NEW YORK INSTITUTE OF TECHNOLOGY
College of Engineering and Computing Sciences

DTSC 710: Machine Learning (Spring 2023)
Project Report

Topic: Breast Cancer Classification and Analysis

Submitted by
Arris Moise (1317678)
Gail Elaine Goveas (1306196)
Patrick Adams (1231065)

Motivation:

Breast cancer is the most common type of cancer in women, except for skin cancers, at a rate of 1 in 3 of all new female cases a year (American Cancer Society). With up to 30% - 40% of breast cancers not being caught during screenings (Pacilè, Serena, et al), being able to find a classifier that might increase those odds seemed like the route to embark on. By comparing different classifiers in how accurately they classify a patient with a malignant or benign tumor, our goal is to leverage technology to improve breast cancer awareness and facilitate early detection, ultimately improving patient outcomes.

Methodology:

For the project, the following steps were completed:

1. Visualize the data and identify any potential correlations and check for outliers.
2. Apply data cleaning techniques, such as handling NaN values and converting text and categorical attributes to numerical form. Implement feature scaling and transformation pipelines if required, and perform Principal Component Analysis (PCA).
3. Create a training and validation set.
4. Train four classifiers - K-Nearest Neighbors (KNN), Decision Trees, Support Vector Machines (SVM), Logistic Regression (LR), and Naive Bayes - on the data.
5. Evaluate the performance of each classifier using cross-validation and select the model with the highest accuracy.
6. Fine-tune the hyperparameters of the chosen model using the Grid Search method.
7. Evaluate the model on a separate test set.
8. Develop a shiny app to recommend whether one should visit a doctor based on the inputs they provide.

1.1 Introduction

To achieve our objective, we have selected datasets consisting of three distinct statuses but sharing identical features and target variables. These statuses pertain to patients undergoing treatment, those who have recovered, and those who have deceased. Each dataset consists of 30 unique features, one of which is the target variable, used for determining whether a tumor is benign or malignant. However, it is not imperative to incorporate all 30 features during the classifier training process to develop a robust and accurate model. The feature selection process was conducted manually as part of the preprocessing phase.

1.2 Preprocessing

The initial step involved merging the three distinct datasets using the pandas concat function. This function facilitates the conjoining of objects along a specific axis, as illustrated in Fig. 1. Following the concatenation, the shape function from pandas was utilized to verify the successful merging as shown in Fig. 1. Upon completion of the concatenation, df.shape was used to ensure the dimensions added up.

```
#Combining the three datasets (under treatment, recovered and death) into one

df_1 = pd.read_excel('undertreatment.xlsx')
df_2 = pd.read_excel('recovered.xlsx')
df_3 = pd.read_excel('death.xlsx')

print(df_1.shape, df_2.shape, df_3.shape)

df = pd.concat([df_1, df_2, df_3], ignore_index=True)

df.shape
```

```
(350, 30) (186, 30) (598, 30)
(1134, 30)
```

Fig. 1: Using pd.concat to concatenate the datasets together

Subsequently, an examination was conducted on the dataset to identify any instances of missing or NaN values. It was discovered that the feature "menopausal_age" contained two missing values. In order to address this issue, the missing values were replaced with zeros, employing the descriptive statistics derived from the column's distribution. Moreover, during the analysis, it was identified that certain features within the dataset exhibited missing values where the "-" symbol was used as a placeholder. Consequently, the rows containing these entries were dropped from the dataset. Addressing missing values and handling erroneous data entries is a crucial step in the preprocessing phase. By effectively managing missing data, such as replacing them based on statistical measures and excluding rows with incorrect entries, the dataset's accuracy and reliability are improved, enhancing the robustness of subsequent analyses and modelling.

It is essential to identify and assess the presence of multicollinearity in the dataset before selecting a model. To evaluate the correlation between the features, a heatmap, as illustrated in Figure 2, was generated. The heatmap revealed significant correlations between the following pairs of features:

1. Gender and pregnancy experience (-0.61)
2. Gender and menstrual age (-0.75)
3. Age and menopausal age (0.51)
4. Pregnancy experience and marital status (0.67)
5. Pregnancy experience and giving birth (0.67)
6. Pregnancy experience and age first giving birth (0.63)
7. Pregnancy experience and menstrual age (0.54)

8. Giving birth and age first giving birth (0.54)

Furthermore, the correlation between these features and the target variable, Benign Malignant Cancer, was examined. The correlations were as follows:

1. Gender (0.08)
2. Pregnancy experience (-0.09)
3. Age (-0.06)
4. Menopausal age (0.04)
5. Marital status (-0.06)
6. Giving birth (-0.03)
7. Age first giving birth (0.01)
8. Menstrual age (-0.05)

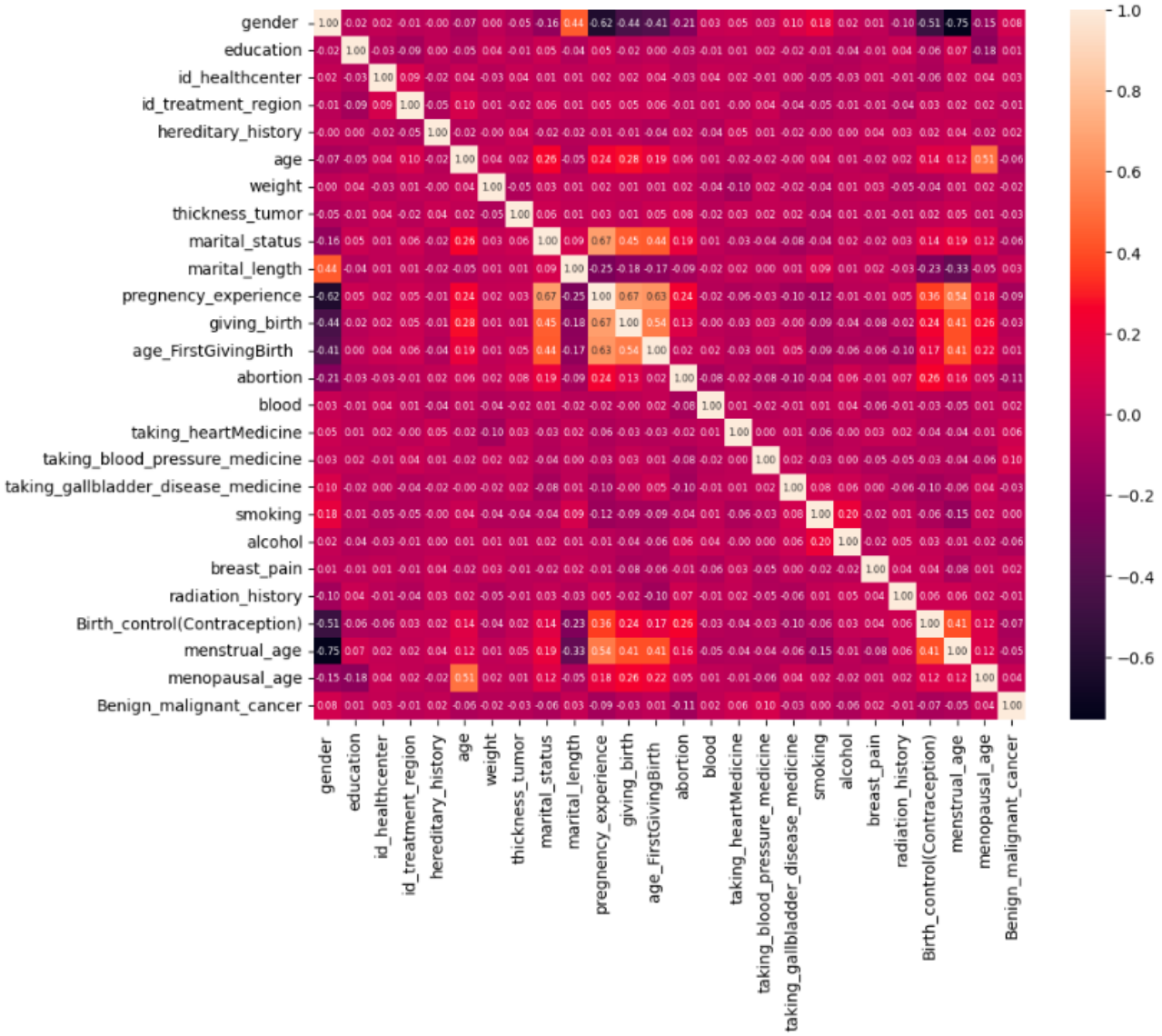

Fig. 2: Heatmap showing the correlation between the features and target.

Features exhibiting strong correlations with each other but weaker correlations with the target variable were subsequently eliminated from consideration. This selection process aimed to mitigate the effects of multicollinearity. Additionaly, the heatmap indicates that there is no strong correlation between the target and any of the features. Therefore, it may not be appropriate to use a linear regression model to analyze this dataset. Other modeling approaches, such as decision trees or support vector machines, which can handle non-linear relationships between variables, may be more suitable for this dataset.

Finally, prior to commencing the model training phase, the dataset underwent a scaling process using the Min-Max scalar function. This transformation was applied to normalize the range of values across the features, ensuring that they all fell within a specific range (typically 0 to 1). Scaling the data in this manner assists in mitigating the potential influence of variables with larger value ranges and facilitates fair comparisons between different features during the modeling process.

1.3 Modelling and Training

Once the preprocessing was finished it was possible to start deciding which classifiers to try and how to model each. The first note that should be mentioned is baseline accuracy. After reading a report on how 30% - 40% of breast cancers are missed during screening, we had a baseline accuracy goal of 60%. With that baseline, we were able to challenge four different classifiers to score above that. The four classifiers we chose to use were Decision Trees, K-Nearest Neighbor, Support Vector Machine, and AdaBoost Ensemble method with Decision Tree as the base estimator. The following will discuss the method and results of each model.

1.3.1 Decision Tree

After fitting the decision tree model with the data and running a 5-fold cross validation, the resulting average accuracy score was 0.5742 and an F1 score of 0.6552.

1.3.2 K-Nearest Neighbor

Following the same idea as when the decision tree was fitted, we also fitted the KNN model with a K value of 3. Average accuracy score of the 5-fold cross validation was 0.5339 and an F1 score of 0.6370.

1.3.3 Support Vector Machine

SVM resulted in the highest average and highest peak accuracy score of all the models we tested. The average was 0.5981 and an F1 score of 0.7351.

1.3.4 AdaBoost Ensemble w/ Decision Tree

AdaBoost Ensemble Method was the final classifier that was tested. Using the decision tree as the base estimator and setting the n_estimators to 100, the average score of the cross validation came out to 0.5849 and an F1 score of 0.661.

The subsequent bar graph Fig. 3., illustrates the performance of each model. Notably, the Support Vector Machine (SVM) demonstrates superior accuracy and F1 score in comparison to the other models, positioning it as the most proficient model for this specific dataset.

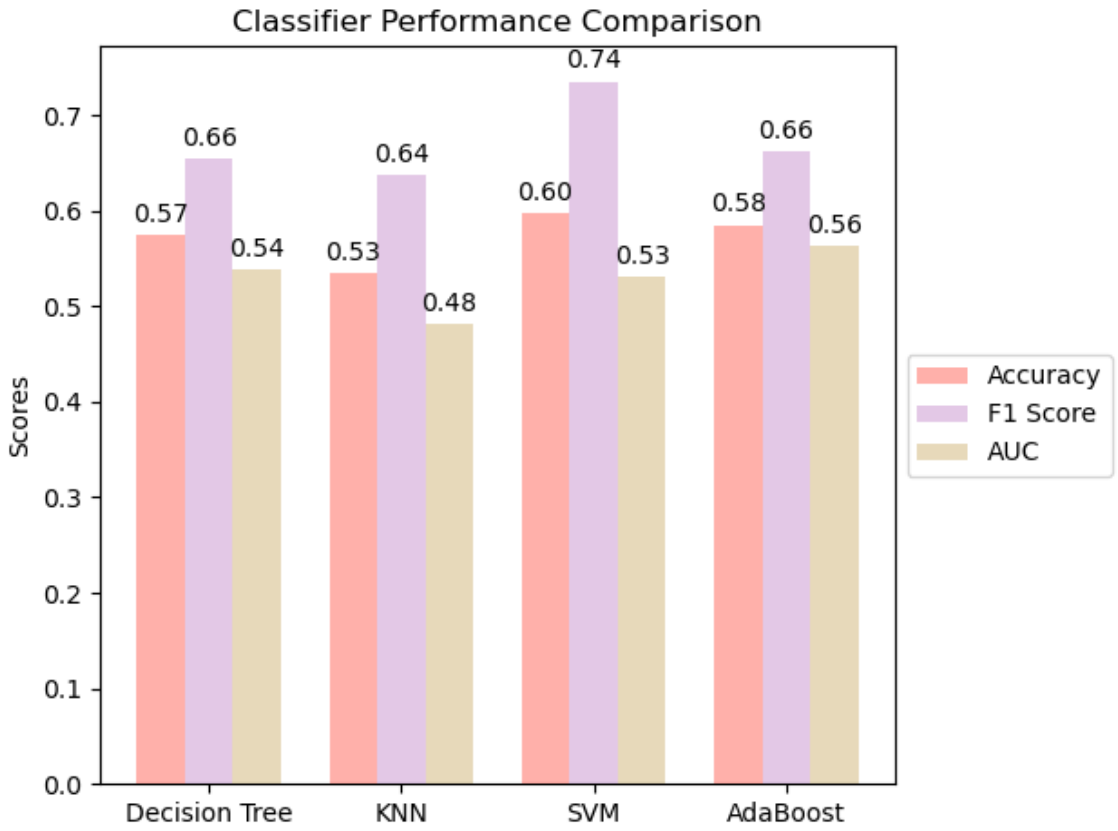

Fig 3: Classifier Performance Comparison

1.3.5 Hypertuning

The results of the experiment showed that SVM had the highest scoring average accuracy. With this in mind, the next step was to try to maximize optimality by running a Grid Search on possible parameters (fig.4). The parameters that performed the best in the Grid Search should, theoretically, increase the accuracy of the model as those are the best parameters the model has to offer. With that being stated, the GridSearchCV showed the best hyperparameters as 'C': 0.1, 'gamma': 0.1, and 'kernal': 'rbf'. Implementing these parameters, resulted with the accuracy score of 0.6161.

```
[28] # define the parameter grid to search over
     param_grid = {'C': [0.1, 1, 10, 100],
                   'gamma': [0.1, 0.5, 1],
                   'kernel': ['linear','rbf']}
```

1.4 Implementation and Analysis

The results of this project left us with hope to say the least. Breast cancer is one of the most common types of cancer within women and quite seldom do they realize they should be receiving a checkup. Our model resulted in a peak average score of 62%. This means our model is almost as accurate as industry standard mammograms and tests run on patients to check if they have a malignant tumor. The implementation of this model would be more so if we were to create an application for a user to test their symptoms. With the model having the ability to correctly identify a malignant tumor in a patient 8/13 times, we believe that if we were to keep finding ways to improve this model, the implementation would be boundless.

Future Work

For future work, we would like to take information and create an application that common people can use to gain more information about their condition using our model. People do not always have their medical records known off the top of their head but they can see when they have certain symptoms that are common with breast cancer. With that being said, we'd also like to make sure certain parameters such as lumps, swelling, or flaky skin can be taken into consideration to identify a model that delivers a higher accuracy. One last goal would be to properly work with industry experts to learn more about breast cancer to increase our ability to push early scanning to unknowing victims.

References

American Cancer Society. "Breast Cancer Statistics: How Common Is Breast Cancer?" *Breast Cancer Statistics | How Common Is Breast Cancer?*,

www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html. Accessed 12 May 2023. Last revised: January 12, 2023

Pacilè, Serena, et al. “Improving Breast Cancer Detection Accuracy of Mammography with the Concurrent Use of an Artificial Intelligence Tool.” *Radiology: Artificial Intelligence*, 4 Nov. 2020, pubs.rsna.org/doi/full/10.1148/ryai.2020190208.