Application of machine learning models to predict nature of cancer tumours

Chiu Ho Sun, Sorawee Chirarattanawilai

Abstract

This project presents the investigation of using machine learning models for a predictive analysis for cancer tumours. The models tested in this project are 1. K-nearest Neighbour Algorithm , 2. Support Vector Machine , 3. Multilayer Perceptron. The dataset used includes the data for features of cancer tumour, and the model is to determine whether the nature of the tumour is benign or malignant. The data set is first pre-processed and scaled, then is trained through each model. The result of the model is assessed on the accuracy of the models in predicting the diagnosis and also the sensitivity and specificity of the result.

The K-nearest Neighbour Algorithm model has a 98% accuracy, with a 100% sensitivity and 94.87% specificity. The Support Vector Machine model has a 98% accuracy, with a 96.77% sensitivity and 100% specificity. The Multilayer Perceptron model has a 93% accuracy, with a 93.55% sensitivity and 92.31% specificity.

Email address: hochi24@student.sdu.dk, sochi24@student.sdu.dk

Contents

Email address: hochi24@student.sdu.dk, sochi24@student.sdu.dk

Introduction

Cancer, as a major incurable disease, has always been a major concern for people. In 2019, cancer, as a major leading cause of death, has caused over 10 million deaths in a year, accounting to 16.4% in all deaths.[1] This means that for every six deaths, one is the result of cancer. Moreover, the amount of cancer patients seems to be rising in recent years. In 2022, there were almost 20 million new cases and 9.7 million cancer-related deaths worldwide. By 2040, the number of new cancer cases per year is expected to rise to 29.9 million and the number of cancer-related deaths to 15.3 million.[2] Therefore, against the rising trend, having a faster and more convenient way to predict the nature of tumours, which are potential cancers, will be very useful in easing the burden in medical resources in the form of a preliminary diagnosis for patients.

With the help of a dataset about the different characteristics of tumours and their nature.[3] This paper makes use of three machine learning models to make a predictive analysis of the nature of the separate datasets. The results of respective models are then analysed and evaluated for each of their performances in predicting the nature of the tumour based on the characteristics of each sample. The machine learning models used in this paper are: K-nearest Neighbour algorithm , Support Vector Machine and Multilayer Preceptron.

The structure of the report includes a collection of related work in section 2. And in section 3, there will be information regarding the data set we used and the pre-processing of the data before training with the models. Section 4 will be the main part of the report with the training process of the data and the method attempted. The results of the data training will be discussed in section 5 and the evaluation of its performance will also take place. And finally a conclusion will be drawn on section 6.

Related Work

Gouda I. Salama , M.B.Abdelhalim , and Magdy Abd-elghany Zeid work on breast cancer dataset. The paper presents a comparison among the different classifiers decision tree (J48), Multilayer Perceptron (MLP), Naive Bayes (NB), Sequential Minimal Optimization (SMO), and Instance Based for K-Nearest neighbor (IBK) on three different databases of breast cancer (Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC)) by using classification accuracy and confusion matrix based on 10-fold cross validation method. [4]

A new, entirely data driven approach based on unsupervised learning methods improves understanding and helps identify patterns associated with the survivability of patients. The results of the analysis can be used to segment the historical patient data into clusters or subsets, which share common variable values and survivability. The survivability prediction accuracy of a MLP is improved by using identified patient cohorts as opposed to using raw historical data. Analysis of variable values in each cohort provides better insights into survivability of a particular subgroup of breast cancer patients.[5]

Email address: hochi24@student.sdu.dk, sochi24@student.sdu.dk

Description of Data

The data set that we use for this paper is the "Cancer Data Set" from Kaggle by Eden Saha[3]. The Data Set contains the characteristics of tumours from patients diagnosed with cancer, accumulating to a total of 31 features for 569 samples accounting to a total of 17,639 entries of features. The features consist of the diagnosis of the nature of the tumour, visual characteristics and measurements of the tumour and average values of these characteristics.
It can be split up into five groups, namely ID, diagnosis, Size, Shape and Texture.

ID

The first column in the data set is the ID of the patient, which is a unique key for each patient.

Diagnosis

The diagnosis column denotes the nature of the tumour of that patient, in which 'M' represents Malignant and 'B' represents Benign. The diagnosis is the target for the predictive analysis for this paper and conclusions will be discussed on the accuracy of different models on predicting the diagnosis. In this paper, the Diagnosis feature is represented by {1,0} for {'M','B'} respectively.

Size

There are three different features related to size {Radius , perimeter, area}. These features are the calculations of the initial dimensions of the tumour in different parameters to give more perspective based on size.

Shape

There are four different features related to Shape {concavity , concave points , symmetry , fractal dimension}. These features are data that illustrates the exterior appearance of the tumour.

Texture

There are two different features related to Texture {texture , compactness}. These features are more abstract characteristics of the content and the inner part of the tumour.

- For every feature under the section 'Size', 'Shape' and 'Texture', each include three attributes:
  - Mean : the mean of the feature from the whole sample
  - SE : the standard error of the feature from the whole sample
  - Worst : the lowest value of the feature out of the whole sample

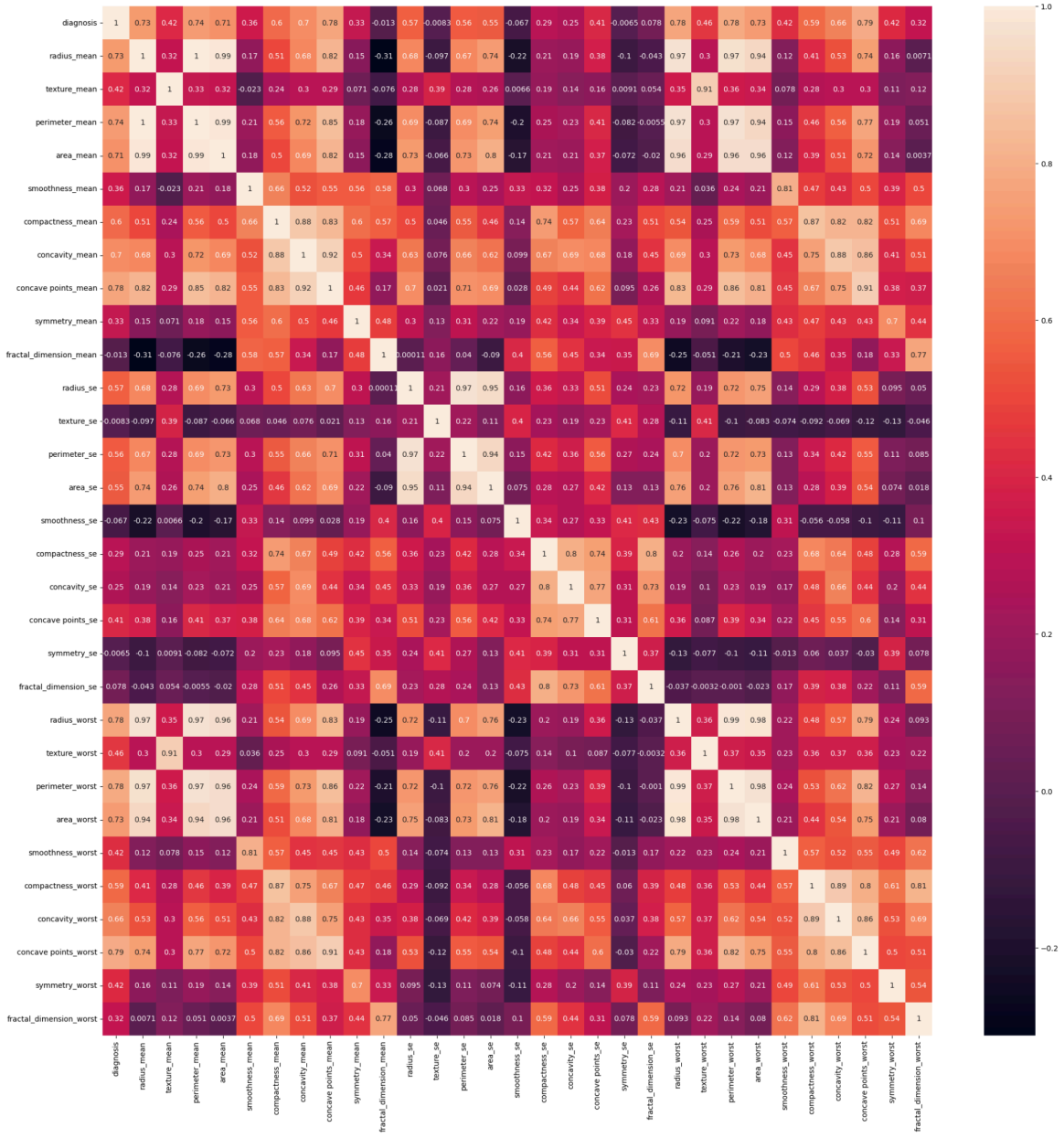
  * The attributes are denoted as the suffix of the main feature, e.g. (radius : {radius_mean , radius_s , radius_worst}) they are separated into three columns in the dataset.

Email address: hochi24@student.sdu.dk, sochi24@student.sdu.dk

Preprocessing and Scaling of Data

Data Selection

To have a better understanding for the relationship between data, so as to facilitate the preprocessing, We plotted a correlation matrix as follows for all of the variables in the data set. The features are then selected to form the final data set to avoid overfitting and underfitting.



We selected features with a relatively higher correlation coefficient i.e. correlation coefficient > 0.5 with the diagnosis. Also, to avoid overfitting, features with high inter correlation are taken out and only the one with the highest correlation coefficient with diagnosis will be retained. For example, the features { radius_mean, area_mean, perimeter_mean } have a high correlation coefficient amongst each other. Thus, only radius_mean is retained as it has the highest correlation coefficient with diagnosis among the three features. We also added some features with a relatively lower correlation

Email address: hochi24@student.sdu.dk, sochi24@student.sdu.dk

coefficient e.g. texture_mean , smoothness_mean , symmetry_mean , etc. to balance the dataset. Based on the above reasons, the list of data for training within the models is reduced to the following :

{ diagnosis , radius_mean , radius_se , radius_worst , concave points_mean, concave points_worst,

texture_mean , smoothness_mean , symmetry_mean , fractal_dimetion_mean }

This removes 21 rows and 11949 entries and the final dataset is reduced to a total of 10 features with 569 samples, accumulating to 5690 entries of features for the final dataset for training and testing.

Data Scaling

As the data from different features are in different formats and units, we performed data scaling to ensure that no single feature dominates the distance calculations in an algorithm, and to help improve the performance of the algorithm.

the following two figures demonstrate the different features in a box diagram before and after data scaling. figure 1 is before the data scaling and figure 2 is after data scaling

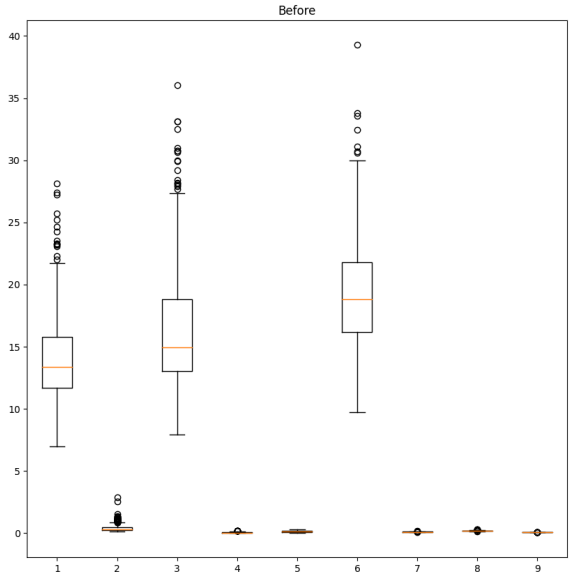


figure 1

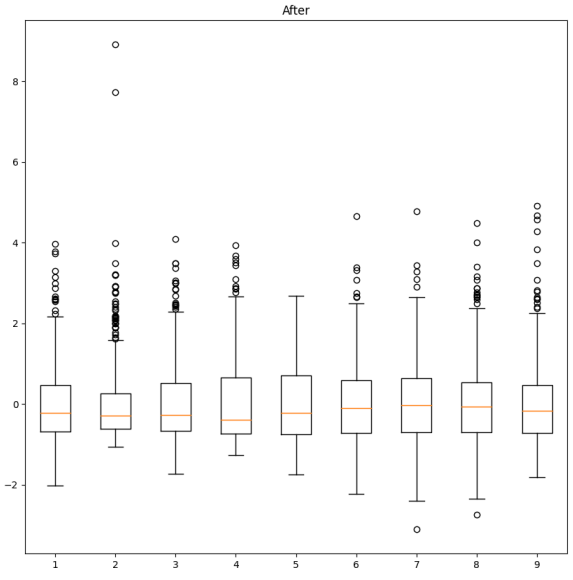


figure 2

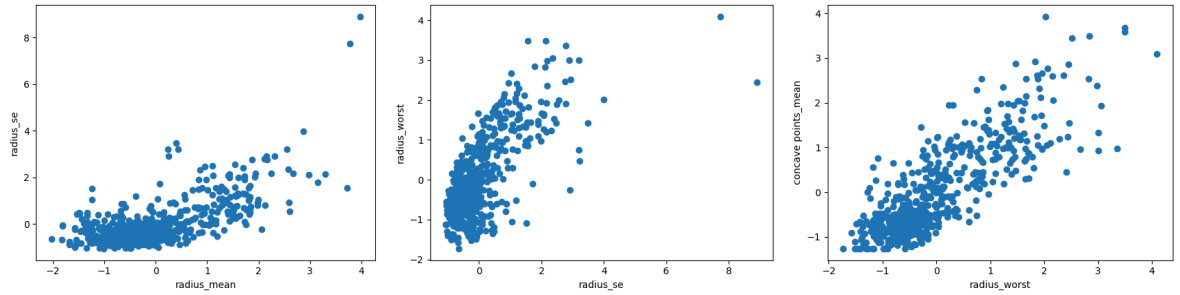
Outlier Detection

In the previous section, we can see on the graph after data scaling that there are some data points that significantly deviate from the other observations in a dataset. Leaving outlier data in the dataset can cause a decrease in model accuracy since the model might learn to fit the noise from the outliers rather than the underlying pattern. This leads to overfitting.

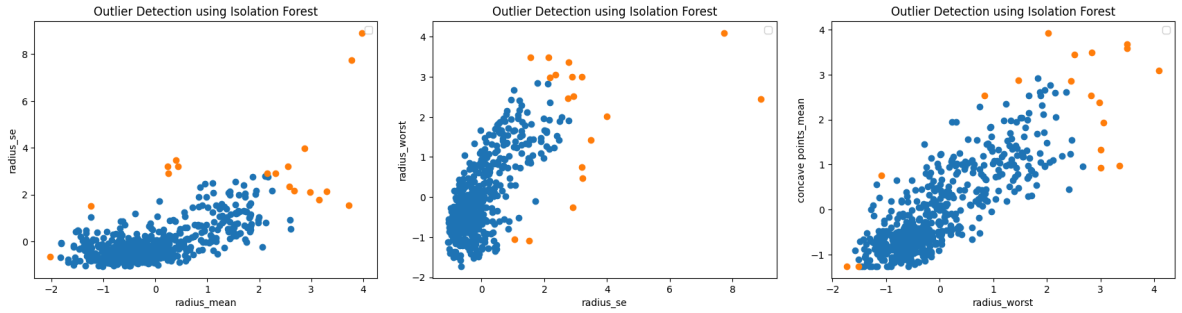
In this case, we use an isolation forest algorithm to detect outliers. The approach is similar to Random Forests, which are built based on decision trees. In the picture below, we built scatter plots, showing

Email address: hochi24@student.sdu.dk, sochi24@student.sdu.dk

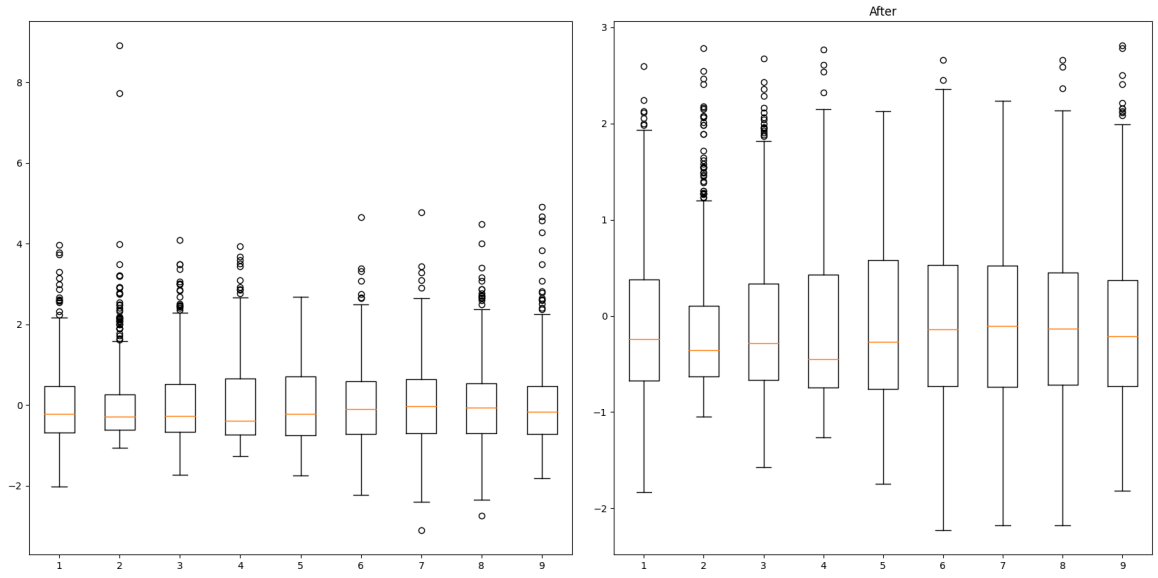
the relationship between each feature. After we run the algorithm, the number of outliers that were detected is 68 data points.



Before detect outlier



After detect outlier



These 2 box plots are compared between before and after removing the outlier.

Email address: hochi24@student.sdu.dk, sochi24@student.sdu.dk

Proposed Method

A total of three Machine Learning Algorithms will be used in this project, KNN Algorithm, Support Vector Machine and Multilayer Perceptron. The details of these algorithms and the finetuning of the parameters during the training process of each algorithm will be mentioned in the later part of this section.

K-nearest Neighbors Algorithm

The K-nearest neighbors algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classification and prediction by grouping the nearest data point to the new data[6]. The algorithm knows what are the closest data points by calculating the distance between the query point and the other data points. There are few distance measures such as Euclidean distance, Manhattan distance, Minkowski distance, and Hamming distance. In this project, we used Euclidean distance:

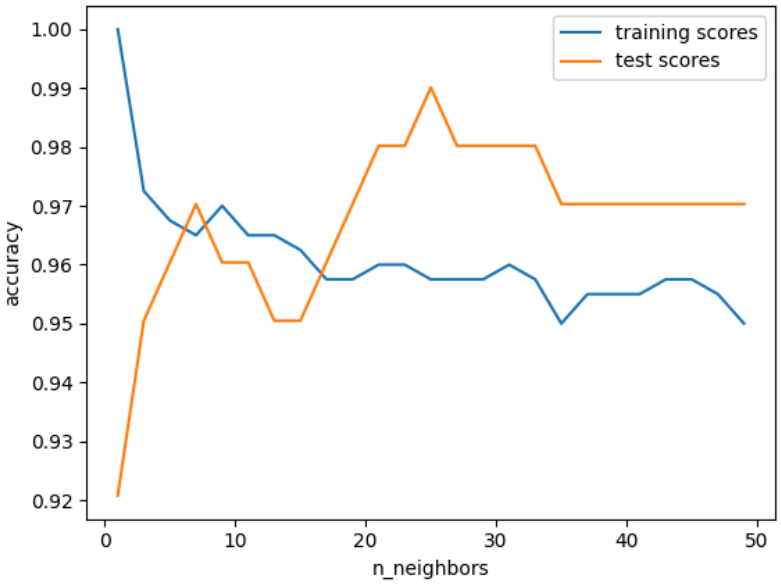
$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

This is the most commonly used distance measure to calculate a straight line between the query point and the other point.

Tuning K-nearest Neighbors Algorithm

For this project, a KNN has only one parameter to tune, which is the number of neighbors. We plotted a graph that shows the training and test scores of the number of neighbors between 1 up to 49, with each number being 2 more than the previous one.

Email address: hochi24@student.sdu.dk, sochi24@student.sdu.dk



The number of neighbors we chose is 21 which is about 98 percent accurate. The reasons why we chose this number is the primary goal is to achieve high accuracy on the test set, as it better reflects the model's performance on unseen data, but we would not choose based solely on the highest training accuracy, as it can lead to overfitting and after the number of neighbors that got the peak accuracy, the accuracy falls immediately which also shows unstable of the model.

Support Vector Machine

A support vector machine (SVM) is a supervised machine learning algorithm that classifies data by finding an optimal line or hyperplane that maximizes the distance between each class in an N-dimensional space[7]. However, when the data is not linearly separable, kernel functions are used to transform the data into a higher-dimensional space to enable linear separation. There are 2 types of SVM classifiers.

First, Linear SVMs are used with linearly separable data; this means that the data do not need to undergo any transformations to separate the data into different classes. Mathematically, this separating hyperplane can be represented as:

$$wx + b = 0$$

where w is the weight vector, x is the input vector, and b is the bias term.

There are two approaches to calculating the margin, or the maximum distance between classes, which are hard-margin classification and soft-margin classification. If we use a hard-margin SVMs, the data points will be perfectly separated outside of the support vectors, or "off the street" to continue with Professor Hinton's analogy. This is represented with the formula,

$$(wx_j + b)\, y_j \geq a$$

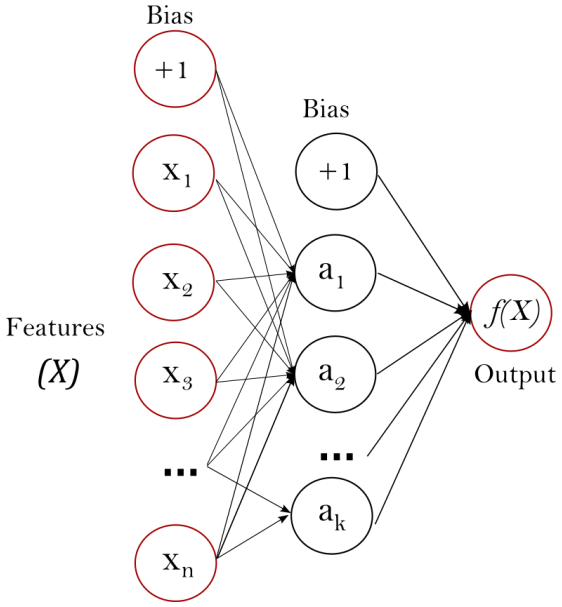
and then the margin is maximized, which is represented as: max ɣ= a / ||w||, where a is the margin projected onto w.

Email address: hochi24@student.sdu.dk, sochi24@student.sdu.dk

Soft-margin classification is more flexible, allowing for some misclassification through the use of slack variables (`ξ`). The hyperparameter, C, adjusts the margin; a larger C value narrows the margin for minimal misclassification while a smaller C value widens it, allowing for more misclassified data.

Multilayer Perceptron

A multi-layer perceptron (MLP) is a type of artificial neural network consisting of multiple layers of neurons[8]. The neurons in the MLP typically use nonlinear activation functions, allowing the network to learn complex patterns in data. MLPs are significant in machine learning because they can learn nonlinear relationships in data, making them powerful models for tasks such as classification, regression, and pattern recognition. In this tutorial, we shall dive deeper into the basics of MLP and understand its inner workings.



Software

All the coding in this project was done on Google Colab with Python version 3.11. Python Libraries Sci-Kit Learn, Numpy, Pandas, Matplotlib and seaborn are utilized for the training and the visualization of the data and training result. Machine Learning Models used in this project were included in the libraries mentioned above.

Email address: hochi24@student.sdu.dk, sochi24@student.sdu.dk

Results and Discussion

Results

We can see that the K-nearest neighbor and support vector machine get the highest test score. Even though these 2 algorithms have the same accuracy value, the predictions are different which cannot be seen in only classification scores.

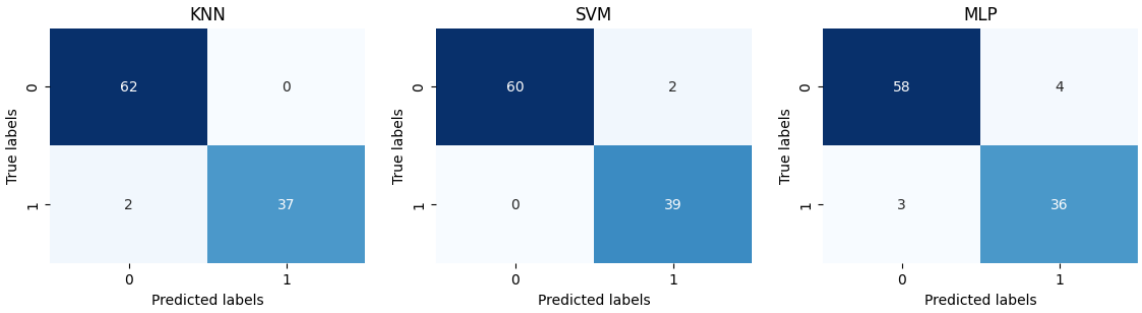
Models	Train Score	Test Score
K-nearest neighbor	96%	98%
Support Vector Machine	96.5%	98%
Multi-layer perceptron	100%	93.1%

Discussion

To evaluate the results of the model, we can use a confusion matrix to show the details.

A Confusion matrix is commonly used to evaluate the performance of a classification model. The table can show you how many are correct of classification and if it classifies wrong, in what class will it go wrong. The confusion matrix compares the actual target values with the machine learning model prediction. the table can be separated into 4 part from left to right and from top to bottom :

1. True positive is the number of malignant cancers that were correctly predicted as positive.
2. False positive is the number of malignant cancers that were incorrectly predicted as positive.
3. False negative is the number of benign cancers that were incorrectly predicted as negative.
4. True negative is the number of benign cancers that were correctly predicted as negative.



After we separated a confusion matrix into 4 parts, we can compute 2 other important measurements. The first is specificity. The specificity can also be known as the true negative rate that can tell us what percentage of patients with benign cancer were correctly identified. This indicator is computed with the following formula:

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

In this case, the highest specificity is 100 percent for k-nearest neighbor model. which can mean that all of the people classified with benign cancer actually have benign cancer.

Email address: hochi24@student.sdu.dk, sochi24@student.sdu.dk

The second is sensitivity which can tell us what percentage of patients with malignant cancer were correctly identified.

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

The highest sensitivity is 100 percent for the support vector machine model. It can mean that 100 percent of people with malignant cancer actually have malignant cancer.

Models	Specificity	Sensitivity
K-nearest neighbor	100%	94.87%
Support Vector Machine	96.77%	100%
Multi-layer perceptron	93.55%	92.31%

We choose these 2 indicators because high specificity is important to reduce the number of healthy individuals being wrongly diagnosed with malignant cancer, thus avoiding unnecessary invasive procedures, stress and medical costs. And, high sensitivity ensures that most cancer cases are detected early, leading to timely treatment and better patient outcomes. Missing a cancer diagnosis could be fatal.

To select what is the best model we should choose for this classification task, we have to weigh between sensitivity and specificity which one we should consider as the first priority. The selection of models will be covered in the Conclusion section.

Conclusion

The objective of this project is to create a suitable and trustworthy predictive analysis model using machine learning. For the three different models we used in this project, we used the specificity and the sensitivity to evaluate its performance. We can first conclude that the Multi-layer perceptron is not a very suitable model for the analysis as both its specificity and sensitivity is relatively lower than the other two. And for the remaining two models, K-nearest Neighbour algorithm has a higher specificity while Support Vector Machine has a higher sensitivity. In this case, as the analysis model is used for cancer detection, having false negatives may cause patients to lose the chance of immediate treatment for their cancer, while having false positives causes a less serious outcome. For diagnosing cancer, sensitivity is typically prioritized to ensure that as many true cancer cases as possible are detected early, thereby improving patient outcomes and survival rates[9] . This indicates that in terms of selecting a model with higher specificity or higher sensitivity, the importance of a high sensitivity trumps over specificity to avoid the chances of false negatives. Thus, Support Vector Machine is a better choice for this predictive analysis.

Though the results in section 5 may show a high accuracy in the different models, both the training and testing size from this dataset is too small for the analysis model to be convincing. To build a more reliable model, a larger sample of data should be used for training. Also, as a complex diagnosis as cancer, we should acquire more distinct features of patients instead of only the tumour for a better and more accurate result.

Despite the deficiencies in the models, it still shows great possibility that diseases like cancer maybe be predicted using machine learning and can yield a good result in the medical field and public health with a faster diagnosis.

Email address: hochi24@student.sdu.dk, sochi24@student.sdu.dk

Acknowledgement

Great Thanks to Erdem Taha for publishing the Benign and Malignant data set for public use and the clear guideline and the description of the data.

Also special thanks for our supervisor Abdolrahman Peimankar and his guidance and teaching on this project.

Reference

[1]Cancers are one of the leading causes of death globally. Are we making progress against cancer? https://ourworldindata.org/cancer

[2]Cancer Statistics
https://www.cancer.gov/about-cancer/understanding/statistics

[3]Benign and Malignant Cancer Data
https://www.kaggle.com/datasets/erdemtaha/cancer-data?resource=download

[4]Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers Gouda I. Salama , M.B.Abdelhalim , and Magdy Abd-elghany Zeid. Computer Science, Arab Academy for Science Technology & Maritime Transport, Cairo, Egypt. September 2012.

[5]Nagesh Shukla, Breast cancer data analysis for survivability studies and prediction
Breast cancer data analysis for survivability studies and prediction - ScienceDirect

[6]What is the KNN Algorithm?
https://www.ibm.com/topics/knn

[7]What are SVMs ?
https://www.ibm.com/topics/support-vector-machine

[8]Multilayer Perceptrons in Machine Learning: A Comprehensive Guide
https://www.datacamp.com/tutorial/multilayer-perceptrons-in-machine-learning

[9]Cancer – WHO
https://www.who.int/health-topics/cancer#tab=tab_1

Email address: hochi24@student.sdu.dk, sochi24@student.sdu.dk