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@ahduni.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.

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 tumour is malignant, i.e. it can grow 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]

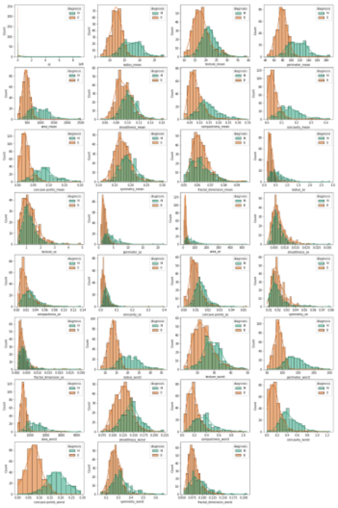
Our project mainly gives analysis of the performance of two algorithms: Logistic Regression and K Nearest Neighbours (KNN Network). Our objective is to predict breast cancer, using classical machine-learning algorithms (Logistic Regression and KNN Network), and find out the most effective algorithm based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

Literature Survey

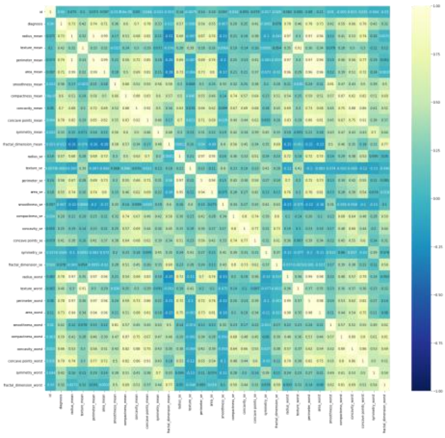
The ability to use a picture of cell to forecast whether a cell is malignant or not will be extremely beneficial to the medical community in terms of reducing medical waste. The prediction directly tends towards machine learning techniques in predicting the state of cell i.e., malignant or benign. Several 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 utilising tumour 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 categorise breast cancer using feature values derived from a digitised picture of a delicate extraction of a breast cell.

Implementation

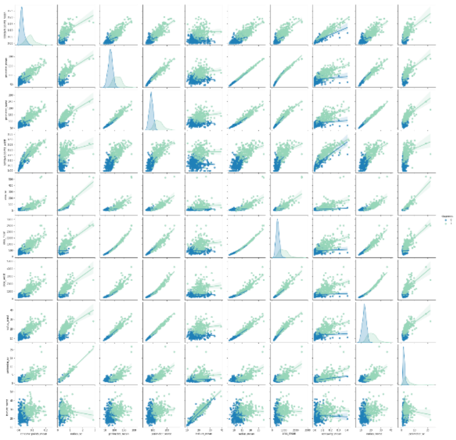
Exploratory Data Analysis - The graph below helps to check which of the parameters are normally distributed and which are skewed distributed.



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

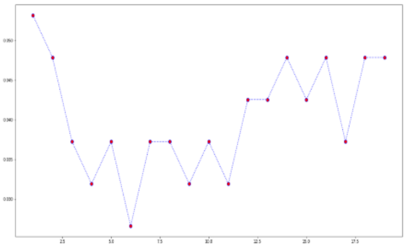


The correlation gives dependence between parameters. So, this helps in performing feature engineering.



Results

***k Nearest Neighbours (kNN)* -** The correlation gives dependence between parameters. So, this helps in performing feature engineering. The K-Nearest neighbours' algorithm is a non-parametric and supervised learning technique. KNN is used 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 neighbours. We used the Euclidean distance in 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 diagnostic value and the independent variables are different 32 features on which the diagnostic value is dependent. And as a result, we got an accuracy of around 98 percent.

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.

References

1. UCI Machine Learning Repository: Breast Cancer wisconsin (diagnostic) data set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+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/6e6d581cf8a96559a91d74274b765b62bde9d4b7 [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]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6647903/. [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].