heatman



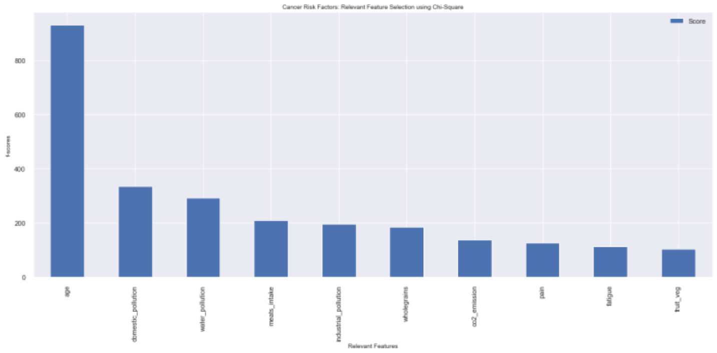
Appendix 6: Decision Tree Decision Making Modelling



Appendix 7: Mutual Information Features Selection



Appendix 8: Chi-square Feature Selection



Appendix 9: Descriptive Statistics of the Variables (Features): Measures of central tendency, dispersion, Correlation, and normality

Skewness

```
 heart_problems             8.661097
lung_problem               7.622609
swallowing_difficulty      7.166333
wheezing                   6.610267
clubbing_of_nails          6.521595
bmi                        5.499917
Loss_weight_or_appetite    5.086085
diabetes                   4.970369
bloody discharge           4.867688
asthma_or_hypertension     4.171486
shortness_of_breath        4.017216
no_cigarette               3.777947
freq_cold                  3.509530
dry_cough                  3.399034
ethnicity                  3.259871
smoke_year                 3.076489
snoring                    2.630502
smoke                      2.261777
fatigue                    2.191886
pain                       2.115719
alcohol                    2.074317
water_pollution            1.608816
family_cancer_history      1.279559
industrial_pollution       1.176916
cancer_status              1.074772
meats_intake               1.011924
age                        0.933836
domestic_pollution         0.845174
co2_emission               0.760033
freq_exercise              0.488558
gender                     0.434977
sugar_intake               0.366492
eclass                     0.329966
wholegrains                0.311987
sun_exposure               0.311381
fruit_veg                  0.147580
intensity_pa              -0.136818
occupational_risk         -0.757916
dtype: float64
```

correlation

```
                           age  gender  eclass  occupational_risk  \
age                       1.00   -0.04    0.26              -0.12
gender                   -0.04    1.00   -0.06               0.03
eclass                    0.26   -0.06    1.00               0.16
occupational_risk        -0.12    0.03    0.16               1.00
intensity_pa             -0.05    0.04    0.22               0.01
freq_exercise             0.05    0.02    0.16              -0.01
fruit_veg                 0.36   -0.07    0.18              -0.13
wholegrains               0.29   -0.09    0.15              -0.16
meats_intake              0.27    0.01    0.13              -0.16
sugar_intake              0.03   -0.03   -0.04              -0.15
alcohol                   0.17    0.11    0.04              -0.13
```

```
smoke                    0.02    0.18   -0.05              -0.04
smoke_year               0.15    0.17   -0.03              -0.09
no_cigarette             0.13    0.12   -0.05              -0.09
pain                     0.14   -0.05   -0.09              -0.16
bloody discharge         0.17   -0.07    0.05              -0.09
fatigue                  0.17   -0.09   -0.06              -0.13
shortness_of_breath      0.13   -0.02   -0.08              -0.05
wheezing                 0.04    0.07   -0.03              -0.01
swallowing_difficulty   -0.04    0.05   -0.01               0.01
clubbing_of_nails        0.03    0.06   -0.03              -0.04
freq_cold                0.13    0.08    0.09               0.03
dry_cough                0.01    0.05   -0.09              -0.13
snoring                 -0.06    0.04   -0.09               0.06
diabetes                 0.14    0.07    0.06              -0.07
heart_problems           0.03    0.08   -0.02               0.05
asthma_or_hypertension   0.24    0.01    0.10              -0.10
lung_problem            -0.04    0.05   -0.03              -0.05
Loss_weight_or_appetite  0.12   -0.09    0.02              -0.12
family_cancer_history   -0.04   -0.07   -0.07              -0.01
sun_exposure             0.07    0.01   -0.16              -0.30
co2_emission             0.26   -0.04    0.00              -0.18
industrial_pollution     0.26   -0.02    0.07              -0.23
domestic_pollution       0.31   -0.14    0.06              -0.26
water_pollution          0.33   -0.06    0.12              -0.21
ethnicity                0.08   -0.08    0.00              -0.03
bmi                      0.11   -0.06    0.07               0.04
cancer_status            0.58   -0.12    0.11              -0.36

                         intensity_pa  freq_exercise  fruit_veg  wholegrains
  \
age                             -0.05           0.05       0.36         0.29

gender                           0.04           0.02      -0.07        -0.09

eclass                           0.22           0.16       0.18         0.15

occupational_risk                0.01          -0.01      -0.13        -0.16

intensity_pa                     1.00           0.64       0.28         0.10

freq_exercise                    0.64           1.00       0.37         0.30

fruit_veg                        0.28           0.37       1.00         0.52

wholegrains                      0.10           0.30       0.52         1.00

meats_intake                     0.10           0.22       0.37         0.52

sugar_intake                     0.06           0.15       0.19         0.41

alcohol                          0.11           0.19       0.18         0.21

smoke                           -0.01          -0.01      -0.06        -0.05

smoke_year                      -0.07          -0.04       0.01        -0.00

no_cigarette                    -0.08           0.00      -0.01        -0.00
```

```
pain                             0.07           0.08       0.10         0.13
bloody discharge                 0.10           0.15       0.19         0.26
fatigue                         -0.06           0.05       0.18         0.28
shortness_of_breath             -0.05           0.02       0.09         0.13
wheezing                         0.01           0.08       0.06         0.03
swallowing_difficulty           -0.04          -0.02      -0.06        -0.05
clubbing_of_nails               -0.02           0.04       0.05         0.04
freq_cold                        0.02           0.09       0.09         0.15
dry_cough                       -0.01           0.01      -0.01        -0.03
snoring                         -0.00          -0.02      -0.05        -0.15
diabetes                        -0.02           0.03       0.04         0.06
heart_problems                   0.04           0.06       0.05        -0.01
asthma_or_hypertension           0.08           0.10       0.17         0.16
lung_problem                     0.05           0.02      -0.03        -0.12
Loss_weight_or_appetite          0.14           0.12       0.12         0.21
family_cancer_history           -0.00           0.02      -0.01        -0.06
sun_exposure                     0.18           0.19       0.21         0.26
co2_emission                     0.12           0.25       0.36         0.45
industrial_pollution             0.18           0.29       0.36         0.43
domestic_pollution               0.09           0.23       0.39         0.60
water_pollution                  0.13           0.23       0.36         0.47
ethnicity                        0.04           0.09       0.05        -0.07
bmi                              0.01          -0.05       0.05        -0.01
cancer_status                    0.06           0.13       0.41         0.44

                         meats_intake  sugar_intake  ...  \
age                              0.27          0.03  ...
gender                           0.01         -0.03  ...
eclass                           0.13         -0.04  ...
occupational_risk               -0.16         -0.15  ...
intensity_pa                     0.10          0.06  ...
freq_exercise                    0.22          0.15  ...
fruit_veg                        0.37          0.19  ...
wholegrains                      0.52          0.41  ...
meats_intake                     1.00          0.55  ...
```

```
sugar_intake                          0.55          1.00  ...
alcohol                               0.33          0.27  ...
smoke                                 0.04          0.12  ...
smoke_year                            0.09          0.14  ...
no_cigarette                          0.08          0.17  ...
pain                                  0.23          0.28  ...
bloody discharge                      0.30          0.22  ...
fatigue                               0.31          0.28  ...
shortness_of_breath                   0.21          0.22  ...
wheezing                              0.15          0.13  ...
swallowing_difficulty                 0.02          0.03  ...
clubbing_of_nails                     0.10          0.09  ...
freq_cold                             0.15          0.13  ...
dry_cough                             0.10          0.17  ...
snoring                              -0.05          0.09  ...
diabetes                              0.08          0.06  ...
heart_problems                        0.03          0.00  ...
asthma_or_hypertension                0.24          0.16  ...
lung_problem                         -0.06          0.01  ...
Loss_weight_or_appetite               0.24          0.18  ...
family_cancer_history                -0.02          0.06  ...
sun_exposure                          0.26          0.34  ...
co2_emission                          0.50          0.45  ...
industrial_pollution                  0.53          0.47  ...
domestic_pollution                    0.59          0.49  ...
water_pollution                       0.52          0.43  ...
ethnicity                             0.01          0.01  ...
bmi                                  -0.02         -0.01  ...
cancer_status                         0.40          0.25  ...

                         Loss_weight_or_appetite  family_cancer_history  \
age                                         0.12                  -0.04
gender                                     -0.09                  -0.07
eclass                                      0.02                  -0.07
occupational_risk                          -0.12                  -0.01
intensity_pa                                0.14                  -0.00
freq_exercise                               0.12                   0.02
fruit_veg                                   0.12                  -0.01
wholegrains                                 0.21                  -0.06
meats_intake                                0.24                  -0.02
sugar_intake                                0.18                   0.06
alcohol                                     0.10                   0.08
smoke                                      -0.01                   0.11
smoke_year                                  0.02                   0.09
no_cigarette                                0.03                   0.11
pain                                        0.33                   0.05
bloody discharge                            0.41                   0.05
fatigue                                     0.23                  -0.04
shortness_of_breath                        -0.00                   0.07
wheezing                                   -0.02                   0.03
swallowing_difficulty                      -0.01                   0.03
clubbing_of_nails                           0.00                   0.03
freq_cold                                  -0.03                   0.02
dry_cough                                   0.03                   0.07
snoring                                    -0.06                   0.16
diabetes                                    0.02                  -0.01
heart_problems                             -0.01                  -0.02
asthma_or_hypertension                     -0.01                   0.01
lung_problem                               -0.01                   0.08
```

```
Loss_weight_or_appetite                 1.00                  -0.05
family_cancer_history                  -0.05                   1.00
sun_exposure                            0.16                   0.00
co2_emission                            0.23                  -0.03
industrial_pollution                    0.26                  -0.01
domestic_pollution                      0.25                   0.03
water_pollution                         0.36                  -0.03
ethnicity                              -0.06                   0.14
bmi                                    -0.07                  -0.02
cancer_status                           0.26                   0.09

                         sun_exposure  co2_emission  industrial_pollution  \
age                              0.07          0.26                  0.26
gender                           0.01         -0.04                 -0.02
eclass                          -0.16          0.00                  0.07
occupational_risk               -0.30         -0.18                 -0.23
intensity_pa                     0.18          0.12                  0.18
freq_exercise                    0.19          0.25                  0.29
fruit_veg                        0.21          0.36                  0.36
wholegrains                      0.26          0.45                  0.43
meats_intake                     0.26          0.50                  0.53
sugar_intake                     0.34          0.45                  0.47
alcohol                          0.14          0.23                  0.31
smoke                           -0.02         -0.05                 -0.00
smoke_year                      -0.00          0.02                  0.04
no_cigarette                     0.03          0.05                  0.04
pain                             0.27          0.22                  0.33
bloody discharge                 0.21          0.26                  0.34
fatigue                          0.13          0.33                  0.29
shortness_of_breath              0.06          0.20                  0.20
wheezing                         0.00          0.08                  0.08
swallowing_difficulty           -0.03          0.02                  0.05
clubbing_of_nails                0.06          0.10                  0.16
freq_cold                        0.08          0.17                  0.15
dry_cough                        0.09          0.05                  0.15
snoring                         -0.05         -0.04                 -0.04
diabetes                         0.06          0.12                  0.17
heart_problems                   0.02          0.04                  0.09
asthma_or_hypertension           0.18          0.24                  0.30
lung_problem                     0.05         -0.02                  0.06
Loss_weight_or_appetite          0.16          0.23                  0.26
family_cancer_history            0.00         -0.03                 -0.01
sun_exposure                     1.00          0.53                  0.52
co2_emission                     0.53          1.00                  0.72
industrial_pollution             0.52          0.72                  1.00
domestic_pollution               0.47          0.78                  0.74
water_pollution                  0.40          0.61                  0.76
ethnicity                       -0.11          0.01                 -0.07
bmi                             -0.02          0.03                  0.01
cancer_status                    0.29          0.41                  0.42

                         domestic_pollution  water_pollution  ethnicity   bm
i  \
age                                    0.31             0.33       0.08  0.1
1
gender                                -0.14            -0.06      -0.08 -0.0
6
eclass                                 0.06             0.12       0.00  0.0
7
```

```
occupational_risk                     -0.26            -0.21      -0.03  0.0
4
intensity_pa                           0.09             0.13       0.04  0.0
1
freq_exercise                          0.23             0.23       0.09 -0.0
5
fruit_veg                              0.39             0.36       0.05  0.0
5
wholegrains                            0.60             0.47      -0.07 -0.0
1
meats_intake                           0.59             0.52       0.01 -0.0
2
sugar_intake                           0.49             0.43       0.01 -0.0
1
alcohol                                0.27             0.25       0.04 -0.0
1
smoke                                 -0.02            -0.02       0.14 -0.0
4
smoke_year                             0.04             0.06       0.18 -0.0
1
no_cigarette                           0.06             0.05       0.22 -0.0
2
pain                                   0.26             0.30       0.05 -0.0
7
bloody discharge                       0.29             0.38      -0.01 -0.0
1
fatigue                                0.36             0.32       0.05 -0.0
4
shortness_of_breath                    0.25             0.19       0.14 -0.0
2
wheezing                               0.10             0.07       0.17  0.0
1
swallowing_difficulty                  0.08             0.12       0.04 -0.0
8
clubbing_of_nails                      0.11             0.15       0.07 -0.0
4
freq_cold                              0.20             0.16       0.07 -0.0
1
dry_cough                              0.08             0.10       0.11 -0.0
5
snoring                               -0.10            -0.09       0.14  0.0
4
diabetes                               0.13             0.17       0.02  0.0
0
heart_problems                         0.06             0.09       0.02 -0.0
4
asthma_or_hypertension                 0.29             0.29      -0.02  0.0
0
lung_problem                          -0.00             0.04       0.10 -0.0
6
Loss_weight_or_appetite                0.25             0.36      -0.06 -0.0
7
family_cancer_history                  0.03            -0.03       0.14 -0.0
2
sun_exposure                           0.47             0.40      -0.11 -0.0
2
co2_emission                           0.78             0.61       0.01  0.0
3
```

```
industrial_pollution                  0.74             0.76      -0.07  0.0
1
domestic_pollution                    1.00             0.73      -0.01 -0.0
3
water_pollution                       0.73             1.00      -0.13 -0.0
5
ethnicity                            -0.01            -0.13       1.00  0.0
3
bmi                                  -0.03            -0.05       0.03  1.0
0
cancer_status                         0.53             0.49       0.06 -0.0
1

                         cancer_status
age                               0.58
gender                           -0.12
eclass                            0.11
occupational_risk                -0.36
intensity_pa                      0.06
freq_exercise                     0.13
fruit_veg                         0.41
wholegrains                       0.44
meats_intake                      0.40
sugar_intake                      0.25
alcohol                           0.27
smoke                            -0.00
smoke_year                        0.07
no_cigarette                      0.10
pain                              0.30
bloody discharge                  0.28
fatigue                           0.30
shortness_of_breath               0.16
wheezing                          0.02
swallowing_difficulty            -0.02
clubbing_of_nails                 0.01
freq_cold                         0.14
dry_cough                         0.10
snoring                          -0.10
diabetes                          0.13
heart_problems                    0.00
asthma_or_hypertension            0.28
lung_problem                      0.05
Loss_weight_or_appetite           0.26
family_cancer_history             0.09
sun_exposure                      0.29
co2_emission                      0.41
industrial_pollution              0.42
domestic_pollution                0.53
water_pollution                   0.49
ethnicity                         0.06
bmi                              -0.01
cancer_status                     1.00

[38 rows x 38 columns]
```

Variance

```
 age                        183.61
gender                       0.26
eclass                       1.33
```

```
occupational_risk            1.59
intensity_pa                 4.08
freq_exercise                4.65
fruit_veg                    5.11
wholegrains                  7.39
meats_intake                 7.26
sugar_intake                 6.49
alcohol                      2.77
smoke                        0.11
smoke_year                   5.30
no_cigarette                 3.50
pain                         4.93
bloody discharge             1.92
fatigue                      4.30
shortness_of_breath          1.65
wheezing                     0.67
swallowing_difficulty        0.48
clubbing_of_nails            0.62
freq_cold                    2.52
dry_cough                    1.38
snoring                      2.11
diabetes                     1.03
heart_problems               0.38
asthma_or_hypertension       1.83
lung_problem                 0.46
Loss_weight_or_appetite      1.68
family_cancer_history        0.18
sun_exposure                 5.53
co2_emission                 5.86
industrial_pollution         5.66
domestic_pollution           7.20
water_pollution              5.21
ethnicity                    3.09
bmi                         88.86
cancer_status                0.19
dtype: float64
```

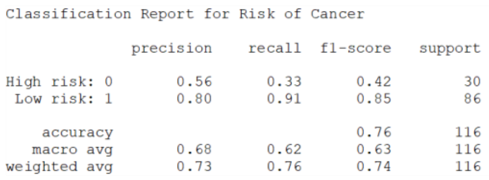
Kutosis

```
 heart_problems             96.153009
lung_problem               74.931717
swallowing_difficulty      64.739681
clubbing_of_nails          49.746113
wheezing                   49.433756
bmi                        45.212508
diabetes                   26.942450
Loss_weight_or_appetite    25.333973
bloody discharge           23.075500
asthma_or_hypertension     18.371149
shortness_of_breath        16.949999
no_cigarette               14.380729
dry_cough                  12.977208
freq_cold                  12.566923
ethnicity                   9.276103
smoke_year                  8.521772
snoring                     7.335806
fatigue                     3.694552
pain                        3.556325
alcohol                     3.534298
smoke                       3.126404
```

```
water_pollution             1.651163
age                         0.827858
industrial_pollution        0.497238
eclass                     -0.261469
intensity_pa               -0.294514
meats_intake               -0.295624
co2_emission               -0.295770
family_cancer_history      -0.363997
freq_exercise              -0.527179
sun_exposure               -0.532736
domestic_pollution         -0.590688
fruit_veg                  -0.628707
cancer_status              -0.847800
sugar_intake               -0.887023
wholegrains                -1.174095
occupational_risk          -1.187165
gender                     -1.360677
dtype: float64
```

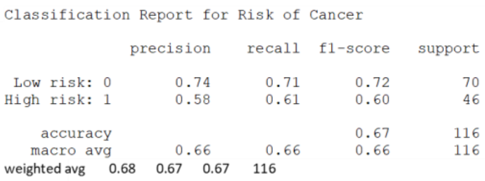
Appendix 10: Machine Learning Models Performance Report

Performance Result for Naïve Bayes Classifier



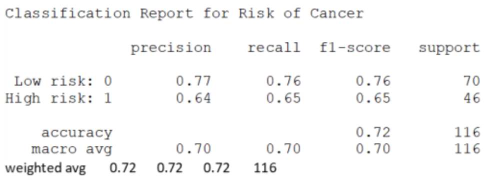
```
Classification Report for Risk of Cancer

              precision    recall  f1-score   support

High risk: 0       0.56      0.33      0.42        30
 Low risk: 1       0.80      0.91      0.85        86

    accuracy                           0.76       116
   macro avg       0.68      0.62      0.63       116
weighted avg       0.73      0.76      0.74       116
```

Performance Result for K-Nearest Neighbour



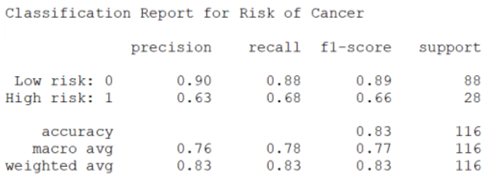
```
Classification Report for Risk of Cancer

              precision    recall  f1-score   support

 Low risk: 0       0.74      0.71      0.72        70
High risk: 1       0.58      0.61      0.60        46

    accuracy                           0.67       116
   macro avg       0.66      0.66      0.66       116
weighted avg  0.68  0.67  0.67  116
```

Performance Result for Logistic Regression



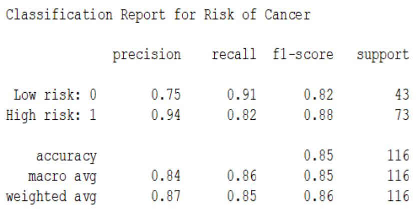
```
Classification Report for Risk of Cancer

              precision    recall  f1-score   support

 Low risk: 0       0.77      0.76      0.76        70
High risk: 1       0.64      0.65      0.65        46

    accuracy                           0.72       116
   macro avg       0.70      0.70      0.70       116
weighted avg  0.72  0.72  0.72  116
```

Performance Result for Support Vector Machine



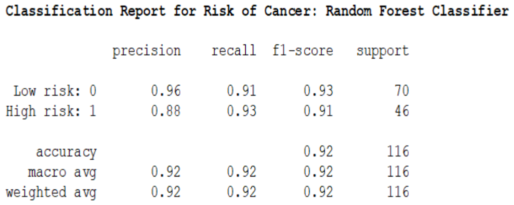
```
Classification Report for Risk of Cancer

              precision    recall  f1-score   support

 Low risk: 0       0.90      0.88      0.89        88
High risk: 1       0.63      0.68      0.66        28

    accuracy                           0.83       116
   macro avg       0.76      0.78      0.77       116
weighted avg       0.83      0.83      0.83       116
```

Performance Result for Decision Tree Classifier



```
Classification Report for Risk of Cancer

              precision    recall  f1-score   support

 Low risk: 0       0.75      0.91      0.82        43
High risk: 1       0.94      0.82      0.88        73

    accuracy                           0.85       116
   macro avg       0.84      0.86      0.85       116
weighted avg       0.87      0.85      0.86       116
```

Performance Result for Random Forest Classifier

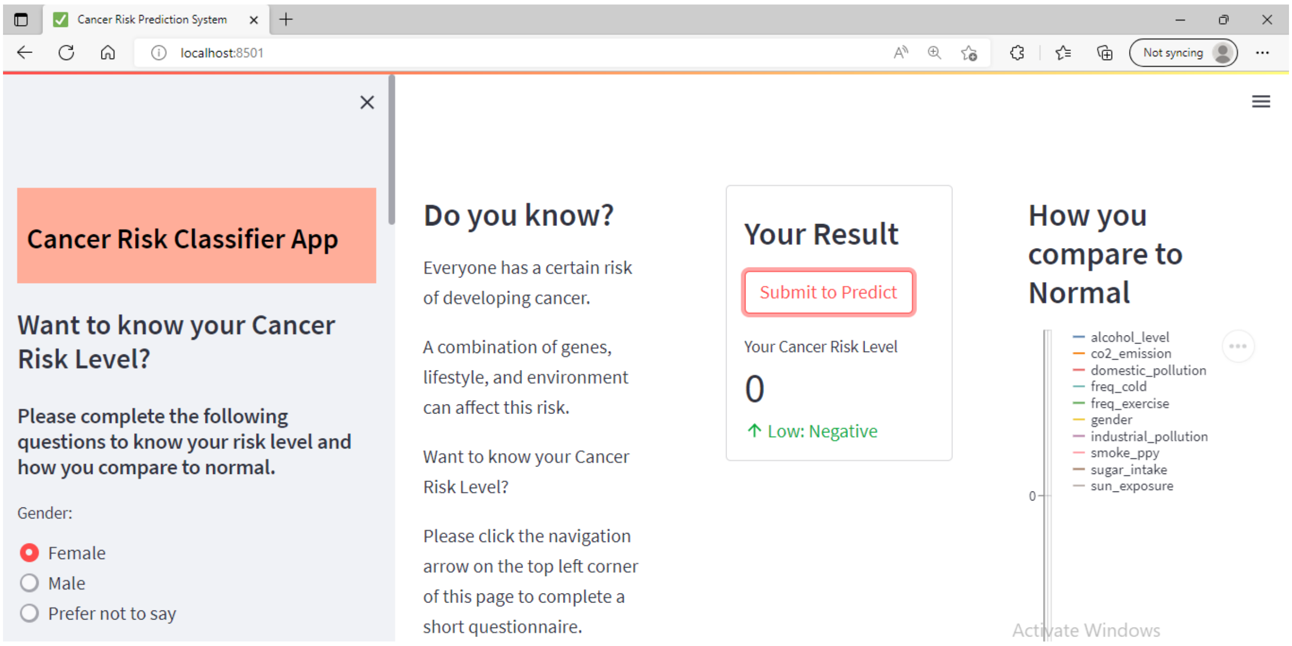


```
Classification Report for Risk of Cancer: Random Forest Classifier

              precision    recall  f1-score   support

 Low risk: 0       0.96      0.91      0.93        70
High risk: 1       0.88      0.93      0.91        46

    accuracy                           0.92       116
   macro avg       0.92      0.92      0.92       116
weighted avg       0.92      0.92      0.92       116
```

Appendix 11: GridSearch Hyperparameter tuning

<table>
<tr><th colspan="3">Grid Search Hyperparameter Tuning</th></tr>
<tr><th>Hyparameters</th><th>Declared dictionary of hyperparameters for tuning</th><th>Best Estimator Hyperparameters</th></tr>
<tr><td>max_depth</td><td>[int(x) for x in np.linspace(10, 120, num = 12)]</td><td>2</td></tr>
<tr><td>min_samples_leaf</td><td>[1,2,4,6,8,9]</td><td>1</td></tr>
<tr><td>min_samples_split</td><td>[2, 6, 10]</td><td>0.5</td></tr>
<tr><td>bootstrap</td><td>[True, False]</td><td></td></tr>
<tr><td>n_estimators</td><td>[5,20,50,100]</td><td></td></tr>
<tr><td>max_features</td><td>['auto', 'sqrt']</td><td></td></tr>
<tr><td>criterion</td><td>["gini", "entropy"]</td><td>entropy</td></tr>
<tr><td>random_state</td><td>[42, 35, 10, 0]</td><td>42</td></tr>
<tr><td colspan="3"><b>Cross-validation scores:</b> [0.82978723 0.93617021 0.85106383 0.87234043 0.89130435 0.76086957 0.82608696 0.84782609 0.84782609 0.7826087]</td></tr>
<tr><td colspan="3"><b>Cross val mean: 0.845 (std:0.048): Best Score: %s 0.977; Number of run was set at 10.</b></td></tr>
</table>

Appendix 12: Landing page of the Cancer Risk Classifier Web App deployed on Streamlit



Cancer Risk Classifier App

Want to know your Cancer Risk Level?

Please complete the following questions to know your risk level and how you compare to normal.

Gender:

● Female
○ Male
○ Prefer not to say

Do you know?

Everyone has a certain risk of developing cancer.

A combination of genes, lifestyle, and environment can affect this risk.

Want to know your Cancer Risk Level?

Please click the navigation arrow on the top left corner of this page to complete a short questionnaire.

Your Result

Submit to Predict

Your Cancer Risk Level

0

↑ Low: Negative

How you compare to Normal

Appendix 13: Research Survey Questionnaire and Evidence of Ethical Approval

Survey - Risk of Cancer Amongst Adults

Introduction and Consent

This survey is addressing a research project which is about creating a system to predict the risk of cancer in adults using machine learning techniques. The project is aimed to understand and identify relevant features of adults that may expose them to the risk of cancer.

Please read the following statements fully and carefully. By proceeding to take the questionnaire, you are giving your consent.

I volunteer to take part in this research questionnaire. I understand that the research aims to collect data on adults **(aged 18 and above)** regarding their risk of having cancer. The data collected in this questionnaire will be used in a dissertation project and furthermore help to identify key variables for cancer risk in adults. The data will be collected ANONYMOUSLY.

1. I understand that my participation in this project is voluntary. I will not be paid for my involvement. I am free to withdraw from the survey at any time, without reason.
2. I have read and understood that all data provided will be treated in confidence, names will not be collected.
3. The ethical clearance of this project has been approved by Solent University.
4. I have read and understood the explanation of the research project provided to me.

By proceeding to take this questionnaire, I agree to take part in this research project and to the above statements.

Contact:

5ajana82@solent.ac.uk

Consent

□ Yes, I am willing to participate

General Questions

Tell us about yourself. Please be informed that this survey has been developed for adults(aged 18 and above) and the results may not be a good reflection of the research if you are under 18.

*What country are you from?

*Your country of residence?

*What is your age?

Enter a value between 0 and 100

*Gender

○ Female ○ Male ○ Prefer not to say

*How much do you weigh (in kg)?

For example, a person weight in kilograms (kg) is: 67

*How tall are you (in cm)?

For example, a person height in centimeter (cm) is: 185

*Which ethnic group are you from?

White

○ British

○ Irish

○ Caribbean

○ African

○ Indian

○ Pakistani

○ Chinese

○ White & Black Caribbean

○ White & Black African

○ White & Asian

○ White & Arab

○ White (any other white background)

Black

○ British

○ Irish

○ Caribbean

○ African

○ Indian

○ Pakistani

○ Chinese

○ White & Black Caribbean

○ White & Black African

○ White & Asian

○ White & Arab

○ White (any other white background)

Asian

○ British

○ Irish

○ Caribbean

○ African

○ Indian

○ Pakistani

○ Chinese

○ White & Black Caribbean

○ White & Black African

○ White & Asian

○ White & Arab

○ White (any other white background)

Arab

○ British

○ Irish

○ Caribbean

○ African

○ Indian

○ Pakistani

○ Chinese

○ White & Black Caribbean

○ White & Black African

○ White & Asian

○ White & Arab

○ White (any other white background)

Mixed

○ British

○ Irish

○ Caribbean

○ African

○ Indian

○ Pakistani

○ Chinese

○ White & Black Caribbean

○ White & Black African

○ White & Asian

○ White & Arab

○ White (any other white background)

Other (please state in the comment field below)

*How would you rate your economic class?

Economic class system is based on earnings per year

○ Lower class

○ Lower-middle class

○ Middle class

○ Upper-middle class

○ Upper Higher class

*Which of the below best describe your occupation or skill level?

Skill 1: Professional occupations that usually require degrees. Skilled work. E.g. Doctors, Teachers, Managers, Professionals.

Skill 2: Technical occupations that usually require a college credential or technical training. Skilled Work. E.g. Electrician, plumber.

Skill 3: Intermediate skilled jobs that usually require high school education and on-the-job training. E.g. Truck Drivers**.**

Skill4: Low Skilled occupations that usually require on-the-job-training.

○ Skill level 1 ○ Skill Level 2 ○ Skill Level 3

○ Skill Level 4

Social and Lifestyle Questions

Please tell us about your level of physical activities, dietary, alcohol consumption and smoking

*How would you describe the level of intensity of your physical activities?

Where physical activities include but not limited to light walking, stretching, lifting hand weights, push-ups against the walls, brisk walking, water aeorbics, tennis (doubles), biking on level ground, sports involving catch and throw (such as volleyball and baseball), jogging, fast swimming, fast dancing, jumping rope, tennis (singles), basket ball, and soccer.

Very low									Very vigorous
1	2	3	4	5	6	7	8	9	10

*How often do you exercise or involve in physical activities in a week?

Rarely or once/week									10 times a week
1	2	3	4	5	6	7	8	9	10

*What is the level of your fruit and veg intake each day on average?

Fresh, frozen and canned all count, as does one portion of pulses and beans, and one glass of juice.

* A portion is 80g — or about the size of your fist

Little or 1 portion per day									10 portion or more per day
1	2	3	4	5	6	7	8	9	10

*How frequently do you consume wholegrains each day on average?

Wholegrains include wholewheat cereals, wholemeal rolls, bread and flour, brown rice, brown pasta, and quinoa

Little or once per day									10 or more times per day
○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	○ 8	○ 9	○ 10

*How often do you eat processed meat each week?

Processed meat includes hog dogs, pepperoni, chorizo, salami, ham, and bacon.

Rarely or once per week									10 or times per week	N/A
○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	○ 8	○ 9	○ 10	○

*How often do you have sugary drinks and processed food high in fat and sugar each week?

Example: cakes, pastries, sweets, pizza, fried chickens, burger, chips, crisps, chocolate, tea/coffee with sugar, sport/energy drinks, and non diet soft drinks

Rarely or once per week									10 or more per week
○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	○ 8	○ 9	○ 10

*How would you define your alcohol consumption per day?

Where one unit of alcohol equals 10ml or 8g of pure alcohol. This is equivalent of half pint or regular beer, lager or cider; 1 small glass of wine; 1 single measure of spirits; 1 small glass of sherry; 1 single measure of aperitifs.

Little or 1 unit per day									10 or more units per day	N/A
○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	○ 8	○ 9	○ 10	○

*Do you smoke?

○ Yes ○ No

*How long have you been smoking (in years)?

1 year or less									10 years or more	N/A
○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	○ 8	○ 9	○ 10	○

*How many cigarettes do you smoke per day?

One cigarette/day									10 or more per day	N/A
○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	○ 8	○ 9	○ 10	○

Personal and Family History

Please tell us about your health history

*Do you have any of the following chronic illness, allergies, or symptoms

- □ Pain (abdominal, back, waist, chest, headaches etc)
- □ Bloody discharge (cough, stool, nose, private parts etc)
- □ Fatigue
- □ Shortness of breath
- □ Wheezing
- □ Swallowing difficult
- □ Clubbing of finger nails
- □ Frequent cold or urination
- □ Dry cough
- □ Snoring
- □ Diabetes
- □ Asthma or hypertension
- □ Heart problem
- □ Lung Problem
- □ Loss of weigh or appetite
- □ Not Applicable

Others (Please specify here)

To what extent would you describe the pain or discomfort you are experiencing in connection with the above illness, allergies, or symptoms

	Never or rarely									Very severe
Pain (abdominal, back, waist, chest, headaches etc)	1	2	3	4	5	6	7	8	9	10
Bloody discharge (cough, stool, nose, private parts etc)	1	2	3	4	5	6	7	8	9	10
Fatigue	1	2	3	4	5	6	7	8	9	10
Shortness of breath	1	2	3	4	5	6	7	8	9	10
Wheezing	1	2	3	4	5	6	7	8	9	10
Swallowing difficult	1	2	3	4	5	6	7	8	9	10
Clubbing of finger nails	1	2	3	4	5	6	7	8	9	10
Frequent cold or urination	1	2	3	4	5	6	7	8	9	10
Dry cough	1	2	3	4	5	6	7	8	9	10
Snoring	1	2	3	4	5	6	7	8	9	10
Diabetes	1	2	3	4	5	6	7	8	9	10
Heart problems	1	2	3	4	5	6	7	8	9	10
Asthma or Hypertension	1	2	3	4	5	6	7	8	9	10
Lung problem	1	2	3	4	5	6	7	8	9	10
Loss of weight or appetite	1	2	3	4	5	6	7	8	9	10

Other (Please state here)

*Have you been diagnosed of cancer recently or in the past?

○ Yes ○ No

* Do you have blood relations (family members) who have been diagnosed of cancer recently or in the past?

□ Yes □ No

*Do you have blood relations (family members) who have suffered the following?

□ Heart Attack
□ Asthma
□ Lung disease
□ Stroke
□ Cardiovascular disease
□ Not Applicable
□ Diabetes
□ High Blood Pressure

Other serious illness: Please specify

Environmental and Location Questions

Please tell us about your exposure to sun, air and water pollution

*How would you describe your level of exposure to the following in your location?

	Rarely or never exposed									I am seriously exposed
Sun exposure between 11am and 3pm	1	2	3	4	5	6	7	8	9	10
Air pollution as a result of Co2 emission from automobiles	1	2	3	4	5	6	7	8	9	10
Air pollution due to industrial activities	1	2	3	4	5	6	7	8	9	10
Air pollution due to domestic sources	1	2	3	4	5	6	7	8	9	10
Water pollution due to oil spillage, sewage or chemical end products	1	2	3	4	5	6	7	8	9	10

Miscellaneous

For SurveyCircle users (www.surveycircle.com), please redeem Survey Code with one click: https://www.surveycircle.com/9WT2-19K9-32GP-K1U6

How did you find out about this survey?

□ SurveyCircle

□ Twitter

□ Word of mouth/referrals

□ Other (Please specify)

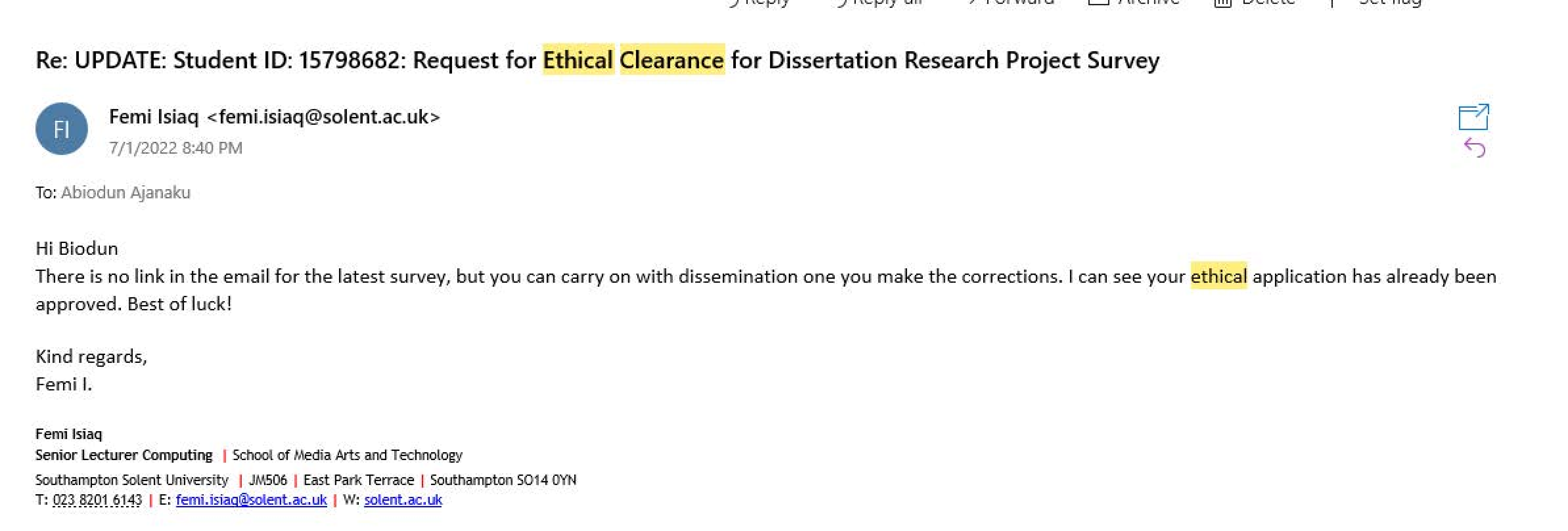
□ PollPool

□ Instagram

□ WhatsApp Group

□ LinkedIn

□ Facebook



Re: UPDATE: Student ID: 15798682: Request for Ethical Clearance for Dissertation Research Project Survey

Femi Isiaq <femi.isiaq@solent.ac.uk>
7/1/2022 8:40 PM

To: Abiodun Ajanaku

Hi Biodun
There is no link in the email for the latest survey, but you can carry on with dissemination one you make the corrections. I can see your ethical application has already been approved. Best of luck!

Kind regards,
Femi I.

Femi Isiaq
Senior Lecturer Computing | School of Media Arts and Technology
Southampton Solent University | JM506 | East Park Terrace | Southampton SO14 0YN
T: 023 8201 6143 | E: femi.isiaq@solent.ac.uk | W: solent.ac.uk