Megan Hoeksema

Data Science Capstone – Midterm

**Introduction**

The Breast Cancer Wisconsin (Diagnostic) Data Set will be utilized in this projet. The data set can be obtained from Kaggle at <https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>. This data set contains 569 breast samples acquired through a fine needle aspirate (FNA). The FNA was then imaged and features defining the cell nuclei contained in the samples were described. Of the 569 breast samples, 357 were confirmed to be benign and 212 were confirmed to be malignant. 10 different features were used to describe the cell nucleus. The features include radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. These 10 features repeat using the mean, SE and “worst” to yield a total of 30 features. The target variable in this dataset will be diagnosis, either benign or malignant.

Of the following algorithms, KNN, SVM, Decision Trees, Random Forest, and Naïve Bayes, which will produce a better model for predicting the diagnosis of breast tissue masses as benign or malignant?

**Data Pre-Processing**

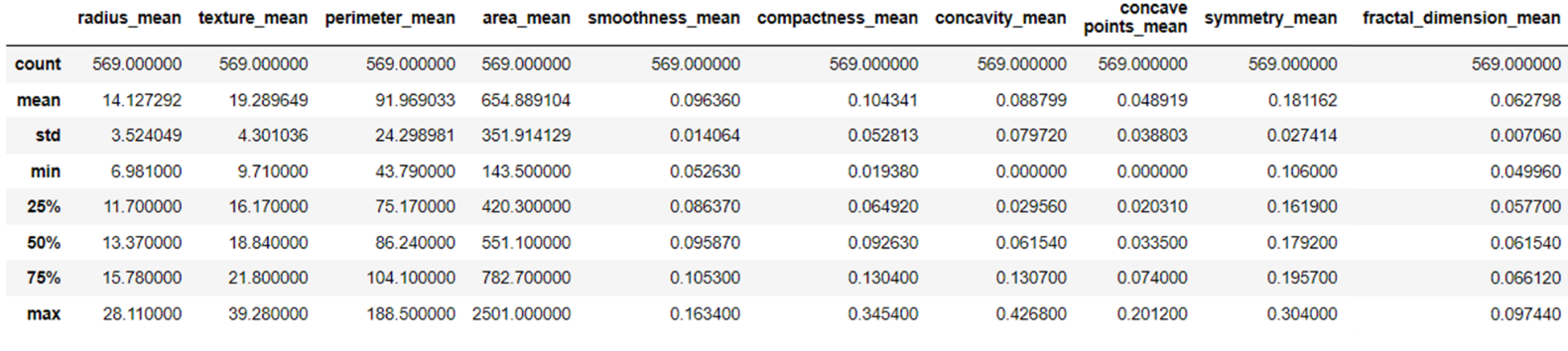
This dataset is very clean, so data pre-processing for this project ended up being pretty minimal. Upon checking the data, two columns could be removed. Column 32 was removed as it contained all null values and the column containing the patient identifier was also removed. Neither of these columns would prove necessary to build a model. After removing these columns, the rest of the columns were checked for null/missing values. When no null values were found, the target variable was transformed into a numerical value, with 0 indicating benign and 1 indicating a malignant diagnosis. Once the target variable was transformed, a MinMaxScalar was applied to all the features. Finally, the transformed dataset was split 70/30 into training and test sets. Below is an example of what the data set looks like after preprocessing.

Graphical user interface

Description automatically generated

**Exploratory Data Analysis**

Using the raw pre-transformed data, several visualizations were created to further understand what the features looked like and which features correlated with each other. First, descriptive statistics were determined which included the mean, std, min, and max.



The mean value of the 10 features were plotted against the diagnosis to try and identify any patterns in these features in terms of diagnosis. In the visualization below, we can see that overall, malignant mean values are higher for almost all the features except the fractal dimension. In this plot, we can also see the outliers in each of the features, however, for the purposes of this project, outliers were not excluded. However, if the models were to perform badly, we could consider removing some or all outliers from the dataset.

Chart, box and whisker chart

Description automatically generated

A heatmap was then generated to look at the correlation between each of the features, however, this created a very large plot. After looking through this plot, the top 5 features were extracted and then re-plotted in a smaller heat map to show a better visualization.

Chart, treemap chart

Description automatically generated

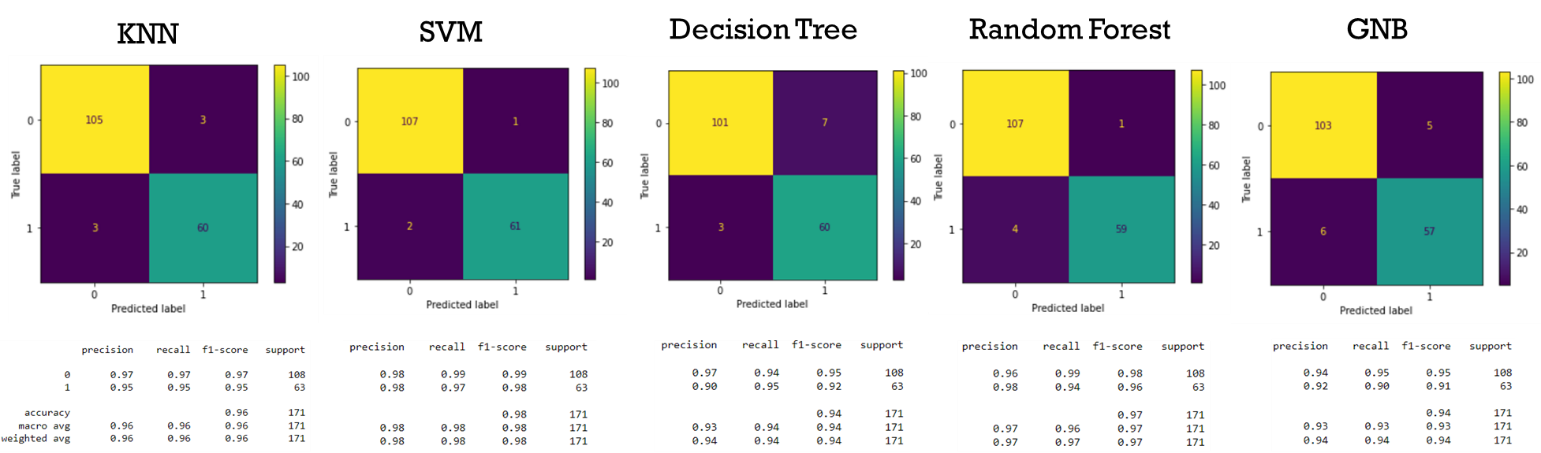
From the visualization above, we can conclude that the top 5 correlated features include the perimeter\_mean, radius\_mean, radius\_worst, perimeter\_worst and the area\_mean. The goal of this exercise would be to help later in the project during the model tuning. If the models perform badly, we can re-run the models with just the top 5 features in hopes of improving the model.

**Model Building and Evaluation**

The five models were first built without using any feature selection or hyperparameter tuning. This would give us a baseline to work with and to improve from should we need to do feature selection or hyperparameter tuning. The five classification models used in the project were KNN Classifier, SVM Classifier, Decision Tree Classifier, Random Forest Classifier, and Gaussian Naïve Bayes Classifier. The packages for each of these models were from sklearn. After the model was set up, the model was used to predict on both the training and test sets. Accuracy scores were printed out for each model on each set.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | KNN | SVM | Decision Tree | Random Forest | GNB |
| Training Set | 0.977 | 0.982 | 1.0 | 1.0 | 0.942 |
| Test Set | 0.965 | 0.982 | 0.941 | 0.970 | 0.935 |

All models perform very well. Based on the accuracy scores on the test set, SVM is the best performer while GNB is the worst. Classification reports and confusion matrices were also printed out to help evaluate the models as well as ROC curves and Calibration plots.



Chart, line chart

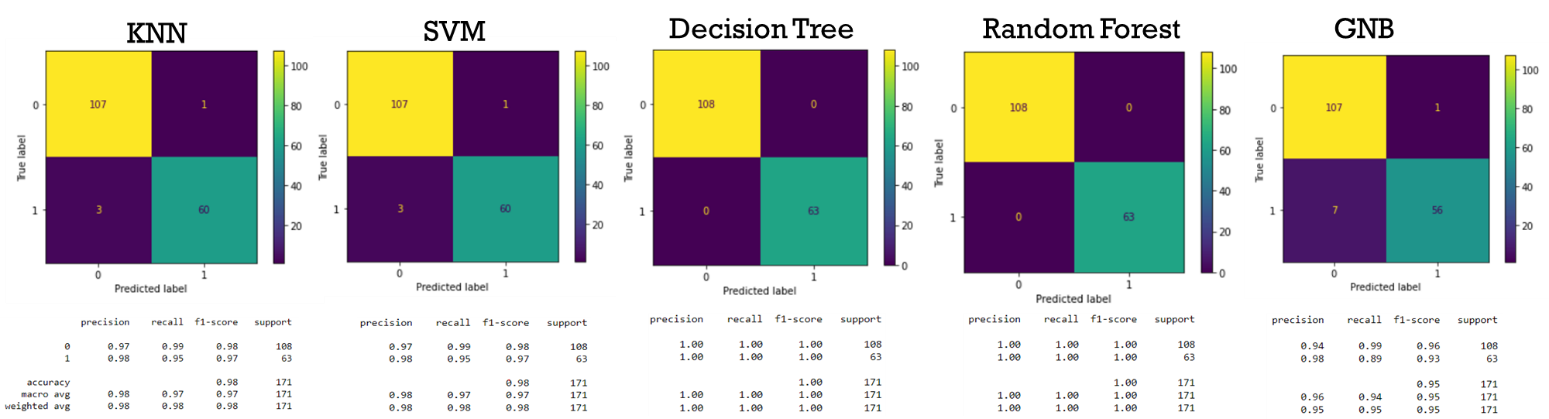
Description automatically generated

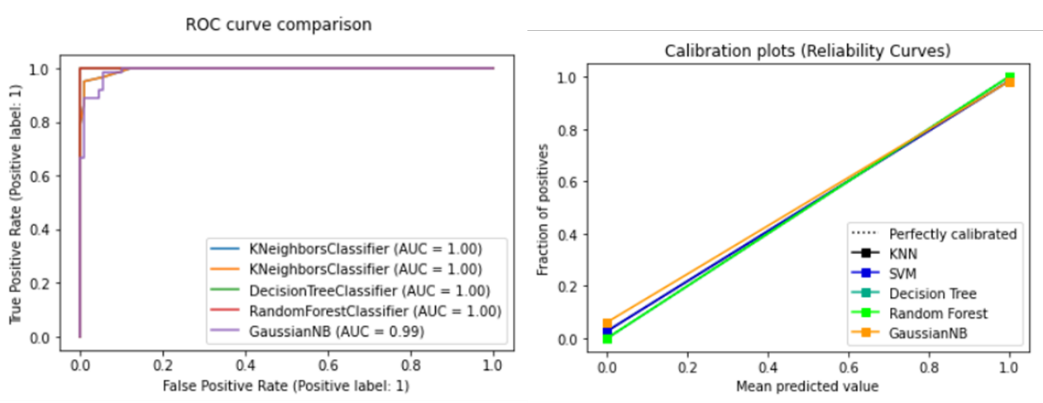
After reviewing both the accuracy scores and the classification table, we can conclude that all models perform very well however, if we absolutely had to choose one that performs the best, based on the accuracy, precision and recall scores, SVM would be the best model to use. Both the calibration plots and the ROC curves also support this. Although all 5 models performed extremely well, we went ahead and tuned the hyper parameters for each of the models and compared their performances as well.

Not all hyper parameters were tuned. A few parameters were chosen for each model. Parameters were tuned using GridSearchCV. For KNN, the only hyper parameter to tune is the number of neighbors. For SVM, the C, gamma and kernel were all tuned, resulting in C = 1, gamma = 1 and kernel = linear. The parameters max depth and max features were tuned for decision trees which yielded a max depth of 11 and a max features of 0.4. Hyper parameter tuning for Random Forest resulted in a max depth of 5 and a max features of 0.2. Finally, Gaussian Naïve Bayes was tuned resulting in var smoothing of 0.28. After tuning the hyperparameters, the tuned models were re-evaluated on the test set. The accuracy scores were again printed out for each model and compared to those of the untuned model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | KNN | SVM | Decision Tree | Random Forest | GNB |
| Training Set | 0.977 | 0.982 | 1.0 | 1.0 | 0.942 |
| Test Set | 0.965 | 0.982 | 0.941 | 0.970 | 0.935 |
| Test (Tuned) | 0.977 | 0.982 | 1.0 | 1.0 | 0.953 |

Comparing the accuracy scores of the un-tuned to the tuned models on the test set reveals slight increases for all models except SVM with the most significant increase being the Decision Tree model. Classification reports, confusion matrices, ROC curves and Calibration plots were all visualized and reviewed.





**Conclusions**

This breast cancer dataset is very clean and minimal preparation was required. After removing two unnecessary columns, data underwent preliminary evaluation. Here, descriptive statistics were performed along with plotting each of the features means against the diagnosis to look for patterns in the data. This resulted in visualizing the outliers as well. For the purposes of this project, outliers were not removed, and proved unnecessary to remove them after building the initial models. A MinMaxScaler was applied to all features, the target was converted from categorical to numerical and the dataset was split 70/30 into a training and test set. The initial 5 models were built without outlier removal, feature extraction or hyper parameter tuning. This was because we wanted to see how the models behaved in order to obtain a baseline and see how we could improve upon those numbers. The initial 5 models proved to perform very well and any one could be used as the final model, however, we went ahead and tuned some of the hyper parameters for each of the models. The tuned models were then ran against the test set and re-evaluated. Each model improved slightly apart from SVM; however, Decision Trees improved the most. After full evaluation, we can conclude that all 5 models would prove to be great performers, however, based on the accuracy scores, ROC curves and classification reports, the best model to choose from would be Decision Trees or Random Forest.

While this project has identified several models capable of predicting malignant over benign tumors, the believability of this from a cancer biology perspective is minimal. From the preliminary data analysis, we can see that most of the features, when plotted against diagnosis, show a higher mean value in the malignant diagnosis over benign indicating that the nuclei found in malignant masses appear significantly different than those found in benign masses. Most likely, the malignant tumors found in this dataset are stage 4 cancer, yielding such large differences, thus creating high performing models. Ideally, we would like to repeat this project on a dataset where the nuclei of the malignant masses resemble those of benign masses, for instance, stage 1 cancer. This would prove to be more difficult for these models to determine malignant over benign. It would also be very interesting to have a dataset with staging information. So, comparing benign vs malignant stage 1, 2, 3, and 4 and try to build a model than can successfully determine not only whether the mass is benign or malignant but what stage the cancer is presenting at.

In conclusion, all 5 models performed very well in determining the diagnosis of breast cancer patients.