IT1244 Project Report: Breast Cancer detection

Team 15
Alicia Ang Xin Yi (A0259603B)
Chooi Chi Kin (A0284666M)
Liew Jun Yang (A0236555Y)
Tay Wan Lin (A0239775L)

1. Introduction

Breast cancer is the most common cancer type among Singaporean women. The number of breast cancer cases is rising by 3.9% each year (Ho et al., 2020), indicating a growing concern for society. Early intervention for breast cancer significantly increases the probability of a cure. In this project, our goal is to develop multiple machine learning models capable of accurately diagnosing breast cancer. With the assistance of these machine learning models, medical professionals can provide appropriate prescriptions and care to patients based on the diagnosis outcome. Various machine learning algorithms have already contributed to the process of breast cancer diagnosis, such as logistic regression, k-nearest neighbour, support vector machine, or even artificial neural network (Yue et al., 2018). Random Forest and Naïve Bayes algorithms are also used in other studies, and their performances are evaluated based on accuracy, precision, recall and F1 scores. (Sharma et al.,2018).

However, the limitations of these works are that they depend largely on the accuracy of data labelling, and any noise in the data would affect the accuracy of machine learning models. Hence, we would like to investigate an algorithm to identify potential mislabelled data so as to increase classification accuracy. We will then explore the algorithms that are taught in the course, including logistic regression, k-means clustering, artificial neural network and other possible machine learning models to examine their effectiveness in acing the classification task.

2. Dataset

2.1 Exploratory Data Analysis (EDA)

There are 30 features in the dataset, data.csv, with our label being the column "diagnosis," where "B" represents benign breast cancer and "M" represents malignant breast cancer. Some algorithms may not be able to process string values as labels; in such cases, we will convert the diagnosis column into a binary variable, with "1" representing "B" and "0" representing "M".

During our exploratory data analysis (EDA), we observed that approximately 61% of the observations are "B", while the remaining 39% are "M". This imbalance could lead to the overrepresentation of "B" samples in the training dataset, potentially affecting the accuracy of predictions. However, upon examining both the training and testing sets, we found similar class representations in both. Therefore, there is no need to perform a stratified train-test-split in this project.

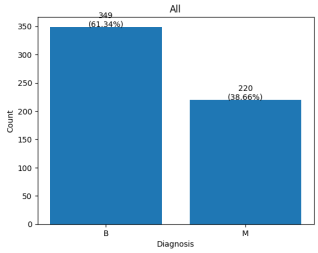

Figure 1: Class distribution of the dataset

Next, we conducted correlation tests between features by generating a correlation heatmap. It is noteworthy that most features exhibit a moderate-to-strong negative correlation with "diagnosis". This suggests that higher values of these features are associated with a greater likelihood of the patient being classified as malignant. It also implies that most of the features are useful predictors for "diagnosis". Consequently, we opted not to remove any features from the original dataset, as only a few weakly-correlated features were identified. These features are not expected to substantially impact the predictions made by the models.

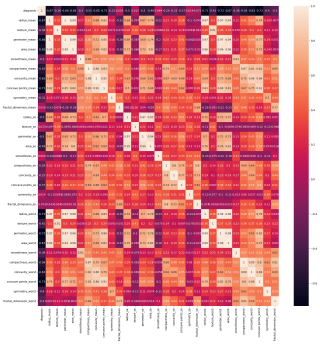

Figure 2: Correlation Heatmap

In addition, outlier detection is a crucial analysis that must be conducted as it can significantly influence outcomes. Upon examining the boxplots for all features, we observed the presence of outliers in certain columns. After verifying that these outliers are not due to data entry errors, we made the decision not to remove them from our dataset. This choice was informed by the fact that these features exhibit a negative correlation with "diagnosis". Therefore, these outliers may represent important observations for the classification of "Malignant" cases.

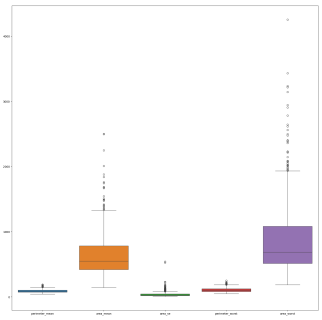

Figure 3: Boxplot to visualize outliers

Lastly, we performed normality tests for all features. For convenience, we conducted the Shapiro-Wilk test on every column. The hypotheses are formulated as shown below:

$$H_0 : \text{the distribution is normal}$$
$$H_1 : \text{the distribution is not normal}$$

We found that the null hypothesis is rejected for all features, concluding that they are not normally distributed. However, since the algorithms that we employed do not require the assumption of normality, we did not transform them to be normally distributed.

2.2 Outliers

In the dataset, there are 28 mislabeled data points identified as "outliers". To discern these mislabeled entries, we merge bootstrap and ML algorithms with the expectation that these models can discern signals from bootstrap samples and employ them to detect noise effectively.

The concept of bootstrapping involves the random selection of iid (independent and identically distributed) small samples from the complete dataset, allowing for the possibility of repeated rows within each sample. In the pursuit of identifying outliers, we aim to utilize ML models to grasp the characteristics of "inliers" (data points with correct labels) and employ these models to predict the correct labels of "outliers". However, it's impractical to solely use "inliers" to train models since we lack any information about "outliers".

Our approach involves utilizing the bootstrap method to generate 28 small samples. For each of these samples, we train five distinct ML models, namely:

1. Logistic Regression
2. Decision Trees
3. Random Forests
4. Support Vector Machine
5. Naive Bayes

The entire dataset serves as input for these five models. Each model produces five different predictions for each data point. Subsequently, we compute the average of these predictions as the predictions for that particular sample.

Given the 28 small samples, we accumulate predictions from each, resulting in 28 sample predictions. Consequently, the final predictions are derived from the average of these 28 sample predictions.

Consider the scenario where we encounter an outlier, denoted as k, incorrectly labeled as 1 but should be 0. We proceed to bootstrap two samples, labeled $Sample1$ and $Sample2$, with k included in $Sample1$ but absent in $Sample2$.

In the case of training ML models on $Sample1$, these models familiarize themselves with the features of k. Predicting k using these models still leads to incorrect class assignments (i.e., the prediction remains 1 instead of 0).

Conversely, in $Sample2$, the models remain unaware of the features of k as it's not part of $Sample2$. Consequently, using these models to predict k is likely to yield the correct class assignment, as these models weren't influenced by the noise.

By averaging the predictions from both scenarios, we consider both possibilities, thereby increasing the likelihood of obtaining the correct labels for the outliers.

Although there are 28 given outliers, this bootstrap method identified 34 outliers, with 26 of them are matched with the given outliers. We then proceed to train our models after replacing these 34 outliers to their correct labels.

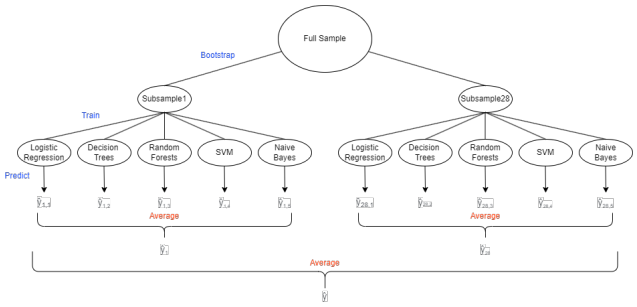

Figure 4: Illustration of Bootstrapping

3. Methods

3.1 Logistic Regression

Logistic regression is a supervised machine learning algorithm that performs binary classification tasks by predicting the probability of an outcome by applying the sigmoid function. In breast cancer classification, logistic regression is preferred over other classifiers due to its successful application, characterised by high accuracy and effectiveness, as evidenced by the findings of Chen et al. (2018), Li et al.

(2019), Mishra et al. (2020), and Sousa et al. (2019) (as cited in Viswanatha et al., 2023).
The limitations of using logistic regression in breast cancer classification are its sensitivity to outliers and assumption of feature independence. Logistic regression can be sensitive to outliers in the dataset as outliers can disproportionately affect the estimated coefficients and reduce the model performance. Moreover, logistic regression assumes that the input features are independent of each other. Since the dataset may contain correlated features, violation of this assumption can result in biased estimates.

3.2 KMeans Clustering

KMeans was used with the aim of clustering the data points into 2 main clusters to represent Benign and Malignant. KMeans was used to visualise the spread of data points within clusters and anomalies from the main cluster. This helps in ensuring the reliability of the clustering. Furthermore, KMeans is efficient and is suitable for high-dimensionality datasets, especially in this case where we have 10 features.

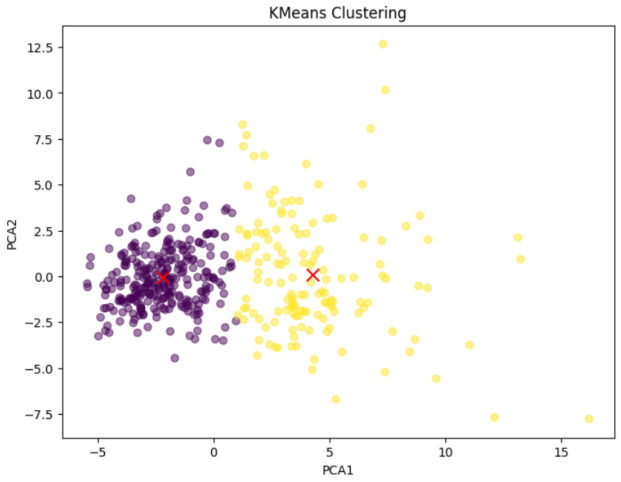

Figure 5: Visualization of 2 clusters

From the chart, a limitation of using KMeans could be identified, where one of the clusters is more widely spread out than the other. When one cluster is more spread out, points further from its centroid may be wrongly assigned to the other cluster, interfering with the accuracy of the prediction.

3.3 Artificial Neural Network (ANN)

Artificial Neural Networks (ANN) represent a powerful machine learning paradigm capable of autonomously adjusting optimal weights and biases across layers through forward-feeding and backpropagation. Notably, studies, such as Amato et al. (2013), have demonstrated the superior sensitivity and specificity of ANN classifiers over conventional statistical methods in cancer diagnosis.
In this project, our objective is to construct an ANN featuring 2 hidden layers tailored for efficient dimensionality reduction and accurate classification. The architectural specifics are as follows:

1. Input Layer: Consisting of 31 nodes, inclusive of one bias node, each node corresponds to a column within the dataset, facilitating comprehensive data representation.
2. First Hidden Layer: Comprising 16 nodes, including one bias node, this layer strategically reduces the dimensionality of the dataset by half. Leveraging Rectified Linear Unit (ReLU) activation function, this layer fosters non-linearity and feature extraction.
3. Second Hidden Layer: With 6 nodes, including the bias node, this layer further diminishes the network's dimensionality while preserving critical features. Employing ReLU activation function, it continues to enhance the network's capacity for intricate pattern recognition.
4. Output Layer: Comprised of a single node, utilizing the sigmoid function as its output function. This configuration facilitates the prediction of the probability of an input being classified as 1 (Benign) or 0 (Malignant), a critical aspect in cancer diagnosis.

By implementing this ANN architecture, we aim to harness the network's capability for automated feature extraction and classification, ultimately advancing the accuracy and efficacy of cancer diagnosis processes.

3.4 Decision Tree (Not covered in IT1244)

Decision tree is a non-parametric supervised learning algorithm and it is a popular machine learning algorithm used for classification tasks. It has a hierarchical tree-like structure where an internal node represents a feature, the branches represent the decision rules, and each leaf node represents the outcome or label. The tree will start from the root node, which is the best predictor among all features.
At each step of building the tree, the algorithm selects the feature that best separates the data into distinct classes. It does this by evaluating different features and splitting criteria, such as Entropy, Gain impurity, information gain, Chi-square and ANOVA to find the feature that maximizes the homogeneity of classes in the resulting subsets.
Once the best feature is selected, the dataset is split into subsets based on the possible values of that feature. This process is repeated recursively for each subset until a stopping criterion is met, such as reaching a maximum tree depth, having a minimum number of samples in a node, or achieving pure nodes (data points in a single class).
To classify a new instance, it traverses the decision tree from the root node down to a leaf node based on the values of its features. The class label associated with the leaf node reached by the instance is then assigned as the predicted class label.
The decision tree is advantageous in the sense that it is easy to interpret, little data preparation is needed, it is non parametric, meaning it does not have to fulfill assumptions, it is versatile, and it is non-linear. However, the model has a risk of overfitting, and it would be highly time-consuming if the dataset has too many features and it is too large. (Kapil, 2022)

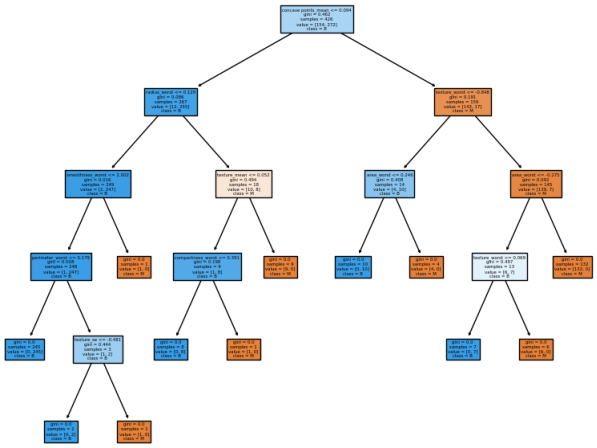

Figure 6: Decision Tree Illustration

In our project, we used the in-built DecisionTreeClassifier from sklearn to help us with the classification.

4. Results

Confusion matrix, accuracy, precision, recall and F1 score were used to evaluate the models.

4.1 Logistic Regression

The confusion matrix of logistic regression is given below:

		Actual	
		Benign	**Malignant**
Predicted	**Benign**	88	2
	Malignant	1	52

Table 1: Logistic Regression Confusion Matrix

1. Accuracy score: **0.979**.
2. Precision score: **0.978**.
3. Recall score: **0.989**.
4. F1 score: **0.983**.

4.2 KMeans Clustering

The confusion matrix of KMeans clustering is given below:

		Actual	
		Benign	**Malignant**
Predicted	**Benign**	86	5
	Malignant	3	49

Table 2: KMeans Clustering Confusion Matrix

1. Accuracy score: **0.944**.
2. Precision score: **0.945**.
3. Recall score: **0.966**.
4. F1 score: **0.956**.

4.3 ANN

The confusion matrix of ANN is given below:

		Actual	
		Benign	**Malignant**
Predicted	**Benign**	87	0
	Malignant	2	54

Table 3: ANN Confusion Matrix

1. Accuracy score: **0.986**.
2. Precision score: **1.0**.
3. Recall score: **0.9775**.
4. F1 score: **0.9886**.

4.4 Decision Tree

The confusion matrix of Decision Tree is given below:

		Actual	
		Benign	**Malignant**
Predicted	**Benign**	85	3
	Malignant	4	51

Table 4: Decision Tree Confusion Matrix

1. Accuracy score: **0.951**.
2. Precision score: **0.966**.
3. Recall score: **0.955**.
4. F1 score: **0.960**.

Based on the results, we concluded that ANN is the most suitable model for the classification. ANN has the highest accuracy score among the models, implying that it is able to correctly classify the most number of cases. ANN also has the highest F1 score, since it has the highest precision score, where all positive predictions made by ANN was correct, and a high recall score. ANN having the highest F1 score indicates that it has the most robust performance in classification tasks among the models.
The reason that ANN outperformed other models is because it possesses the ability to "self-learn from mistakes". By minimizing the error with respect to the weight of each variable through backpropagation, ANN manages to find the optimal weights for all variables that output the most accurate predictions.

References

Artificial Neural Network for medical diagnosis. (2014). Medical Diagnosis Using Artificial Neural Networks, 85–94. https://doi.org/10.4018/978-1-4666-6146-2.ch007

Ho, P.J., Lau, H.S.H., Ho, W.K. *et al.* Incidence of breast cancer attributable to breast density, modifiable and non-modifiable breast cancer risk factors in Singapore. *Sci Rep* 10, 503 (2020). https://doi.org/10.1038/s41598-019-57341-7

Kapil, A. R. (2022, October 1). Advantages and disadvantages of Decision Tree in machine learning. Blogs Updates on Data Science, Business Analytics, AI Machine Learning. https://www.analytixlabs.co.in/blog/decision-tree-algorithm/

S. Sharma, A. Aggarwal and T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, 2018, pp. 114-118, doi: 10.1109/CTEMS.2018.8769187.

Viswanatha, V., Ramachandra, A., C., Bhagat, A. Shekhar, S. (2023). Breast cancer classification using logistic regression. *High Technology Letters*, *29*(8). http://www.gjstx-e.cn

Yue, W., Wang, Z., Chen, H., Payne, A., Liu, X. (2018, May 9). Machine learning with applications in breast cancer diagnosis and prognosis. MDPI. https://www.mdpi.com/2411-9660/2/2/13