IDENTIFICATION OF SIGNIFICANT FEATURES AND DATA MINING TECHNIQUES IN PREDICTING BREAST CANCER

YEW KWANG YONG

FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

IDENTIFICATION OF SIGNIFICANT FEATURES AND DATA MINING TECHNIQUES IN PREDICTING BREAST CANCER

YEW KWANG YONG

RESEARCH REPORT SUBMITTED TO THE
FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA, IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF DATA SCIENCE

2019

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: YEW KWANG YONG (I.C/Passport No: 940227-01-6489)

Registration/Matric No: WQD180091

Name of Degree: Master of Data Science

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

IDENTIFICATION OF SIGNIFICANT FEATURES AND DATA MINING TECHNIQUES IN PREDICTING BREAST CANCER

Field of Study:

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

05/12/2019

Candidate's Signature Date

Subscribed and solemnly declared before,

Witness's Signature Date

Name:
Designation:

IDENTIFICATION OF SIGNIFICANT FEATURES AND DATA MINING TECHNIQUES IN PREDICTING BREAST CANCER

ABSTRACT

Breast cancer is one of the greatest critical illness for women in worldwide. In term of statistics, it is anticipated that roughly 252,710 new records of breast cancer were identified in women in 2017 alone, where 40,610 women died from the disease. The efficiency of machine learning algorithms has successfully supported many medical officers in breast cancer analysis and prediction. Accuracy of a model has been interpreted as the most significant performance metric of a machine learning model. However, the predictive models from previous studies is far from perfect as there is room for improvement in term of accuracy. An accurate predictive model is beneficial in assisting medical officers to diagnosis the breast cancer survivability precisely when the veteran oncologists are absent and to help patients to escape complicated surgeries, futile therapies, and soaring health expenses. The objective of this research is to propose a fine-tuned classification model with improved performance metrics such as accuracy, sensitivity, specificity and area under the curve (AUC). Four leading machine learning classification models are studied and compared to determine the best predictive model for breast cancer survivability prediction, they are: Random Forest, XGBoost, Graph-based semi supervised learning and CatBoost. The findings of the research indicate that the CatBoost model is the best performing model as it demonstrated the best overall performance metrics in term of accuracy, specificity and AUC. The proposed CatBoost model is a decent for medical officers to apply without requiring extra time and attempt to fine tune the model. Most importantly the proposed model provides better overall performance as compared to the machine learning suggested in previous researches.

Keywords: Random Forest, CatBoost, XGBoost, Semi Supervised Learning, Breast Cancer

PENGENALAN PENTING CIRI-CIRI DAN DATA MINING TEKNIK DALAM MERAMAL BREAST CANCER

ABSTRAK

Kanser payudara merupakan penyakit kritikal yang paling besar untuk wanita di seluruh dunia. Dari segi statistik, kira-kira 252,710 rekod baru kanser payudara telah dikenal pasti di kalangan wanita pada tahun 2017 sahaja, di mana 40,610 wanita mati akibat penyakit ini. Kecekapan algoritma pembelajaran mesin telah berjaya disokong ramai pegawai perubatan dalam analisis kanser payudara dan ramalan. Ketepatan model yang telah ditafsirkan sebagai prestasi yang paling penting metrik model pembelajaran mesin. Namun, model ramalan daripada kajian sebelumnya adalah jauh dari sempurna kerana terdapat ruang untuk penambahbaikan dari segi ketepatan. Model ramalan tepat bermanfaat dalam membantu pegawai perubatan untuk diagnosis yang kemandirian kanser payudara dengan tepat apabila pakar onkologi veteran tidak hadir dan membantu pesakit untuk megelakkan pembedahan rumit, terapi, dan perbelanjaan kesihatan yang tidak penting. Objektif kajian ini adalah untuk mencadangkan model pengelasan diperhalusi dengan metrik prestasi yang lebih baik seperti ketepatan, kepekaan, kekhususan dan kawasan di bawah lengkung (AUC). Empat model klasifikasi pembelajaran mesin dikaji dan dibandingkan untuk menentukan model ramalan terbaik ramalan kanser payudara, mereka adalah: *Random Forest, Graph -based semi supervised learning, XGBoost* dan *CatBoost*. Dapatan kajian menunjukkan bahawa model CatBoost adalah model terbaik kerana ia menunjukkan yang prestasi terbaik keseluruhan dari segi ketepatan, kekhususan dan AUC. Model CatBoost adalah paling sesuai bagi pegawai perubatan untuk mengguna. Model yang dicadangkan memberikan prestasi paling baik berbanding dengan pembelajaran mesin yang disyorkan dalam kajian sebelumnya.

Kata Kunci: Random Forest, CatBoost, XGBoost, Semi Supervised Learning, Breast Cancer

ACKNOWLEDGEMENTS

Upon completing the research project as one of the required courses for the Master of Data Science, I would like to extend my deepest gratitude to Dr. Hoo Wai Lam, my research project supervisor for being helpful in assisting me to solve the challenges encountered throughout the journey of completing my project. He is also patient in giving me guidance and he understands well the conditions of students where they have other commitment in other subjects. I would like to thank my internal examiner Dr. Ong Sim Ying, for providing valuable suggestions to improve my project. The knowledges shared by them has enhanced my understanding in problem solving. Lastly, I would like to thank The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI) for providing the assess to the cancer database for extracting the required dataset for this research project, without the valuable dataset, this research project would not be completed smoothly.

Table of Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

ANN	:	Artificial Neural Network
AUC	:	Area Under the Curve
CART	:	Classification and Regression Tree
GBDT	:	Gradient Boosting Decision Tree
FDT	:	Fuzzy Decision Tree
FP	:	False Positive
FN	:	False Negative
NCI	:	National Cancer Institute
RF	:	Random Forest
SEER	:	Surveillance, Epidemiology, and End Results
SSL	:	Semi Supervised Learning
SVM	:	Support Vector Machine
TP	:	True Positive
TN	:	True Negative
XGBoost	:	Extreme Gradient Boosting

Chapter 1 : INTRODUCTION

1.1 Research Background

Breast cancer is one of the greatest critical illness for women in worldwide. In term of statistics, it is anticipated that roughly 252710 new records of breast cancer were identified in women in 2017 alone, where 40610 women died from the disease as published by Harbeck and Gnant (2017). Breast cancer could be categorised into benign and malignant. Medical officers will provide different treatment plans for therapy depending on the category of breast cancer diagnosed. If there are any mis-assessment, it will result in inappropriate therapies and reduce the best moment for curing the patients, which will run into terrible outcomes. Thus, the choice of model for forecasting the category of breast tumour cell is very significant in preventing the disease from becoming critical. During the lifespan of a woman, one in twelve women is anticipated to be plagued by this illness. However, the breast cancer is treatable if identified and cured in its initial stages before becoming critical, Unfortunately, a great number of the woman patients decease because of the delayed discovery of breast cancer.

The most important medical challenge linked with breast cancer is to foresee the conclusion (survival or death) when a patient is diagnosed with breast cancer disease. In many situations, medically distinct metastases have happened when the main tumour is identified. The cancer therapies such as chemotherapy is deemed to slow down the propagation of breast cancer cells by diminishing the remote metastases, by significant amount. Nevertheless, experiments proves that 70% of patients accepting these treatments can stay alive without undergoing the unnecessary therapy, therefore the treatment seems unnecessary for them to survive (Sun et al., 2007). Consequently, the capability to foresee illness consequences further precisely would grant physicians to make correct judgments on the possible need of therapy. This possibly will breed the advancement of discretely customized therapies to boost therapy competence (Khan, Shin, Choi, & Kim, 2007). Prognosis facilitates the establishment of a therapy scheme by forecasting the consequence of the illness. Three categories of predictive focal points of cancer prognosis, they are:(1) forecast of cancer vulnerability (risk assessment), (2) forecast of cancer repetition (regeneration of cancer after therapy), and (3) forecast of cancer survivability. Researchers concentrate on forecasting the consequence in the aspects of life expectancy, survivability, evolution, or tumour-drug sensitivity after the identification of the disease. This research concentrates on the survivability forecast for breast cancer patients. Delen et al. (2005) suggested that the endurance analysis is a component of clinical prognosis and includes the utilization of approaches and methods for forecasting the endurance of a certain patient on the foundation of past data of patients. In overall, “survival” can be described as the lasting life of a patient for a particular phase after the identification of the disease. As proposed by Delen et al. (2005) and Brenner et al. (2002) , if the patient stays alive for 1825 days (5 years) after the time of analysis, then the patient is deemed to have survived.

Feature selection method is a dimensionality reduction method, that is frequently applied in data pre-processing step, of choosing a subgroup of pertinent features for constructing a good predictive model (Efitorov, Burikov, Dolenko, Laptinskiy, & Dolenko, 2015). In most of the actual world challenges, lowering dimensionality of a problem is an important stage before proceeding to data modelling process. The typical benchmark for lowering the dimension of the dataset without sacrifing most of the important information in the original data corresponding to some optimality standards. Feature selection seeks to enhance the productivity in analytic process and lowering the inaccuracy percentage in the consequence. Diminishing the quantity of unrelated and redundant features significantly lowers the training and prediction time of a predictive model and generates a more conventional notion. This improves in securing a clearer understanding into the fundamental theory of a real-world categorization challenge (Kohavi & Sommerfield, 1995).

Due to the multiple benefits it offers, feature selection has been widely applied in numerous researches. The selection of significant features from mammogram images for breast cancer diagnosis conducted by Eltoukhy and Faye (2014) using multiresolution representation. On the other hand, the identification of significant metabolic features related to breast cancer pathogenesis by summarizing of RNA metabolites with SVM based feature selection is applie by Henneges et al. (2009). Huang, Hung, and Chen (2010) employed neural network algorithm with feature selection based on entropy on breast cancer analysis for Wisconsin Breast Cancer Dataset. However, such method is yet to be applied on SEER breast cancer data set for survival prediction to identify the significant features in the dataset.

Data driven decision making is the essential motive for the escalating stress on precise and less- intrusive customized prognostic models founded on machine learning methods. This methodology can permit many patients to escape complicated surgical surgeries, pointless therapies, and soaring health expenses. Furthermore, in circumstances where veteran oncologists are absent, prognostic models designed by the use of machine learning methods can assist physicians' decision-making with satisfactory precision (Amir, 2003).

The latest machine learning models can foresee the survivability of breast cancer patients are compared. In this research, the focal point of study is to improve the classification algorithm that forecasts whether the cancer patient fit in to the cluster of those who stay alive after a particular phase. The research intends to build a precise and stable classification model. The establishment of the model could grant health oncologists to produce useful judgments for curing breast cancer patients. Impacts of the study comprise of its validity and detailed experimental research. Three distinct classification models are studied: Random Forest model (RF), Catboost model, XGBoost model and graph-based semi supervised learning (SSL) model. The surveillance, epidemiology, and end results (SEER) cancer incidence database, which is the most complete resource of data on cancer occurrence and survival in the United States are utilized to achieve the goal of this research (SEER et al.,2010).

1.2 Statement of Problem

This research concentrates on building a better machine learning model for the forecast of the survivability of a breast cancer patient. Meanwhile, the significant attributes will be identified from the dataset and play a considerable position in the prediction accuracy of the model. This is because there were no previous works focus on identification of important features for SEER breast cancer dataset. Current findings revealed that machine learning models applied in breast cancer detection are open for further improvement in term of prediction accuracy. Gradient Boosted Decision Trees (GBDT) has been crowned as the most popular classification algorithm due to its high accuracy they are speedy for making predictions, interpretable and have small memory foot print (Saberian, Delgado, & Raimond, 2019). For many years, it has persisted as the leading approach for understanding problems with diverse heterogenous attributes, and complicated colonies: web search, and etc (Caruana, 2006; Roe, Yang, Zhu, Liu, & Stancu, 2005; Q. Wu, Burges, & Svore, 2010) . Catboost and XGBoost have been deemed as the *best* performing prediction models among the GBDT algorithms due to its fast learning speed and low logloss value. These GBDT algorithms has excelled over other traditional machine learning models in several aspects (Prokhorenkova, Gusev, Vorobev, Dorogush, & Gulin, 2018) and are the leading models in most of the Kaggle competitions in recent years, however such method is yet to be applied in breast cancer.

1.3 Research Objective

- To identify the significant feature for forecast of breast cancer.
- To propose a classification algorithm with higher performance.
- To evaluate classification algorithms

1.4 Research Questions

- What are the important attributes that affect the forecast of the survivability of a breast cancer patient?
- Can CatBoost and XGBoost classifier models (GBDT algorithms) outperform classical models in term of superior forecast accuracy of the survivability of a breast cancer patient?

1.5 Research Gap, Research Significance, and Expected Outcome

Even though there are many previous studies of machine learning models on the forecast of the survivability of breast cancer patient. There are room of improvement in term of the accuracy of the prediction models. In a recent research conducted by Park et al. (2013), a research that focuses on the comparison of three major machine learning models was conducted to study the prediction of breast-cancer survivability: support vector machines (SVM), artificial neural networks (ANN), and semi-supervised learning models (SSL). They conclude the semi-supervised learning model is the best performing model among the three models, however the performance of a GBDT model is not studied and compared. Meanwhile, in a research conducted by Endo, Shibata, & Tanaka (2018), Random Forest algorithm is identified as the best performing classification model as compared to other six algorithms. However, the research did not identify the key features using feature selection method, which may improve the prediction accuracy. Studies have revealed that feature selection method is significant for the following reasons:

- It allows the machine learning models to learn speedier.
- It decreases the intricacy of a model and allows to be interpreted easily.
- It advances the accuracy of a model if the appropriate subclass is selected.
- It decreases overfitting.

Therefore, this research aims to propose an improved predictive model for the survivability of breast cancer patient and identify the key important attributes that impact to the survivability of a breast cancer patient.

1.6 Scope of the Study

This research will be focusing on building a prediction model that has improved the prediction accuracy and the determination of the most critical key attributes in predicting the survivability of breast cancer patient. After that, performance metrics will be employed to evaluate the accomplishment of each prediction models.

Chapter 2 : LITERATURE REVIEW

2.1 Introduction

Cancer is a serious health disease in the worldwide. Researchers concluded figure of additional cancer records in 2012 alone have reached 1,639,910 cases. The quantity of deaths caused by cancer is approximately 577,190 cases (Siegel et al., 2012). Out of all the cancer records, breast cancer was frequently identified cancer amongst the female, standing at 29% of newly assessed woman cancer records (790,740 cases) (Siegel et al., 2012). Correctly identifying the tumours is a critical issue to be solved in the healthcare industry. With the advancement of data science, innovative solutions offer us different methods to acquire enormous amount of tumour dataset generated by the medical equipment and sensors. Conventionally, breast cancer is detected by relying on the mammography by radiologists and physicians. (Elmore, Wells, Lee, Howard, 1994). The worth of applying mammograms was confirmed, the inconsistency of the radiologists' understandings triggered a low precision of prediction. It is confirmed that almost 90% of radiologists identified a small number of cancers.

Problem arises when medical doctor needs to study large amount of cancer information out of the hefty volume of cancer records. As a result, data science is beneficial to assist medical doctor in analysing the cancer. To surge the precision and manage the intensely escalating tumour raw data, scholars applied machine learning methods for detecting breast cancer. The present tumour raw data are categorized into malignant and benign categories discretely. By building a model to isolate the categories of tumours, a novel suspected cancer case can be forecast, based on the past tumour data.

2.2 Machine learning

It is a subbranch of data science that dedicated on a major problem: By what means to build systems that instinctively learn from the history? This subfield of data science is significant to address the underlying scientific questions (Jordan & Mitchell, 2015). Machine learning has advanced intensely to the real-world commercial application. It is created as a technology to build a weak artificial intelligence software for face recognition, recommendation system, robot regulator, and etc. Numerous artificial intelligence scientist has identified. The machine learning has been extensively applied throughout a variety of data driven projects, such as forecasting, the anomaly detection in complicated systems, and natural language processing. Machine learning is the hybrid intersection of information technology, a diversity of other specialties and statistics relevant to automated process and making decisions.

2.2.1 Algorithms of Machine Learning

Machine learning models are divided into many categories. They are classified by identifying the types of anticipated output of a model (Ayodele, 2010). Two major categories of models are:

- **Supervised learning** --- the algorithm produces a function that projects inputs to anticipated outputs, for example the classification problem: the system is compulsory to study the estimated behaviour of a case study which projects a vector into one of numerous classes by observing at numerous labelled dataset.

- **Unsupervised learning** --- It is the opposite of supervised learning, where the labelled attributes are not existing.

2.2.1.1 Supervised Learning Approach

Supervised learning is a subbranch of machine learning algorithm that specialize in studying the function that correlate the relationship between an input and an output

(Caruana, 2006). It formulates a customized function from labelled dataset. Supervised classification is the machine learning task frequently conducted by the Intelligent Systems. supervised learning is applied to establish a succinct algorithm of the dispersal of class labels. Supervised learning is a frequently applied model in classification complications as it aims to allow the system to study a classification problem that has been issued.

Figure 2.1 shows the structures of both supervised learning and unsupervised Learning. Supervised learning is suitable for the datasets where the inputs are offered. On the contrary if the input labels are incomplete, Outputs are impossible to be generated and deduced. For unsupervised learning approach, dormant variables trigger all the records, which signifies the observations of unlabelled classes could be deduce.

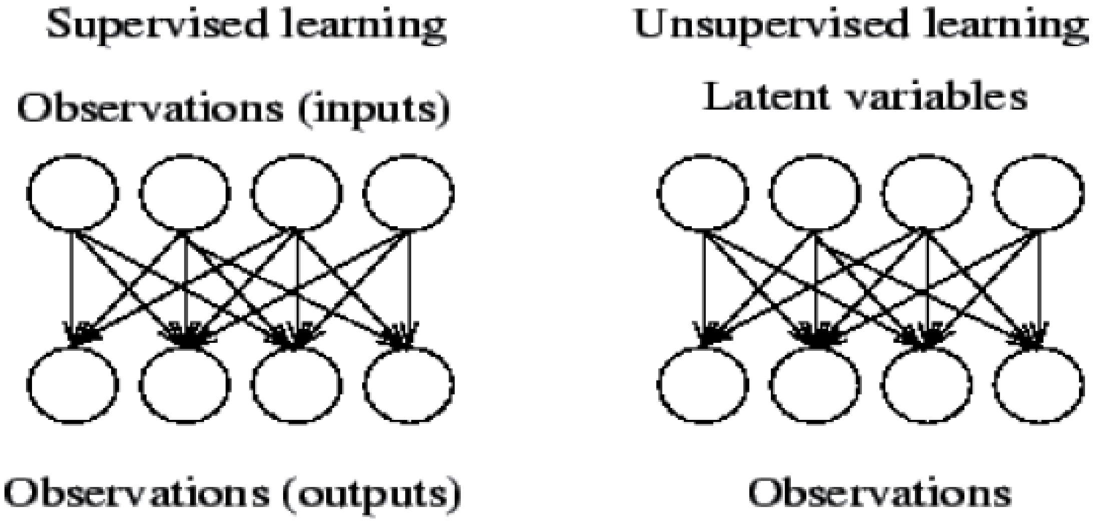


Figure 2.1: The Structures of Supervised and Unsupervised Learning

2.2.1.2 Unsupervised learning

Unsupervised learning's aim is the teaching of the computer to conduct tasks that are not unambiguously programmed. Unsupervised learning has two methods in general. The primary method is to train the model by not providing the class labels to each record in the dataset, instead, a form of reward system is given to designate accomplishment. This reward system aims to make judgments based on incentives maximization, instead of generating a classification system. The method well simplifies to the reality cases, in which the algorithm will be rewarded or punished depending on the actions taken. The actions of the algorithm rely on the preceding prizes and penalties devoid of essentially even studying any data regarding the precise behaviours that how the world is affected the actions taken. The information becomes unnecessary as the algorithm's action is determined by studying a reward function, the algorithm basically distinguishes the necessary actions without the need of undergoing any process. This is for the reason that it identifies the precise reward it anticipates for accomplishing the correct actions. This will be extremely advantageous in situations where probability calculation is time exhausting. Unsupervised learning is a tedious learning algorithm to train by undergoing the lengthy trial and error process, however, unsupervised learning is useful in situation where there are no pre-determined classification samples.

2.2.2 Algorithm Types

For this research, four main classification algorithms that are going to be discussed are:

- Random Forest
- Catboost
- Graph-based semi supervised learning (SSL)
- XGBoost

2.2.2.1 Random Forest

Random forests is a classic machine learning technique that utilized decision tree as its underlying mechanism for classification and regression that works by the construction a multitude of decision trees during the training time and the class that is the highest frequency of the classes or the average prediction of the individual trees are outputted(Ho, 1995). The shortcoming of decision tree, that is overfitting issue, is successfully mitigated in Random Forest algorithm (Mining, 2017).

The debut of random forests algorithm is invented by Tin Kam Ho, by employing the casual subspace technique, it is a technique to employ the "stochastic discrimination" method to classification (Kleinberg, 2000).

The universal technique of random forests was first suggested by Ho in 1995(Ho, 1995). He built a forest of trees splitting with tilted hyperplanes able increase precision by cultivating without having the problem of overfitting like decision tree.

The thorough analysis of random forests proper was completed in a study conducted by Breiman (2001). This paper defines the construction of a forest of unrelated trees by employing a CART process, blended with bagging and randomized node optimization. In addition, several crucial ingredients are blended that formulate the foundation of the current random forest algorithm, they are: estimation of the generality miscalculation by employing the out-of-bag error and application of permutation for the measurement of significant attribute.

2.2.2.2 CatBoost

CatBoost is a gradient boosting based on the decision trees (GBDT) machine learning model. CatBoost is built by Yandex scientists and software engineers. It is extensively employed in Kaggle competition, forecasting and countless other duties at huge companies, including CERN, Cloudflare, Careem taxi (Prokhorenkova et al., 2018). The advantages of CatBoost includes:

(1) Great quality with default parameter settings

Time consumed on parameter tuning is greatly reduced, as it delivers terrific outcomes by using the default settings.

(2) Categorical features support

Training outcomes with CatBoost can be improved as it accepts the usage of categorical attributes, extra data pre-processing is not needed which may be time-consuming and energy converting it to digits.

(3) Rapid and ascendable GPU usage

Prediction model can be trained on a rapid execution of gradient-boosting model for GPU.

(4) Superior precision

Overfitting is greatly reduced if building the prediction algorithms with a unique gradient-boosting architecture.

(5) Speedy prognostication

Trained model can be applied swiftly and competently even to latency-sensitive duties using Cat Boost's model.

Two key algorithmic innovations presented in CatBoost are the enactment of ordered boosting, a mutation-propelled replacement to the classic algorithm, and a pioneering algorithm for treating categorical features. Both techniques were designed to fight a prediction shift triggered by a unique kind of target leakage appear in all presently existing executions of gradient boosting algorithms. The figure below shows the pseudocode and technical details of CatBoost algorithm:

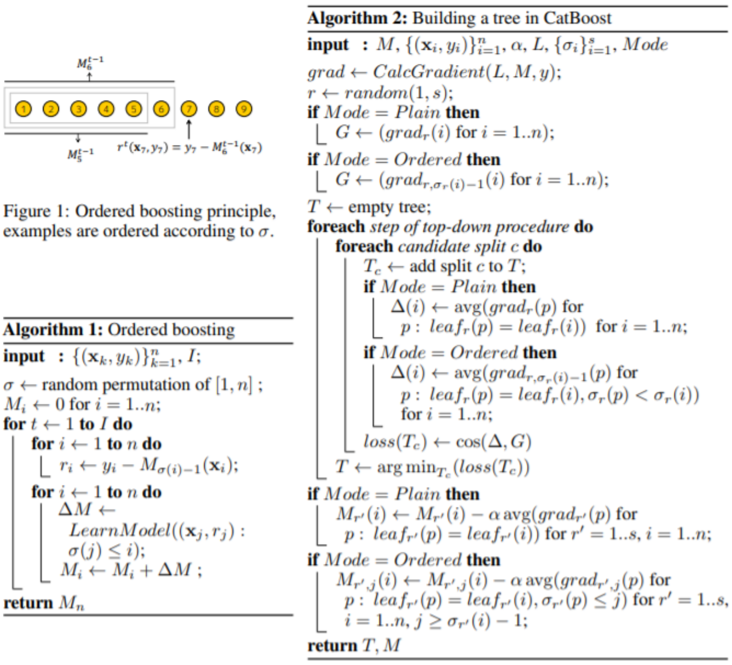

M_6^{t-1}
1
2
3
4
5
6
7
8
9
M_5^{t-1}
$r^t(\mathbf{x}_7, y_7) = y_7 - M_6^{t-1}(\mathbf{x}_7)$

Figure 1: Ordered boosting principle, examples are ordered according to σ.

Algorithm 1: Ordered boosting

```
input  : ${(\mathbf{x}_k, y_k)}_{k=1}^n$, $I$;
$\sigma \leftarrow$ random permutation of $[1, n]$ ;
$M_i \leftarrow 0$ for $i = 1..n$;
for $t \leftarrow 1$ to $I$ do
    for $i \leftarrow 1$ to $n$ do
        $r_i \leftarrow y_i - M_{\sigma(i)-1}(\mathbf{x}_i)$;
    for $i \leftarrow 1$ to $n$ do
        $\Delta M \leftarrow$
          $LearnModel((\mathbf{x}_j, r_j) : \sigma(j) \leq i)$;
        $M_i \leftarrow M_i + \Delta M$ ;
return $M_n$
```

Algorithm 2: Building a tree in CatBoost

```
input  : $M$, ${(\mathbf{x}_i, y_i)}_{i=1}^n$, $\alpha$, $L$, ${\sigma_i}_{i=1}^s$, $Mode$
$grad \leftarrow CalcGradient(L, M, y)$;
$r \leftarrow random(1, s)$;
if $Mode = Plain$ then
    $G \leftarrow (grad_r(i)$ for $i = 1..n)$;
if $Mode = Ordered$ then
    $G \leftarrow (grad_{r,\sigma_r(i)-1}(i)$ for $i = 1..n)$;
$T \leftarrow$ empty tree;
foreach step of top-down procedure do
    foreach candidate split $c$ do
        $T_c \leftarrow$ add split $c$ to $T$;
        if $Mode = Plain$ then
            $\Delta(i) \leftarrow$ avg$(grad_r(p)$ for
              $p : leaf_r(p) = leaf_r(i))$ for $i = 1..n$;
        if $Mode = Ordered$ then
            $\Delta(i) \leftarrow$ avg$(grad_{r,\sigma_r(i)-1}(p)$ for
              $p : leaf_r(p) = leaf_r(i), \sigma_r(p) < \sigma_r(i))$
              for $i = 1..n$;
        $loss(T_c) \leftarrow \cos(\Delta, G)$
    $T \leftarrow \arg\min_{T_c}(loss(T_c))$
if $Mode = Plain$ then
    $M_{r'}(i) \leftarrow M_{r'}(i) - \alpha$ avg$(grad_{r'}(p)$ for
      $p : leaf_{r'}(p) = leaf_{r'}(i))$ for $r' = 1..s$, $i = 1..n$;
if $Mode = Ordered$ then
    $M_{r',j}(i) \leftarrow M_{r',j}(i) - \alpha$ avg$(grad_{r',j}(p)$ for
      $p : leaf_{r'}(p) = leaf_{r'}(i), \sigma_{r'}(p) \leq j)$ for $r' = 1..s$,
      $i = 1..n$, $j \geq \sigma_{r'}(i) - 1$;
return $T, M$
```

Figure 2.2: CatBoost Technical Details

A short-range weather condition predictive algorithm based on wavelet denoising and Catboost is recommended by (Diao, Niu, Zang, & Chen, 2019). The application of heatmap, recursive style feature removal and tree algorithms are fused for the selection of important features. The validation outcomes reveal that the recommended paradigm can attain quicker union time and greater accuracy assessing with several other machine learning techniques such as LSTM, Seq2Seq and Random Forest.

A Catboost-based structure is recommended to foretell social media reputation by Kang et al. (2019). Catboost is implemented as the regression paradigm which is trained by using post-associated, user associated and supplementary user data. Moreover, to fully utilize the dataset for model preparation, a dataset augmentation approach founded on pseudo labels is recommended. The recommended technique accomplishes the runner up position in the Grand Challenge of Social Media Prediction.

2.2.2.3 XGBoost

XGBoost (Extreme Gradient Boosting) is a type of gradient boosting decision tree (GBDT) machine learning algorithm with gradient boosting structure. It is a gradient boosting library in python based on decision trees. XGBoost is designed by Tianqi (Tianqi & Carlos, 2016) that the system features and optimizations in XGBoost have deliver it 10 times quicker than the most sought after machine learning solutions.

XGBoost is a boosting algorithm extensively utilized in various machine learning competitions to enhance model accuracy and robustness. Its goal is to exceed the maximum of machines computation in order to deliver a compact, adaptable and precise large scaling tree boosting algorithm. XGBoost is the most prominent machine learning algorithm today because it can be utilized in broad range of applications including both classification and regression, runs efficiently on various operating system and supports major programming languages.

2.2.2.4 Graph-based Semi Supervised Learning

It is defined as a unique category of machine learning algorithm that employ unlabelled data for model training – normally a minor quantity of labelled records and a substantial quantity of unlabelled records (Chapelle & Schölkopf, 2009). It is categorized as a fusion of both unsupervised learning and supervised learning. Numerous data scientist discovered that unlabelled records, significant improvement of learning accuracy can be achieved if minor quantity of labelled records are applied. The possession of labelled records from a data source frequently needs a competent expert (e.g. to transliterate an audio portion) or to ascertain the 3D composition of a protein or verifying whether there is oil at a specific site). The expense linked with the classifying procedure causes a completely labelled training set impractical, the possession of unlabelled data is reasonably cost-effective. In such circumstances, semi-supervised learning can prove to be handy.

2.3 Confusion Matrix

Confusion matrix, is also called the error matrix, it is defined as a detailed table outline portraying the vision of the output generated by a classification-based model(Stehman, 2018). The table row exemplifies the cases in a predicted class, meanwhile the column symbolizes cases in an actual class (or conversely). The major advantage it brings is to make a simple visualization to observe whether the model is mystifying two classes (i.e. generally mislabelling the classes oppositely)(Powers, 2007).

Confusion matrix is a unique visualization tool that is made of two elements (actual outcome and predicted outcome. It is applied to explain the performance output of a classification algorithm when handling a dataset, where the actual class labels are recognized. In addition, it functions as a review of prediction findings on a classification challenge. Each and every square of the confusion matrix represents the number of mis predicted records and correctly predicted records in an aggregated number. It reveals the ways how our classification model is mystified when it is predicting the outcome of test dataset. It delivers the vision into the mistakes caused by a model but more crucially which types of mistake the model has conducted. It accepts easy recognition of misunderstanding in the middle of classes. For example, one class is frequently mislabelled as the another. A variety of performance metrics can calculated from the confusion matrix (Fawcett, 2006). They are:

Classification Rate/Accuracy:

It is the proportion of quantity of accurate predictions to the complete quantity of input cases. It presumes identical expenses for both types of inaccuracies. A 99% accuracy of the model can be defined outstanding, decent, average, inferior or horrible model depending the nature of the problem.

Recall:

Recall is defined as the proportion of the full amount of properly categorized positive cases divided by the overall quantity of positive cases. High recall implies the class is accurately identified (small number of FN).

Precision:

Precision is defined as the overall amount of properly categorized positive cases is divided by the overall number of predicted positive cases. High Precision reveals a case labelled as positive is undeniably positive (small number of FP).

F-measure:

F score is defined as a type of measurement in term of a model's accuracy. It is the product of precision and the recall. It considers both the precision and the recall of the test to compute the score. There are two types of elements (Precision and Recall) which aids to have a formula that symbolizes both elements. An F-measure is calculated which consumes Harmonic Mean in place of Arithmetic Mean as it penalizes the extreme values more.

2.4 Adoption of Feature Selection on Breast Cancer Survivability Detection

There have been numerous previous studies that conducted feature selection technique on breast cancer diagnosis by different authors. The selection of significant features from mammogram images for breast cancer diagnosis conducted by Eltoukhy and Faye (2014) using multiresolution representation. The feature abstraction is conducted using multiresolution representations (wavelet,curvelet) from the electronic mammograms images. The technique begins with both conducting wavelet and curvelet transform on the mammogram images. The recommended approach is able to locate an suitable feature set that produce substantial enhancement in categorization precision.The acquired findings were reasonable and the accomplishments of both wavelet and curvelet are produced and contrasted.

On the other hand, the identification of significant metabolic features related to breast cancer pathogenesis by summarizing of RNA metabolites with SVM based feature selection is applied by Henneges et al. (2009). A useable set of urinary metabolites is chosen by rejection of all entrants with inferior linearity in the analytic configuration. The bioinformatic instrument of Oscillating Search Algorithm for Feature Selection (OSAF) is utilized to repeatedly enhance features for preparation of Support Vector Machines (SVM) to improve classification accuracy. The permutation of group spectrometric assessment and successive SVM-based feature selection symbolizes an encouraging apparatus for the advancement of a non-intrusive prognostication model.

Huang, Hung, and Chen (2010) employed neural network algorithm with feature selection based on entropy on breast cancer analysis for Wisconsin Breast Cancer Dataset. An ensemble of the condensed dataset patterns grounded on feature selection is applied to model an artificial neural network (ANN) by means of the Levenberg–Marquardt (LM) and the Particle Swarm Optimization (PSO) algorithms to develop the suitable ANN preparation weighting constraints, and then build a good Neural Network algorithm to advance the Wisconsin Breast Cancers' categorization precision and effectiveness. Research findings indicate that the AROC and accuracy progressed emphatically, and the top performance in accuracy and AROC are 98.83% and 0.9971, correspondingly.

2.5 Adoption of Machine Learning in Breast Cancer Survivability Detection

The authors of (Delen et al., 2005) employed two classical machine learning models, namely ANN and decision tree, collectively with a conventional statistical approach, logistic regression, to build models for breast cancer survivability prediction. The SEER breast cancer dataset from 1973 to 2000 is utilized to train the models. The decision tree is concluded to be the best predictive model out of the rest of the models by attaining the greatest performance of 93.62% accuracy.

In a study conducted by Cruz and Wishart (2006), the researchers performed an extensive examination of distinct machine learning algorithms, reviewing matters associated to the kinds of data integrated and the implementation of these methods in breast cancer detection and forecasting. This analysis offers comprehensive justifications guiding to first-rate study standards for the usage of machine learning algorithms to cancer prediction.

An improvement is made on the decision trees to enhance the prediction accuracy of breast cancer survivability is portrayed in (Khan et al., 2007). The researches recommend a fusion predictive method based on weighted fuzzy decision trees (FDT). The resultant outcome of the study concludes the values for AUC are 0.69 for fuzzy decision trees, and 0.77 for weighted fuzzy decision trees. However other important performance metrics such as F1 score, accuracy and recall value were not measured.

Support vector machine (SVM) algorithm are utilized to solve the breast cancer survivability prediction problem. In (Thongkam, Xu, Zhang, & Huang, 2009), the researchers recommended a fusion method to produce a superior quality dataset from Srinagarind hospital database, to build advanced breast cancer survival predictive models. The plan has two key steps: (a) application of an outlier-cleaning method based on C-support vector classification to eliminate anomalies from the datasets; and(b) over-sampling. The results indicated that this fusion scheme increased the performance of SVM significantly

An experimental work on breast cancer survivability detection can be discovered in (Society, 2013). In their research, a statistical algorithm known as the proportional hazards' regression model, is employed to breast cancer dataset to distinguish if the patient has endured the cancer successfully.

Wang, Zheng, Won, & Sang (2017) suggested a support vector machine-based ensemble classification model in prediction of breast cancer. To confront the restriction of individual model performance, this research concentrates on breast cancer analysis that utilizes an SVM- centred machine learning model to lessen the identification discrepancy and enhance identification precision. Twelve distinct SVMs, founded on the recommended Weighted Area Under the Receiver Operating Characteristic Curve Ensemble (WAUCE) method, are crossbred. The outcome of the study indicate that the WAUCE model accomplishes a greater accuracy with a substantially softer variation for breast cancer analysis contrasted to five other methods and two ordinary troupe methods, i.e., adaptive boosting and bagging classification tree. The recommended SVM-WAUCE model diminishes the discrepancy by 97.89% and boosts accuracy by 33.34%, when contrasted to the SVM model on the SEER dataset.

Kim and Shin (2013) developed an artificial neural network (ANN) based prediction model for breast cancer detection. The ANN structural design known as multi-layer perceptron with back-propagation algorithm are employed as the foundation of the predictive model. The ANN includes three categories of layers: the input layer, hidden layers, and the output layer. The nodes in the input layer deliver incoming waves to the nodes in the concealed layer through weighted connections. The total outcome of the model is signified by the nodes in the output layer which transmit output signs on the core of a transfer function. The accuracy of the model heavily dependent on the structure changes. For example, the number of hidden nodes, and the preliminary weights coupled with the links among the nodes. Commonly, the quantity of hidden nodes is chosen by trial-and-error manner and the primary weights are arbitrarily selected.

Li & Chen (2018) studied and evaluated the accomplishment of five different classification algorithms on the breast cancer detection, they are: Random Forest, Logistic Regression, Decision Tree, Support Vector Machine, and Artificial Neural Network.

Three performance metrics used are accuracy, F-measure and AUC values. Their study concluded that random forest algorithm is the best performing algorithm. They also suggested that random forest can be merged with other data mining tools to achieve better precision and useful outcomes in the future research.

In a research organized by Endo et al. (2018), comparison of seven classification algorithms such as logistic regression model, artificial neural network, Naïve Bayes, Bayes Net, Decision Trees with naïve Bayes, Decision Trees (ID3) and Decision Trees (J48), are conducted to determine the breast cancer survivability of the patients. In their study, the concluded that Logistic Regression model displayed the greatest accuracy. the J48 had the greatest sensitivity and the ANN produced the greatest specificity. The Decision Trees models have a tendency to exhibit great sensitivity. Meanwhile, the Bayesian models were apt to demonstrate the accuracy going up. They also discovered that the optimum algorithm may differ based on the predicted objects and dataset.

In a study conducted by Edeki and Shardul (2012), classification models such as decision tree, logistic regression, artificial neural network. Support vector machine and Random Forest are applied to predict the breast cancer survivability using SEER breast cancer dataset. The result shows Random Forest is the best performing model in term of overall accuracy. However, the study is not conducted with GBDT models and the specificity score is relatively low.

Park et al. (2015) proposed a robust predictive model for assessing the breast cancer survival rate by comparing three major machine learning models for the diagnosis of breast-cancer survivability. They are: support vector machine (SVM), artificial neural network (ANN) and semi-supervised learning (SSL) models. They proposed that semi-supervised learning model is a good contender that medical specialists easily utilize without exhausting the time and work for parameter exploring for a specific model. This is because SL performed the best in term of overall accuracy, sensitivity, specificity ad AUC. The simplicity of usage and swifter time to outcomes of the predictive model will ultimately lead to the precise and less-intrusive prognosis for breast cancer patients. However, the performance of their proposed model is not compared with GBDT algorithms.

Table 2.1:Past Publications on SEER Breast Cancer Survivability Prediction

Publication	Machine Learning Method	Accuracy	Validation method	Important features	Limitation of the Study
(Burke et al., 1997)	Artificial Neural Network	73%	Not specified	Not specified	Tested on small dataset with few attributes
(Endo et al., 2010)	Logistic Regression	85%	10-fold cross-validation	Not specified	Low specificity of the model
(Edeki & Shardul, 2012)	Random Forest	75%	10-fold cross-validation	Not specified	Only use accuracy as performance metric
(Park et al., 2015)	Graph-Based SSL	71%	5-fold cross validation	Not specified	Only 5 -fold cross validation is used
(Shukla, Hagenbuchner, Win, & Yang, 2018)	Artificial Neural Network + Clustering	69%	10-fold cross-validation	Not specified	Clustering method caused significant data compression

2.6 Discussion

Based on our literature review, it was found that there are research gaps to be studied for data mining techniques on breast cancer detection in previous studies:

- The performance of Gradient Boosting Decision Tree algorithm in breast cancer survivability prediction is yet to be investigated.
- There is no identification of significant features in past researches.

Among all the classification algorithm studied in past paper, the Random Forest algorithm, and semi supervised learning are the best performing algorithm of breast cancer survivability detection, however their study did not identify the significant attributes that greatly affect the performance of the prediction model. Meanwhile, the performance of Gradient Boosting Decision Tree algorithm in breast cancer survivability prediction is yet to be investigated as past researches mostly focus on the study of classic models such as Decision Tree, Naïve Bayes, Logistic regression and etc. The performance of Gradient Boosting Decision Tree algorithms has excelled over other classical classification models such as artificial neural network in particle identification (Roe et al., 2005) and brain tumour (B, Zeng, Sotiras, & Rathore, 2016). However, the performance of GBDT algorithm in breast cancer survivability is yet to be identified.

Meanwhile the prediction accuracy of the proposed model will be compared against models proposed by previous studies. The dataset that is used in current research is relatively complete and is retrieved from SEER database.

Chapter 3 : METHODOLOGY

3.1 Introduction

To recap the purpose of this study:

- To review the previous common practice of machine learning methods and to propose an improved prediction model.
- To identify the key factor for causing breast cancer survivability through feature importance score.

3.2 Research Workflow

The figure below illustrates the experiment workflow for building an improved prediction model. It starts from data collection from the patients. Then raw data is pre-processed and cleaned before proper analysis can be done. The data analysis is a crucial preliminary step for us to understand the dataset better. After that, data modelling process is the core study of our research to build a better breast cancer survivability prediction model. Then the models will be validated by feeding with test data. Performance metrics such as accuracy, sensitivity and f- measure will be utilized to compare the performance of every models.

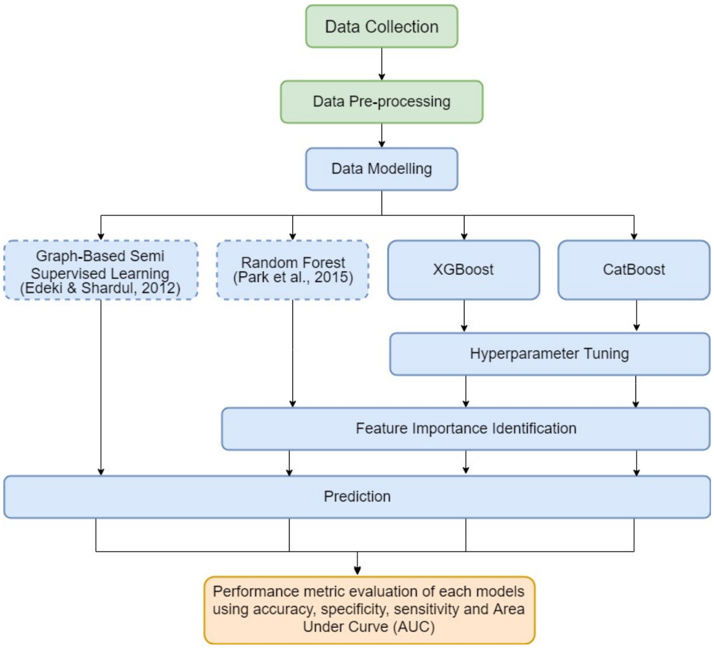

Data Collection
Data Pre-processing
Data Modelling
Graph-Based Semi Supervised Learning (Edeki & Shardul, 2012)
Random Forest (Park et al., 2015)
XGBoost
CatBoost
Hyperparameter Tuning
Feature Importance Identification
Prediction
Performance metric evaluation of each models using accuracy, specificity, sensitivity and Area Under Curve (AUC)

Figure 3.1: Research Workflow of the Experiment

The following procedure describes the step by step process from data collection to classification outcome.

Step 1: Data collection from SEER breast cancer database

Step 2: Data Pre-processing on missing values, data format and outlier detection

Step 3: Data modelling using four different classification models (default settings)

Step 4: Identification of important features from decision tree-based algorithms

Step 5: Performance assessment of predictive models using testing test using performance metrics such as accuracy, specificity, sensitivity and Area Under Curve (AUC).

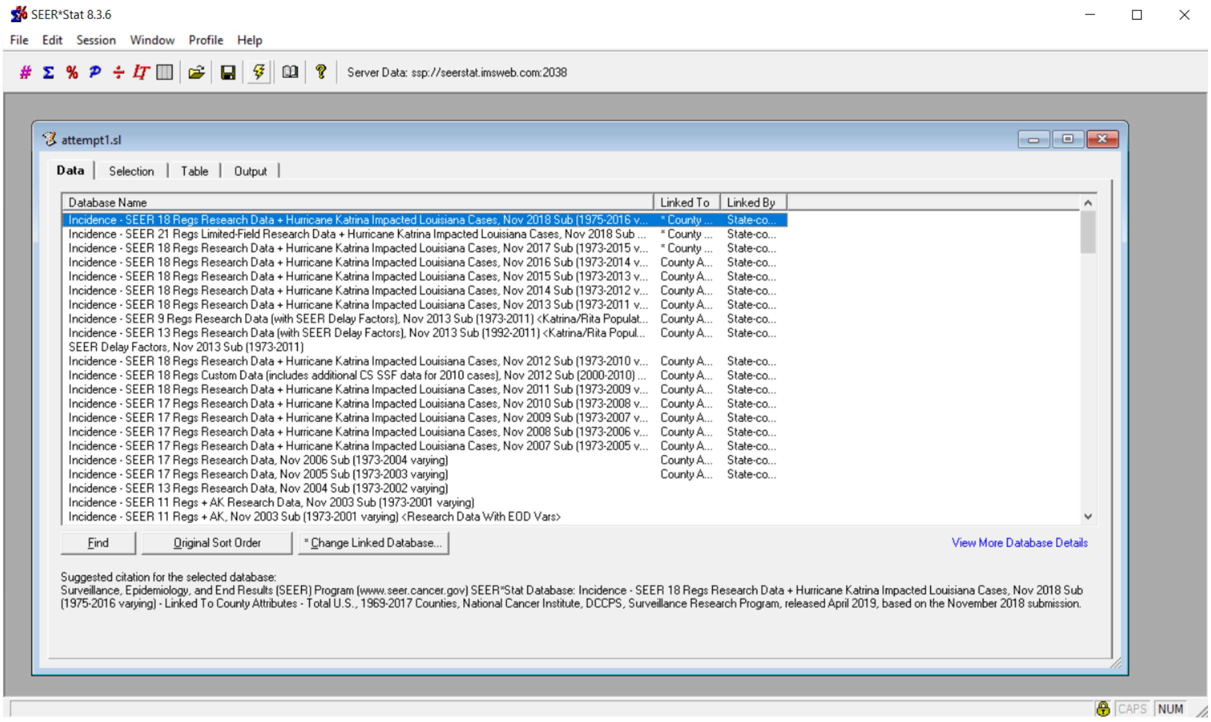
Step 6: Repeat steps 3-5 for Gradient Boosting Decision Tree algorithms (CatBoost and XGBoost) using optimized parameters selected with random search method.

3.2.1 Data Collection

The Surveillance, Epidemiology, and End Results (SEER) program is a trustworthy resource for cancer data on cancer incidence and survival in the United States. The SEER program offers an extensive database on cancer statistics to lower the cancer liability among the U.S. citizen. SEER is endorsed by the Surveillance Research Program (SRP) in National Cancer Institute (NCI)'s Division of Cancer Control and Population Sciences (DCCPS). SEER presently gathers and distributes the cancer incidence and survival data from different cancer archives involving nearly 34.6 percent of the U.S. population. SEER analysis comprises 31.9 percent of Whites, 30.0 percent of African Americans, 44.0 percent of Hispanics, 49.3 percent of American Indians and Alaska Natives, 57.5 percent of Asians, and 68.5 percent of Hawaiian/Pacific Islanders (SEER, 2019).

The SEER Program archives are regularly gathering data on patient demographics, major tumour spot, tumour morphology and phase at diagnosis, first treatment period, and updates for essential condition check-up. The SEER Program is the only comprehensive resource of population-based database in the United States that incorporates phase of cancer at the moment of diagnosis and patient survival statistics. The death figures stated by SEER are supplied by the National Centre for Health Statistics. The population data applied in evaluating cancer rates is attained intermittently from the Census Bureau. The registry and database are modernized yearly and offered as a public service in print or electronic copies, SEER data are utilized by thousands of researchers, public health officials, policymakers, and the public to understand the cancer statistics and perform data analysis on it.

To access the SEER database and extract required dataset for the research project, a signed SEER Research Data Agreement application is essential to gain access to the SEER database. The SEER program will handle the application within 2 business days of accepting the signed agreement. Once successful application is accepted by SEER, a username and password is created and user is required to download their statistical software, called SEER*Stat. This statistical software is powerful for users to filter and extract all the necessary data they needed. The Figure 3.2 below shows the screenshot of the user interface of SEER*Stat software.


SEER*Stat 8.3.6
File Edit Session Window Profile Help
attempt1.sl
Data Selection Table Output
Find
Original Sort Order
* Change Linked Database...
View More Database Details
CAPS NUM

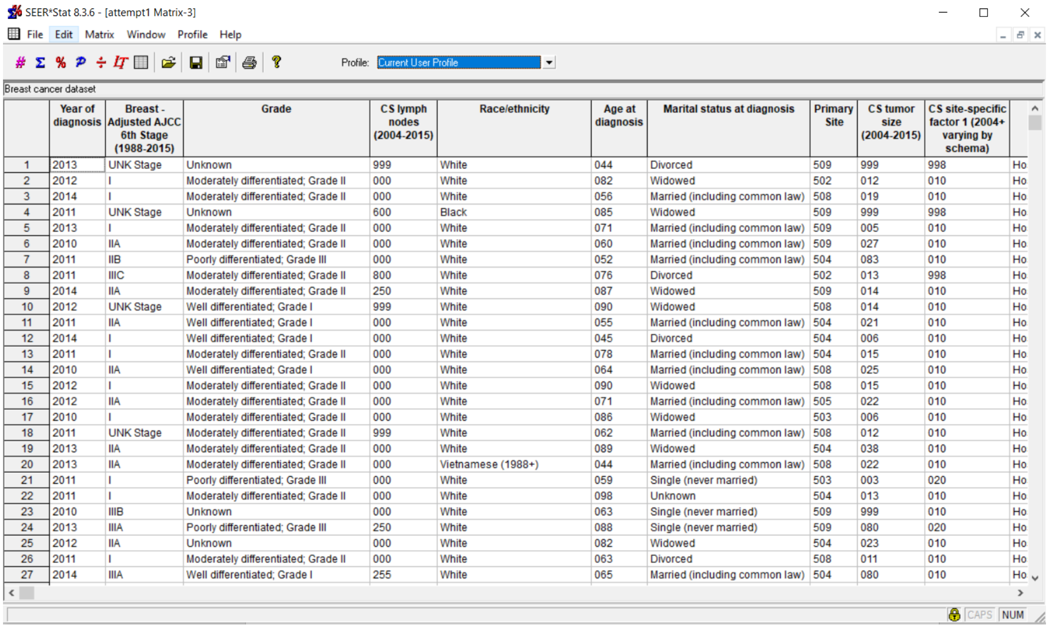
Server Data: ssp://seerstat.imsweb.com:2038

Database Name	Linked To	Linked By
Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2018 Sub (1975-2016 v...	* County ...	State-co...
Incidence - SEER 21 Regs Limited-Field Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2018 Sub ...	* County ...	State-co...
Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2017 Sub (1973-2015 v...	* County ...	State-co...
Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2016 Sub (1973-2014 v...	County A...	State-co...
Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2015 Sub (1973-2013 v...	County A...	State-co...
Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2014 Sub (1973-2012 v...	County A...	State-co...
Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2013 Sub (1973-2011 v...	County A...	State-co...
Incidence - SEER 9 Regs Research Data (with SEER Delay Factors), Nov 2013 Sub (1973-2011) <Katrina/Rita Populat...	County A...	State-co...
Incidence - SEER 13 Regs Research Data (with SEER Delay Factors), Nov 2013 Sub (1992-2011) <Katrina/Rita Popul...	County A...	State-co...
SEER Delay Factors, Nov 2013 Sub (1973-2011)		
Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2012 Sub (1973-2010 v...	County A...	State-co...
Incidence - SEER 18 Regs Custom Data (includes additional CS SSF data for 2010 cases), Nov 2012 Sub (2000-2010) ...	County A...	State-co...
Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2011 Sub (1973-2009 v...	County A...	State-co...
Incidence - SEER 17 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2010 Sub (1973-2008 v...	County A...	State-co...
Incidence - SEER 17 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2009 Sub (1973-2007 v...	County A...	State-co...
Incidence - SEER 17 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2008 Sub (1973-2006 v...	County A...	State-co...
Incidence - SEER 17 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2007 Sub (1973-2005 v...	County A...	State-co...
Incidence - SEER 17 Regs Research Data, Nov 2006 Sub (1973-2004 varying)	County A...	State-co...
Incidence - SEER 17 Regs Research Data, Nov 2005 Sub (1973-2003 varying)	County A...	State-co...
Incidence - SEER 13 Regs Research Data, Nov 2004 Sub (1973-2002 varying)		
Incidence - SEER 11 Regs + AK Research Data, Nov 2003 Sub (1973-2001 varying)		
Incidence - SEER 11 Regs + AK, Nov 2003 Sub (1973-2001 varying) <Research Data With EOD Vars>		

Suggested citation for the selected database:
Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2018 Sub (1975-2016 varying) - Linked To County Attributes - Total U.S., 1969-2017 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2019, based on the November 2018 submission.

Figure 3.2: User Interface of SEER*Stat

The Figure 3.3 below shows the examples of attributes selected and the number of records.


SEER*Stat 8.3.6 - [attempt1 Matrix-3]
File Edit Matrix Window Profile Help
Profile: Current User Profile
CAPS NUM

Breast cancer dataset

	Year of diagnosis	Breast - Adjusted AJCC 6th Stage (1988-2015)	Grade	CS lymph nodes (2004-2015)	Race/ethnicity	Age at diagnosis	Marital status at diagnosis	Primary Site	CS tumor size (2004-2015)	CS site-specific factor 1 (2004+ varying by schema)	
1	2013	UNK Stage	Unknown	999	White	044	Divorced	509	999	998	Ho
2	2012	I	Moderately differentiated; Grade II	000	White	082	Widowed	502	012	010	Ho
3	2014	I	Moderately differentiated; Grade II	000	White	056	Married (including common law)	508	019	010	Ho
4	2011	UNK Stage	Unknown	600	Black	085	Widowed	509	999	998	Ho
5	2013	I	Moderately differentiated; Grade II	000	White	071	Married (including common law)	509	005	010	Ho
6	2010	IIA	Moderately differentiated; Grade II	000	White	060	Married (including common law)	509	027	010	Ho
7	2011	IIB	Poorly differentiated; Grade III	000	White	052	Married (including common law)	504	083	010	Ho
8	2011	IIIC	Moderately differentiated; Grade II	800	White	076	Divorced	502	013	998	Ho
9	2014	IIA	Moderately differentiated; Grade II	250	White	087	Widowed	509	014	010	Ho
10	2012	UNK Stage	Well differentiated; Grade I	999	White	090	Widowed	508	014	010	Ho
11	2011	IIA	Well differentiated; Grade I	000	White	055	Married (including common law)	504	021	010	Ho
12	2014	I	Well differentiated; Grade I	000	White	045	Divorced	504	006	010	Ho
13	2011	I	Moderately differentiated; Grade II	000	White	078	Married (including common law)	504	015	010	Ho
14	2010	IIA	Well differentiated; Grade I	000	White	064	Married (including common law)	508	025	010	Ho
15	2012	I	Moderately differentiated; Grade II	000	White	090	Widowed	508	015	010	Ho
16	2012	IIA	Moderately differentiated; Grade II	000	White	071	Married (including common law)	505	022	010	Ho
17	2010	I	Moderately differentiated; Grade II	000	White	086	Widowed	503	006	010	Ho
18	2011	UNK Stage	Moderately differentiated; Grade II	999	White	062	Married (including common law)	508	012	010	Ho
19	2013	IIA	Moderately differentiated; Grade II	000	White	089	Widowed	504	038	010	Ho
20	2013	IIA	Moderately differentiated; Grade II	000	Vietnamese (1988+)	044	Married (including common law)	508	022	010	Ho
21	2011	I	Poorly differentiated; Grade III	000	White	059	Single (never married)	503	003	020	Ho
22	2011	I	Moderately differentiated; Grade II	000	White	098	Unknown	504	013	010	Ho
23	2010	IIIB	Unknown	000	White	063	Single (never married)	509	999	010	Ho
24	2013	IIIA	Poorly differentiated; Grade III	250	White	088	Single (never married)	509	080	020	Ho
25	2012	IIA	Unknown	000	White	082	Widowed	504	023	010	Ho
26	2011	I	Moderately differentiated; Grade II	000	White	063	Divorced	508	011	010	Ho
27	2014	IIIA	Well differentiated; Grade I	255	White	065	Married (including common law)	504	080	010	Ho

Figure 3.3: The Extracted Data Matrix

The Table 3.1 illustrates each type of attributes and their respective attribute type and descriptions:

Table 3.1: SEER Breast Cancer Dataset

No	Attributes	Description	Level
1	Stage	Identified by volume of cancer tumour and its distribution	Categorical
2	Grade	Look of tumour and its resemblance to more-or less-aggressive tumours	Categorical
3	Lymph node involvement	None, (1–3) negligible, (4–9) substantial, etc	Categorical
4	Race	Ethnicity	Categorical
5	Age at diagnosis	Real age of patient	Numerical
6	Marital Status	Marital status of patient	Categorical
7	Primary site	Presence of tumor at site in body. Topographical classification of cancer	Categorical
8	Tumour size	2–5 cm; at 5cm diagnosis deteriorates	Numerical
9	Site-specific surgery	Evidence on operation during first course of therapy	Categorical
10	Radiation	None, beam radiation, radioisotopes, refused, recommended, etc	Categorical
11	Histological type	Form and structure of tumour	Categorical
12	Behaviour code	Normal or aggressive tumour behaviour	Categorical
13	Number of positive nodes examined	When lymph nodes are participating in cancer, they are called positive	Numerical
14	Number of nodes examined	Overall nodes(positive/negative) analysed	Numerical
15	Number of primaries	Number of primary tumours (1–6)	Categorical
16	Clinical extension of tumour	Describes propagation of tumour relative to breast	Categorical
17	Survivability	Target binary variable identifies grade of survival of patient: '+1' if survived longer than five years, '-1' otherwise	Categorical

The SEER dataset consists of 17 attributes and 162500 records. Out of the 17 attributes, 13 attributes are categorical variables and 4 attributes are numerical variables. Our target variable is the 'Survivability' attribute, where a binary classification prediction model will be built upon it.

3.2.2 Data Pre-processing

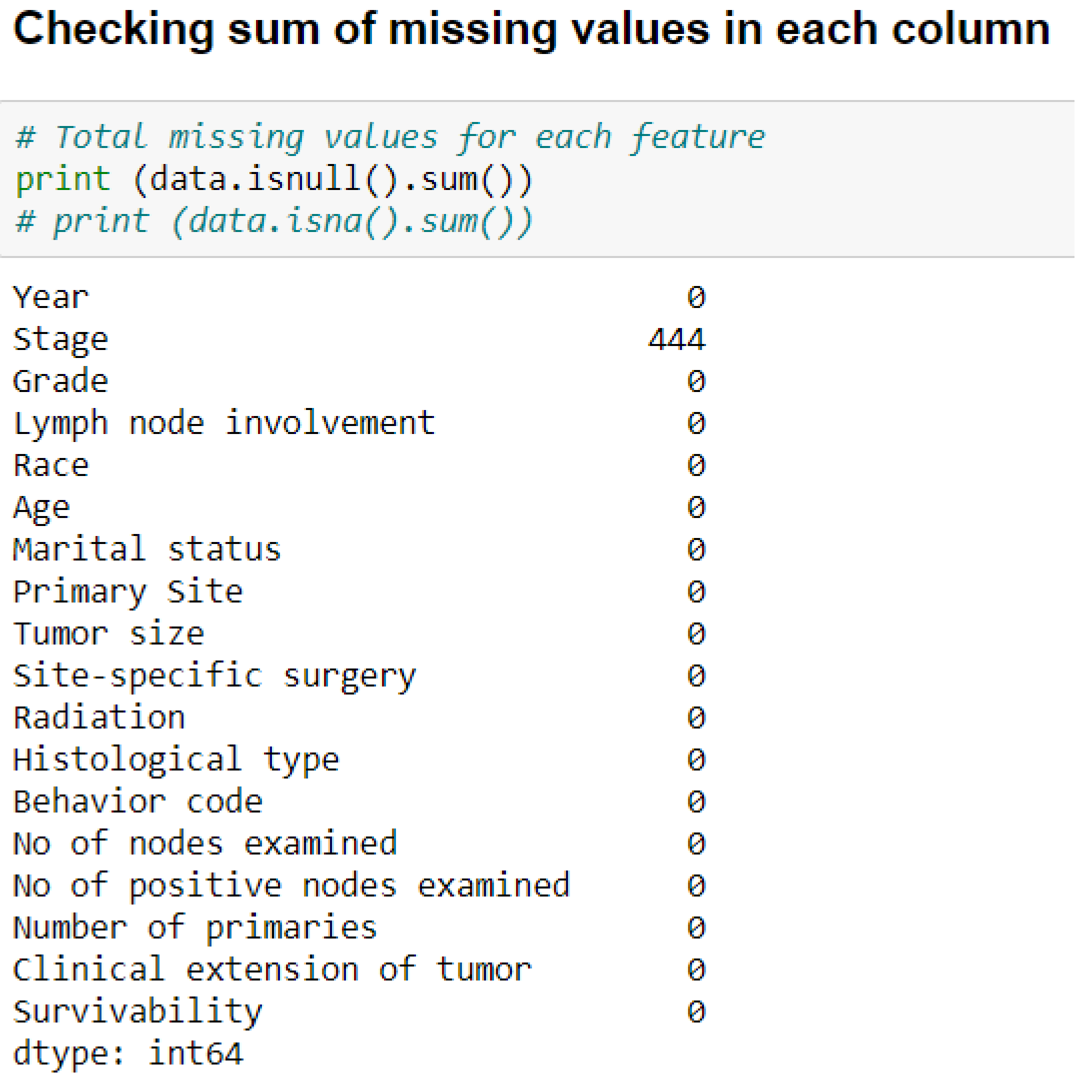
Steps Involved in Data Pre-processing:

1. Data Cleaning:

The raw data may have many irrelevant, missing values or noisy data. To handle this issue, a proper data cleaning process has to be completed to ensure cleaned quality data ready to be analysed. The steps involved in data cleaning are:

- **(a). Missing Data:**

  Occurred when there is missing value or incomplete data rows. It can be handled by using imputation



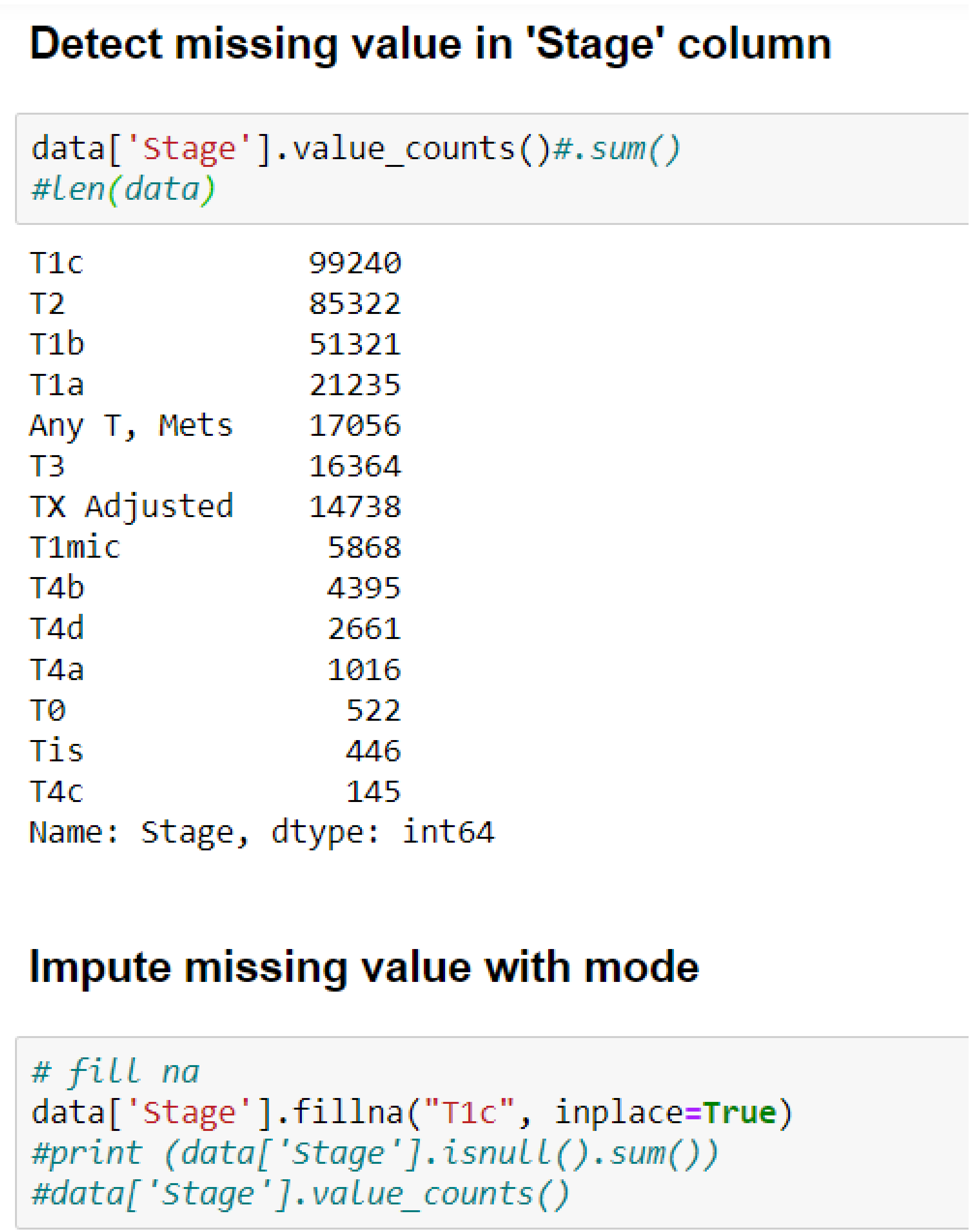
Checking sum of missing values in each column

```
# Total missing values for each feature
print (data.isnull().sum())
# print (data.isna().sum())
```

```
Year                               0
Stage                            444
Grade                              0
Lymph node involvement             0
Race                               0
Age                                0
Marital status                     0
Primary Site                       0
Tumor size                         0
Site-specific surgery              0
Radiation                          0
Histological type                  0
Behavior code                      0
No of nodes examined               0
No of positive nodes examined      0
Number of primaries                0
Clinical extension of tumor        0
Survivability                      0
dtype: int64
```

Figure 3.4: Missing Values of Dataset

The code in figure 3.4 above shows the checking of the number of missing values in each column. It is noticed that the dataset is clean except for the ‘Stage’ attribute, with 444 missing values.



Detect missing value in 'Stage' column

```
data['Stage'].value_counts()#.sum()
#len(data)
```

```
T1c             99240
T2              85322
T1b             51321
T1a             21235
Any T, Mets     17056
T3              16364
TX Adjusted     14738
T1mic            5868
T4b              4395
T4d              2661
T4a              1016
T0                522
Tis               446
T4c               145
Name: Stage, dtype: int64
```

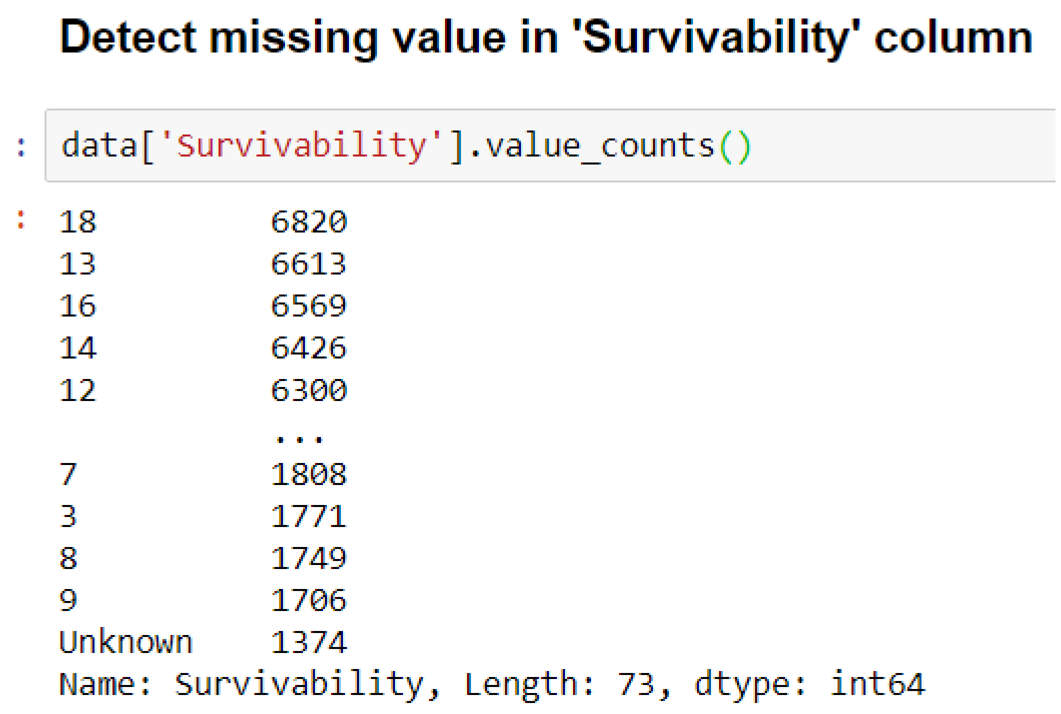
Impute missing value with mode

```
# fill na
data['Stage'].fillna("T1c", inplace=True)
#print (data['Stage'].isnull().sum())
#data['Stage'].value_counts()
```

Figure 3.5: Imputation with Mode

The figure 3.5 shows the highest frequency of the value in ‘Stage’ attribute is ‘T1C’. Therefore, the missing value of the ‘Stage’ attribute is imputed with mode instead of the mean as this attribute is categorical variable.

The figure below shows the distribution values of our target variable, ‘Survivability’ attribute. The attribute shows there are 1374 of unknown values. Therefore, the unknown values are imputed with mean since our target variable is numerical variable.

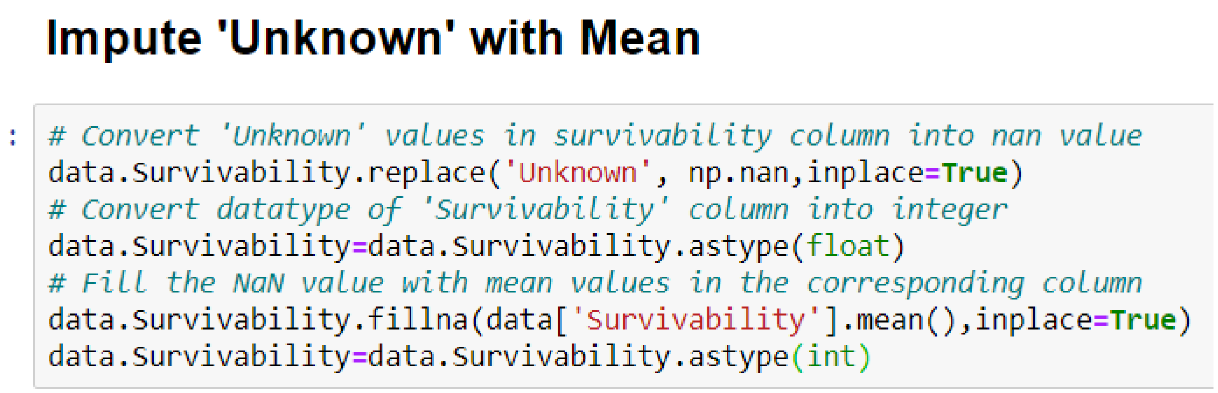


Detect missing value in 'Survivability' column

```
data['Survivability'].value_counts()
```

```
18         6820
13         6613
16         6569
14         6426
12         6300
           ...
7          1808
3          1771
8          1749
9          1706
Unknown    1374
Name: Survivability, Length: 73, dtype: int64
```

Figure 3.6: Missing Values in 'Survivability' Attribute



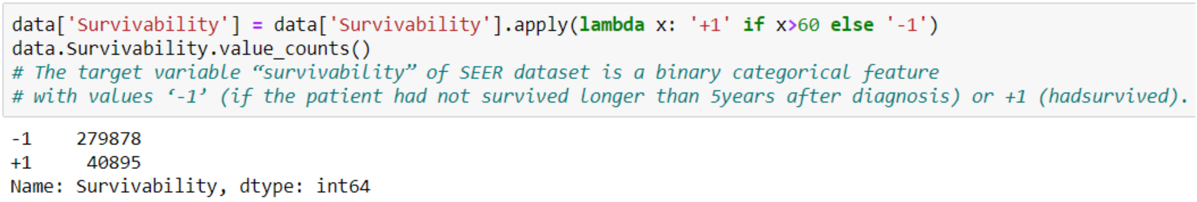
Impute 'Unknown' with Mean

```
# Convert 'Unknown' values in survivability column into nan value
data.Survivability.replace('Unknown', np.nan,inplace=True)
# Convert datatype of 'Survivability' column into integer
data.Survivability=data.Survivability.astype(float)
# Fill the NaN value with mean values in the corresponding column
data.Survivability.fillna(data['Survivability'].mean(),inplace=True)
data.Survivability=data.Survivability.astype(int)
```

Figure 3.7: Imputation with Mean

Imputation of ‘unknown’ is first converted to nan value, and the data type of ‘Survivability’ attribute is converted to float type before mean is calculated. At the end, the data type of ‘Survivability’ attribute is converted to the right data type, which is integer type.

The ‘Survivability’ attribute consists of the number of months each patient managed to survive before they deceased. Therefore, the data in the attribute need to be transformed into binary outcome, where if the patient managed to survive more than 5 years (60 months), Survivability=1, else Survivability =0 if survived less than 5 years. The figure below shows the data value transformation.



```
data['Survivability'] = data['Survivability'].apply(lambda x: '+1' if x>60 else '-1')
data.Survivability.value_counts()
# The target variable “survivability” of SEER dataset is a binary categorical feature
# with values ‘-1’ (if the patient had not survived longer than 5years after diagnosis) or +1 (hadsurvived).
```

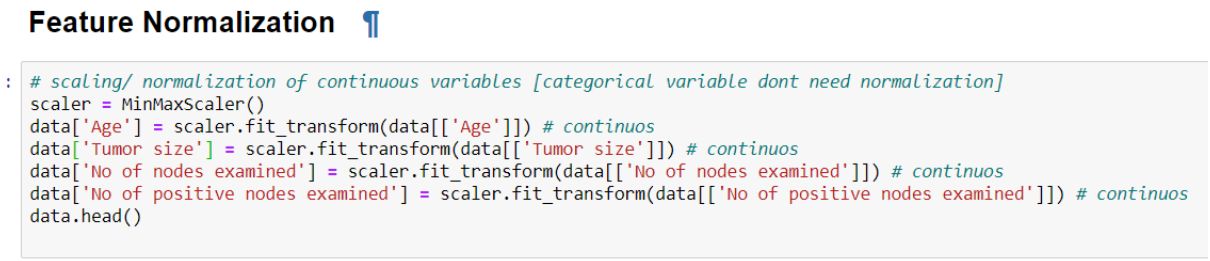
```
-1    279878
+1     40895
Name: Survivability, dtype: int64
```

Figure 3.8: Data Value Transformation

2. Feature Normalization

The objective of normalization is to shift the values of numeric attributes in the dataset to a mutual scale, without altering changes in the scales of values. Most of the times, your dataset will include features highly contrasting in scales, units and scope. Due to the fact that machine learning algorithms utilize Eucledian distance among two data points in their calculations, this would cause biased result in prediction model. The algorithms which use Euclidean Distance measure are sensitive to Magnitudes. Here feature normalization aids to weigh all the features uniformly.

If left unmanaged, these algorithms will only consider the magnitude of attributes overlooking the units. The outcomes would fluctuate greatly between different units, 5kg and 5000gms. The features with high magnitudes will carry more weightage in the distance computations than features with low magnitudes.



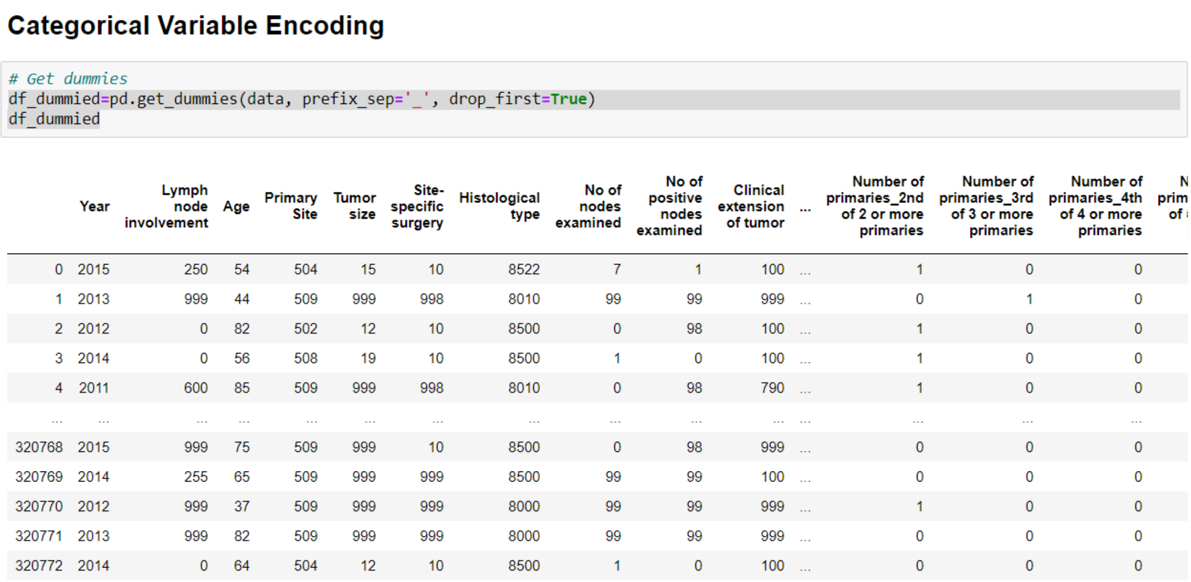
Feature Normalization ¶

```
# scaling/ normalization of continuous variables [categorical variable dont need normalization]
scaler = MinMaxScaler()
data['Age'] = scaler.fit_transform(data[['Age']]) # continuos
data['Tumor size'] = scaler.fit_transform(data[['Tumor size']]) # continuos
data['No of nodes examined'] = scaler.fit_transform(data[['No of nodes examined']]) # continuos
data['No of positive nodes examined'] = scaler.fit_transform(data[['No of positive nodes examined']]) # continuos
data.head()
```

Figure 3.9: Feature Normalization of Attributes

3. Categorical Variable Encoding

In general, machine learning models does not work on labelled data immediately. They expect all input attributes and output attributes to be numerical. Therefore, this constrains the execution of machine learning models rather than hard drawbacks on the algorithm itself.



Categorical Variable Encoding

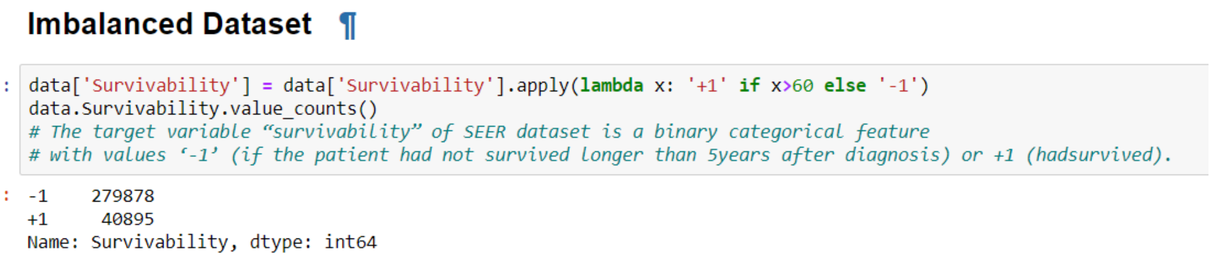
```
# Get dummies
df_dummied=pd.get_dummies(data, prefix_sep='_', drop_first=True)
df_dummied
```

	Year	Lymph node involvement	Age	Primary Site	Tumor size	Site-specific surgery	Histological type	No of nodes examined	No of positive nodes examined	Clinical extension of tumor	...	Number of primaries_2nd of 2 or more primaries	Number of primaries_3rd of 3 or more primaries	Number of primaries_4th of 4 or more primaries	N prim of
0	2015	250	54	504	15	10	8522	7	1	100	...	1	0	0	
1	2013	999	44	509	999	998	8010	99	99	999	...	0	1	0	
2	2012	0	82	502	12	10	8500	0	98	100	...	1	0	0	
3	2014	0	56	508	19	10	8500	1	0	100	...	1	0	0	
4	2011	600	85	509	999	998	8010	0	98	790	...	1	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
320768	2015	999	75	509	999	10	8500	0	98	999	...	0	0	0	
320769	2014	255	65	509	999	999	8500	99	99	100	...	0	0	0	
320770	2012	999	37	509	999	999	8000	99	99	999	...	1	0	0	
320771	2013	999	82	509	999	999	8000	99	99	999	...	0	0	0	
320772	2014	0	64	504	12	10	8500	1	0	100	...	0	0	0	

Figure 3.10: Categorical Variable Encoding

The figure above shows the encoded dataset where all the categorical variables are transformed from character values to integer values. This causes expansion of the dataset from 17 attributes to 83 attributes.

4. Imbalanced Dataset



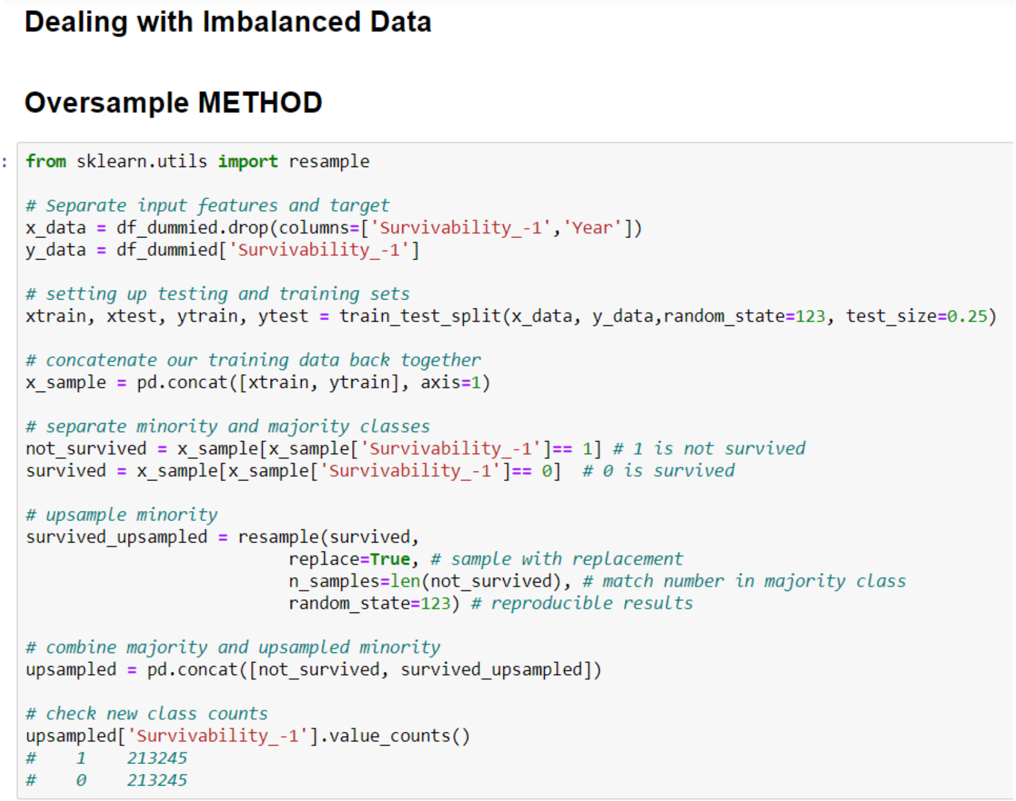
Imbalanced Dataset ¶

```
: data['Survivability'] = data['Survivability'].apply(lambda x: '+1' if x>60 else '-1')
  data.Survivability.value_counts()
  # The target variable “survivability” of SEER dataset is a binary categorical feature
  # with values ‘-1’ (if the patient had not survived longer than 5years after diagnosis) or +1 (hadsurvived).
```

```
: -1     279878
  +1      40895
  Name: Survivability, dtype: int64
```

Figure 3.11: Imbalanced Class Label

Imbalanced dataset is a general challenge in machine learning classification problem where there is an unequal ratio of observations in each class of the target variable. Class imbalance can be found in many different fields such as medical diagnosis, spam filtering, and fraud detection. The fatal issue arise from imbalanced dataset is the inaccuracy of prediction model because the algorithm will be trained better when the number of samples in each class is almost equal. If there is one minor class in the target variable, the algorithm will learn less about the minor class and less smart in predicting the minor class. There are two ways to resolve imbalanced data such as oversampling or under sampling. One issue with under sampling is it seeks to lessen the number of majority samples to stabilize the class distribution. Since it is eliminating observations from the original data set, some useful information may be discarded. On the other hand, over-sampling improves the number of minority class samples in the training set. The benefit of over-sampling is the perseverance of information in the original training set, as all records in the minority and majority classes are maintained. However, oversampling is susceptible to overfitting. This issue can be prevented by using stratified cross validation in later stage.



Dealing with Imbalanced Data

Oversample METHOD

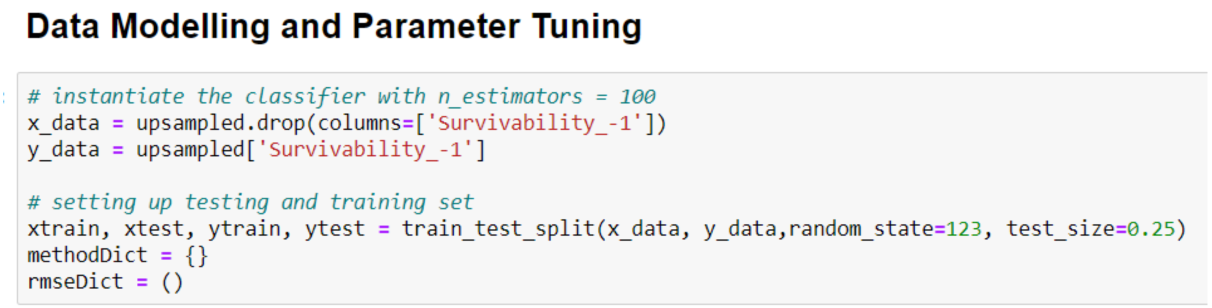
```
from sklearn.utils import resample

# Separate input features and target
x_data = df_dummied.drop(columns=['Survivability_-1','Year'])
y_data = df_dummied['Survivability_-1']

# setting up testing and training sets
xtrain, xtest, ytrain, ytest = train_test_split(x_data, y_data,random_state=123, test_size=0.25)

# concatenate our training data back together
x_sample = pd.concat([xtrain, ytrain], axis=1)

# separate minority and majority classes
not_survived = x_sample[x_sample['Survivability_-1']== 1] # 1 is not survived
survived = x_sample[x_sample['Survivability_-1']== 0]  # 0 is survived

# upsample minority
survived_upsampled = resample(survived,
                          replace=True, # sample with replacement
                          n_samples=len(not_survived), # match number in majority class
                          random_state=123) # reproducible results

# combine majority and upsampled minority
upsampled = pd.concat([not_survived, survived_upsampled])

# check new class counts
upsampled['Survivability_-1'].value_counts()
#    1     213245
#    0     213245
```

Figure 3.12: Up sampling Procedure

In this SEER dataset, the target variable, 'Survivability' attribute is made of 279878 records with the value '-1' if the patient did not survive more than 5 years, and 40895 records with the value '+1' if the patient survived more than 5 years. The dataset is very imbalanced with the minor class accounts for almost 12% of the target variable. Therefore, oversampling method is utilized to make the dataset balance.

3.2.3 Data Modelling and Parameter Tuning

3.2.3.1 Dataset Splitting: Training Set and Test Set



Data Modelling and Parameter Tuning

```
# instantiate the classifier with n_estimators = 100
x_data = upsampled.drop(columns=['Survivability_-1'])
y_data = upsampled['Survivability_-1']

# setting up testing and training set
xtrain, xtest, ytrain, ytest = train_test_split(x_data, y_data,random_state=123, test_size=0.25)
methodDict = {}
rmseDict = ()
```

Figure 3.13: Training Set and Test Set

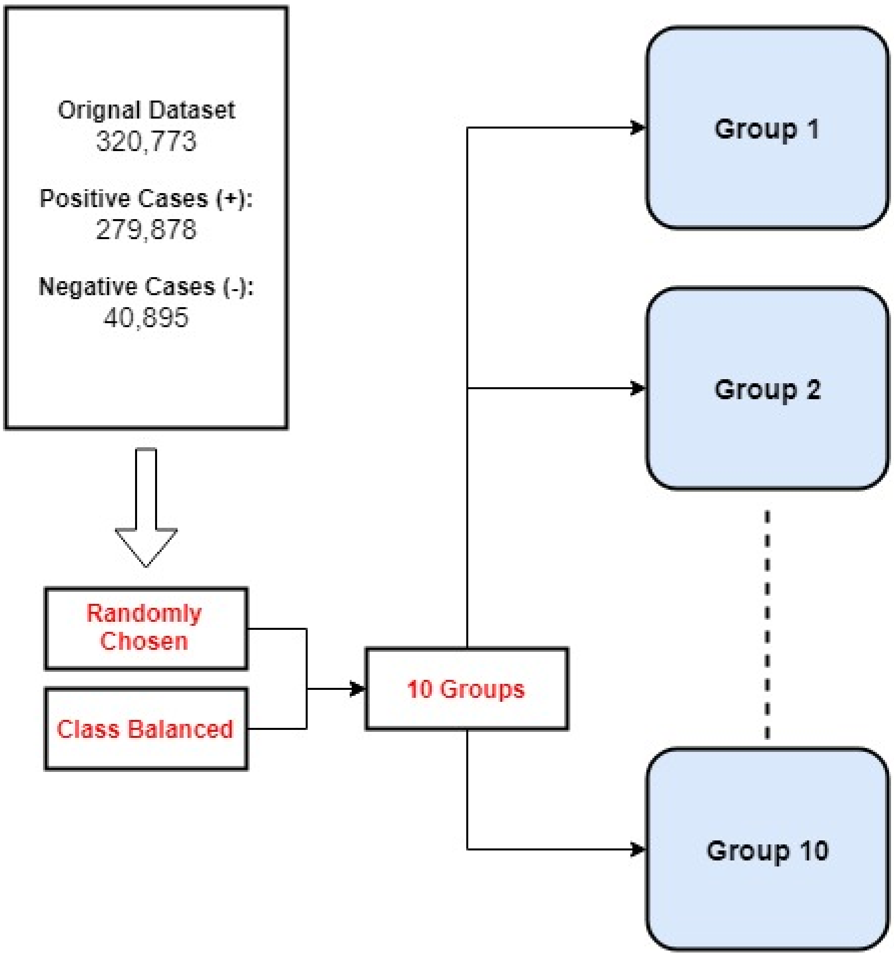
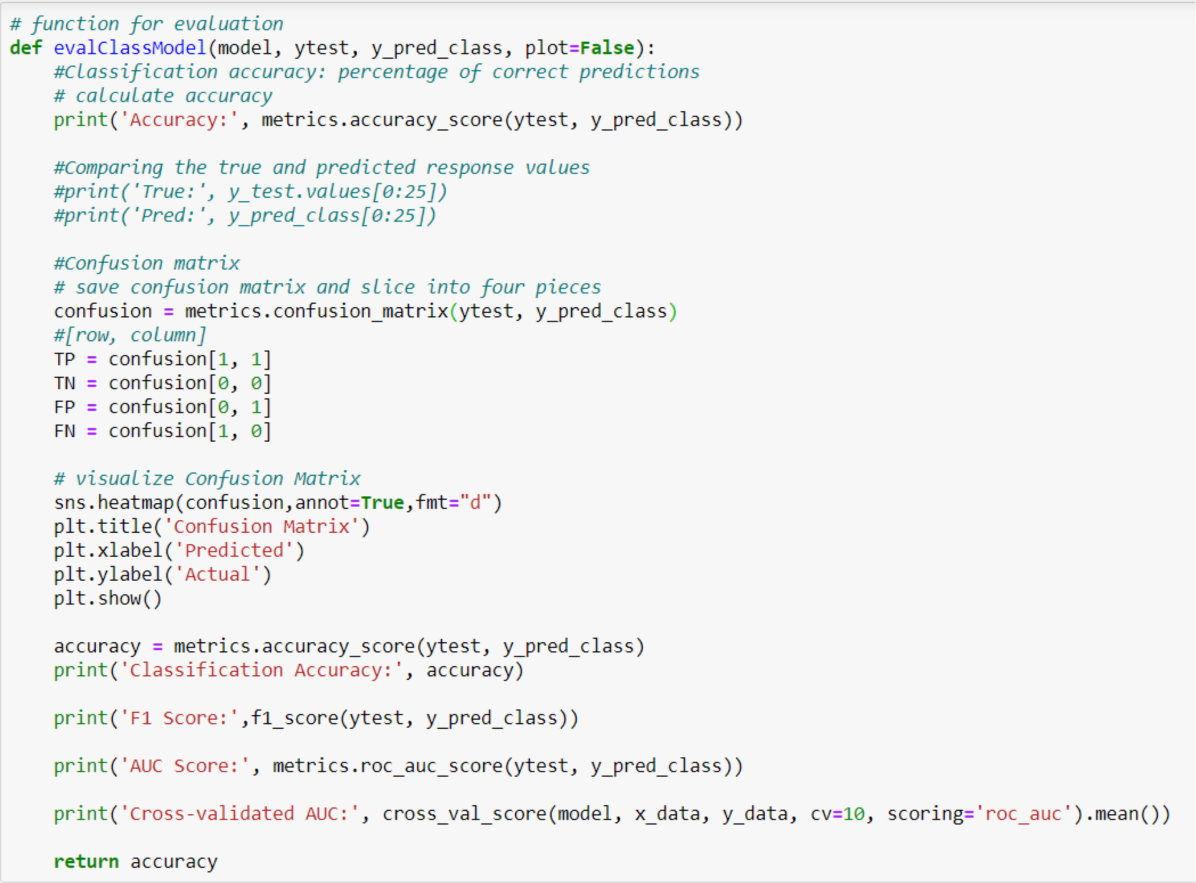


Figure 3.14: Dataset Splitting

Before the prediction model can be built, the SEER dataset contains 162,500 records with 128,469 positive cases and 34,601 negative cases. A subset of 50,000 records with class balanced (25,000 positive cases and 25,000 negative cases) must be split randomly into 10 groups of 5000 cases with two parts using 10-fold cross validation, namely: 80% training set and 20% test set. The purpose of training set is to train our prediction model algorithm so that it becomes better at prediction. Performing 10-fold cross validation is crucial to prevent the problem of overfitting when training our predictive model (Kohavi, 1995). In k-fold cross-validation, the initial sample is arbitrarily subdivided into k identical sized subsamples. Out of the total k subsamples, a particular subsample is reserved as the validation data for assessing the model, and the left-over k − 1 subsamples are utilized as training data. The cross-validation procedure is then repetitive k times, with each of the k subsamples utilized precisely once as the validation data. The k outcomes are averaged to generate a single estimate. On the other hand, the test set serves the purpose of validating the predicting performance of the model in term of accuracy, sensitivity, AUC etc. The target variable is required to be stored into y_data and the rest of attributes are stored in x_data.

3.2.3.2 Defining Function for Performance Metric Evaluation

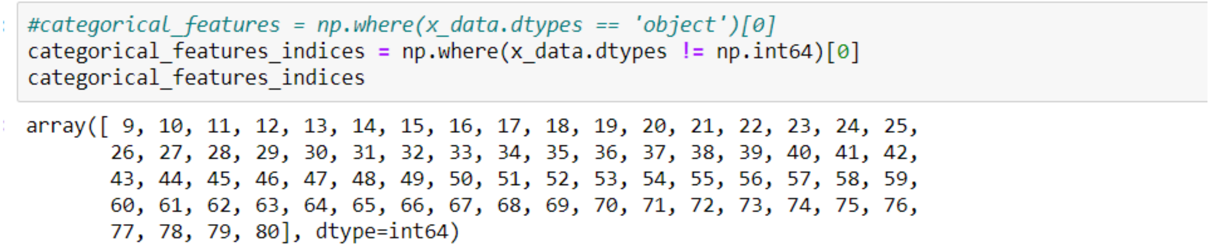


```
# function for evaluation
def evalClassModel(model, ytest, y_pred_class, plot=False):
    #Classification accuracy: percentage of correct predictions
    # calculate accuracy
    print('Accuracy:', metrics.accuracy_score(ytest, y_pred_class))

    #Comparing the true and predicted response values
    #print('True:', y_test.values[0:25])
    #print('Pred:', y_pred_class[0:25])

    #Confusion matrix
    # save confusion matrix and slice into four pieces
    confusion = metrics.confusion_matrix(ytest, y_pred_class)
    #[row, column]
    TP = confusion[1, 1]
    TN = confusion[0, 0]
    FP = confusion[0, 1]
    FN = confusion[1, 0]

    # visualize Confusion Matrix
    sns.heatmap(confusion,annot=True,fmt="d")
    plt.title('Confusion Matrix')
    plt.xlabel('Predicted')
    plt.ylabel('Actual')
    plt.show()

    accuracy = metrics.accuracy_score(ytest, y_pred_class)
    print('Classification Accuracy:', accuracy)

    print('F1 Score:',f1_score(ytest, y_pred_class))

    print('AUC Score:', metrics.roc_auc_score(ytest, y_pred_class))

    print('Cross-validated AUC:', cross_val_score(model, x_data, y_data, cv=10, scoring='roc_auc').mean())

    return accuracy
```

Figure 3.15: Function for Performance Metric Evaluation

The figure above illustrates a user defined function to evaluate the prediction model by using a few performance matrixes, called evalClassModel function. Inside this function, the confusion matrix of the model is calculated and visualize. Other than that, the value of each performance metric such as accuracy, F1 score and AUC score are evaluated.

3.2.3.3 CatBoost Modelling and Parameter Tuning

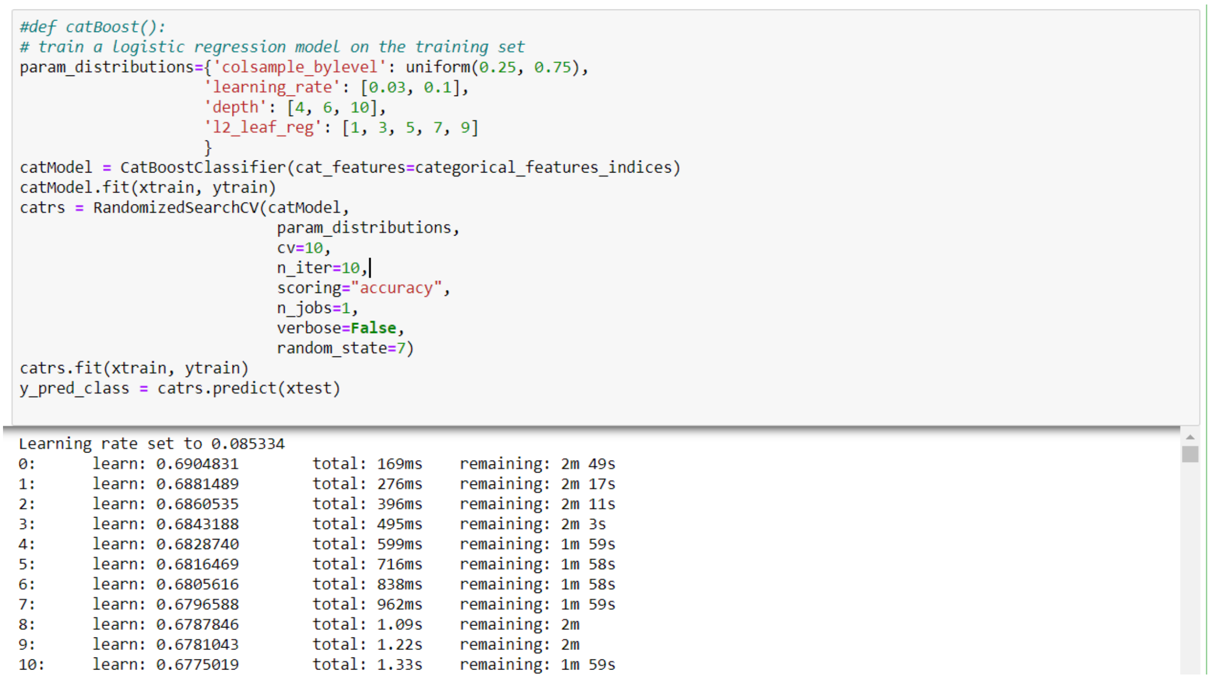


```
#categorical_features = np.where(x_data.dtypes == 'object')[0]
categorical_features_indices = np.where(x_data.dtypes != np.int64)[0]
categorical_features_indices
```

```
array([ 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25,
       26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42,
       43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59,
       60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76,
       77, 78, 79, 80], dtype=int64)
```

Figure 3.16: Defining the Index of Categorical Attribute

Before the model can run, the parameters inside the Catboost classifier need to be properly defined. First, the index of the categorical attribute in the dataset need to be specified so that Catboost knows which attributes are categorical as shown in figure 3.16 above.



```
#def catBoost():
# train a logistic regression model on the training set
param_distributions={'colsample_bylevel': uniform(0.25, 0.75),
                     'learning_rate': [0.03, 0.1],
                     'depth': [4, 6, 10],
                     'l2_leaf_reg': [1, 3, 5, 7, 9]
                     }
catModel = CatBoostClassifier(cat_features=categorical_features_indices)
catModel.fit(xtrain, ytrain)
catrs = RandomizedSearchCV(catModel,
                            param_distributions,
                            cv=10,
                            n_iter=10,
                            scoring="accuracy",
                            n_jobs=1,
                            verbose=False,
                            random_state=7)
catrs.fit(xtrain, ytrain)
y_pred_class = catrs.predict(xtest)
```

```
Learning rate set to 0.085334
0:      learn: 0.6904831        total: 169ms    remaining: 2m 49s
1:      learn: 0.6881489        total: 276ms    remaining: 2m 17s
2:      learn: 0.6860535        total: 396ms    remaining: 2m 11s
3:      learn: 0.6843188        total: 495ms    remaining: 2m 3s
4:      learn: 0.6828740        total: 599ms    remaining: 1m 59s
5:      learn: 0.6816469        total: 716ms    remaining: 1m 58s
6:      learn: 0.6805616        total: 838ms    remaining: 1m 58s
7:      learn: 0.6796588        total: 962ms    remaining: 1m 59s
8:      learn: 0.6787846        total: 1.09s    remaining: 2m
9:      learn: 0.6781043        total: 1.22s    remaining: 2m
10:     learn: 0.6775019        total: 1.33s    remaining: 1m 59s
```

Figure 3.17: Catboost Modelling and Parameter Tuning

The Catboost model demands to be appropriately fine-tuned to give the best performance. Therefore, parameter tuning is crucial to determine the best parameters' values so that a best prediction model can be built. To achieve this purpose, the parameter tuning method that is applied is random search.

Hyperparameter Tuning is the exploring progress for the right set of hyperparameter to accomplish high precision and accuracy. Optimising hyperparameters represent one of the trickiest parts in constructing the machine learning models. The crucial purpose of hyperparameter tuning is to discover the sweet point for the model's parameters so that a better performance is achieved.

Random search is a method where random permutations of the hyperparameters are utilized to discover the best answer for the built model. It attempts arbitrary permutations of a array of values. To improve with random search, the function is assessed at some number of random designs in the parameter space.

The four main parameters that are tuned are:

a. Colsample_bylevel
b. Learning_rate
c. Depth
d. L2_leaf_reg

Inside the RandomizedSearch function, 10-folds cross validation method will be applied to confirm the accomplishment of the models. The whole dataset is separated into 10 subsets and then treated 10-times. 9 subsets are applied as training data sets and the residual 1 subset is applied as a testing data set. Lastly, the results will be shown by be an average of each 10 repetitions. The subclasses are split utilizing stratified sampling, indicating that every subset will carry similar class proportion of the major dataset. The modelling process is then executed to build the optimal model for the dataset.

3.2.4 Feature Importance Selection

CatBoost and XGBoost are tree-based algorithm that relies on entropy based splitting method to form multiple decision trees (weak learners) and combine the prediction outcome of each individual weak learners to form a strong learner. Entropy is the impurity measurement of a features. Entropy regulates how a tree algorithm split up the data. Information gain (IG) calculates the amount of "information" a feature offers about the class. Information Gain is computed for a split by deducting the weighted entropies of each node from the initial entropy of the class feature .Tree algorithms will attempt to maximize Information gain, where the most important feature is the feature with highest information gain and it will be the root node of a tree, followed by subsequent important features that form branch node beneath the root node. The following shows the formula for entropy and information gain:

$$Entropy, E = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$Information\ Gain(class, feature) = Entropy(class) - Entropy(class, feature)$$

3.2.5 Performance Metric Evaluation

The performance of the classification model will be measured using three performance metrics: accuracy, f-measure and precision. Accuracy is the fraction of properly predicted cases among all cases. F-measure is the weighted average of the precision and recall. Precision is the fraction of correct predictions for the positive class. To classify the significant attributes, these three performance metrics are used, whereas to detect data mining method to produce the best performing models, the accuracy and precision methods were applied. For detection of significant attributes, the three performance metrics provide a well understanding of the general behaviour of the various blend of features. Meanwhile, evaluation of data mining procedures emphases on the best performing models that can yield high accuracy in breast cancer prediction because accuracy and precision are the most instinctive evaluation metrics on performance.

Chapter 4 : RESULTS

4.1 Exploratory Data Analysis

Exploratory data analysis (EDA) is a method to analysing data sets to condense their main traits, often with visual approaches. EDA is to unveil what the data can inform us beyond the proper modelling or hypothesis testing task (Paul & David, 2004).

The data presented in graphical method so that the information hidden in the numbers can be grasped easier and clearer. Visualization aids researchers to understand statistics and establish next course of action. In addition, it makes complex dataset easier to be digest and utilized visualisation approaches of presenting complex information in easy to digest manner.

It is commonly applied in data science field for researchers to understand their dataset in visual sense and interpret what are the possible anomalies and outliers of the data values. It could assist to uncover trend, patterns, produce insights and develop judgment (Matthew N.O.Sadiku, 2016).

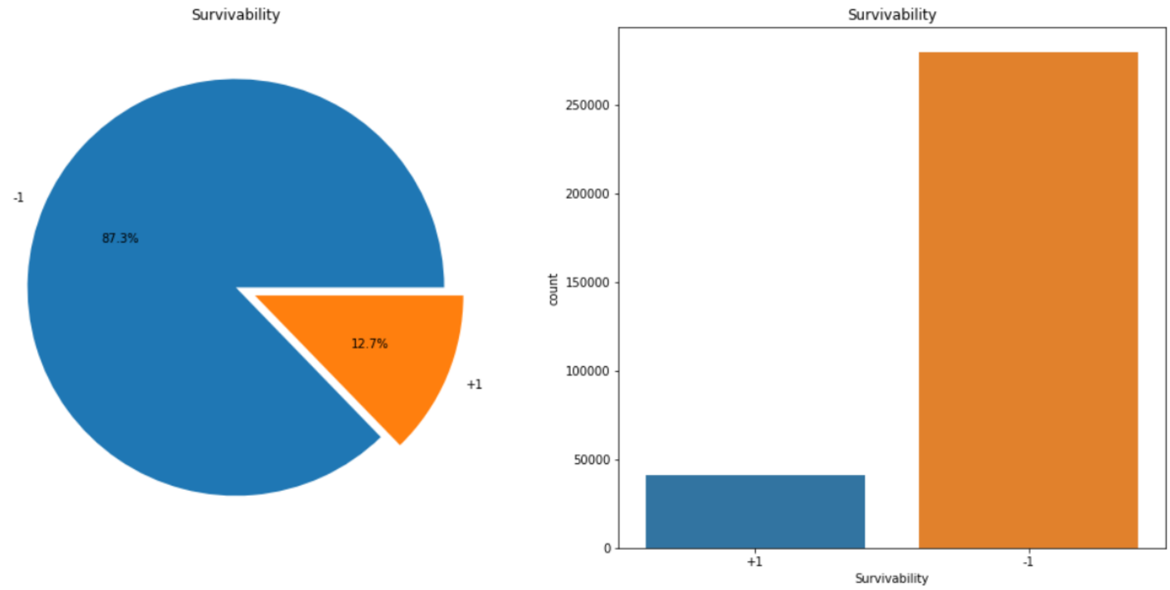


Figure 4.1: Target Variable Composition

According to the bar chart and pie chart above, we can conclude that our dataset is highly imbalance. This is because our target variable, 'Survivability' attribute has disproportionate ratio of class label. The class label of ‘-1’ signifies the patient did not survive more than 5 years , whereas class label of ‘+1’ signifies the patient did not survive more than 5 years The class label of ‘-1’ consists of 87.3% of the variable, whereas class label of ‘+1’ consist of 12.7% of the target variable. This prompts the necessity to perform oversampling on minority sample to achieve a balanced dataset.

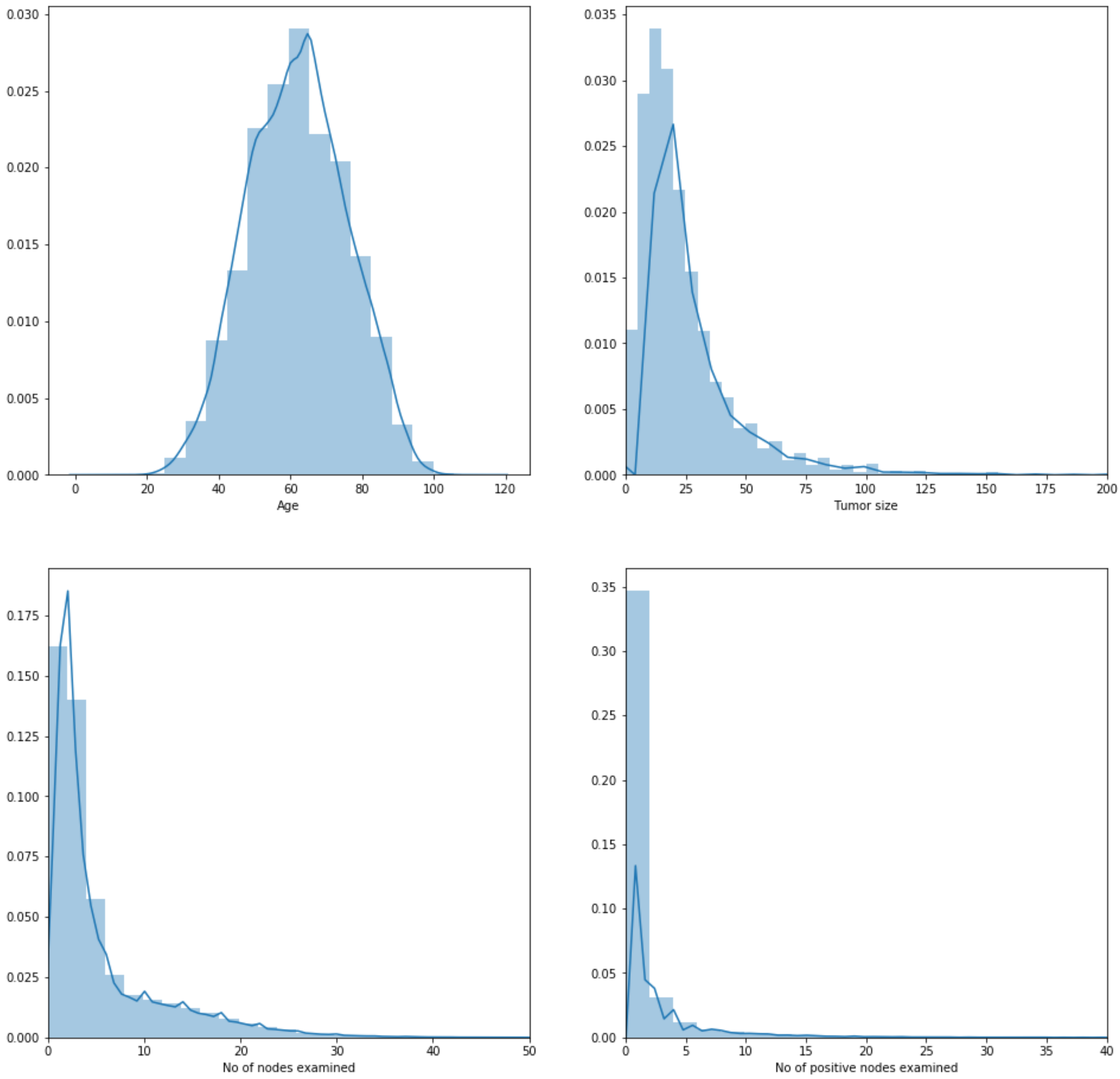


Figure 4.2: Histogram of Numerical Variable Distribution

According to the figure above, the patient’s age is uniformly distributed, ranging from age 20 until age 85. The mode of the age is approximately 60 years old. The histogram shows the distribution of tumour size is a right skewed distribution. A right-skewed distribution will have the mean to the right of the median and mode at the left of the median. Same pattern is observed in the histograms for ‘No. of nodes examined’ attribute and ‘No. of positive nodes examined’ attribute, where the histograms are right skewed distribution.

4.2 Hyperparameter Tuning

The method of random search is applied to find the best hyperparameter for each model Random Forest, XGBoost, CatBoost), specifically for this SEER breast cancer dataset. The table below shows the best parameter tuned from using the random search method.

Table 4.1: Best Parameters for Each Models

Models	**Best Hyperparameter for selected model**
Random Forest	• max_depth=None • max_features='auto' • max_leaf_nodes=None • min_samples_leaf=8 • min_samples_split=2 • n_estimators=30 • random_state=1
XGBoost	• colsample_bytree= 0.6258403477449535, • learning_rate=0.16, • max_depth=9 • subsample= 0.6249118756191699 • n_jobs=1
CatBoost	• colsample_bylevel= 0.8430726336825362 • depth = 10

- l2_leaf_reg =3
- learning_rate=0.1

4.3 Feature Importance

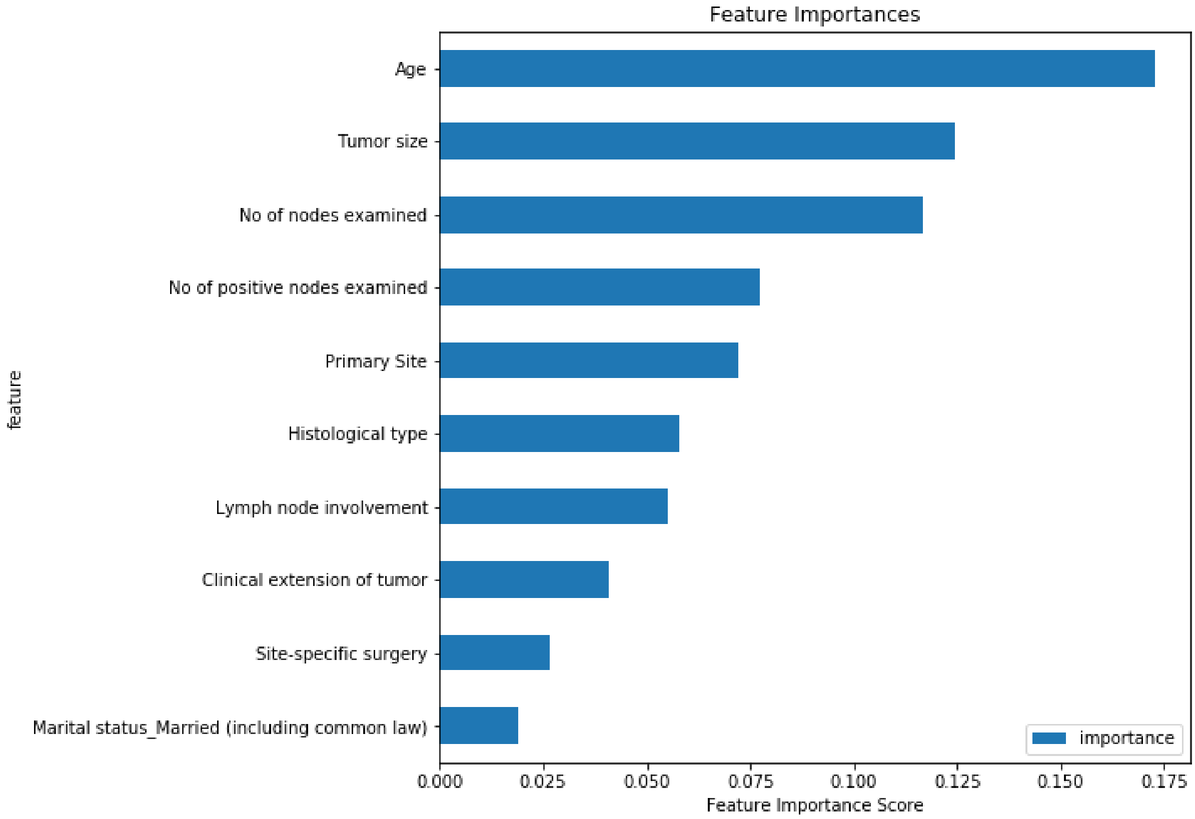


Figure 4.3: Feature Importance Score by Random Forest

Figure 4.3 shows that the age of patient is the most important factors affecting breast cancer survivability for a patient, followed by tumor size and number of nodes examined. The result is sensible this is because it is aligned with the statistics released by National Cancer Institute (NIH). According to NIH (2019), the danger associated with cancer rises as we age. About 80% of females identified with breast cancer are 45 years old or more, and roughly 43% are 65 years old or more.

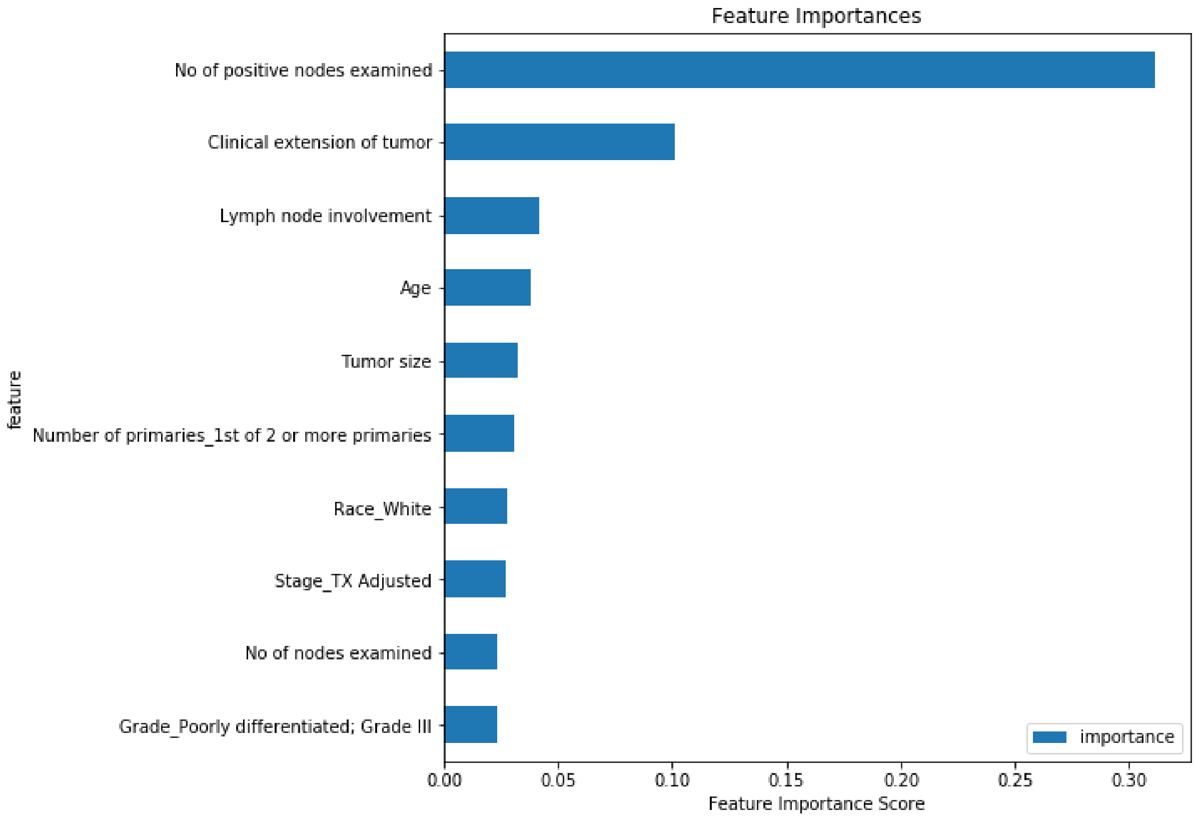


Figure 4.4: Feature Importance Score by XGBoost

Figure 4.4 shows that the number of positive nodes examined is the most important factors affecting breast cancer survivability for a patient, followed by clinical extension of tumor and lymph node involvement. Whereas the age attribute which is the most important feature in Random Forest only ranked number four in XGBoost. This explains why XGBoost has a considerably lower prediction accuracy than Random Forest, as ‘age’ attribute has been confirmed by National Cancer Institute as one of the most important factor affecting breast cancer survivability of a patient.

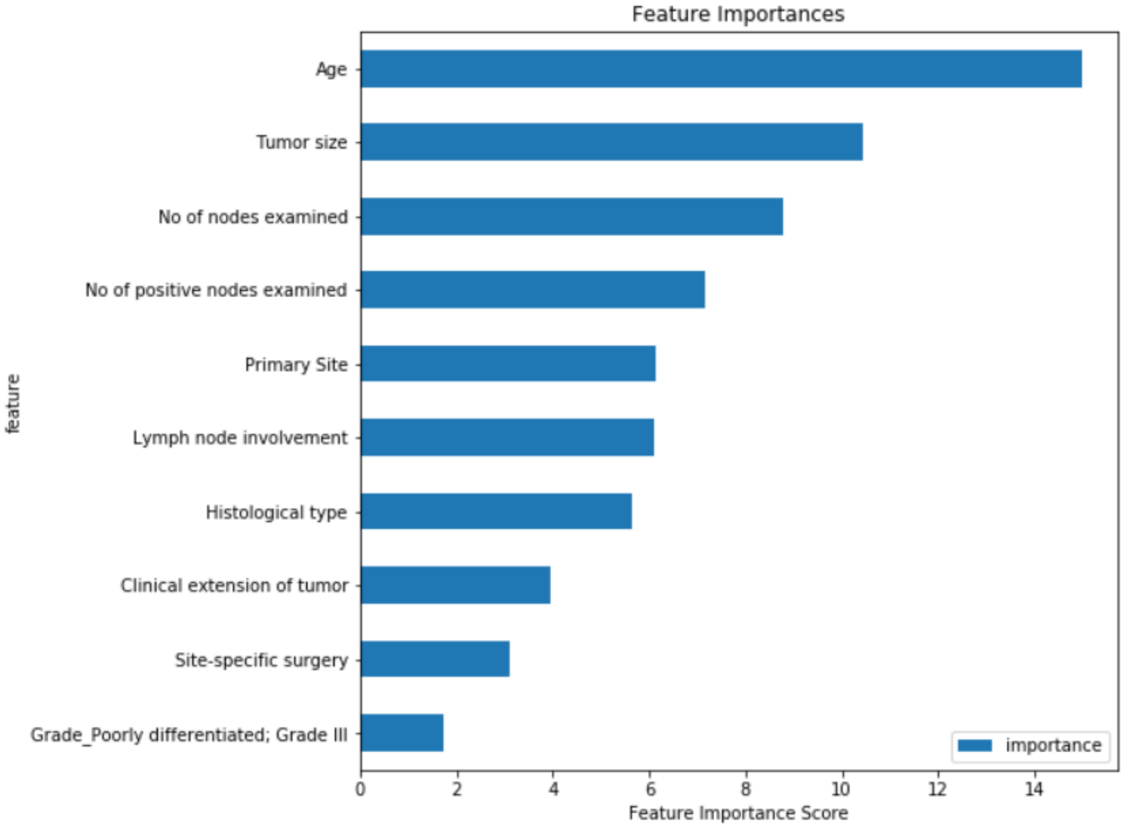


Figure 4.5: Feature Importance Score by CatBoost

Figure 4.5 shows that the age of patient is the most important factors affecting breast cancer survivability for a patient, followed by tumor size and number of nodes examined. The top 9 important factors calculated by CatBoost is identical with the feature importance score calculated by Random Forest. This explains why CatBoost and Random Forest perform comparatively better than XGBoost in term of accuracy.

In this study, the one of the objectives to be accomplished is to discover the significant features that play a vital role on the accuracy of the prediction model. By using features importance score calculated by each prediction models, a list of top 10 most significant attributes is recorded in a descending order in term of feature importance score. Figure above shows each feature and feature importance score. Based on the feature importance score generated by the models, Random Forest and CatBoost have the identical top five significant attributes. They are: age, tumour size, No. of nodes examined, No. of positive nodes examined and primary site.

The ‘age’ attribute is the most significant feature that will alter the prediction accuracy of Catboost model, and Random Forest followed by the “tumour size” is second most significant feature. On the other hand, the top five significant attributes calculated by the XGBoost is different than both Random Forest and CatBoost. They are: No. of positive nodes examined, Clinical extension of tumour. Lymph node involvement, age and tumour size.

4.4 Performance Evaluations

4.4.1 Performance of Random Forest

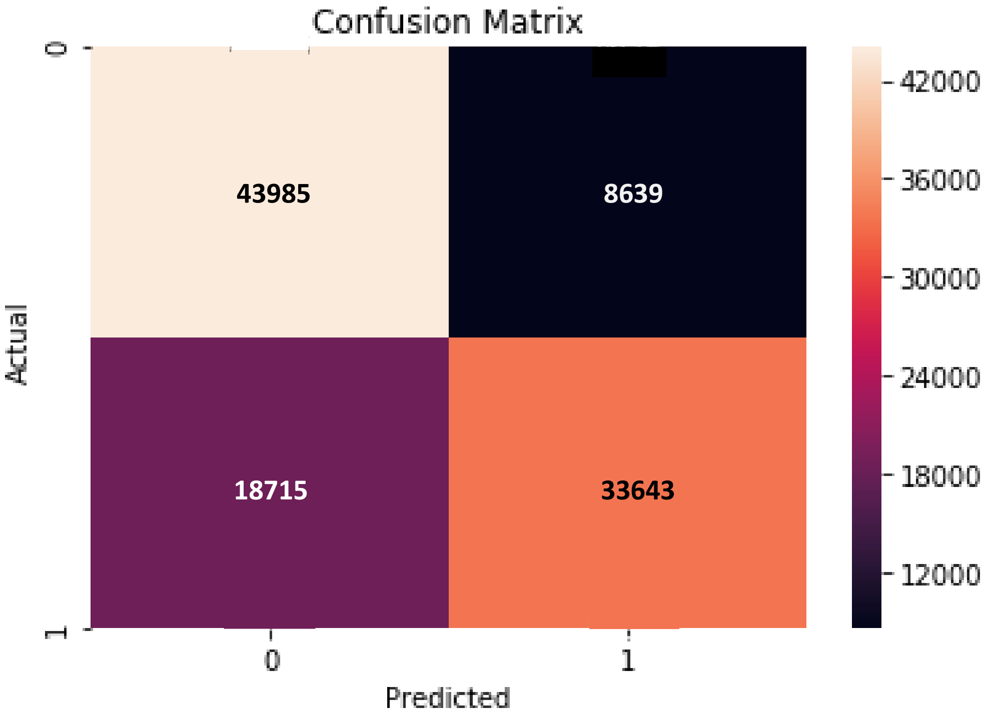


Figure 4.6: Confusion Matrix of Random Forest Model

Figure 4.6 shows the confusion matrix for Random Forest model. As illustrated, Random Forest predicted 43985 True Negative cases, 33643 True Positive cases, 18715

False Negative cases and 8639 False Positive cases. This signifies that Random Forest perform well in predicting most of the True Positive and True Negative cases correctly.

Table 4.2: Performance Evaluations of Random Forest

Random Forest	
Performance Metrics	Percentage (%)
Accuracy	73.94
Precision	79.60
Recall	64.3
F1 score	71.10
Specificity	83.60
AUC	73.92

Table 4.2 indicates different performance metrics of Random Forest. The performance is evaluated after the hyperparameter tuning is conducted to obtain the most optimal parameters for the model. In summary, Random Forest perform best in specificity and precision, but not so well in recall.

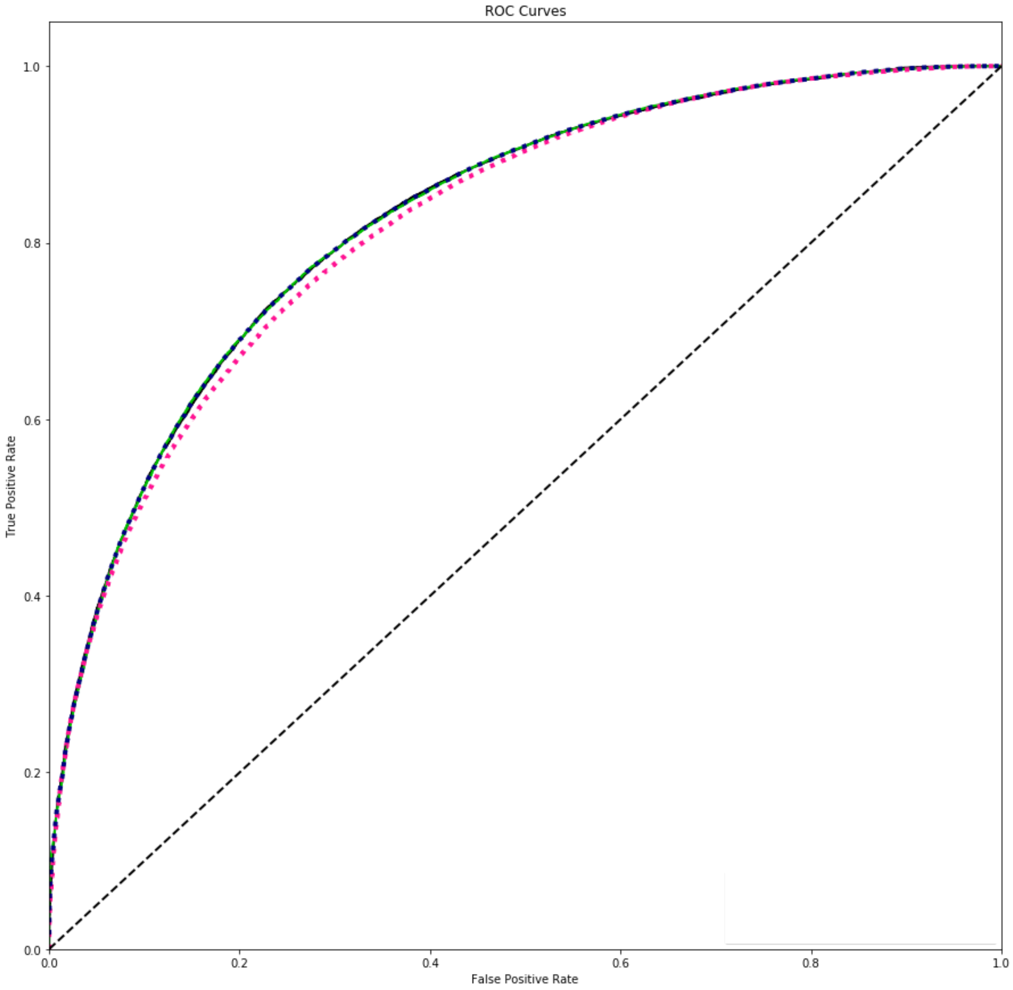


Figure 4.7: ROC Curve of Random Forest Model

Figure 4.7 illustrates the Receiver Operating Characteristics (ROC) curve for Random Forest model. The ROC curve is generated by charting the true positive percentage against the false positive percentage. Its purpose is to demonstrate the analytical competence of a binary classifier system. For Random Forest, it achieves a score of 73.92 Area Under Curve of the ROC.

4.4.2 Performance of XGBoost

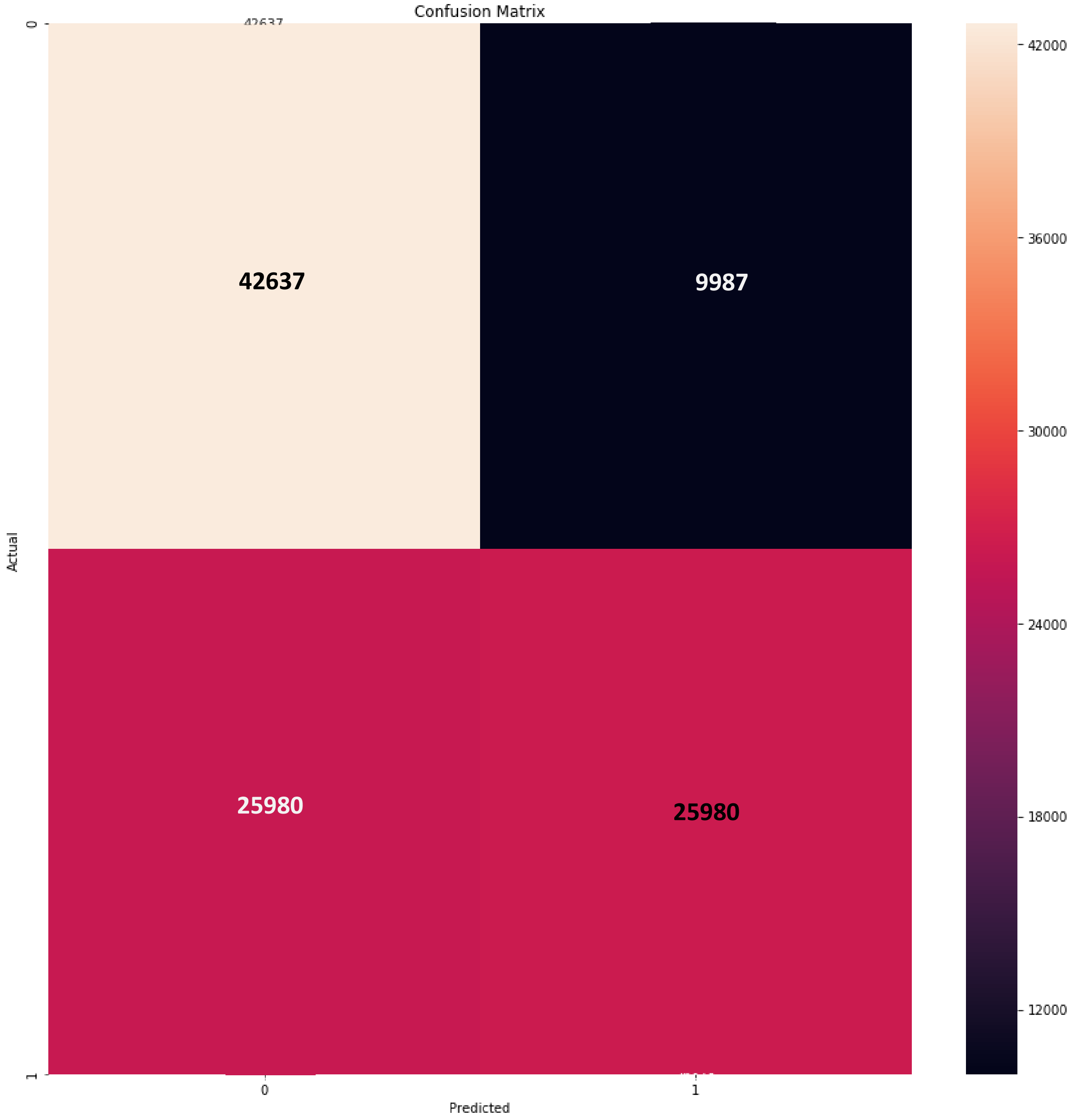


Figure 4.8: Confusion Matrix for XGBoost

Figure 4.8 shows the confusion matrix for Random Forest model. As illustrated, XGBoost predicted 42637 True Negative (TN) cases, 25980True Positive (TP) cases, 25980 False Negative (FN) cases and 9987 False Positive (FP) cases. XGBoost perform comparatively worse than Random Forest as it predicted less TN and TP cases, but more FN and FP cases.

Table 4.3: Performance Evaluations of XGBoost

XGBoost	
Performance Metrics	Percentage (%)
Accuracy	65.74
Precision	72.50
Recall	50.40
F1 score	59.50
Specificity	81.00
AUC	65.70

Table 4.3 indicates different performance metrics of XGBoost. The performance is evaluated after the hyperparameter tuning is conducted to obtain the most optimal parameters for the model. In summary, XGBoost perform best in specificity and precision, but worse in recall. When compared with Random Forest, XGBoost perform poorer in term of overall performance metrics.

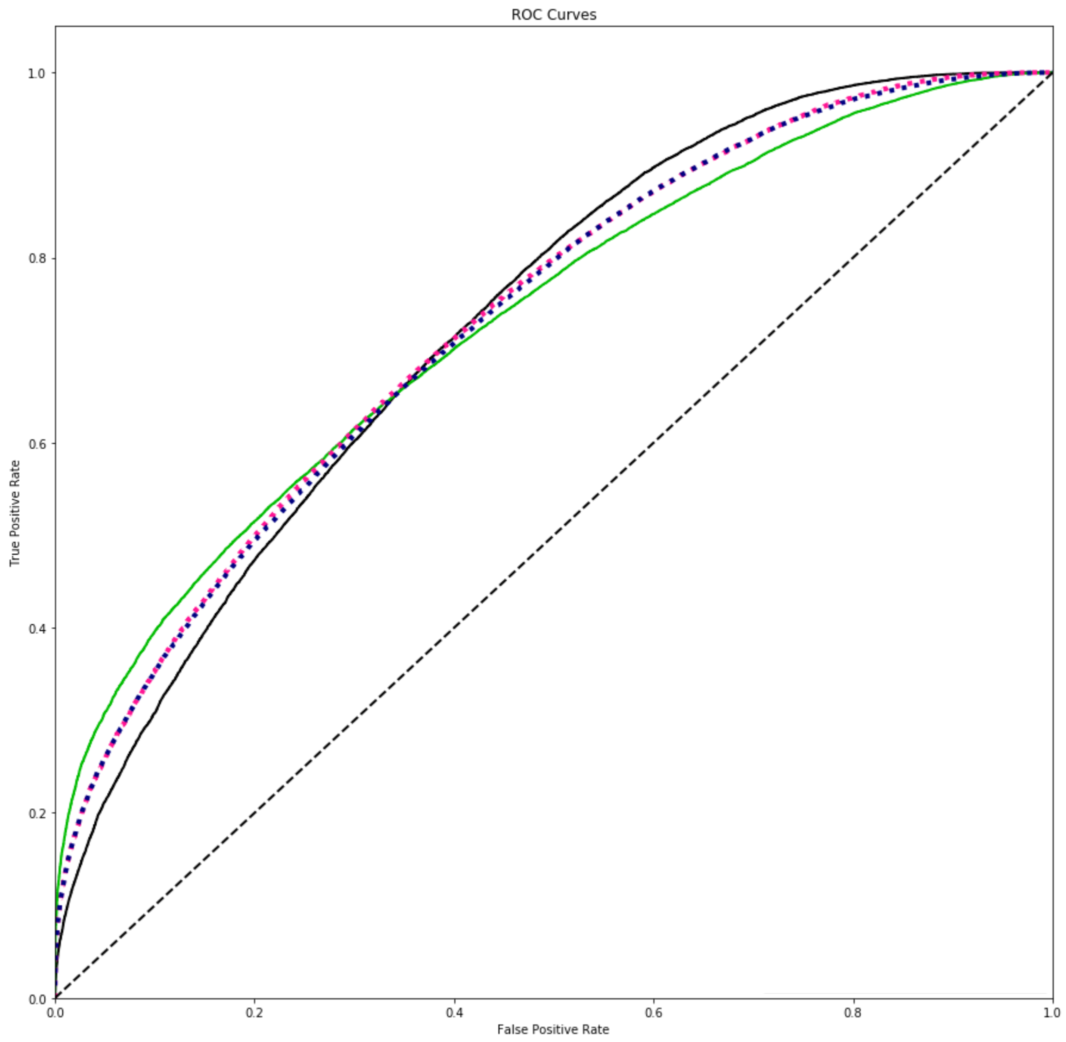


Figure 4.9: ROC Curve for XGBoost Model

Figure 4.9 illustrates the Receiver Operating Characteristics (ROC) curve for XGBoost model. The ROC curve is generated by charting the true positive percentage against the false positive percentage. Its purpose is to demonstrate the analytical competence of a binary classifier system. For XGBoost, it achieves a score of 65.70 Area Under Curve of the ROC, which is comparatively lower than Random Forest model.

4.4.3 Performance of CatBoost

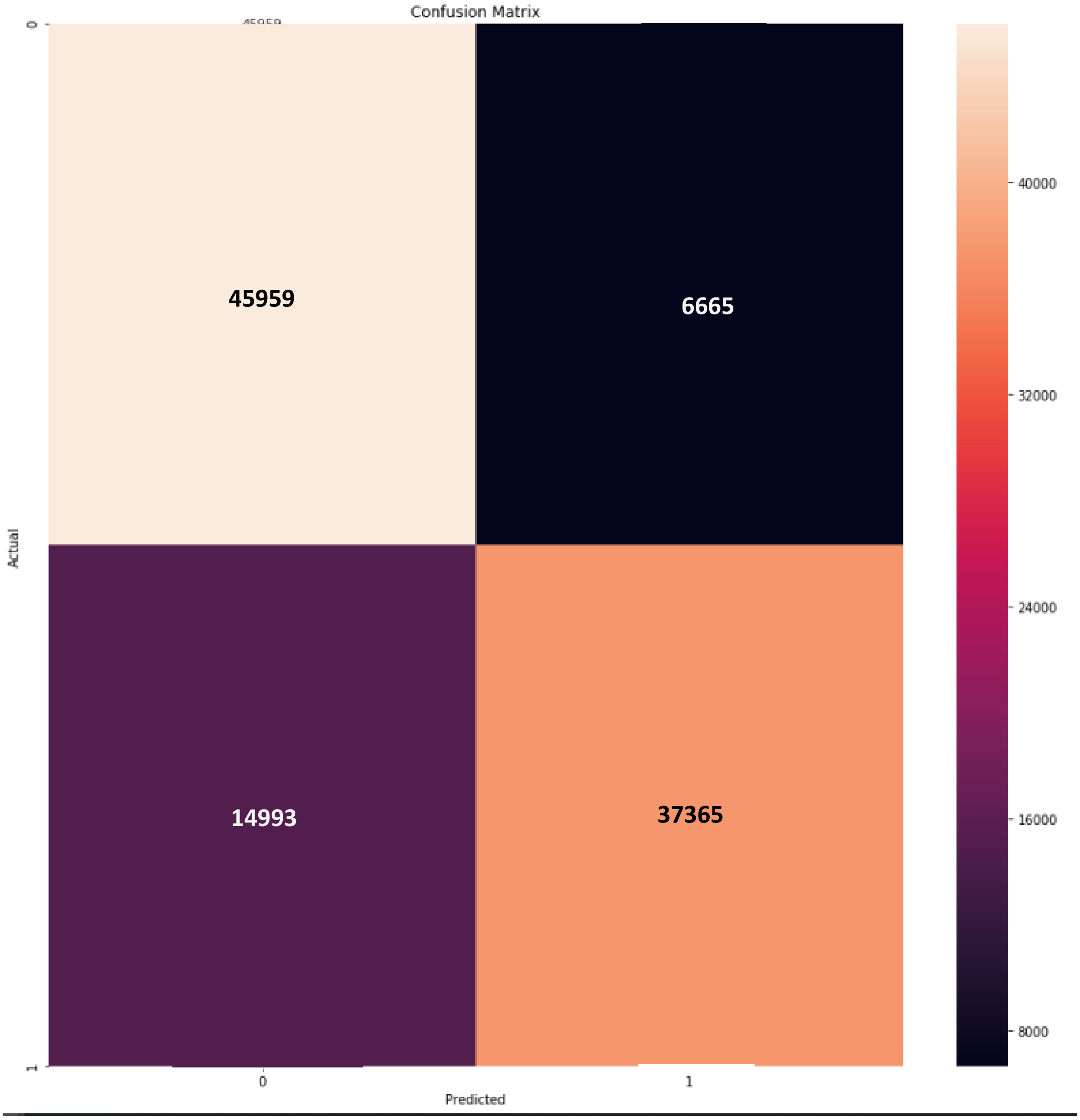


Figure 4.10: Confusion Matrix of CatBoost Model

Figure 4.10 shows the confusion matrix for CatBoost model. As illustrated, CatBoost predicted 45959 True Negative (TN) cases, 37365 True Positive (TP) cases, 14993 False Negative (FN) cases and 6665 False Positive (FP) cases. CatBoost the best when compared with Random Forest and XGBoost as it predicted most number of TN and TP cases, but lesser FN and FP cases.

Table 4.4: Performance Evaluation of CatBoost

Catboost	
Performance Metrics	Percentage (%)
Accuracy	79.37
Precision	84.90
Recall	71.40
F1 score	77.50
Specificity	87.30
AUC	79.35

Table 4.4 indicates different performance metrics of CatBoost. The performance is evaluated after the hyperparameter tuning is conducted to obtain the most optimal parameters for the model. In summary, CatBoost perform best in specificity and precision, but lower in recall. When compared with Random Forest and XGboost, CatBoost perform the best in term of overall performance metrics.

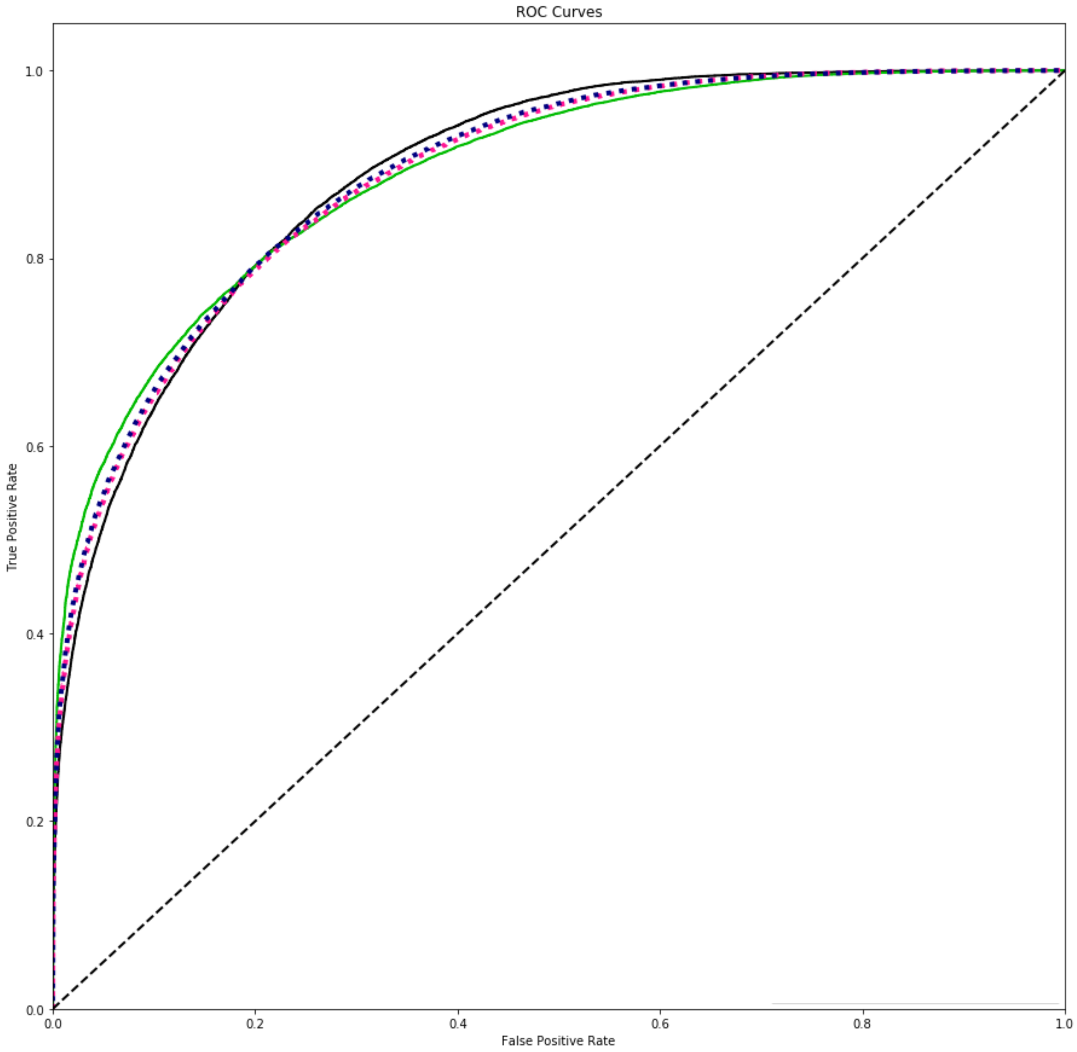


Figure 4.11: ROC Curve for CatBoost Model

Figure 4.11 illustrates the Receiver Operating Characteristics (ROC) curve for CatBoost model. The ROC curve is generated by charting the true positive percentage against the false positive percentage. Its purpose is to demonstrate the analytical competence of a binary classifier system. For CatBoost, it achieves a score of 79.35 Area Under Curve of the ROC, which is comparatively higher than Random Forest and XGBoost model.

Chapter 5 : DISCUSSION

5.1 Significant Attribute Score by Random Forest, XGBoost and CatBoost

In this study, the one of the objectives to be accomplished is to discover the significant features that play a vital role on the accuracy of the prediction model. By using features importance score calculated by each prediction models, a list of top 10 most significant attributes is recorded in a descending order in term of feature importance score. Based on the feature importance score generated by the models, Random Forest and CatBoost have the identical top five significant attributes. They are: age, tumour size, No. of nodes examined, No. of positive nodes examined and primary site.

The 'age' attribute is the most significant feature that will alter the prediction accuracy of Catboost model, and Random Forest followed by the "tumour size" is second most significant feature. On the other hand, the top five significant attributes calculated by the XGBoost is different than both Random Forest and CatBoost. They are: No. of positive nodes examined, Clinical extension of tumour. Lymph node involvement, age and tumour size.

5.2 Performance Comparison Between Default and Tuned GBDT Models

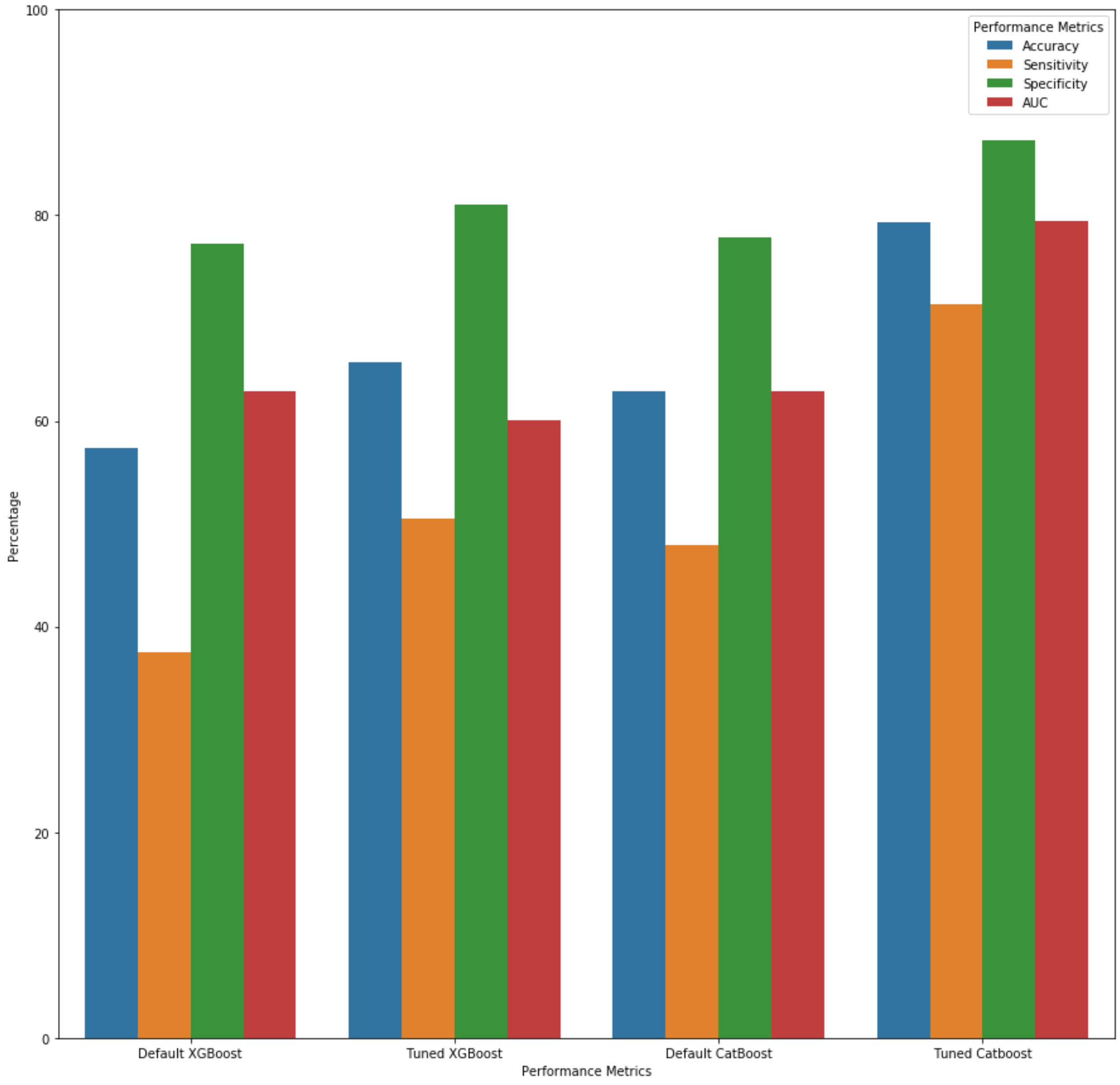


Figure 5.1: Performance Comparison between Default and Tuned GBDT Models

The figure above shows the significance of hyperparameter tuning to truly unleash the hidden power of gradient-boosting based decision tree models. Before hyperparameter tuning, the default XGBoost and CatBoost models perform worse than Random Forest (Edeki & Shardul, 2012)and Graph Based semi supervised learning (Park et al., 2015)classification models proposed by previous studies. However, after hyperparameter tuning using random search, the tuned XGBoost and CatBoost models gain significant boost in overall performance, surpassing the default models and the classification models from previous work. The performance of tuned XGBoost and CatBoost models are compared and discussed with classification models proposed by previous studies in next section.

5.3 Performance Comparison with Previous Works

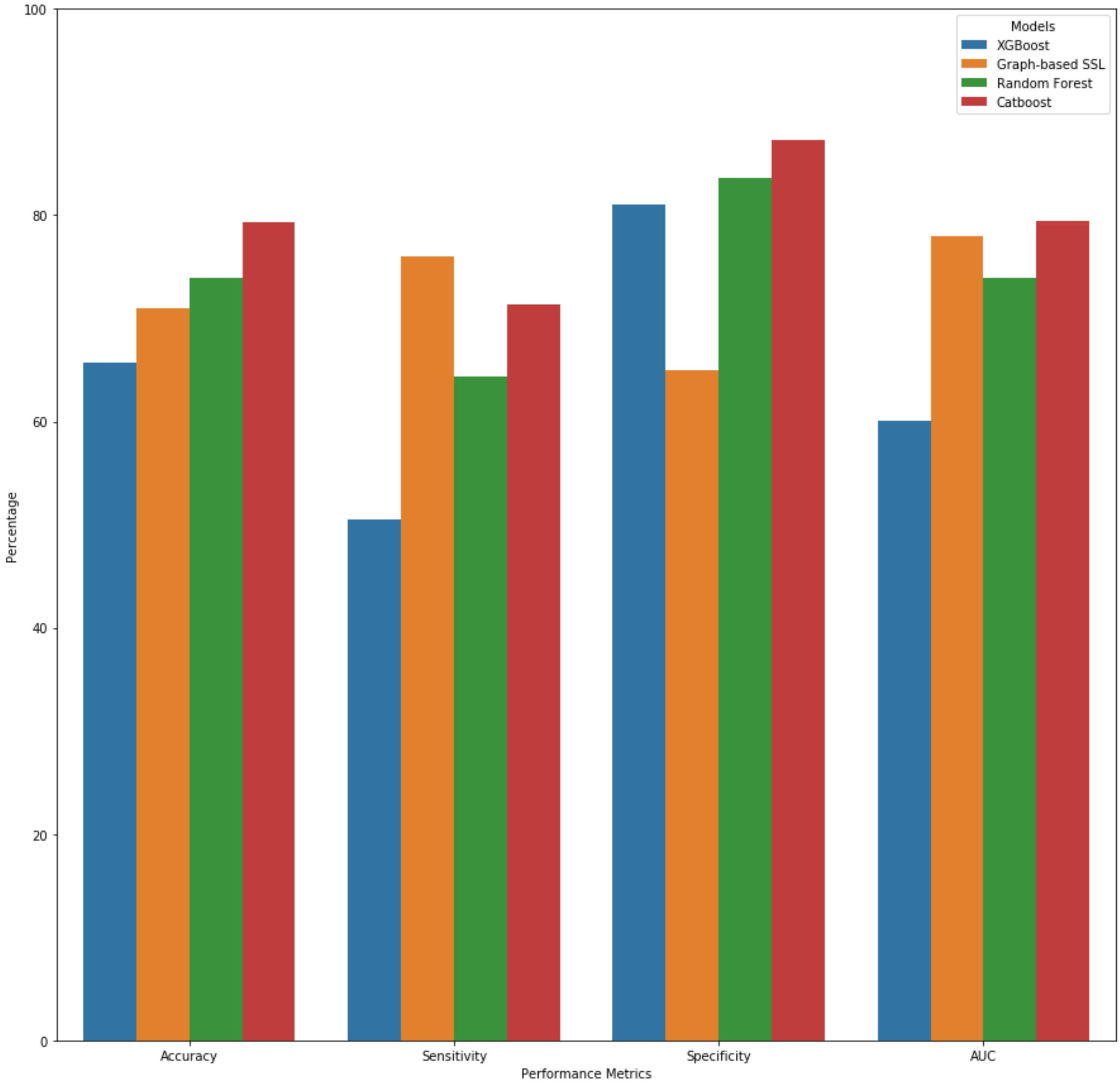


Figure 5.2: Performance Comparison Between the Four Classification Models

The figure 5.2 shows the summary of performance comparison between the four classification models. Among the four classification models, two gradient boosting based decision tree models (XGBoost and CatBoost) are compared against two classical models that are proposed by in past study, they are Random Forest model (Edeki & Shardul, 2012) and graph-based semi supervised learning (Park et al., 2015).

In term of accuracy, specificity and AUC, the proposed CatBoost is the best performing model with highest values. This is indicating that CatBoost performs the best in term of prediction of correct cases for both True Positive and True Negative out of the total number of input records. Meanwhile, CatBoost also perform the best in term of specificity, that is 'True Negative Rate' (Percentage of real negatives that are accurately classified). CatBoost also scores the highest Area Under Curve (AUC) score among the four models, this signifies that CatBoost has a better ability to differentiate between two diagnostic classes (true positive rate and false positive rate).

Whereas the Graph- based -SSL method proposed by (Park et al., 2015) performs the best in term of sensitivity. This indicates the graph-based semi supervised learning is better performer in identifying the whole related cases. Meanwhile, the Random Forest method proposed by Edeki and Shardul (2012) ranks the second best model in term of overall performance just behind the CatBoost. Lastly the XGBoost model perform the worse in overall. Therefore, the best performing models in descending orders are: CatBoost, Random Forest, Graph-based SSL and lastly the XGBoost.

Chapter 6 : CONCLUSION

6.1 Conclusion

Breast cancer is the most widespread cancer among women that triggers large number of fatal casualties across global. the malignant tumour must be removed if a patient is identified with breast cancer. During this stage, physicians must examine and consider the prediction of the disease. This is the prediction of the anticipated movement of the disease, which is called prognosis. Prognosis is vital because the nature and strength of the medications or treatments prescribed are based on the seriousness of the cancer stage (Gupta, Kumar, & Sharma, 2011).

Survival assessment is a field in medical prognosis that manages with application of different techniques to data warehoused in health database to forecast the survival probability of a specific patient suffering from a disease over a particular time interval (Delen et al., 2005) .The compilation of enormous quantities of health data has suggested an occasion to build prediction models for survival by the health researchers.

Machine learning models are gaining attention by the health researchers to build a prediction model to forecast the survivability of a breast cancer patient. The medical officers are able to provide proper medication and treatment based on the health condition of a breast cancer patient, to extend their chances of survivability. In this study, the top ten most crucial attributes in determining the survivability of a cancer patients are determined by each decision tree-based models.

These highlighted important attributes based on the feature importance score allows the medical officer to identify some precaution actions that a woman should take on to reduce the probability of getting breast cancer. For instance, age is the most important attribute of the SEER dataset. Therefore, medical officers could conclude that woman should get more frequent breast check-up as they grow older in order to detect the cancer earlier. This is because prevention is better than cure. Ultimately the number of deaths caused by breast cancer could be drastically reduced.

The dataset used in this research is extracted from the SEER cancer database, is a trustworthy resource for cancer data on cancer incidence and survival in the United States. This dataset is treated with several data pre-processing method such as data cleaning and data transformation to improve the data quality so that a high-quality prediction model can be built. After performing the data pre-processing measures, the processed data is randomly split into 10 groups of training set and testing set. Four classifications model (XGBoost, Graph-based SSL, Random Forest and CatBoost) are utilized to train the model. The performance of the models is assessed by using the confusion matrix and AUC score. The results produced are compared with the previous work produced by Park et al. (2015), using graph-based semi supervised learning (SSL) and also (Edeki & Shardul, 2012), that proposed the Random Forest model. The comparison suggests that the proposed CatBoost model is the best performing model compared to the rest in term of classification accuracy, AUC score and specificity. Therefore, the best performing models in descending orders are: CatBoost, Random Forest, Graph-based SSL and lastly the XGBoost.

The production of breast cancer prediction model not only allows the medical officers to prescribe proper medication and treatment to the patients based on the result of prediction, but also create awareness to the woman to often undergo breast check-up as a preventive measure. In summary, the number of deaths caused by breast cancer could further reduced by using the proposed model as it gives high prediction accuracy.

6.2 Recommendations and Future Work

For future work, the Catboost model can be further improved by using grid search for the better hyperparameter tuning, as there are many more parameters to be explored. This is because by default, the proposed model performs mediocre on the prediction of breast cancer survivability of patient with default parameter setting. In this project, only random search is applied to fine tune the proposed model to unleash its potential. The grid search method can unleash the full potential of the proposed model but is not applied in this research project due to time constraint and lack of powerful computational machine to cope with the exhaustive search of the best parameters. Unlike random search that only randomly search for the combinations of parameters to train the model, grid search tries every possible combination of parameters and computational time grows exponentially with the increase in number of parameters to be tuned. For instance, exploring 15 distinct parameter values for a total of 4 parameters demand 50,625 tests of cross-validation. This is equivalent to 506,250 model fits and 506,250 forecasts if 10-fold cross validation is utilized. Meanwhile, data pre-processing method can be further improved by using methods such as feature engineering, principal component analysis (PCA) or feature elimination method to further shrink the data dimension of the dataset. This is especially useful in high dimensionality dataset with high number of attributes. The data reduction technique helps to decrease the computational complexity and improve the prediction accuracy of the model by eliminating interrelated attributes and noises in the data.

REFERENCES

Amir, E. (2003). *Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme*. 807–814.

Ayodele, T. O. (2010). Types of Machine Learning Algorithms. *New Advances in Machine Learning. IntechOpen*.

B, S. B., Zeng, K., Sotiras, A., & Rathore, S. (2016). *GLISTRboost : Combining Multimodal MRI Segmentation , Registration , and Biophysical Tumor Growth Modeling with Gradient Boosting Machines for Glioma Segmentation*. 144–155. https://doi.org/10.1007/978-3-319-30858-6

Breiman, L. E. O. (2001). *Random Forests*. 5–32.

Brenner, H., Gefeller, O., & Hakulinen, T. (2002). *A computer program for period analysis of cancer patient survival*. *38*, 690–695.

Burke, H. B., Goodman, P. H., Rosen, D. B., Henson, D. E., Weinstein, J. N., Harrell, F. E., … Bostwick, D. G. (1997). Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*, *79*(4), 857–862. https://doi.org/10.1002/(SICI)1097-0142(19970215)79:4<857::AID-CNCR24>3.0.CO;2-Y

Caruana, R. (2006). *An Empirical Comparison of Supervised Learning Algorithms*. 161–168.

Chapelle, S. L. O., & Schölkopf, B. (2009). *Semi-supervised learning*. *20*(3), 2015975.

Cruz, J. A., & Wishart, D. S. (2006). *Applications of Machine Learning in Cancer Prediction and Prognosis*. 59–77. https://doi.org/10.1177/117693510600200030

Delen, D., Walker, G., & Kadam, A. (2005). *Predicting breast cancer survivability : a* *comparison of three data mining methods*. https://doi.org/10.1016/j.artmed.2004.07.002

Diao, L., Niu, D., Zang, Z., & Chen, C. (2019). Short-term weather forecast based on wavelet denoising and catboost. *Chinese Control Conference, CCC, 2019-July*(2018), 3760–3764. https://doi.org/10.23919/ChiCC.2019.8865324

Edeki, C., & Shardul, P. (2012). *Comparison of Data Mining Techniques used to Predict Cancer Survivability*. *10*(6).

Efitorov, A., Burikov, S., Dolenko, T., Laptinskiy, K., & Dolenko, S. (2015). Significant Feature Selection in Neural Network Solution of an Inverse Problem in Spectroscopy. In *Procedia Computer Science* (Vol. 66). https://doi.org/10.1016/j.procs.2015.11.012

Elmore, J. G., Wells, C. K., Lee, C. H., Howard, D. H., & F. (1994). Variability in radiologists interpretations of mammograms. *New England Journal of Medicine,* 1493–1499.

Eltoukhy, M. M., & Faye, I. (2014). An optimized feature selection method for breast cancer diagnosis in digital mammogram using multiresolution representation. *Applied Mathematics and Information Sciences, 8*(6), 2921–2928. https://doi.org/10.12785/amis/080629

Endo, A., Shibata, T., & Tanaka, H. (2010). *Comparison of SevenAlgorithms toPredict Breast Cancer Surviva*. 11–16.

Fawcett, T. (2006). *An introduction to ROC analysis*. *27*, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Gupta, S., Kumar, D., & Sharma, A. (2011). *DATA MINING CLASSIFICATION TECHNIQUES APPLIED FOR BREAST CANCER DIAGNOSIS AND*

PROGNOSIS. *2*(2), 188–195.

Harbeck, N & Gnant, M. (2017). Breast Cancer. *The Lancet*, (389), 1134–1150.

Henneges, C., Bullinger, D., Fux, R., Friese, N., Seeger, H., Neubauer, H., … Kammerer, B. (2009). Prediction of breast cancer by profiling of urinary RNA metabolites using Support Vector Machine-based feature selection. *BMC Cancer*, *9*, 1–11. https://doi.org/10.1186/1471-2407-9-104

Ho, T. K. (1995). Random Decision Forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *47*, 4–8.

Huang, M. L., Hung, Y. H., & Chen, W. Y. (2010). Neural network classifier with entropy based feature selection on breast cancer diagnosis. *Journal of Medical Systems*, *34*(5), 865–873. https://doi.org/10.1007/s10916-009-9301-x

Jordan, M. I., & Mitchell, T. M. (2015). *Machine learning: Trends, perspectives, and prospects*. *349*(6245).

Kang, P., Lin, Z., Teng, S., Zhang, G., Guo, L., & Zhang, W. (2019). Catboost-based framework with additional user information for social media popularity prediction. *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia*, 2677–2681. https://doi.org/10.1145/3343031.3356060

Khan, U., Shin, H., Choi, J. P., & Kim, M. (2007). *wFDT - Weighted Fuzzy Decision Trees for Prognosis of Breast Cancer Survivability*. 141–152.

Kim, J., & Shin, H. (2013). *Breast cancer survivability prediction using labeled , unlabeled , and pseudo-labeled patient data*. 613–618. https://doi.org/10.1136/amiajnl-2012-001570

Kleinberg, E. M. (2000). *On the Algorithmic Implementation of Stochastic Discrimination*. *22*(5), 473–490.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference of Artificial Intelligence*, (June).

Kohavi, R., & Sommerfield, D. (1995). Feature subset selection using the wrapper method: Overfitting and dynamic search space topology. *First International Conference on Knowledge Discovery and Data Mining*, 192–197. Retrieved from http://www.aaai.org/Papers/KDD/1995/KDD95-049.pdf

Li, Y., & Chen, Z. (2018). *Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction*. *7*(4), 212–216. https://doi.org/10.11648/j.acm.20180704.15

Mining, D. (2017). *Springer Series in Statistics The Elements of*.

Park, K., Ali, A., Kim, D., An, Y., Kim, M., & Shin, H. (2015). Robust predictive model for evaluating breast cancer survivability. *Engineering Applications of Artificial Intelligence*, *26*(9), 2194–2205. https://doi.org/10.1016/j.engappai.2013.06.013

Paul, F. V., & David, C. H. (2004). *Applications, Basics, and Computing of Exploratory Data Analysis By*.

Powers, D. M. W. (2007). *Evaluation : From Precision , Recall and F-Factor to ROC , Informedness , Markedness & Correlation*. (December).

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). *CatBoost : unbiased boosting with categorical features*. (Section 4), 1–11.

Roe, B. P., Yang, H., Zhu, J., Liu, Y., & Stancu, I. (2005). *Boosted decision trees as an alternative to artificial neural networks for particle identification*. *543*, 577–584. https://doi.org/10.1016/j.nima.2004.12.018

Saberian, M., Delgado, P., & Raimond, Y. (n.d.). *Gradient Boosted Decision Tree Neural Network*. 1–3.

Shukla, N., Hagenbuchner, M., Win, K. T., & Yang, J. (2018). Breast cancer data analysis for survivability studies and prediction. *Computer Methods and Programs in Biomedicine*, *155*, 199–208. https://doi.org/10.1016/j.cmpb.2017.12.011

Siegel, R., Naishadham, D., & Jemal, A. (2012). *Cancer Statistics , 2012*. https://doi.org/10.3322/caac.20138.Available

Society, I. B. (2013). *Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data*. *34*(1), 57–67.

Stehman, S. V. (2018). *Selecting and Interpreting Measures of Thematic Classification Accuracy*. *4257*(July). https://doi.org/10.1016/S0034-4257(97)00083-7

Sun, Y., Goodison, S., Li, J., Liu, L., & Farmerie, W. (2007). *Gene expression Improved breast cancer prognosis through the combination of clinical and genetic markers*. *23*(1), 30–37. https://doi.org/10.1093/bioinformatics/btl543

Thongkam, J., Xu, G., Zhang, Y., & Huang, F. (2009). Toward breast cancer survivability prediction models through improving training space. *Expert Systems With Applications, 36*(10), 12200–12209. https://doi.org/10.1016/j.eswa.2009.04.067

Tianqi, C., & Carlos, G. (2016). XGBoost: A Scalable Tree Boosting System. *The Journal of the Association of Physicians of India, 42*(8), 665.

Wang, H., Zheng, B., Won, S., & Sang, H. (2017). *A Support Vector Machine-Based Ensemble Algorithm for Breast Cancer Diagnosis A support vector machine-based ensemble algorithm for breast cancer diagnosis*. (December). https://doi.org/10.1016/j.ejor.2017.12.001

Wu, Q., Burges, Æ. C. J. C., & Svore, Æ. K. M. (2010). *Adapting boosting for information retrieval measures*. 254–270. https://doi.org/10.1007/s10791-009-9112-1