Breast Cancer Classification

Anurag Khanra
Computer Science
PES University
Bangalore, India
anurag.khanra306@gmail.com

Hanuraag Ravilla Baskaran
Computer Science
PES University
Bangalore, India
rbhanuraag01@gmail.com

Harshavardhan Veeranna Navalli
Computer Science
PES University
Bangalore, India
harshavardhannavalli3@yahoo.com

Prakruti Pruthvi Kumar
Computer Science
PES University
Bangalore, India
ppk20142001@gmail.com

***Abstract*— The goal of this project is to use various methods of statistical analysis, such as clustering, random forest, logistic regression and Naïve-Bayes classifier, so as to classify breast cancer into whether it is malignant or benign. This classification is done based on three factors of multiple criteria of tumor size, namely the mean, squared error and the worst values of the tumor. This information has been compiled from various scans of breast cancer tumor, into the Wisconsin Breast Cancer Dataset, which is considered the go-to dataset for breast cancer classification.**
This research into breast cancer classification is very important, as breast cancer is the most invasive form of cancer, and the second highest cause of death for women all around the world.

Keywords—breast cancer, prediction, machine learning, classification, tumor, clustering, k-means

I. Introduction

Breast cancer is a form of cancer that forms from the abnormal growth of the cells in the breast. After skin cancer, breast cancer is the most common form of cancer that is diagnosed in women. As with all forms of cancer, the earlier the detection of cancer, the higher chance of survival. With the automation of classical knowledge gained by previous research into breast cancer detection, the amount of time it takes to detect breast cancer drops massively. This helps with the timely detection of cancer, as the stage in which we detect cancer is the most important step of the process. The earlier we detect cancer, the earlier the stage of cancer we catch it in, and the earlier the stage, the higher the chance for survival. Not only that, but as an added benefit, we have the option to correct any potential misdiagnoses that may accidentally occur. There are two types of classifications for a tumor, namely malignant and benign. Malignant tumors indicate that the tumor is cancerous and that it has the potential to spread to the other cells around it. These are the more dangerous type of tumor. By contrast, benign tumors only form in one place and will not spread around to other cells.

With the rise in the use of computers in various industrial applications even for the smallest use cases, it is inevitable that the medical industry will deploy large scale applications of CAD (computer aided diagnoses) for time sensitive diseases such as cancer. A lot of research has been conducted into the diagnosis of cancer using different datasets compiled from various sources, with existing information. The work done by compiling datasets for future generations to work with to improve computer aided diagnosis cannot be underestimated, as this research could be a step towards saving millions of lives. This specific dataset is considered as the go-to dataset for breast cancer detection into malignant and benign. This dataset has been compiled from information extracted from scans of breast cancer, and we classify the tumor into malignant or benign based on the criteria in the dataset.

By the development of computer aided diagnosis, we have the potential to save millions of lives. Most cases of cancer go undetected in the early stages, and often it is too late to make any inroads while treating cancer after detecting it. This research is very important and can go a long way in increasing survival rates.

II. Previous work

The given dataset is divided into the test data set and the training data set, with a 3:7 ratio, i.e., 30 percent test data and 70 percent training data. RF is employed to build the classification model, and a baseline is registered. Following this, the other models are compared (SVM, DT, LR and NN) to the baseline set by the RR model. The baseline set by the measure of accuracy and the F-measure metric. The higher F-measure metric signifies a higher efficiency of the model. When the results are plotted onto a histogram, the RF model shows the highest F-measure metric, and shows the highest accuracy as well. For this reason, RF is chosen as the primary analytical model. To verify this, the ROC curve is plotted, and the AUC is taken, which verifies RF model is the most accurate classifier. [1]

A baseline is registered using RF, and the other models are compared (SVM, DT, LR and NN). With comparative plotting on a histogram, we can see RF has the best accuracy, and to verify this, the ROC curve is plotted and AUC is taken for the same, which proves the above result. [2]

Three datasets are generated, which have all features, highly correlated features and low correlation features. The Machine Learning algorithms used on these datasets were logistic regression, kNN, support vector machines, naive Bayes, decision tree, random forest and rotation forest. Accuracy results showed Logistic Regression had the highest accuracy at about 98%. [3]

In this paper, the researchers have concentrated on creating a prediction system that can predict the incidence of breast cancer at an early stage by analyzing the smallest set of attributes that has been selected from a clinical set, and achieved an accuracy of 99.28 percent by using correlation-based measures. They have used Correlation-Based measures to get a good feature subset that contains features highly correlated with predictions of the class. The performance of the proposed system was evaluated by considering the actual and the predicted classification. They compared the results achieved with different classifiers used, such as KNN Accuracy, Linear regression Accuracy and SVM Accuracy, among which KNN Accuracy had the highest score. [4]

In this paper the researchers have chosen Colon cancer as their main issue, this is due to very less presence of early symptoms. The main method chosen here is Naive Bayes classification. This is a technique prediction based on simple probabilistic and on the application of the Bayes theorem. This model has an accuracy classification of 95.24% achieved. [5]

Testing for this dataset had been conducted based on the following models: Multi-layer Perceptron (MLP), K nearest Neighbors, Classification and Regression tree, Support Vector Machines and Gaussian Naïve Bayes. Different activation functions were used, such as tanh, rectifier and maxout. These were used to compare to other learning algorithms such as naïve bayes, SVM and random forest as a baseline to understand how these models would perform. After all analysis has been done, it is seen that multi-layered perceptron and support vector machines have the highest accuracy. [6]

III. DATASET AND PREPROCESSING

From the dataset that we have taken, we can conclude whether the tumor in a patient is malignant or benign. The dataset consists of 569 rows and 32 columns. There are 18208 values in total. It contains the variables like radius mean, texture mean and smoothness mean that help in the prediction of the tumor.

An exploratory data analysis was done on the dataset, the null values were removed, the data was cleaned. Our target variable was a diagnosis that was Benign and Malignant. There were 357 and 212 tumors that were diagnosed as benign and malignant respectively.

For preprocessing our data, we first decided to normalize all values. To accomplish this, we used the MinMaxScaler which is provided by the sk-learn module in python. Post normalizing our data, we applied Principal Component Analysis (PCA) on our dataset to perform dimensionality reduction on our data. It is important to notice that unlike feature selection, which is mostly done manually and completely removes a feature and its impact on the dataset, PCA only performs feature reduction which maps the data to lower dimensional through some projection of all original dimensions therefore still retaining a fraction of variance that the other features provide. After a couple of trial and error we decided on reducing down to 7 features from our original 30 features dataset.

IV. OUR APPROACH

A. Background

We try to classify the tissue as malignant or benign by using unsupervised learning with a labelled dataset. Although supervised learning classifiers perform better in most of the cases, they tend to overfit on the training set and produce poor results when applied on test set. Supervised learning methods struggle with over-fitting especially when the training data is not sufficiently large. Supervised learners fail to identify the noise and anomalies in the data. It is really hard to find the right balance between bias and variance.

Unsupervised learning methods explore the underlying data and try to find patterns and trends in the data. Clustering algorithms are unsupervised learning algorithms where they try to group similar data into clusters where each cluster explains the characteristics of the underlying similar set of data. Hence using unsupervised learning on a labelled dataset can produce some good classifications.

B. Basic Idea

In this study, we are provided with a labeled dataset. We use the data in this dataset without the labels and model it using k-means clustering. The k-means clustering algorithm groups the data into clusters where each data potentially explains a class of the data. Since each cluster contains similar data and potentially explains a single class of data, our job is to identify which cluster belongs to which class of data(classification). We consult the labels given in the dataset for this. We compare the data points in a given cluster and their corresponding labels provided in the original dataset. The class label which is most prevalent among the datapoints in that cluster is assigned as the class of the cluster. We use mode to determine which class label is prevalent. So that class of highest occurrence in that cluster is assigned as the class of the cluster. After doing this for each cluster, we now have clusters with each cluster representing a classification. Note that more than 2 clusters can belong to the same classification, as it depends on class prevalent in that cluster. If the number of clusters is more than the number of classes, 2 or more clusters belong to the same class.

Now for classification of the new unseen data, we could use the k-means clusters to identify the cluster that the data belongs to and assign the label of that cluster as the label of the unseen data. Or else we could model the dataset with the clustered classifications we just found, with a supervised learner. By doing this, all the noise and the anomalies will have been removed by the clustering, and we would a more insightful data classification. The supervised learner can now learn the new classifications as there is no risk of overfitting anymore.

C. Implementation of this idea on our dataset:

In this study, the Breast Cancer Wisconsin (Diagnostic) Data set with 569 samples has been chosen which describes characteristics of the cell nuclei present in the digitized images of a fine needle aspirate of breast mass.

We preprocess the attributes of the dataset by scaling the data to a common scale as it is important to scale the data when clustering so that all features are given equal weightage. We use MinMaxScaler for this task. For feature selection the methodology used here is PCA which transforms the given features into the reduced feature space where the reduced feature space contains the features in the form of principle components that contribute the most in explaining the variance in data.

Then we cluster the preprocessed data using k-means clustering. k-means clustering is a hard and non-probabilistic approach to clustering where each point belongs to at most one cluster. We try out clustering for different values of k and find out that k=2 performs the best clustering. This is

indicative of the underlying trend of the data that the 2 classes (malignant and benign) have 2 different feature characteristics.

After this we find out the class to which each cluster belongs to. We do it by comparing each data point in a cluster to the corresponding classification label provided in the dataset. The class of highest occurrence in the cluster is assigned as the classification of the cluster. We use mode to find the class of highest occurrence.

Now we assign each datapoint in the dataset with the new classification labels provided by the clusters which they belong to. A datapoint is assigned the label of the cluster it belongs to.

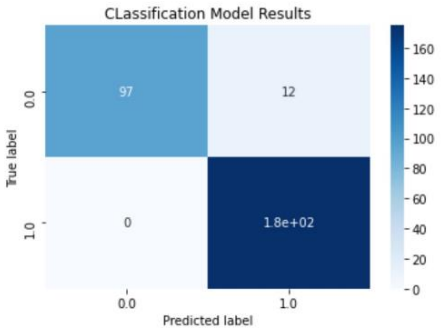
At last, we use the new classification labels to predict the new unseen datapoints. We predict it using the k-means clusters that we formed as a part of classification. We assign the label as the label of the cluster to which the new data point is closer to.

V. Results and Analysis

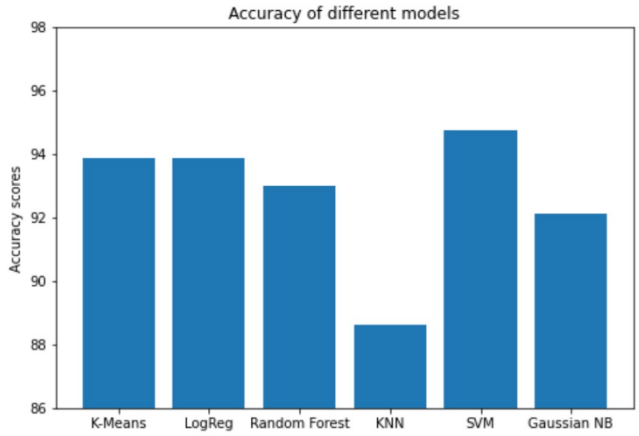
Given below is the confusion matrix and classification report for our k-means clustering classification:

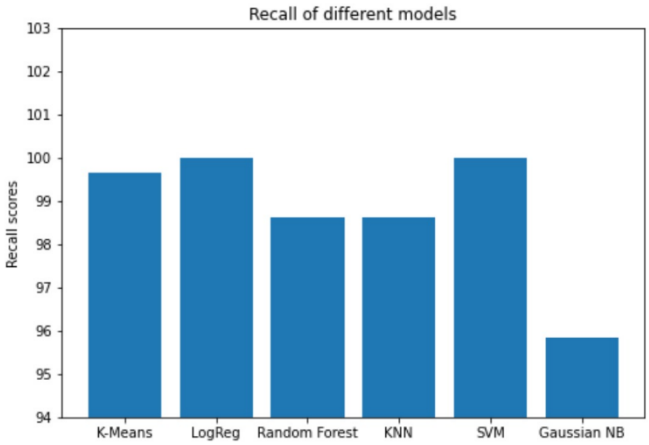
```
K-Means
Accuracy Score: 95.77%
```

```
CLASSIFICATION REPORT:
                   0.0         1.0  accuracy   macro avg  weighted avg
precision     1.000000    0.935829  0.957746    0.967914      0.960458
recall        0.889908    1.000000  0.957746    0.944954      0.957746
f1-score      0.941748    0.966851  0.957746    0.954299      0.957216
support     109.000000  175.000000  0.957746  284.000000    284.000000
```



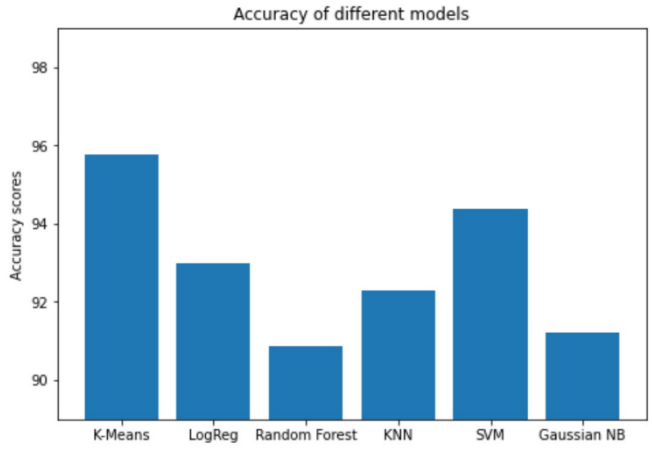
After averaging the test results for 10 iterations for the test-train split of 20-80%, we observe the following accuracies and recall scores for various classifiers:

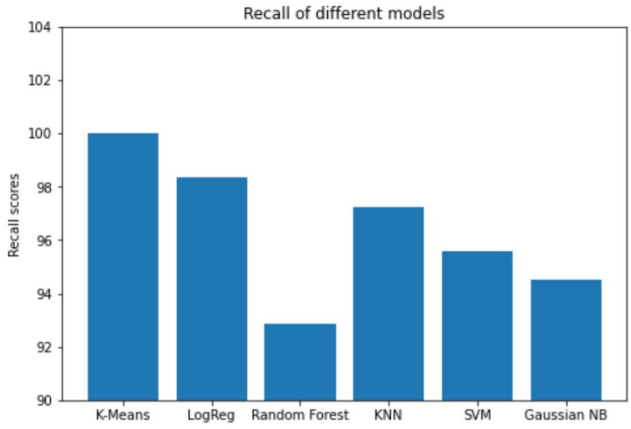




We see that the accuracy and recall of the k-means clustering classification is comparable to other classification accuracies.

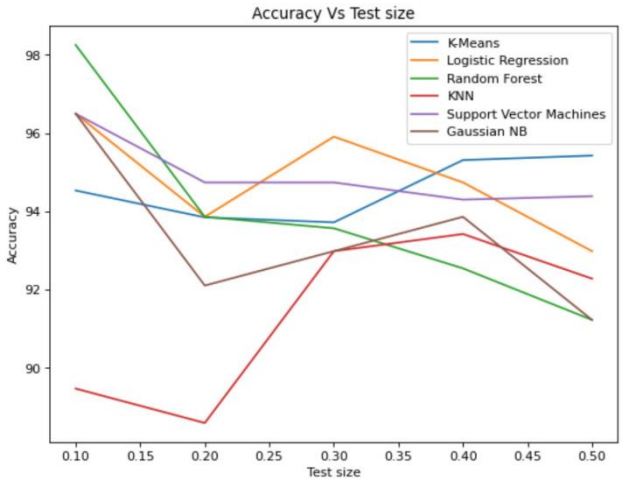
When we compare the accuracies and recall on a test-train split of 50-50% we observe the following results for various classifiers:

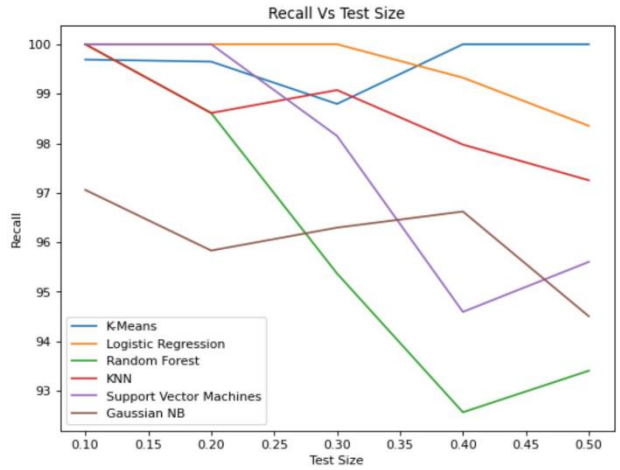




We observe that the k-means clustering classifier performed better than other classifiers both in terms of recall and accuracy. This indicates that the k-means clustering classifier performs better on small training datasets.

On comparing the accuracy and recall of the classifiers over various test-train split sizes, we observe the following graph:





This shows that k-means clustering performed even better on small train sizes to give a superior recall and accuracy. The accuracy and recall of other classifiers constantly decreased on increasing the test size (or decreasing the train size) whereas the accuracy and recall of k-means clustering classifier slightly increased on increasing the test size (decreasing the train size).

VI. Conclusion

From the above analysis we conclude that k-means clustering when used for classification over the given dataset produces comparable results with other classifiers. The k-means clustering algorithm performs better compared to other classifiers when the there is insufficient data i.e., when the data used for training is small enough. This is indicative of the fact that clustering analyses the underlying data for patterns. And when we classify based on these clusters, we get more generalized results whereas other supervised learners fail to identify the patterns and simply tend to overfit. The above results show the potential of unsupervised learners in aiding supervised learning.

The k-means clustering show a consistently higher recall than other classifiers. This could be helpful in cancer detection as we don't want to miss out any patients who actually have cancer. Detecting a false positive is better than detecting a false negative. This test could be used as a preliminary test before performing further analysis on the patient. Even if preliminary test results in a false positive, further clinical tests can dismiss the chance of it being a false positive.

In the above analysis, the clustering method used was k-means which is a hard clustering method. Hard clustering suffers from low accuracies at the boundary of each cluster. It fails to identify the gradient at the boundaries of each cluster Use of soft clustering methods like gaussian mixture models can more efficiently cluster the data. More over k-means provides only spherical clusters whereas gaussian mixture models provides clusters of almost any shape. Thus, by using soft clustering algorithms like gaussian multivariate models the performance of the model can be further improved. In the above analysis the cluster label assignment criteria were based on the mode, i.e. the label of highest occurrence in a given cluster. We could improve this labelling using more complex labelling criteria for better performance. With better feature selection methods, we could achieve a better performance.

Thus, from the above analysis we conclude that unsupervised learners can used for supervised learning to produce comparable results and can be very effective on producing classification results with very limited amount of data.

VII. References

[1] 'Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction', Yixuan Li, Zixuan Chen, 2018

[2] 'Breast Cancer Prediction Using Dominance-based Feature Filtering Approach: A Comparative Investigation in Machine Learning Archetypes', Noreen Fatima, Li Liu, Sha Hong, Haroon Ahmed, 2020

[3] Muhammet Faith Ak, 'A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications' , 26 April 2020.

[4] Madhu Kumari and Vijendra Singh 'Breast Cancer Prediction System', 2018.

[5] Nafizatus Salmi and Zuherman Rustam, Naive Bayes Classifier Models for predicting colon cancer, 2019

[6] Noreen Fatima, Li Liu, Sha Hong and Haroon Ahmed, 'Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis', 2020