CANCER PREDICTION BASED ON DATA MINING USING DECISION TREE ALGORITHM

LOH KIN MING

Bachelor Of Computer Science (Software Engineering) With Honors

UNIVERSITI MALAYSIA PAHANG

UNIVERSITI MALAYSIA PAHANG

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name : LOH KIN MING

Date of Birth

Title : CANCER PREDICTION BASED ON DATA MINING USING DECION TREE ALGORITHM 2022/2023
Academic Session

I declare that this thesis is classified as:

☐ CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☐ RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☑ OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Universiti Malaysia Pahang reserves the following rights:

1. The Thesis is the Property of Universiti Malaysia Pahang
2. The Library of Universiti Malaysia Pahang has the right to make copies of the thesis for the purpose of research only.
3. The Library has the right to make copies of the thesis for academic exchange.

Certified by:

(Student's Signature)

New IC/Passport Number
Date: 16/02/2023

(Supervisor's Signature)

TS. DR. KOHBALAN A/L MOORTHY
SENIOR LECTURER
FACULTY OF COMPUTING
COLLEGE OF COMPUTING & APPLIED SCIENCES
UNIVERSITI MALAYSIA PAHANG
26600 PEKAN, PAHANG DARUL MAKMUR
TEL : 09-424 4661 FAX : 09-424 4606

Name of Supervisor
Date: 16/02/2023

NOTE : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

THESIS DECLARATION LETTER (OPTIONAL)

Librarian,
Perpustakaan Universiti Malaysia Pahang,
Universiti Malaysia Pahang,
Lebuhraya Tun Razak,
26300, Gambang, Kuantan.

Dear Sir,

CLASSIFICATION OF THESIS AS RESTRICTED

Please be informed that the following thesis is classified as RESTRICTED for a period of three (3) years from the date of this letter. The reasons for this classification are as listed below.

Author's Name
Thesis Title

Reasons (i)

(ii)

(iii)

Thank you.

Yours faithfully,

_____
(Supervisor's Signature)

Date:

Stamp:

Note: This letter should be written by the supervisor, addressed to the Librarian, *Perpustakaan Universiti Malaysia Pahang* with its copy attached to the thesis.

SUPERVISOR'S DECLARATION

I/We* hereby declare that I/We* have checked this thesis/project* and in my/our* opinion, this thesis/project* is adequate in terms of scope and quality for the award of the degree of *Doctor of Philosophy/ Master of Engineering/ Master of Science in ……..

(Supervisor's Signature)

Full Name : TS. DR. KOHBALAN A/L MOORTHY
SENIOR LECTURER
FACULTY OF COMPUTING
COLLEGE OF COMPUTING & APPLIED SCIENCES
Position : UNIVERSITI MALAYSIA PAHANG
26600 PEKAN, PAHANG DARUL MAKMUR
TEL : 09-424 4661 FAX : 09-424 4666

Date : 16/02/2023

(Co-supervisor's Signature)

Full Name :

Position :

Date :

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang or any other institutions.

LOH

(Student's Signature)

Full Name : LOH KIN MING

ID Number : CB19091

Date : 3 JUN 2022

CANCER PREDICTION BASED ON DATA MINING USING DECISION TREE ALGORITHM

LOH KIN MING

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Doctor of Philosophy/Master of Science/Master of Engineering

Bachelor Of Computer Science (Software Engineering) With Honors

UNIVERSITI MALAYSIA PAHANG

JUNE 2022

ACKNOWLEDGEMENTS

I can't thank my committee enough for their unwavering support and encouragement for DR. KOHBALAN A/L MOORTHY, my supervisor; Dr. DANAKORN NINCAREAN A/L EH PHON, course lecturer. I'd want to express my gratitude for the educational possibilities offered by my committee.

Without the help of my friends, Foong Kin Hong, Chew Min Wei, and Teo Voon Chuan, I would not have been able to finish this job, thank you for allowing me time away from you to research and write. You guys really deserve every good day. I would like to show my thanks to my parents as well, Mr. and Mrs. Loh. The countless times you kept support me during my difficult time will not be forgotten.

ABSTRAK

Kanser adalah salah satu penyakit yang paling membawa maut di dunia pada masa kini. Beberapa faktor keturunan, pembolehubah persekitaran, dan gaya hidup kontemporari hari ini semuanya menyumbang kepada perkembangan kanser. Di sesetengah negara maju, kanser telah menjadi punca utama kematian. Pengesanan awal kanser adalah pendekatan yang paling berkesan untuk mencegah kematian akibat barah. Diagnosis kanser adalah sukar, tetapi jika didapati cukup awal, ia boleh dirawat. Banyak kerja telah dilakukan dalam meramalkan kanser. Kaedah perlombongan data yang berbeza telah muncul dengan algoritma yang berkaitan telah diterima pakai oleh penyelidikan yang berbeza. Setiap kerja mempunyai beberapa batasan seperti kekurangan ramalan pintar, dan struktur yang tidak cekap yang mendorong untuk menangani masalah ini dan melaksanakan ramalan kanser. Saya telah mencadangkan satu penyelidikan iaitu ramalan kanser menggunakan perlombongan data. Matlamat utama penyelidikan ini adalah untuk menyediakan maklumat dan ramalan yang berguna ke arah kanser bukan sahaja untuk orang ramai, tetapi juga untuk industri sains dan industri berkaitan.

ABSTRACT

One of the most fatal diseases nowadays in the world will be cancer. Some hereditary factors, environmental variables, and today's contemporary lifestyle all contribute to the development of cancer. In some developed countries, cancer has become the leading cause of mortality. To prevent cancer mortality, early detection of cancer is the most efficiency way to approach. Cancer diagnosis is hard, but if it is found early enough, it may be treated. Many studies have been conducted in order to predict cancer. Different data mining method has showed up with its related algorithms were adopted by different research. Each work has some shortcomings like lack of intelligent prediction and an inefficient structure which eventually prompted me to take on this subject and develop a cancer prediction system. I have proposed a research which is cancer prediction using data mining. The main aim of this research is to provide a helpful information and prediction towards cancer not only for the public, but also for the science industry and related industry.

TABLE OF CONTENT

DECLARATION

TITLE PAGE

ACKNOWLEDGEMENTS ..... **ii**

ABSTRAK ..... **iii**

ABSTRACT ..... **iv**

TABLE OF CONTENT ..... **v**

LIST OF TABLES ..... **viii**

LIST OF FIGURES ..... **ix**

LIST OF SYMBOLS ..... **x**

LIST OF ABBREVIATIONS ..... **xi**

CHAPTER 1 INTRODUCTION **1**

1.1 INTRODUCTION 1

1.2 PROBLEM BACKGROUND 2

1.3 PROBLEM STATEMENTS 2

1.4 OBJECTIVE 3

1.5 RESEARCH SCOPE 3

1.6 SIGNIFICANT OF RESEARCH 3

1.7 SUMMARY 4

CHAPTER 2 LITERATURE REVIEW **5**

2.1 OVERVIEW 5

2.2 SUPPORT VECTOR MACHINE (SVM) 5

2.3 NAÏVE BAYES CLASSIFIER 5

2.4 DECISION TREE 6

2.5 COMPARATIVES ANALYSIS 6

2.6 SUMMARY 8

CHAPTER 3 METHODOLOGY 10

3.1 OVERVIEW 10

3.2 RESEARCH FLOWCHART 10

3.3 PROPOSED METHOD 12

3.4 DATASET 12

3.5 PERFORMANCE MEASUREMENTS 13

3.6 HARDWARE & SOFTWARE REQUIREMENTS 15

3.7 SUMMARY 15

CHAPTER 4 17

IMPLENTATION AND DISCUSSION 17

4.1 INTRODUCTION 17

4.2 BASIC STEP AND CONTRIBUTION 17

4.2.1 DATA PREPARATION 17

4.2.2 MISSING VALUE IMPUTATION AND PRE-PROCESSING 17

4.2.3 MODELLING, HYPER PARAMETER TUNNING NAD COMPARISON AFTER AND BEFORE TUNING 18

4.2.4 DECISION TREE CLASSIFIER PLOT 18

4.3 MISSING VALUE IMPUTATION 20

4.4 TRAINING AND TESTING in GOOGLE COLAB 22

4.5 MODELLING 22

4.5.1 DEFINING MODEL 26
4.6 HYPERPARAMETER TUNING 28
4.7 COMPARISON BEFORE AND AFTER HYPERPARAMETER TUNING 30
4.8 DECISION TREE CLASSIFIER PLOT AND FEATURE IMPORTANCE 31
4.9 RESULT AND DISCUSSION 32
4.9.1 MODELLING 32
4.9.2 HYPER PARAMETER TUNNING 35
4.9.3 COMPARISON BEFORE AND AFTER TUNING 36
4.9.4 FEATURE IMPORTANCE 36
4.10 SUMMARY 37

CHAPTER 5 CONCLUSION 38

5.1 INTRODUCTION 38
5.2 RESEARCH CONTRAINS AND CHALLENGES 39
5.3 FUTURE WORK 39
REFERENCES 40

LIST OF TABLES

Table 1 Comparative analysis study of existing Data mining based cancer prediction methods 7
Table 2 Table Confusion Matrix 14
Table 3 Hardware and Software 15
Table 4 CV Summary 34
Table 5 Parameter Tuning 35
Table 6 Comparison Score Before and After 36

LIST OF FIGURES

Figure 1 Flowchart 11
Figure 2 Breast Cancer Wisconsin (Diagnostic) Dataset 13
Figure 3 Flowchart of Basic Step 19
Figure 4 Training and Testing Accuracy for KNN Classifier 32
Figure 5 Training and Testing Accuracy for Decision Tree Classifier 33
Figure 6 Features Importance 36

LIST OF SYMBOLS

LIST OF ABBREVIATIONS

SVM	Support Vector Machine
ANN	Artificial Neural Network
CFS	Cluster-Based Feature
SSD	Solid-State Drive
GHz	Gigahertz
MHz	Megahertz
Hz	Hertz
DT	Decision Tree
RF	Random Forest
NN	Neural Network
TN	True Negative
FN	False Negative
FP	False Positive
TP	True Positive

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Cancer is diagnosed when some blood cells grow out of control and spreads to the rest of the body part. It could diagnose in both men and women if they manage having a bad habit like smoking, drinking and stay up late night etc (A.Priyanga, S.Prakasam, 2013). However, with today's technology, early cancer detection can aid in forecasting, resulting in a higher likelihood of survival, since it can enable patients receive prompt therapeutic therapy.

Data mining is the process of extracting and identifying useful information and knowledge from large data sets using statistical, mathematical, and artificial intelligence techniques. It involves methodologies at the confluence of machine learning, statistics, and database systems. Once the data and patterns have been discovered, they may be leveraged to make business choices. It can turn raw data into valuable data in a variety of study domains and identify crucial patterns to forecast future medical trends (V.Krishnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra, 2013). Prediction is the process of identifying data points based on the description of another associated data value.

Data mining that used in healthcare does help the medical technology improved. It has a lot of potential when it comes to improving health-care systems. Data mining identifies best practices for improving treatment and lowering costs by combining data and analytics. The number of patients in each group may be forecasted using data mining (Ahmad Al-Aiad, Salsabil Abualrub,Yazan Alnsour, Mohammad Alsharo, 2020). In general, it identifies the links between disorders and treatment efficacy. Data mining aids in the identification of successful medical treatment patterns for various illnesses.

Classification is the most commonly utilized data mining approach for real-world healthcare problems since it provides categories to a set of data to help in more accurate predictions and analysis. Classification is a sort of machine learning that employs supervised learning, such as while using classification method to predict the type of disease based on the symptoms.

The Decision Tree algorithm is a data mining approach for predicting data. Decision tree algorithm is different from other supervised learning algorithms, the algorithm may be used to solve issues involving regression and classification. Develop a training model that can predict the class or value of the target variable using simple decision rules drawn from training data is the final goal of the decision tree algorithm. Start at the top of the tree when using Decision Trees to forecast a record's class label. The values of the root attribute and the attribute of the record are then compared. Then, based on the comparison, follow the branch that corresponds to that value and proceed to the next node.

1.2 PROBLEM BACKGROUND

Cancer research has seen a steady change throughout the last few decades. Several approaches had used by scientists such as early-stage screening, to detect cancer forms before they start showing symptoms. They've also developed new methods for predicting cancer treatment results in the early stages. As a result of the advent of new technologies in the field of medicine, large amounts of cancer data have been accumulated and made available to the medical research community. However, one of the most intriguing and difficult challenges for clinicians is accurately predicting the fate of an illness. The prediction may expect multiple outcomes.

1.3 PROBLEM STATEMENTS

The problem that currently facing of this research is:

1.) Having difficulty for making early detection because of many cancer types.

2.) Lack of existing system related to cancer prediction that can be research.

3.) Low prediction accuracy due to small data set.

1.4 OBJECTIVE

There are few objectives in this research which are:

1.) To analyse existing system related to cancer prediction (breast cancer).

2.) To develop a cancer prediction method for cancer prediction to improve the accuracy for breast cancer.

3.) To validate the propose solution with previous work.

1.5 RESEARCH SCOPE

User Scope:

i. People that in a middle age who care about they health.

ii. Doctor who needs this research in case to prevent cancer.

System Scope:

i. Covered topic in subject medical sciences, medical and etc.

ii. Covered cancer prediction by using data mining algorithm.

Development Scope:

i. Contains a lot of research based of disease prediction such as breast cancer.

ii. Using Python language in Google Colab.

1.6 SIGNIFICANT OF RESEARCH

Prediction of cancer by using data mining methods will provide whole new insights into the medicinal benefit.

Through this research, the community will further realize that predicting cancer is helpful as the public can know earlier and get control over it. People and medical

institutions may also consider this prediction as it is a serious stuff which it may affect our life.

Furthermore, the findings of this study will contribute to future research into the numerous medicinal benefits of cancer prediction in a range of circumstances.

1.7 SUMMARY

This research is use to predict cancer as early as possible so that the patient can get his/her treatment earlier before the cancer get any worse. The purpose of this study is to identify which traits are the most helpful in cancer prediction and to look for general trends that might assist us choose models and hyper parameters. In this case, data mining method are applied to match a function that can predict discrete class of certain new input.

CHAPTER 2

LITERATURE REVIEW

2.1 OVERVIEW

Many research has been conducted on cancer prediction. The research works applied on different method of machine learning algorithms. Some of the previous research will be bring up in this section. Prior studies of Support Vendor Machine (SVM), K-Means Clustering, and decision tree are few widely used method in data mining that will be discuss. This research is to focus on prior studies of cancer prediction by using K-Means Clustering, Support Vector Machine (SVM), Decision Tree data mining methods.

2.2 SUPPORT VECTOR MACHINE (SVM)

The Support Vector Machine is a learning machine that uses a hypothesis linear function space in a high-dimensional feature space and is taught using a learning approach based on statistical learning theory's optimization theory. Concept of the Support Vector Machine method is to find the hyperplane that differentiates the two class which are positive and negative (Ade Jamala, Annisa Handayania, Ali Akbar Septiandria, Endang Ripmiatina, Yunus Effendib, 2018).

2.3 NAÏVE BAYES CLASSIFIER

A probabilistic graphical model such as the Naive Bayes classifier is based on Bayes' Theorem which offers an explanation of the probability of associating certain classes at certain instances (Maimon, O., and Rokach, L, 2005). The Naive Bayes classifier model has an assumption that features are conditionally independent. (Haifeng Wang, Sang Won Yoon, 2015)

2.4 DECISION TREE

Decision trees are a type of prediction model that enables firms to successfully collect data. This algorithm allows users to quickly observe how the data inputs affect the outputs. Random forest is a predictive analytics model that is made up of many decisions tree models and it also referred as black box machine learning techniques because the random forest's outputs are not always easy to understand based on their inputs.

2.5 COMPARATIVES ANALYSIS

The first research I look for is titled as "Breast Cancer Using Data Mining Method" state that the author used data mining technique like Naive Bayes classifier, AdaBoost tree, Artificial Neural Network (ANN) and Support Vector Machine (SVM). Eventually based on the clinical data of a variety of patients, this study was able to develop an accurate model for predicting breast cancer. Strength and limitation of this paper is that it gives better performance but they may consume much amount of time.

Coming up is research which titled as "Predicting Breast Cancer using effective Classification with Decision Tree and K Means Clustering technique". The authors state that they used data mining technique like Decision tree and K-means Clustering. Eventually they managed to predict cancer for a patient. Advantages of this paper is two techniques used in this research, which are classification and clustering, with effectively used these two techniques, it can give better performance. Disadvantage of this paper is that the algorithm may be used more time than usual to perform.

The third research is titled as "Optimized machine learning model using Decision Tree for cancer prediction". The authors stated that they have used the technique like Random Forest, Decision Tree and Support Vector Machine (SVM). Eventually they bring up the result of the research which every technique has high accuracy to cancer prediction.

Classification method has been used by many papers. Research "Breast Cancer Using Data Mining Method" has used data mining classification method to assess and

find an accurate model for predicting breast cancer incidence while paper "Predicting Breast Cancer using effective Classification with Decision Tree and K Means Clustering technique" has used data mining classification and clustering method to predict breast cancer that will majorly help the medical researchers to analyse and then do a prediction about the cancer for a patient. Paper "Optimized machine learning model using Decision Tree for cancer prediction" has used classification method and the authors test the experiment in three stages, first stage with all features of breast cancer; second stage included decision tree and the final stage feature selection. The experimental results of their used method all has a very high accurate percentage. The comparative analysis study of existing Data Mining based on cancer prediction methods is shown as below:

Table 1 Comparative analysis study of existing Data mining based cancer prediction methods

Technique used	**Result**	**References**
1. Support Vector Machine (SVM) 2. Artificial Neural Network (ANN) 3. Naive Bayes classifier 4. AdaBoost tree	The goal of this study is to compare and find an accurate model for predicting the occurrence of breast cancer based on the clinical data of diverse patients. The findings of this study show a complete trade-off between various tactics, as well as a full evaluation of the models, clinicians, and patients who can benefit from the feature identification outcome in the prevention of breast cancer.	Breast Cancer Using Data Mining Method (Haifeng Wang, 2015)
1. Data mining	They gather data from the UC Irvine Machine Learning Repository dataset in order to develop a suggested model. The assessment	Predicting Breast Cancer using effective

2. Decision tree 3. K-means Clustering	of two discrete machine learning algorithms for breast cancer prediction can assist medical researchers such as scientists and physicians in analysing and predicting any cellular cancer present for a patient who falls into the category contained in the data sets.	Classification with Decision Tree and K Means Clustering technique (Samiksha Marne, Shweta Churi, Maheshwari Marne, 2020)
1. Decision Tree (DT) 2. Random Forest (RF) 3. Support Vector Machine (SVM) 4. Neural Network (NN)	They put the classifiers to the test in three phases, the first of which included all breast cancer characteristics. The second step contained a decision tree (DT)-based cluster, followed by feature selection in the third stage. Random Forest (RF) achieves 99.10 percent accuracy in all features, 99.37 percent accuracy in decision tree (DT) based cluster Random Forest (RF), and 99.50 percent accuracy in cluster-based feature selection (CFS) with just four features. The experimental findings of their CFS technique, as well as DT, RF, SVM, and NN classifiers, are provided, along with their performance.	Optimized machine learning model using Decision Tree for cancer prediction (T. Chandrasegar, Sai Brahma Nikhilesh Vutukuri, 2019)

2.6 SUMMARY

Many studies have been conducted to predict cancer risk. For the identification of cancer risk, several writers have proposed various methodologies. Throughout the

research of the related prior studies of cancer prediction, machine learning like Support Vector Machine (SVM), decision tree, classification method etc. does a lot of help to the medical industry for preventing cancer. Cancer prediction is unquestionably a complicated and nondeterministic task, and there are a plethora of cancer prediction tests available.

CHAPTER 3

METHODOLOGY

3.1 OVERVIEW

This section provides an introduction to the methodology and data mining techniques that is going to use in this research. Various approaches are used in data mining. Diverse methods serve different goals, each with its own set of benefits and drawbacks. One of the most major techniques in data mining is classification. It converts data into pre-determined objectives. Because the objectives are predetermined, it is supervised learning. In this research, one of the classification algorithms will use which is the C4.5 decision tree.

3.2 RESEARCH FLOWCHART

It's a knowledge representation structure made up of nodes and branches in a shape of tree with each internal non-leaf node labelled with attribute values. The values of the characteristics of an internal node are indicated on the branches that emerge from it. Every node has a class that corresponds to a target attribute value. Induction modelling is commonly done with tree-based models that contain categorization. Data mining is best served by decision tree models. They are simple to build, understand, and connect with database systems, and they offer equivalent or higher accuracy in many applications.

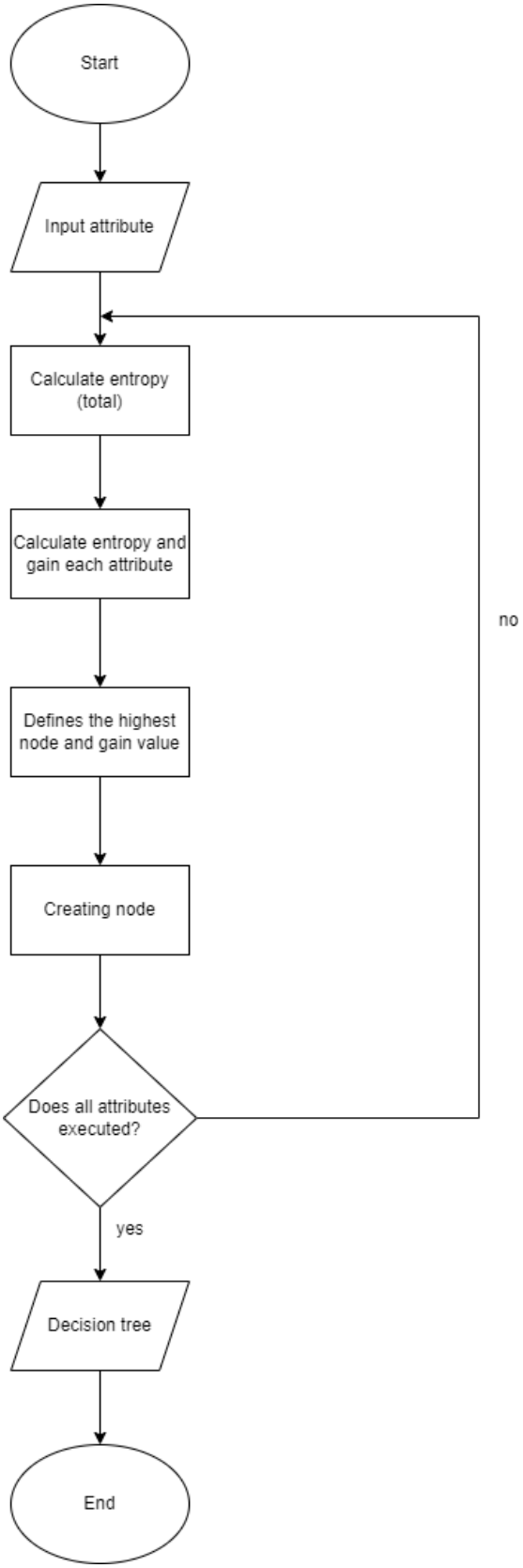


Figure 1 Flowchart

3.3 PROPOSED METHOD

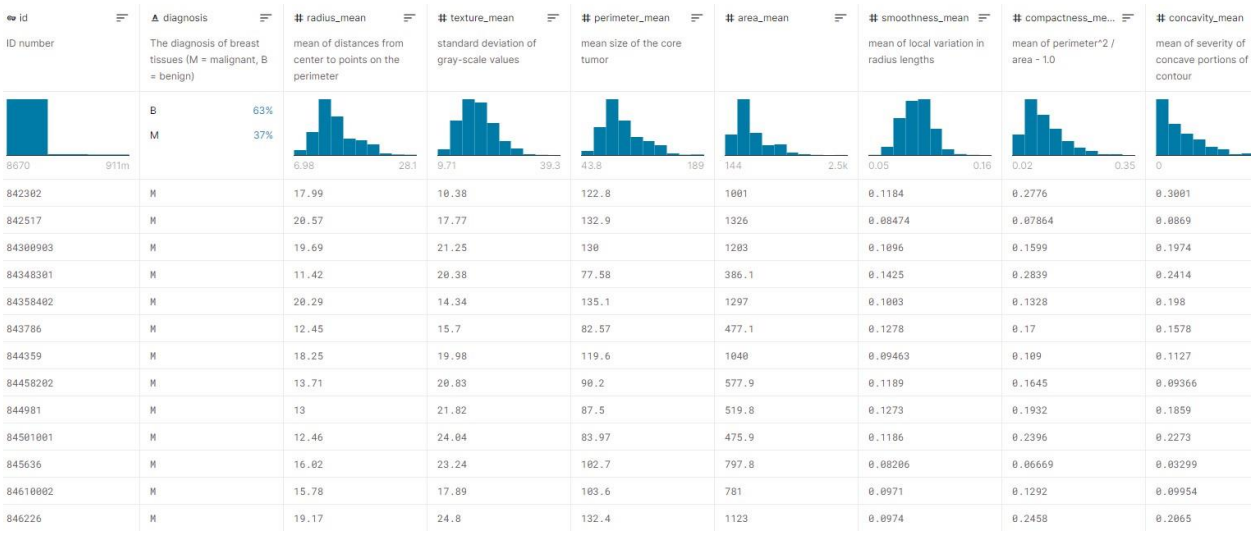
I have proposed C4.5 decision tree algorithm in this research as this strategy is the most common and efficient algorithm in a decision tree-based method. The decision tree algorithm builds a tree model using only one attribute's values at a time. (Masud Karim, Rashedur M. Rahman, 2013). The algorithm will sort the dataset on the attribute's value first. Then the algorithm will continue searches the dataset for regions that clearly contain only one class and labels them as leaves. After that it picks another attribute for the remaining areas with one class or even more class and continues branching with only the number of instances in those regions until it has no leaves left or no attribute that can be utilized to produce leaves in the conflicting regions.

3.4 DATASET

Breast Cancer Wisconsin (Diagnostic) Dataset is a dataset that predict whether the cancer is benign or malignant. A digitized picture of a fine needle aspirate (FNA) of a breast mass was used to create this dataset's characteristics. They describe the properties of cell nuclei as Figure 2.

Only a few attributes are shown in the table: Each cell nucleus has an ID number, a diagnosis (M = malignant, B = benign), and 10 real-valued features: radius (average distance between centre and points on the perimeter), area, perimeter, texture (standard deviation of grey-scale values), concavity (severity of concave portions of the contour), smoothness (local variation in radius lengths), symmetry, compactness (perimeter2 / area - 1.0), concave points (number of concave portions of the contour), fractal dimension ("coastline approximation" -1).

For each picture, the mean, standard error, and "worst" or worst (mean of the three largest values) features were computed, yielding 30 features. For example, field 3 represents Mean Radius, field 13 represents Radius SE, and field 23 represents Worst Radius. All feature values have four significant digits recoded. There are no missing attribute values, and the distribution of classes is 357 benign and 212 malignant.



id ID number	diagnosis The diagnosis of breast tissues (M = malignant, B = benign)	radius_mean mean of distances from center to points on the perimeter	texture_mean standard deviation of gray-scale values	perimeter_mean mean size of the core tumor	area_mean	smoothness_mean mean of local variation in radius lengths	compactness_me... mean of perimeter^2 / area - 1.0	concavity_mean mean of severity of concave portions of contour
8670 911m	B 63% M 37%	6.98 28.1	9.71 39.3	43.8 189	144 2.5k	0.05 0.16	0.02 0.35	0
842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001
842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869
84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974
84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414
84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198
843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578
844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127
84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366
844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859
84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273
845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299
84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954
846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065

Figure 2 Breast Cancer Wisconsin (Diagnostic) Dataset

3.5 PERFORMANCE MEASUREMENTS

This section provides an introduction to the methodology and data mining techniques that is going to use in this research. A variety of methods are used in data mining. Different approaches are used for various goals, each with its own set of benefits and drawbacks. One of the most significant strategies for data mining is categorization. It converts data into pre-determined objectives. It is guided learning since the objectives are predetermined. [Mrs.Sagunthaladevi.S, Dr. Bhupathi Raju Venkata Rama Raju, 2016] In this research, one of the classification algorithms will use which is the C4.5 decision tree.

Accuracy

The accuracy of a classifier is calculated as the ratio of the total number of correctly predicted samples by the total number of samples [Venu Gopal Kadamba, 2021].

Accuracy = Total number of correctly predicted samples/ Total number of samples (3.1)

The accuracy measure used to evaluate the classifier when the data set is balanced, and at the same time the accuracy measure should not be applied when the data set is unbalanced. Consider a data set with two target classes and 100 samples, 95 of which are from class 1 and 5 from class 2. When trying to create a classifier for the data set above, the classifier will be affected towards class 1 and will predict that all of the samples will be class 1. This will result in a 95% accuracy, which is incorrect. To prevent making this error, only balanced data sets should be utilized for accuracy metrics.

Confusion Matrix

A confusion matrix is a method of describing a classification algorithm's performance. If you have an unbalanced amounts of observations in each class or if your dataset has more than two classes, classification accuracy alone might be deceptive. Consider the case of a binary classification task i.e. the number of target classes are 2. A typical confusion matrix with two target classes (say “Yes” and “No”):

Table 2 Table Confusion Matrix

	Predicted: NO	Predicted: YES
Actual: NO	True Negative (TN)	False Positive (FP)
Actual: YES	False Negative (FN)	True Positive (TP)

In a confusion matrix, there are four key terms which are True Positive (TP) where the cases are predicted to “Yes” and actually belonged to class “Yes”; True Negative (TN) where the cases are predicted to “No” and actually belonged to class “No”; False Positive (FP) where the cases are predicted to “Yes” and actually belonged to class “No”; False Negative (FN) where the cases are predicted to “No” and actually belonged to class “Yes” (Jason Brownlee, 2016); The accuracy of the classifier of the confusion matrix can be calculated using the below formula:

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN) \quad (3.2)$$

3.6 HARDWARE & SOFTWARE REQUIREMENTS

Table 3 Hardware and Software

Hardware	**Software**
CPU: Intel Core i5 8400 @ 2.80GHz RAM: 16.0GB Dual-Channel Unknown @ 1197MHz Motherboard: ASUSTeK COMPUTER INC. PRIME H310M-K (LGA1151) Graphics: PHL 242E1GJ (1920x1080@144Hz), 4096MB ATI Radeon RX 570 Series (MSI) Storage: 223GB KINGSTON SUV500240G (SATA-2 (SSD)), 447GB KINGSTON SA400S37480G (SATA-2 (SSD))	Microsoft Word 2019 Microsoft Excel 2019 Google Chrome Version 101.0.4951.67 (64-bit) (Official Build)

3.7 SUMMARY

In this paper one of the classification algorithms will proposed which is the decision tree technique throughout the research. With the Breast Cancer Wisconsin (Diagnostic) Dataset, I can measure the performance in way of accuracy and confusion

matrix. Cancer prediction is unquestionably a complicated and nondeterministic task, and there are a plethora of cancer prediction tests available.

CHAPTER 4

IMPLENTATION AND DISCUSSION

4.1 INTRODUCTION

This chapter is briefly discussing the implementation of the research. The flow of this chapter is started from the data imputation step for comparing the 2 algorithm which is the KNN algorithm and decision tree algorithm. The process training and testing dataset are getting from the results, the discussion of the performance and lastly the comparison of the results with the previous method. About 70% from dataset was used for training and another 30% was used for testing purpose. The implementation is defined to meet the objectives that were stated to show the result is relevance to the research. Hyper Parameter Tuning was used in the experiment where the dataset was divided into training and testing phase.

4.2 BASIC STEP AND CONTRIBUTION

4.2.1 DATA PREPARATION

In order to achieve a successful result, data preparation is essential. The dataset must be prepared before the network design methods can begin. The dataset is organised and the normalisation procedure is completed using Microsoft Excel.

4.2.2 MISSING VALUE IMPUTATION AND PRE-PROCESSING

Missing value imputation is an own contribution and its used to impute missing data with mean imputation. Then, we will see the data can be successfully impute or not. If not, we will remove the rows and columns that contain null values.

Pre-Processing. This step is used to define target data that if where the cancer type was Benign, the value will be 0; while if the cancer type was Malignant, the value will be 1.

4.2.3 MODELLING, HYPER PARAMETER TUNNING NAD COMPARISON AFTER AND BEFORE TUNING

The modelling step is used to test and train the accuracy by using 0.3 as default score for test_size and X.shape for random.state so the data will be divided equally. Hyper Parameter Tuning. This is a process of searching the ideal model architecture to get the best score and best parameter of both algorithms. Comparison Between Before and After Tuning. This step is used to compare both algorithm before and after tuning score using Decision Tree Classifier.

4.2.4 DECISION TREE CLASSIFIER PLOT

This step is using Decision Tree Classifier to plot a tree figure and show the result of the parameter for the algorithm.

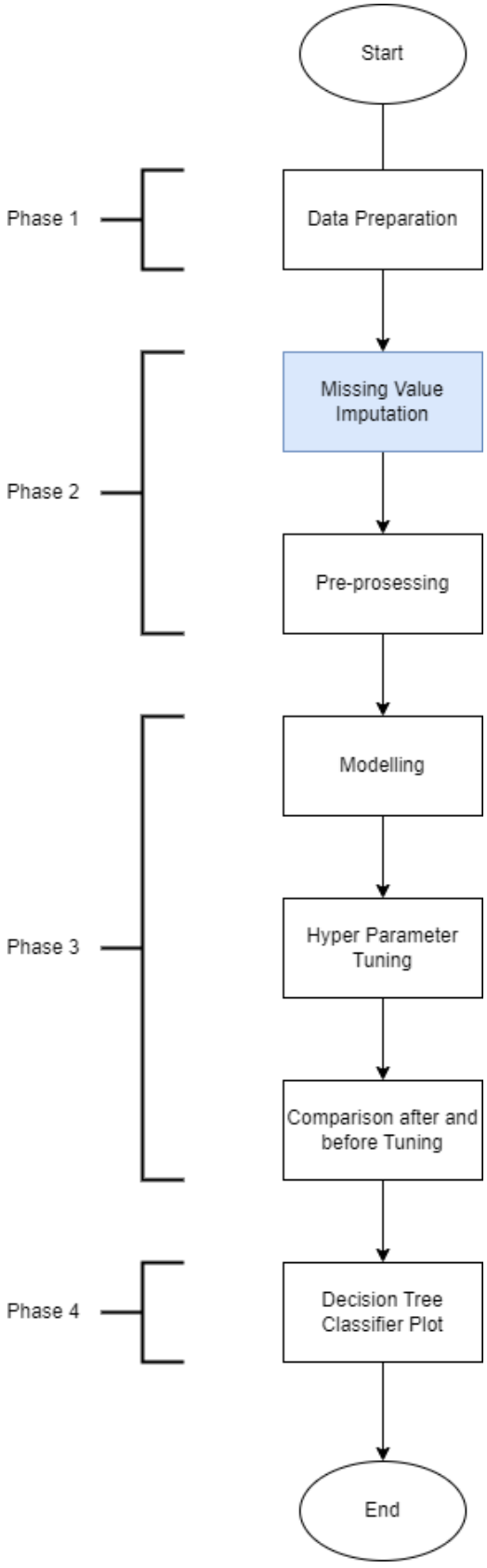


Figure 3 Flowchart of Basic Step

4.3 MISSING VALUE IMPUTATION

This line of code is a method to detect any missing values in the dataset. By getting the total of missing value, in each column, sum() method is implemented in it. The missingno package (msno) is to generate a visualize bar graph that represent the missing value.

```
import missingno as msno



cancer.isnull().sum()

msno.matrix(cancer)
```

To check whether the missing value imputation is in a good function, SimpleImputer is implemented. SimpleImputer function has a parameter which is strategy that has four options, one of the options is mean imputation method. Mean method is used to replace missing value using the mean of the column.

```
# Mean Imputation

from sklearn.impute import SimpleImputer



cancer_mean = cancer.copy()

mean_imputer = SimpleImputer(strategy='mean')

cancer_mean['radius_mean'] = mean_imputer.fit_transform(cancer_mean['radi
us_mean'].values.reshape(-1,1))
```

As a result, the line of code from mean imputation show that that is no missing value exist in this dataset as there is no value could be input. In this case, dropna() function is used to drop the column that contains NaN in order to proceed with the training and testing process.

```
cancer.drop(columns=['id', 'Unnamed: 32'], inplace = True)
```

4.4 TRAINING AND TESTING in GOOGLE COLAB

These codes showed the test data for X.shape which is test_size is equalled to 30% while the random state is 3030.

```
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y,
                                                    stratify = y,
                                                     test_size = 0.3,
                                                    random_state = 3030)
```

4.5 MODELLING

The script is about to plot a graph of the accuracy for testing and training according to the size of the training and testing function for KNN Classifier. The result of the output of the code indicates that the training and testing accuracy both has slight decrease for the KNN Classifier.

```
k = range(1,100,2)

testing_accuracy = []

training_accuracy = []

score = 0



for i in k:

    knn = KNeighborsClassifier(n_neighbors = i)

    knn.fit(X_train, y_train)
```

```
    y_predict_train = knn.predict(X_train)
    training_accuracy.append(accuracy_score(y_train, y_predict_train))


    y_predict_test = knn.predict(X_test)
    acc_score = accuracy_score(y_test,y_predict_test)
    testing_accuracy.append(acc_score)


    if score < acc_score:
        score = acc_score
        best_k = i
```

```
sns.lineplot(k, training_accuracy)

sns.scatterplot(k, training_accuracy)

sns.lineplot(k, testing_accuracy)

sns.scatterplot(k, testing_accuracy)

plt.legend(['training accuracy', 'testing accuracy'])

print('This is the best K for KNeighbors Classifier: ', best_k, '\nAccura
cy score is: ', score)
```

The script is about to plot a graph of the accuracy for testing and training according to the size of the training and testing function for Decision Tree Classifier. The result of the output of the code has shown that the testing accuracy has an unstable and inconsistent for the Decision Tree Classifier.

```
depth = range(1,25)

testing_accuracy = []

training_accuracy = []

score = 0



for i in depth:

    tree = DecisionTreeClassifier(max_depth = i, criterion = 'entropy')

    tree.fit(X_train, y_train)



    y_predict_train = tree.predict(X_train)

    training_accuracy.append(accuracy_score(y_train, y_predict_train))



    y_predict_test = tree.predict(X_test)

    acc_score = accuracy_score(y_test,y_predict_test)

    testing_accuracy.append(acc_score)
```

```
    if score < acc_score:
        score = acc_score
        best_depth = i


sns.lineplot(depth, training_accuracy)
sns.scatterplot(depth, training_accuracy)
sns.lineplot(depth, testing_accuracy)
sns.scatterplot(depth, testing_accuracy)
plt.legend(['training accuracy', 'testing accuracy'])
```

4.5.1 DEFINING MODEL

By using the KNN Classifier with best K score and Decision Tree Classifier with best depth score, it will show the first 5 cv score in the model and by using the 5 cv score , the mean, standard and recall score are calculated. The model of showing the cv score, mean score, std score and recall score is showing with the script below.

```
def model_evaluation(model, metric):

    model_cv = cross_val_score(model, X_train, y_train, cv = StratifiedKF
old(n_splits = 5), scoring = metric)

    return model_cv




knn_cv = model_evaluation(knn, 'recall')

tree_cv = model_evaluation(tree, 'recall')
```

```
for model in [knn, tree]:

    model.fit(X_train, y_train)
```

```
score_cv = [knn_cv.round(5), tree_cv.round(5)]

score_mean = [knn_cv.mean(), tree_cv.mean()]

score_std = [knn_cv.std(), tree_cv.std()]

score_recall_score = [recall_score(y_test, knn.predict(X_test)),

            recall_score(y_test, tree.predict(X_test))]
```

```
method_name = [ 'KNN Classifier', 'Decision Tree Classifier']

cv_summary = pd.DataFrame({

    'method': method_name,

    'cv score': score_cv,

    'mean score': score_mean,

    'std score': score_std,

    'recall score': score_recall_score

})

cv_summary
```

4.6 HYPERPARAMETER TUNING

For the hyperparameter tunning, there are few option that fits for each of the parameter which is “criterion”, “splitter”, “max_depth”, “min_sample_leaf”, “class_weight” and “random_state”. Model parameters are learned from data and hyper-parameters are tuned to get the best fit. The script are shown as below

```
tree = DecisionTreeClassifier(max_depth = 3, random_state = 3030)



hyperparam_space = {

    'criterion': ['gini', 'entropy'],

    'splitter': ['best', 'random'],

    'max_depth': [3, 5, 7, 9, 11],

    'min_samples_leaf': [3, 9, 13, 15, 17],

    'class_weight': ['list', 'dict', 'balanced'],

    'random_state': [3030]

}
```

```
grid = GridSearchCV(

                tree,

                param_grid = hyperparam_space,

                cv = StratifiedKFold(n_splits = 5),
```

```
                scoring = 'recall',

                n_jobs = -1)
```

```
grid.fit(X_train, y_train)
```

```
print('best score', grid.best_score_)

print('best param', grid.best_params_)
```

4.7 COMPARISON BEFORE AND AFTER HYPERPARAMETER TUNING

Hyperparameter tuning is choosing a set of optimal hyperparameters for Decision Tree Classifier. A hyperparameter is a model argument whose value is set before the learning process begins. In these few line of codes, the comparison state by the method and score.

```
tree.fit(X_train, y_train)

tree_recall = (recall_score(y_test, tree.predict(X_test)))



grid.best_estimator_.fit(X_train, y_train)

grid_recall = (recall_score(y_test, grid.predict(X_test)))
```

```
score_list = [tree_recall, grid_recall]

method_name = ['Decision Tree Classifier Before Tuning', 'Decision Tree C
lassifier After Tuning']

best_summary = pd.DataFrame({

    'method': method_name,

    'score': score_list

})

best_summary
```

4.8 DECISION TREE CLASSIFIER PLOT AND FEATURE IMPORTANCE

The first coding is how to plot a decision tree figure of the cancer dataset. The second coding show that creating a table by calculating each of the imp value.

```
plt.figure(figsize=(15,8))


plot_tree(grid.best_estimator_, feature_names = list(X), class_names = ['
Benign','Malignant'], filled = True)


plt.title('Tree Plot')


plt.show()
```

```
importance_table = pd.DataFrame({


    'imp': grid.best_estimator_.feature_importances_


}, index = X.columns)


importance_table.sort_values('imp', ascending = False)
```

```
importance_table.sort_values('imp', ascending = True).plot(kind = 'barh',
 figsize = (15,8))
```

4.9 RESULT AND DISCUSSION

4.9.1 MODELLING

4.9.1.1 KNN CLASSIFIER

The training accuracy is at its peak 100% when the depth is 0. Then its decrease consistently while the depth was increased. So as the testing accuracy, it has a start off 95% at first then it reached the peak which has a 95.9% of accuracy, since after the depth is increased, the accuracy slightly drops a little and when the depth reached 100, the accuracy drops below 90%. The best accuracy score for KNN Classifier is 95.9%. This model shows underfitting because the training accuracy and testing accuracy are both decreased.

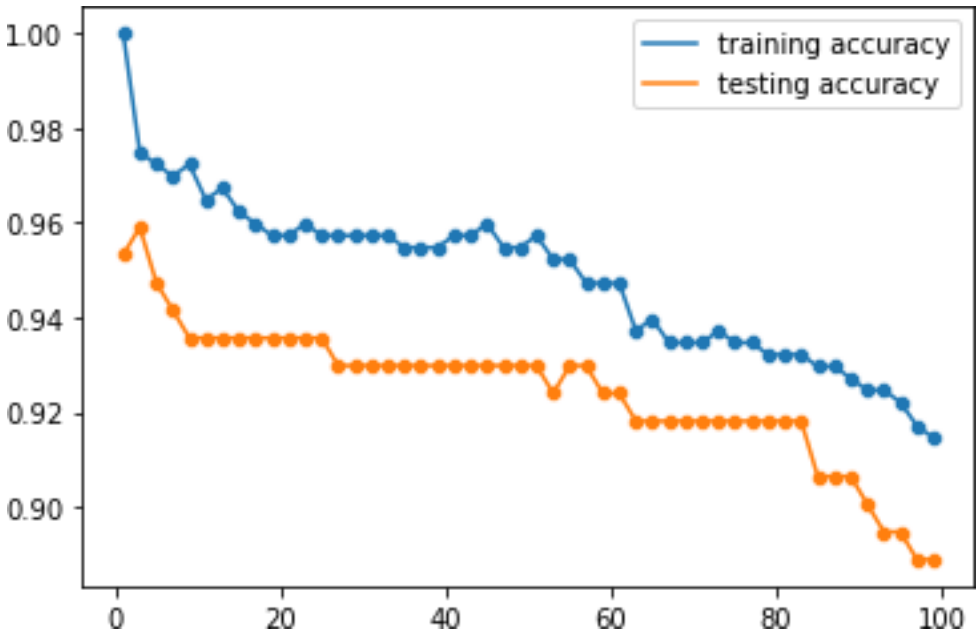


Figure 4 Training and Testing Accuracy for KNN Classifier

4.9.1.2 DECISION TREE CLASSIFIER

The training accuracy has reached 93.5% when the depth is 1. Then its has raise consistently while the depth was increased and eventually reached and maintained in 100% accuracy. Testing accuracy has below 88% at first then it has a sudden raise form below 88% to 97.5%, then it has a drop to 94% when the depth is 5. The testing accuracy is not stable until the depth increased until 25 and its accuracy is 93%. The best accuracy score for Decision Tree Classifier is 97.07%. This model shows overfitting because the training accuracy is overwhelming but the testing accuracy are not stable.

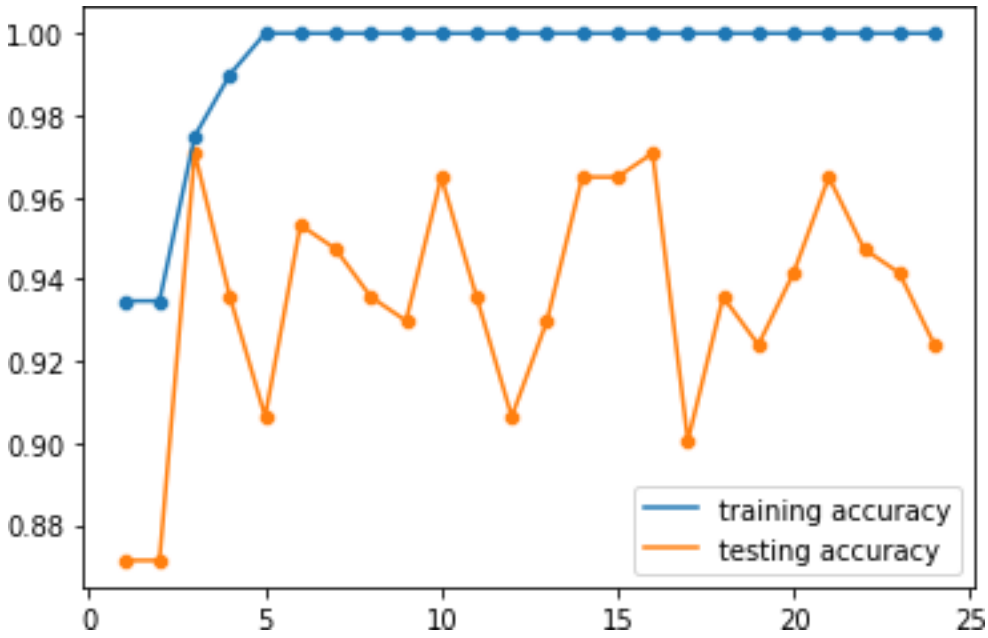


Figure 5 Training and Testing Accuracy for Decision Tree Classifier

4.9.1.3 DEFINE MODEL

As the output, the mean score, std score and recall score for KNN Classifier has a higher average than Decision Tree Classifier which is 0.905287, 0.049522, and 0.890625. From the cross validation and model evaluation processes, Decision Tree Classifier is still being choose for hyper parameter tunning even the score is indicated overfitting.

Table 4 CV Summary

method	cv_score	mean score	std score	recall score
KNN Classifier	[0.83333, 0.93333, 0.96667, 0.93103, 0.86207]	0.905287	0.049522	0.890625
Decision Tree Classifier	[0.9, 0.93333, 0.73333, 0.93103, 0.86207]	0.871954	0.073970	0.921875

4.9.2 HYPER PARAMETER TUNNING

After the tunning process, the best score for Decision Tree Classifier has been increased to 95% and the best parameter for each hyperpara has been selected to fit its best. The table below indicated that the parameter and the best fit parameter (highlighted).

Table 5 Parameter Tuning

	Parameter				
Criterion	Gini			Entropy	
Splitter	best			random	
Max_depth	3	5	7	9	11
Min_samples_leaf	3	9	13	15	17
Class_weight	List		Dict		Balanced
Random_state	3030				

4.9.3 COMPARISON BEFORE AND AFTER TUNING

Before tuning the recall score was 0.921875, after tuning the recall score has increased to 0.9375. This is because after tuning the algorithm find the best parameter that fit to the algorithm so that it will process with its best form. The score has shown as below table:

Table 6 Comparison Score Before and After

method	score
Decision Tree Classifier Before Tuning	0.921875
Decision Tree Classifier After Tuning	0.937500

4.9.4 FEATURE IMPORTANCE

The feature importance process is to check again the data from the dataset whether the data is important to predict. By using this function, out of 30 features, only 4 are important to this prediction.

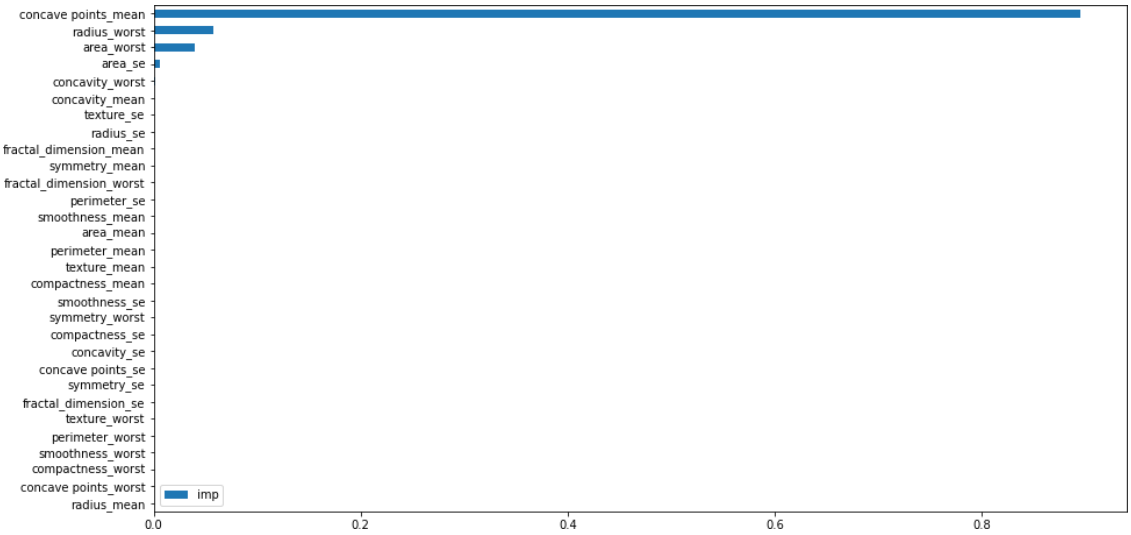


Figure 6 Features Importance

4.10 SUMMARY

Chapter 4 is mainly focus on design and implementation of the algorithm and coding. In this chapter, missing value imputation was used to check whether the dataset have any missing value, if it exists any missing value but if it is NaN, we drop the column to continue to comparison process. The modelling of Decision Tree classifier and KNN classifier is explained in detail. Eventually the Decision Tree Classifier is choosing from the cross validation and model evaluation processes even it is overfitting. After implement the Hyper Parameter Tuning with Decision Tree Classifier, there are some improvements in the accuracy compare to the original code before tuning.

CHAPTER 5

CONCLUSION

5.1 INTRODUCTION

Decision Tree classifier is one of the methods in Data Mining. Many methods of data mining have been used widely in other classification or prediction in other field problem. In this research, decision tree method is used in classification for cancer prediction. From the result obtained from comparison between Decision Tree and KNN classifier, Decision Tree has the highest score with 0.92. Even the Tree model indicated overfitting, I still choose to use this score to continue the process.

There are few essential phases in comparison Decision tree classifier and KNN classifier's training and testing network design which is data preparation, contribute missing value imputation, pre-processing, modelling, Hyper parameter tunning, Comparison after and before tunning, and Decision tree classifier pot. The training and testing basic procedure applied using Google Colab platform to complete. Data preparation is a necessary in the process in order to give a good result. From the cross-validation process, the KNN model has the highest score with 0.9 but after model evaluation using recall metric, the Tree model has the highest score with 0.92. Even the Tree model indicated overfitting, it still be choose to use this score to continue the process. Finally, when the choosing algorithms being tuning after the process, the result is then generated.

Several tests were conducted by using the Breast Cancer Wisconsin (Diagnostic) Data Set to get a better result for cancer prediction using data mining methods. The decision tree algorithm is not able to produce a good result compare to KNN algorithm.

5.2 RESEARCH CONTRAINS AND CHALLENGES

Throughout the whole research, there are few limitations were encountered. Firstly, the high-performance experiments are not supported by the hardware. Google Colab is used in the experiment on cancer prediction but Google Colab slows down the laptop. To get around this, the minimum requirement for conducting the research might use hardware with a lot of CPU cores and RAM.

Aside from that, the outcome of each workout isn't particularly satisfying. This could be due to the benchmark dataset's small size. Although small data sets are quick and easy to classify, the results of validation may be influenced by the small dataset. Validation is not accurate when the data split is small. As a result, the larger dataset should be applied to overcome this constraint.

5.3 FUTURE WORK

The benefits of using data mining in any other application will be significant. As a result, the accuracy has improved in the cancer prediction using decision tree algorithm can be explored in the future.

For the detection of breast cancer, the effectiveness of the decision tree approach was evaluated and explored. Throughout the implementation phase, only the numerical values of particular breast cancer features are assessed. The experimental findings reveal that the Decision Tree classifier has an over 90% of accuracy rate. The performance of Decision Tree is superior than the other method for the specified dataset.

Cancer is potentially fatal disease. Detecting cancer is still challenging for the doctors in the field of medicine. Even now the actual reason and complete cure of cancer is not invented. Detection of cancer in earlier stage is curable.

REFERENCES

A.Priyanga, S.Prakasam. (2013). Effectiveness of Data Mining - based Cancer Prediction System (DMBCPS), SCSVMV University Enathur,Kanchipuram

Haifeng Wang, Sang Won Yoon. (2015). Breast Cancer Prediction Using Data Mining Method. State University of New York at Binghamton Binghamton, NY 13902

Masud Karim, Rashedur M. Rahman. (2013). Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing. North South University, Dhaka, Bangladesh

Ade Jamala, Annisa Handayania, Ali Akbar Septiandria, Endang Ripmiatina, Yunus Effendib. (2018). Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction.

V.Krishnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra. (2013). Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques.

Ahmad Al-Aiad, Salsabil Abualrub,Yazan Alnsour, Mohammad Alsharo. (2020). Data Mining Algorithms Predicting Different Types of Cancer: Integrative Liter e Literature Review

Maimon, O., and Rokach, L. (2005). Data Mining and Knowledge Discovery Handbook (Vol. 2), Springer, New York.

Venu Gopal Kadamba. (2021). Evaluation Metrics for Classification Problems with Implementation in Python

Jason Brownlee. (2016). What is a Confusion Matrix in Machine Learning. https://machinelearningmastery.com/confusion-matrix-machine-learning/

Mrs.Sagunthaladevi.S, Dr. Bhupathi Raju Venkata Rama Raju. (2016). Classification Technique in Data Mining: An Overview

L. Breiman, J. Friedman, R. Olshen, and C. Stone. (1984). Classification and Regression Trees. Wadsworth, Belmont, CA.

T. Hastie, R. Tibshirani and J. Friedman. (2009). Elements of Statistical Learning.

F.J.Shaikh, D.S.Rao. (2021). Prediction of Cancer Disease using Machine Learning Approach.

Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos. (2014). Machine Learning application in cancer prognosis and prediction. Molecular Oncology Unit, Greece.

Elinor Nemlander, Andreas Rosenblad, Eliya Abedi, Simon Ekman, Jan Hasselstrom. (2022). Lung Cancer Prediction Using Machine Learning on Data from a Sympthon e-questionnaire for never smokers, formers smokers and current

Smokers.

Vipul Bhardwaj, Arundhiti Sharma, Xi Zhang, PeiWu Qin. (2022). Machine Learning For Endometrial Cancer Prediction and Prognostication. Shenzhen Bay Laboratory, Shenzhen, China