CSE 523 - Machine Learning
Mid-Semester Project Report
The Curators: Breast Cancer Detection

Harsh Patel
AU1940114
harsh.p7@ahduni.edu.in

Kavan Desai
AU1940126
Kavan.d@ahdni.edu.in

Sarthak Bharad
AU1940176
Sarthak.b@ahduni.edu.in

Abstract

Prediction of breast cancer using machine learning can be a great gist of importance for detecting cancer. We used data available on UCI Machine Learning Repository [1]. This report aims to provide the reader an insight into several regression and classification algorithms used for breast cancer detection that involves the concepts of classical Machine Learning.

Keywords - Exploratory Data Analysis (EDA), Correlation, KNN, Logistic Regression, Misclassification rate, Random forest, Feature Importance, PCA, Feature Elemination

Introduction

Breast Cancer begins when healthy cells in the breast undergo change and grow out of control, thus forming a mass or sheet of cells called a tumour. A tumour can be cancerous or benign. A cancerous tumor is malignant, i.e., it cangrow and spread to other parts of the body, the cancer spreads to other parts of the body when the cancerous cells move through the blood vessels and/or lymph vessels, this process is called metastasis. A benign tumour can grow but will not spread. [2]

Breast cancer has currently overtaken lung cancer as the most commonly diagnosed cancer in women worldwide (according to statistics released by the International Agency for Research on Cancer (IARC) in December 2020). [3]

Our project mainly gives analysis of the performance of two algorithms: Logistic Regression and K Nearest Neighbours (KNN Network). Our objective is to predict breast cancer, using classical machine-learning algorithms (Logistic Regression and KNN Network), and find out the most effective algorithm based on the performance of each classifier in terms of accuracy.

Literature Survey

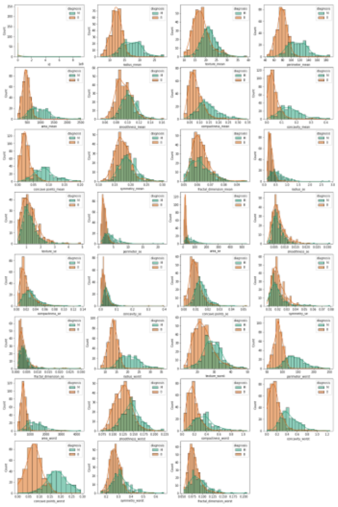
The ability to use a picture of the cell to forecast whether a cell is malignant or not will be extremely beneficial to the medical community in terms of reducing medical waste. The prediction directly tends towards machine learning techniques in predicting the state of the cell i.e., malignant or benign. Several studies employing several classifications and feature engineering strategies are undertaken on medical data sets. The literature has a lot of information on breast cancer datasets. Many of them have a high level of categorization precision. To discover the best classifier, Sivaprakasam et al. [4] examined the performance of C4.5, Nave Bayes, Support Vector Machine (SVM), and K- Nearest Neighbour (KNN), and found that SVM was the most accurate, with an accuracy of 96.99 percent. ML techniques have been used to breast cancer survival forecasting, diagnostic ultrasonography, and breast cancer diagnosis utilizing tumor tissue imaging in a number of recent research [5]. When compared to other methodologies, the Trees Random Forest (TRF) technique produced better outcomes in this investigation (Nave Bayes, 1NN, AD, SVM and RBFN, MLP) [6]. As a result, the goal of this research was to see how machine learning may be used to categorize breast cancer using feature values derivedfrom a digitized picture of a delicate extraction of a breast cell. We dived deep into random forest as we need to reduce dimensionality. As We want some important features to be derived from the set of 32 feature. We need to ensure the importance of the feature so here comes the random forest which provide the best features to be selected. Moreover, the Principal component analysis also comes into account by providing the best principle, components which helps us to predict the diagnosis based on the prime principal components.

Implementation

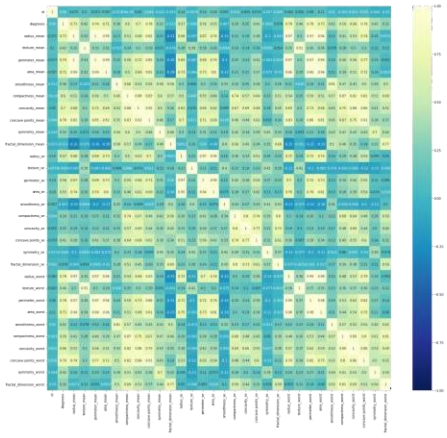
Introduction to dataset – Breast cancer image-based data is available on UCI Machine Learning Repository [1]. Basically, the dataset is a tabular data derived from 569 images of breast cell. Data consist of diagnosis in terms of binary classification that is malignant and benign. The table has certain features derived from the radius, perimeter, area even the grey scale image is been preprocessed to obtain the following dataset.

Problem Statement – Our aim is to build a model which predicts the state of cancer provided the data of the features mentioned below that is if the cancer cell is malignant or benign.

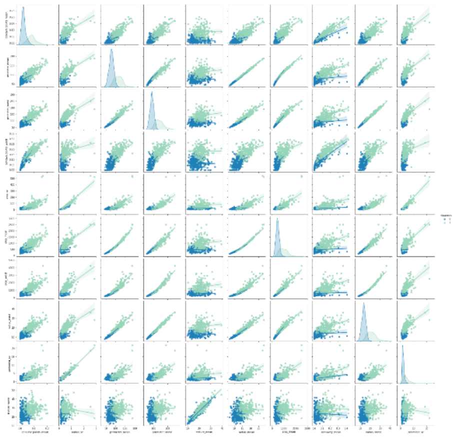
Exploratory Data Analysis - The graph below helps to check which of the parameters are normally distributed and which are skewed distributed.



The heatmap below shows how the different variables arerelated to each other and up to what extent.



The correlation gives dependence between parameters. So, this helps in performing feature engineering.



Classification algorithms –

1) K-Nearest neighbors: - The K-Nearest neighbours' algorithm is a non-parametric and supervised learning technique. We used both Euclidean and Manhattan distance and found that Manhattan distance partially overfits the data whereas Euclidean distance performs apparently well compared to other.

2) Logistic regression: - Based on past observations of a data set, logistic regression is a statistical analytic approach for predicting a binary result, such as yes or no. The diagnostic value is the dependent variable, and the independent variables are the 32 attributes on which the diagnostic value is based.

Feature Importance - We have used threshold value to evaluate the feature based on the correlation of the feature. The threshold value was tempted to be kept 0.8 which specifies high threshold value. The negative class is in minority in comparison to negative class i.e., benign. Therefore, Feature Elimination in Recursive Mode Cross-Validation identifies the most essential qualities and ranks them in order of significance. This allows us to create a model with the best dimensions possible.

1) Random Forest: -

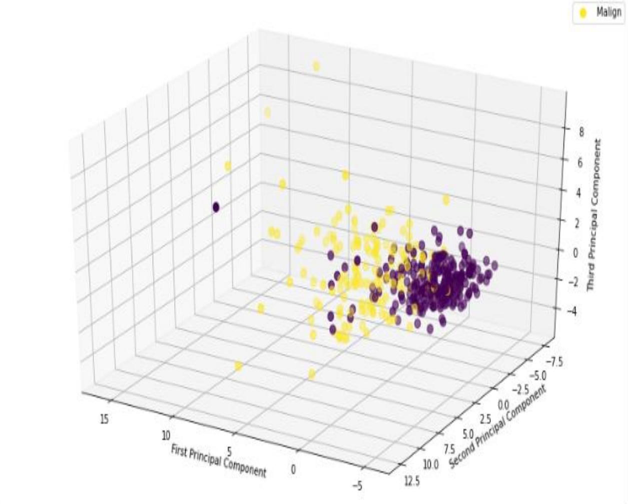
In the Eda plots, we were able to find that some of the features were of the same importance. We have used the random forest classification to cross-check this fact and eliminate the features. We selected a few features and tried to find their importance using random forest classification. We also keep the eye on the accuracy of the algorithm. We have also used univariate feature selection and Random Forest classification to select best K which removes all but the k highest scoring features.

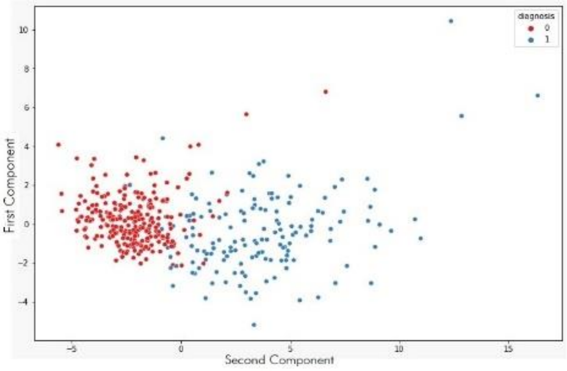
We have also tried to increase the accuracy of our algorithm by implementing the recursive feature elimination with random forest classification. Here the goal behind increasing the accuracy was to minimize the scope of error in comparing the feature importance of cell features. Now in recursive feature elimination, the features whose importance is less are eliminated. This process is done in recursive until the highest importance feature is only left.

Finding only the best feature will not work. We have to also find how many features will be required for more accuracy. For this purpose, we have used recursive feature elimination with cross-validation and random forest classification.

2) Principal component analysis: -

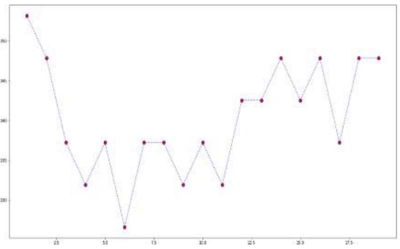
We may use the explained variance to determine the number of Principal Components to evaluate. The proportion of variation explained by each primary component is known as explained variance. The variance of the ith main component is equal to i as previously stated. Unfortunately, this remarkable capacity of dimensionality reduction comes at the expense of being able to comprehend what these components represent quickly. Here we can notice from the below graph that we can simply classify the two components using a linear classification model.





Results

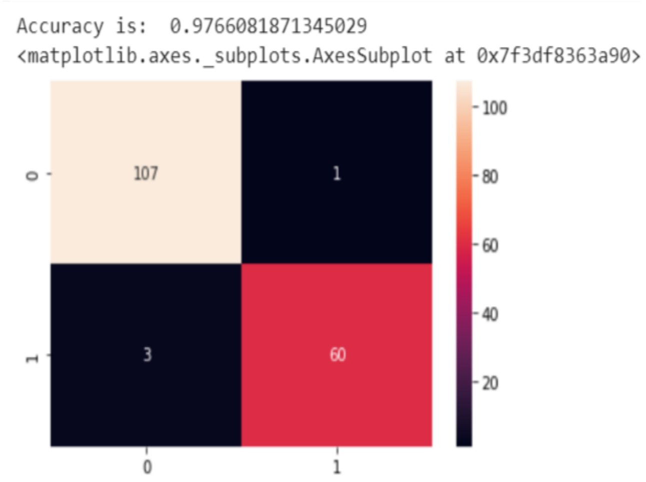
k Nearest Neighbours (kNN) - The correlation gives dependence between parameters. So, this helps in performingfeature engineering. KNN isused for both classification and regression. Here we have used KNN as the features of the cell are corelated with each other and the cancerous cell can be classified based on the neighbors. We used the Euclidean distance in the implementation of KNN algorithm. For the best suitable K, we implemented the mean squared analysis over a range of values of K from 1-20. And as a result, we could conclude that the least error valued K gives approximately 97% accuracy.



Logistic Regression **-** The correlation gives dependence between parameters. So, this helps in performing feature engineering. Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. Here the dependent variable is the diagnostic value and the independent variables aredifferent 32 features on which the diagnostic value is dependent. And as a result, we got an accuracy of around 98 per cent.

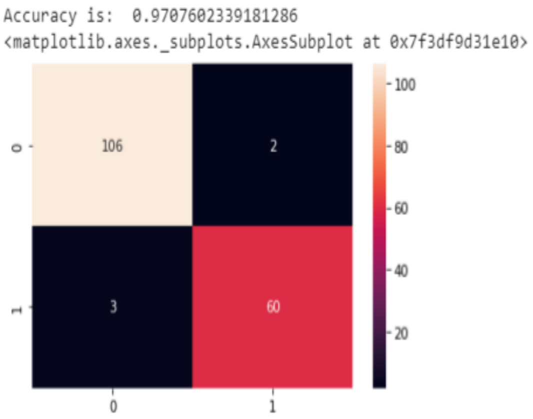
Random forest classification:

Univariate Feature Selection:

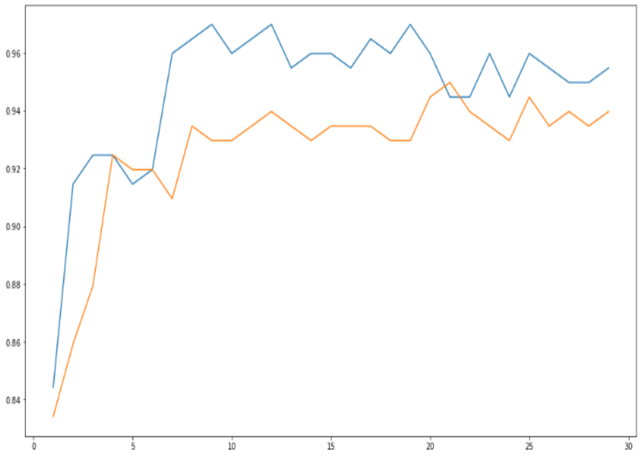


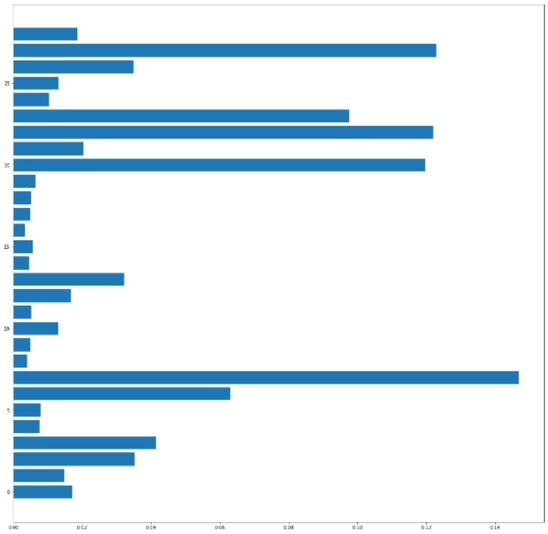
```
Accuracy is:  0.9766081871345029
<matplotlib.axes._subplots.AxesSubplot at 0x7f3df8363a90>
```

Recursive Feature Elimination:

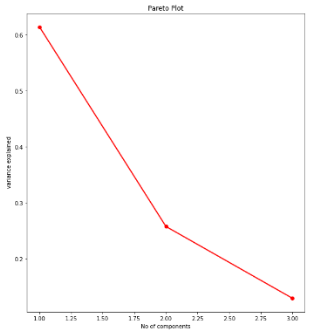


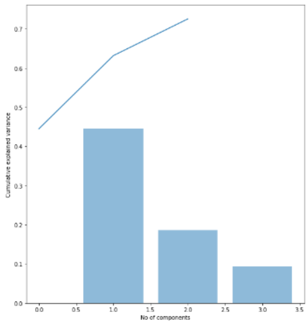
```
Accuracy is:  0.9707602339181286
<matplotlib.axes._subplots.AxesSubplot at 0x7f3df9d31e10>
```

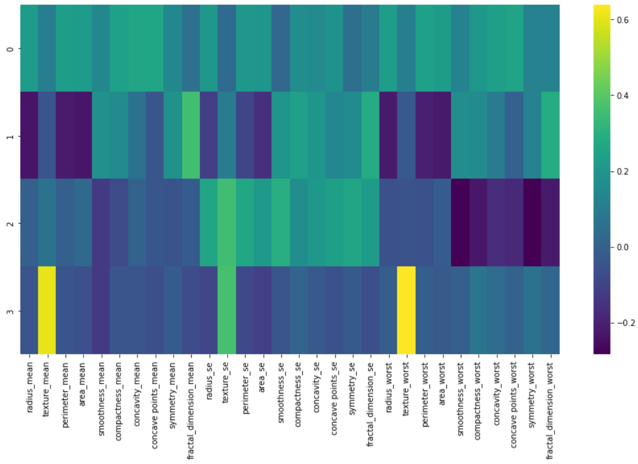




PCA:-







Conclusion

We hereby conclude that Breast cancer can be predicted well based on the image characteristics of the cell. In terms of accuracy, we can conclude that Logistic regression is a better option than KNN classification. With further feature engineering we aim to improve the performance of our current models. Further we are planning to implement new algorithms and will try to implement different distance metrics on our current models. We implemented the random forest which helps use find the best features throughout the set of 32 feature. We use K fold cross validation to intrinsically identify the best validation set. In our data the dimensionality of the data is high, which redirects the problem to reduce dimensionality where the PCA and random forest comes into play. PCA was also implemented to identify 4 best principal components. The inference to be drawn is that the principal component and the features are nearly non correlated. We use K fold cross validation to intrinsically identify the best validation set which performs better that other, So we used the Cross validation process to get the best features suitable for the dataset. Using PCA we create 3 components and we found that we get the best fit from the relation between first and second component analysis. We found that the variation in correlation between features and Principal components is varied in nature where as it is positive near to zero which helps us differentiate two class

References

1. UCI Machine Learning Repository: Breast Cancer wisconsin (diagnostic) data set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cance r+Wisconsin+%28Diagnostic%29. [Accessed: 20- Mar-2022].

2. “Breast cancer - introduction,” Cancer.Net, 31-Dec- 2020. [Online]. Available: https://www.cancer.net/cancer-types/breast-cancer/introduction. [Accessed: 20-Mar-2022].

3. M. A. Naji, S. E. Filali, K. Aarika, E. L. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, “Machine learning algorithms for breast cancer prediction and diagnosis,” Procedia Computer Science, 08-Sep-2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921014629. [Accessed: 20-Mar-2022].

4. Upadhayay, A., 1970. Empirical comparison by data mining classification algorithms (C 4 . 5 &amp; c 5 . 0) for Thyroid Cancer Data Set . Semantic Scholar. Available at: https://www.semanticscholar.org/paper/Empirical-Comparison-by-data-mining-Classification-Upadhayay/6e6d581cf8a96559a91d74274b765b62b de9d4b7 [Accessed March 20, 2022].

5. R. Turkki, D. Byckhov, M. Lundin, J. Isola, S. Nordling, P. E. Kovanen, C. Verrill, K. von Smitten, H. Joensuu, J. Lundin, and N. Linder, “Breast cancer outcome prediction with tumour tissue images and machine learning,” Breast cancer research and treatment, Aug-2019. [Online].Available: R. Turkki, D. Byckhov, M. Lundin, J. Isola, S. Nordling, P. E. Kovanen, C. Verrill, K. von Smitten, H. Joensuu, J. Lundin, and N. Linder,“Breast cancer outcome prediction with tumour tissue images and machine learning,” Breast cancer research and treatment, Aug-2019. [Online].https://www.ncbi.nlm.nih.gov/pmc/articles/PMC66 47903/. [Accessed: 20-Mar-2022].. [Accessed: 20- Mar-2022].

6. M. Montazeri, M. Montazeri, M. Montazeri, and A. Beigzadeh, “Machine learning models in breast cancer survival prediction,” Technology and Health Care, 01-Jan-2016. [Online]. Available: https://content.iospress.com/articles/technology-and-health-care/thc1071. [Accessed: 20-Mar-2022].