**Logo, company name

Description automatically generated**

**Project Report**

Evaluation of the Supervised Classification Algorithms on Wisconsin Breast Cancer Dataset

**Introduction**

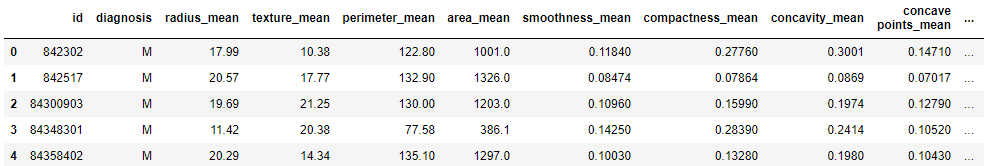
Cancer is the second leading cause of death globally, accounted for an estimated 9.6 million deaths, or one in six deaths according to the World Health Organization (WHO) in 2018. According to the GLOBOCAN 2018 data, one of every four cancer cases diagnosed in women worldwide is breast cancer, and it ranks fifth among the causes of death worldwide.

Despite the increase in the number of medical studies and technological developments that contribute to the treatment of cancer, there are still some problems in the diagnosis of cancer. It is important to make an accurate diagnosis of tumors. Most tumors are the result of benign (non-cancerous) changes within the breast, but if a malignant tumor is diagnosed as benign it will cause serious problems. Early detection of breast cancer and getting modern cancer treatment are the most important strategies to prevent deaths from breast cancer. It is easy to treat early, small, and non-spreading breast cancer successfully. The most reliable way to find breast cancer early is by having regular screening tests.

In this study, public data about breast cancer tumors from University of Wisconsin Hospital were taken and used for data visualization, classification, and machine learning algorithms, which included logistic regression, random forest, and k-nearest neighbor classification algorithms.

**Dataset**

In this study, Wisconsin Breast Cancer (Diagnostic) dataset is used. The dataset is obtainable from UCI machine learning repository.



**Fig.1**. Snapshot of part of the Wisconsin Breast Cancer (Diagnostic) dataset

The table above displays the features of the breast cancer. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

*The initial two columns indicate attribute information:*

1) ID number

2) Diagnosis (M = malignant, B = benign)

*Ten real-valued features are computed for each cell nucleus:*

1) radius (mean of distances from center to points on the perimeter)

2) texture (standard deviation of gray-scale values)

3) perimeter

4) area

5) smoothness (local variation in radius lengths)

6) compactness (perimeter2 / area - 1.0)

7) concavity (severity of concave portions of the contour)

8) concave points (number of concave portions of the contour)

9) symmetry

10) fractal dimension ("coastline approximation" - 1)

Ten real-valued features are computed for each cell nucleus:

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

*There are also important conclusions we can draw by just simply analyzing the table:*

1) Attribute ID is no use for our classification algorithms.

2) Diagnosis Column indicates the labels of our tumors.

3) Unnamed32 Column only has NaN values, hence will be removed.

**Exploratory Data Analysis**

Initially, the statistical analysis (count, mean, standard deviation, min, max and etc.) of the data will be presented:

Table

Description automatically generated

**Fig.2**. Snapshot of part of the statistical analysis of data

The initial analysis shows that dataset that has multiple features spanning varying degrees of magnitude and range. This is a significant obstacle as some machine learning algorithms are highly sensitive to these features.

But we can do better with the visualization. Firstly, the count and the percentage of benign and malignant tumors will be displayed.

Chart, bar chart

Description automatically generated

**Fig.3**. Snapshot of part of the count of the data

The visualization revealed that there are 212 malignant (37.3%) and 357 benign (62.7%) tumors in this dataset.

Now, the next step is to scale the data due to huge differences in the magnitude of the features before visualizing the distribution.

Diagram, histogram

Description automatically generated

**Fig.4**. Snapshot of the distribution of the data

From this snapshot, it is possible to draw some conclusions about the data. For instance, the radial means of benign and malignant tumors vary significantly whereas fractal dimensional mean is almost the same. Another observation is that the area\_se feature has the largest value of outliers. Similar observational conclusions can be made by analyzing the distribution of the data.

**Feature Engineering**

Multicollinearity is one of the primary issues we need to address in this classification of tumors. Plotted heatmap revealed that some of the features are highly correlated between each other. Thus, in this project, we will apply our classification algorithms on both multicollinear and non-multicollinear dataset and compare the accuracy scores and confusion matrices.

* The multicollinearity limit is defined to be 0.8 (Pearson’s Coefficient). If the limit is exceeded, feature will be removed from the dataset.

After the removal of multicollinear features, 14 features remained in the dataset.

**Application of Supervised Learning Algorithms**

**Logistic Regression Mathematical Model**

Logistic regression is used to study the relationship between a categorical dependent variable (Y) and a set of independent variables (X). The dependent variable in logistic regression is expected to have two values, such as Yes or No, 1 or 0. In case of multinomial logistic regression, the dependent variable will have more than two values (married, divorced, widowed or single).

Logistic regression models the probability of the default class (e.g. the first class).

Let us start with an example. We are modelling whether a tumor is malignant or benign based on a number of features. To make it easier to understand, only one feature – radius of the tumor is chosen to represent the association between dependent and independent variables. The default (first) class is selected to be malignant and the logistic regression can be written as the probability of a tumor being malignant given its radius.

Or more formally,

Above equation describes the probability that an input (X) belongs to the default class (Y=1). If p is the proportion of observations with an outcome of 1, then 1-p is the probability of an outcome of 0. The ratio p/(1-p) is called the odds and the *logit* is the logarithm of the odds, or just log odds. Mathematically, the logit transformation is written:

The logistic transformation is the inverse of the logit transformation.

The equation above is the formula of sigmoid function. It is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

A picture containing chart

Description automatically generated

**Fig.5**. Sigmoid Function

The coefficients (Beta values b) of the logistic regression algorithm must be estimated from your training data. This is done using maximum-likelihood estimation.

We will obtain the beta values by training the data and then apply logistic regression based on the values of dependent variable (in this care radius of tumor) to obtain the probability of a tumor being malignant. If the probability:

**Result of Logistic Regression**

Accuracy Score of Logistic Regression on Dataset:

Removing multicollinearity decreases the number of features, thus impacts the accuracy score of our model. However, it also helps the model to satisfactorily distinguish between the individual effects of the independent variables on the dependent variable.

A screenshot of a cell phone

Description automatically generatedChart, treemap chart

Description automatically generated

**Fig.6**. Confusion Matrices of Logistic Regression

**Random Forest Mathematical Model**

1. For b = 1 to B:
2. Draw a bootstrap sample Z∗ of size N from the training data.
3. Grow a random-forest tree Tb to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size nmin is reached.
4. Select m variables at random from the p variables.
5. Pick the best variable/split-point among the m.
6. Split the node into two daughter nodes.
7. Output the ensemble of trees

Classification: Let be the class prediction of the bth random forest tree. Then, = majority vote .

**Bootstrap sample**: Practice of random sampling with replacement.

Chart, line chart

Description automatically generated

**Fig.7**. Bootstrapping

After bootstrapping, we have n samples, represented by the circles containing dots in the image above. From each of the samples, a decision tree is constructed. Each decision tree generates its own predictions, which are then aggregated to produce the overall predictions.

The decision tree taken from the ensemble (random forest) will assist to visualize the concept better:

Diagram, engineering drawing

Description automatically generated

**Fig.8**. Decision Tree from Random Forest

While designing the decision tree, the features possessing the least value of the Gini Index would get preferred. Gini impurity calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly.

**Results of Random Forest Classification**

Accuracy Score of Logistic Regression on Dataset:

Chart, treemap chart

Description automatically generatedA screenshot of a cell phone

Description automatically generatedRemoving multicollinearity decreases the number of features, thus impacts the accuracy score of our model. However, it also helps the model to satisfactorily distinguish between the individual effects of the independent variables on the dependent variable.

**Fig.9**. Confusion Matrices of Random Forest Classification

**Naïve Bayes Mathematical Model**

Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Another assumption made in this model is that all the predictors have an equal effect on the outcome. This is not the best assumptions to work with, especially in case of our classification problem.

The Naïve Bayes Classifier is based on the Bayesian Theorem:

The variable y is the class variable(malignant or benign tumor). Variable X represent the parameters/features.

Given the:

Where x represents the features/parameters, by expanding by using the chain rule:

By some additional mathematical work, the equation can be modified to provide:

**Results of Naïve Bayes Classifier**

Chart, treemap chart

Description automatically generatedThe accuracy of the Naïve Bayes Classifier is lower for non-multicollinear dataset. This is expected to be because of the reduced number of features as this classifier assumes that the predictors are independent (which in our case they are not) and have the equal impact on the outcome (again not the appropriate assumption in case of our model).

Chart, treemap chart

Description automatically generated

**Fig.10**. Confusion Matrices of Naïve Bayes Classifier

**K-Nearest Neighbor Mathematical Model**

In the classification setting, the K-nearest neighbor algorithm essentially boils down to forming a majority vote between the K most similar instances to a given “unseen” observation. Similarity is defined according to a distance metric between two data points. A popular choice is the Euclidean distance given by:

Diagram

Description automatically generated

**Fig.11**. KNN Classifier Visualization

Given a positive integer K, an unseen observation x and a similarity metric d, KNN classifier performs the following two steps:

* It runs through the whole dataset computing d between x and each training observation. We will refer the K points in the training data that are closest to x the set A. Note that K is usually odd to prevent tie situations.
* It then estimates the conditional probability for each class, that is, the fraction of points in A with that given class label. (Note I(x) is the indicator function which evaluates to 1 when the argument x is true and 0 otherwise).

X gets assigned to the class with the largest probability.

Another important step is choosing the K which is usually preferred to be an odd number. A method referred as elbow method is used in most of the cases:

A picture containing chart, shape

Description automatically generated

**Fig.12**. Elbow Method in KNN Classifier

Chart, treemap chart

Description automatically generatedA screenshot of a cell phone

Description automatically generated**Results of KNN Classifier**

**Fig.13**. Confusion Matrices of KNN Classifier

**References**

1. Evaluating the Performance of Machine Learning Techniques in the Classification of Wisconsin Breast Cancer, December 2018, International Journal of Engineering and Technology 7(4.36):160-166
2. Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection, Lucas Borges, October 2015,
3. Analysis of k-means clustering approach on the breast cancer Wisconsin dataset, Ashutosh Kumar Dubey, Umesh Gupta & Sonal Jain, International Journal of Computer Assisted Radiology and Surgery volume 11, pages 2033–2047(2016)
4. <https://www.who.int/news-room/fact-sheets/detail/cancer>
5. <https://machinelearningmastery.com/logistic-regression-for-machine-learning/>
6. <https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Logistic_Regression.pdf>
7. <https://medium.com/data-science-group-iitr/logistic-regression-simplified-9b4efe801389>
8. <https://www.math.mcgill.ca/yyang/resources/doc/randomforest.pdf>
9. <https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9>
10. <https://medium.com/analytics-steps/understanding-the-gini-index-and-information-gain-in-decision-trees-ab4720518ba8>
11. <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>
12. <https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c>
13. <https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/>
14. <https://bradleyboehmke.github.io/HOML/knn.html>