Breast Data Science Project

Abstract— As much as scientific research plays a pivotal role everywhere, healthcare also finds its prominent application. Breast Cancer is the top-rated type of Cancer amongst women, taking away 627,000 lives alone. This high death rate due to breast cancer does need attention for early detection so that prevention can be done in time. Data extraction finds a multi-fold application in predicting Brest cancer as a potential contributor to state-of-art technology development. This task focuses on various classification techniques implementation for data mining in predicting malignant and benign breast cancer. Breast Cancer Wisconsin data set from the UCI repository has been used as an experimental dataset, while attribute clump thickness is used as an evaluation class. The performances of this Random Forest algorithm are analyzed on this data set.

Keywords— Data Mining, Classification Techniques, UCI repository, Breast Cancer,

1. INTRODUCTION

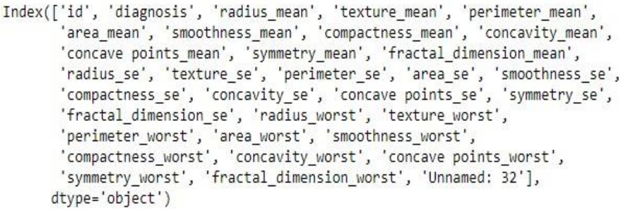
Cancer is a disease characterized by uncontrollable growth and the spread of abnormal cells. According to cancer research UK, there are more than 200 types of cancers. Most common cancers include lungs, breast, colorectal, cervical, and stomach. (World Cancer Report 2014). Malignant tumors begin to form in either specific or multiple parts of the body that cause detrimental effects. If left untreated, these tumors slowly begin to amass, leading to complications such as the production of lumps, internal bleeding, ductal blockage, and cell toxicity. The eventual outcome is death. Cancer can be caused by exposure to environmental factors such as tobacco, infectious organisms, radiations, carcinogens, unhealthy diet, unhygienic living standards, and biological factors such as inherited genetic Mutations, hormones, and immune conditions. A combination of these factors or perhaps a sequential order can trigger cancer growth. Cancer causes DNA alteration leads to the loss of normal function of a cell as well as uncontrolled cell growth and often metastasis (Kiyuhara et al., 2006). Many genes which are lost or mutated themselves have been identified to induce cell division under provided conditions like ras and myc proto-oncogenes and genes which are programmed to stop proliferation damages cells like TP53 and RB1 tumor suppresser genes. (Kiyuhara et al., 2006).

According to an epidemiological study conducted by World Health Organization (WHO), Cancer is the leading cause of death in developed countries, while for developing countries, heart diseases are the leading cause of death. Although numerically, the deaths caused by Cancer in developing countries are higher, the percentage of death in developed countries is 16.2% greater than in developing countries (World Health Organization, 2004). In Canada, Cancer is the leading cause of death, followed by cardiovascular diseases (statistics Canada, 2015). In the United States, it is the second deadliest disease. However, according to the CDC, the difference between death rates caused by Cancer and cardiovascular disease varies by only 3.8% (CDC/ National Center for Health Statistics, 2016).

2. Data and Analytical Questions

2.1 Data

The dataset is Breast Cancer **Wisconsin (Diagnostic) Data Set** available at UCI Machine Learning Repository. The dataset is a free source and can be downloaded for the analysis and making models for the predictions. The dataset contains several independent features based on each tumor (cancerous cell) is labeled to be benign or malignant. This is the target column to be predicted system design, and here are some attributes of the dataset.



```
Index(['id', 'diagnosis', 'radius_mean', 'texture_mean', 'perimeter_mean',
       'area_mean', 'smoothness_mean', 'compactness_mean', 'concavity_mean',
       'concave points_mean', 'symmetry_mean', 'fractal_dimension_mean',
       'radius_se', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se',
       'compactness_se', 'concavity_se', 'concave points_se', 'symmetry_se',
       'fractal_dimension_se', 'radius_worst', 'texture_worst',
       'perimeter_worst', 'area_worst', 'smoothness_worst',
       'compactness_worst', 'concavity_worst', 'concave points_worst',
       'symmetry_worst', 'fractal_dimension_worst', 'Unnamed: 32'],
      dtype='object')
```

Fig: 1.0 Dataset columns

2.2 Analytical Approach:

To investigate the dataset following plan was derived. However, many steps were revisited, and several iterations were taken to find the most comprehensive result.

1. Read the dataset and take a look at the dataset in the data frame
2. Exploring the dataset and performing some EDA
3. Visualization and analysis of the dataset
4. Checking the correlation of the dataset
5. Removing some unnecessary columns from the dataset
6. Splitting the dataset into train and test
7. Applying the model and making the predictions from the test dataset
8. Evaluating the Accuracy as well Confusion Matrix of the model.

3. Findings and Critical Reflections:

3.1 Findings:

3.1.1Malignant: Malignant Tumors retain a high metastasis potential. Cancer cells possess the ability to invade the neighboring tissues, and regenerate secondary growth at distant sites.

Malignant tissue varies in shape, size and replicative potential. These tissues can grow in a dynamic way and also cause damage to neighboring tissues blood, vessels or lymphatic vessels. They can interfere with body functions and metastasize to distant locations in the body after dislodging from primary tumor size. Metastasis cancer expands from its locations of origin to new different regions of the body. Malignant tumors are able to return after they are isolated.

3.1.2 Benign:

Benign tumors retain a relatively limited mitotic potential. They rarely produce life-threatening problems unled they occur in a vital organ or press the nearby tissues. Benign tumors cannot spread in the body; if once removed by surgery they cannot be reoccurred. Once benign tumors are removed by surgery, they do not usually recur. (Kelsey *et al*., 1993)

3.1.3 causes of breast cancer:

4. Uncontrolled growth of cells in breast cancer is the leading cause of breast cancer. The abnormal growth of cells can be malignant or benign. In addition, 5%-10% of cancers are due to abnormality that is inherited from the parents, and about 90% of breast cancer are due to genetic abnormality that happens as a result of the aging process (Christos Stergiou *et al*,.2013).

5. We know that breast cancer is a highly complex disease caused by multiple risk factors like age, genetic factors, heredity, and overweight, lack of exercise, and smoking (Breast Cancer Organization).

3.1.1 Bar Chart of Target attribute

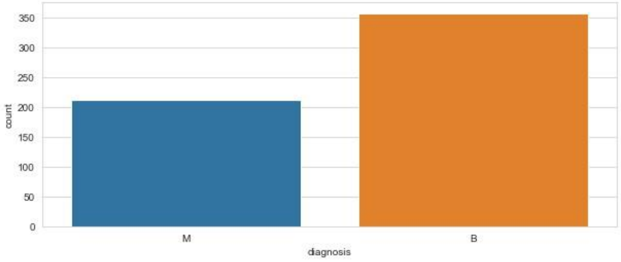


Fig 3.0: Bar Chart of number target values

Figure 3.1 is for radius mean of the target attribute

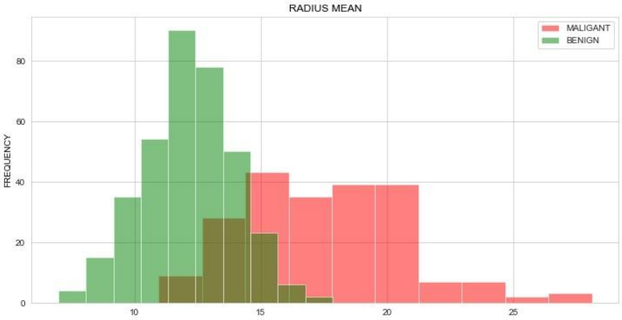


Figure 3.2 is for the texture mean of the target attribute

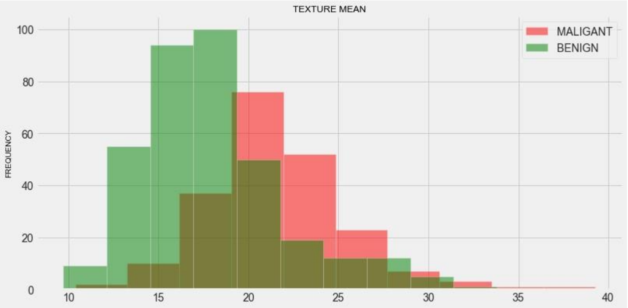


Figure 3.3 is for the perimeter mean of the target

Figure 3.4 is for the area mean of the target attribute

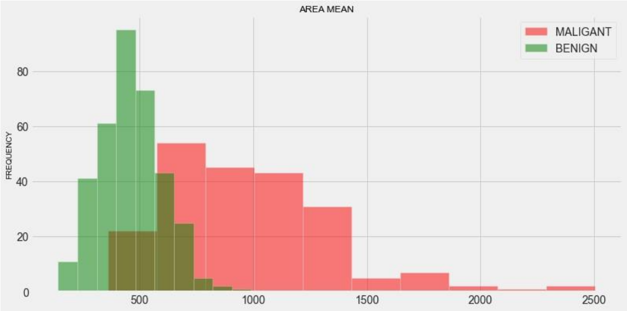


Figure 3.5 is for the smoothness mean of the target attribute.

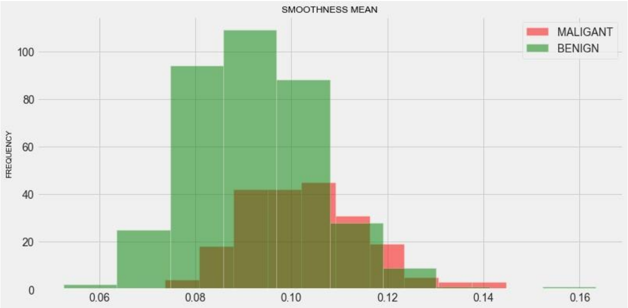


Figure 3.6 is for the compactness mean of the target attribute

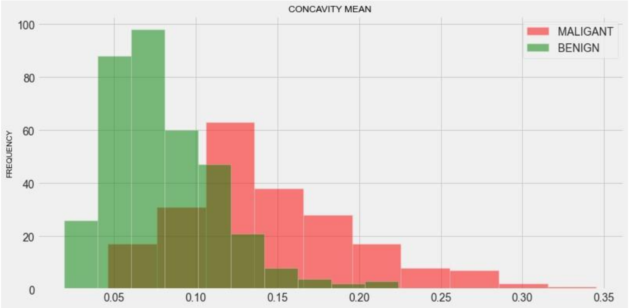


Figure 3.7 is for the concavity mean of the target attribute

attribute

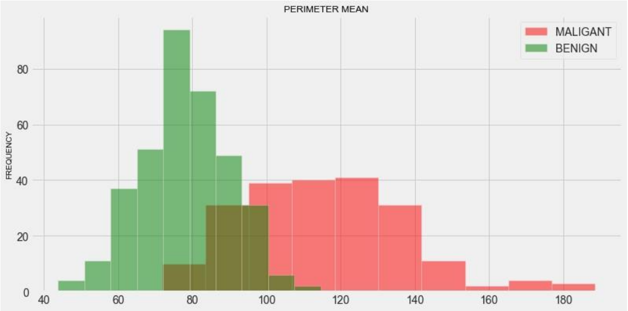


Figure 3.8 is for the concave mean of the target attribute.

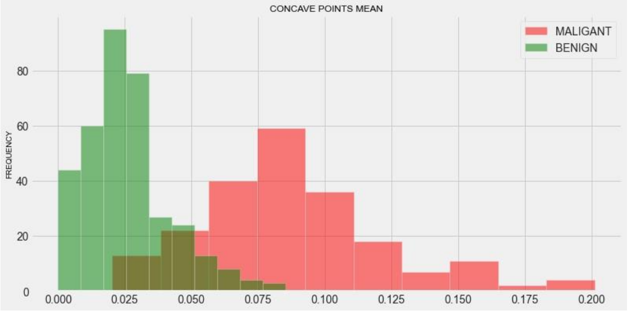


Figure 3.9 is for the Symmetry Mean of the target attribute.

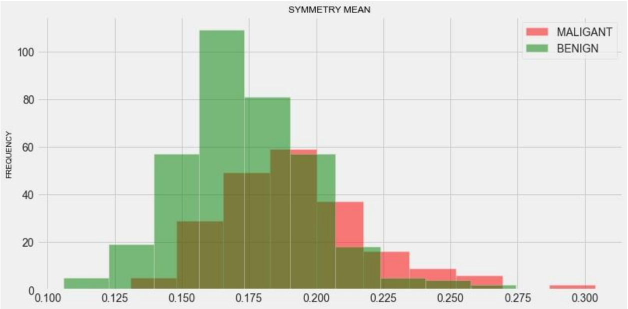
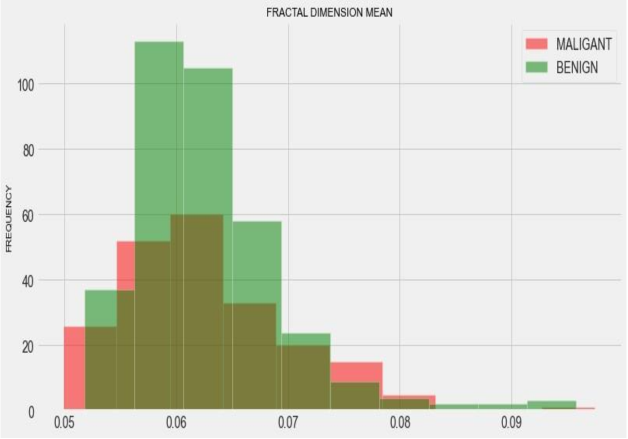


Figure 3.10 is for the Fractal Dimension Mean of the target attribute



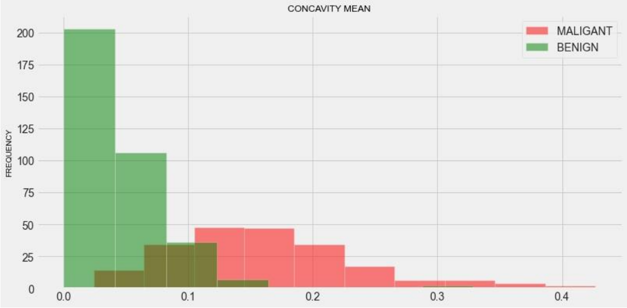


Figure 3.11 is for the Correlation chart of the target attribute

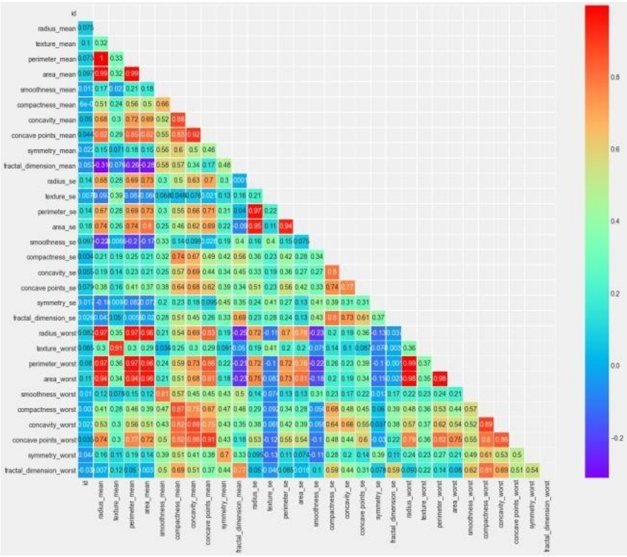
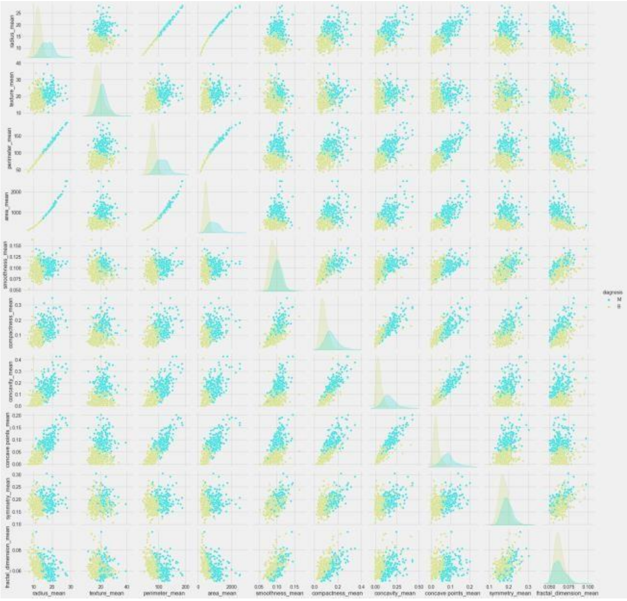
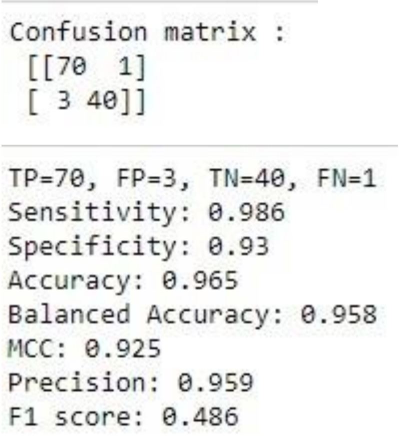


Figure 3.12 is for the Mean Correlation chart of the target attribute



4. Random Forest Analysis

Random Forest was carried out to inspect the coefficients of the key features as a tool to determine feature importance. This is done by looking at how the features predict if a patient has benign or malignant.



```
Confusion matrix :
 [[70  1]
 [ 3 40]]

TP=70, FP=3, TN=40, FN=1
Sensitivity: 0.986
Specificity: 0.93
Accuracy: 0.965
Balanced Accuracy: 0.958
MCC: 0.925
Precision: 0.959
F1 score: 0.486
```

Fig:4.0 Accuracy and confusion matrix of the model

5. Findings and Reflections:

Using a prediction model to classify breast cancer cases is statistical in nature. In this paper, Random Forest machine learning techniques for breast cancer diagnosis were used. The results are promising, and scores above 96%. As an extension of this work, future work includes the implementation of artificial neural net and deep learning for predictive model development with a larger and unstructured data set. This will use unsupervised learning algorithms such as dimensionality reduction PCA, SVM, etc. to first dimensionality reduction PCA, SVM etc. to first label the data and distributing them over training set, cross-validation set and test set.

6. Word Count

Introduction – 141 words

Data, Question Analytics Approach – 600 words

Findings – 200 words

7. REFERENCES

[1]. The Cancer Atlas- http://canceratlas.cancer.org/theburden/ breast-cancer/

[2]. American Institute of Cancer Research Statistics - https://www.wcrf.org/dietandcancer/cance r-trends/breast-cancer-statisticsC. Kang and A. Goldman. (2016, Dec) In Washington pizzeria attack, fake news Brought real guns. [Online]. Available: https://www.nytimes.com/2016/12/05/business/media/ comet-ping-pong-pizzashooting-fake-news-consequences.html

[3]. World Health Organization Cancer Report: https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

[4]. Kumar, V., Tiwari. P, Mishra. B.K., Kumar. S.: Implementation of n-gram methodology for rotten tomatoes review dataset sentiment analysis. International Journal of Knowledge

Discovery in Bioinformatics (IJKDB), 7(1), pp. 30–41. DOI: https://doi.org/10.4018/ijkdb.2017010103 (2017).

[5]. Kumar V., Verma A., Mittal N., Gromov S.V.: Anatomy of Preprocessing of Big Data for Monolingual Corpora Paraphrase Extraction: Source Language Sentence Selection. In: Emerging Technologies in Data Mining and Information Security. Advances in Intelligent Systems and Computing, Vol 814, pp. 495-505, Springer Nature, Singapore. DOI: https://doi.org/10.1007/978-981-13-15015_43 (2019).

[6].https://github.com/adbc602/IN3061_INM430_coursework_2021/blob/master/Documents/coursework/Coursework.ipynb