UNIVERSITY of HOUSTON

CULLEN COLLEGE of ENGINEERING

Breast Cancer Classification Dataset

Presented by

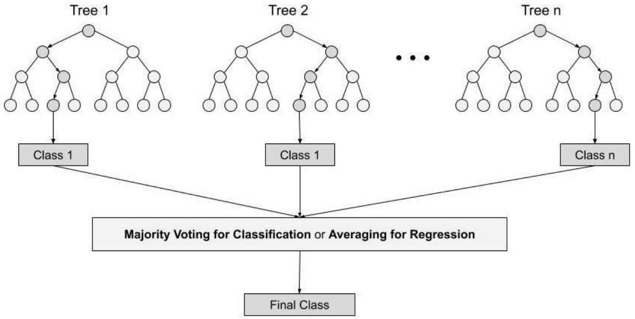
Kaushal Arvindbhai Dhanani

About The Dataset

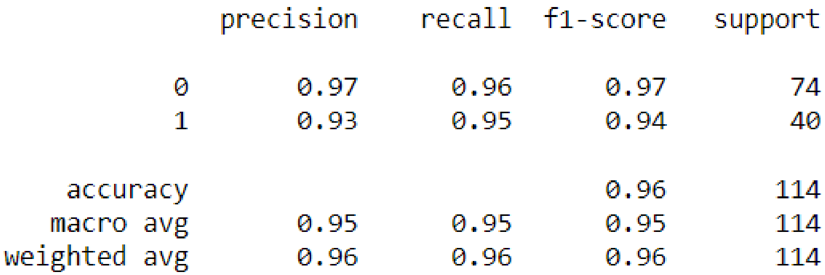
- The dataset is about Breast Cancer classification that is about the diagnosis of cancer into Malignant or Benign categories.
- By processing the given dataset using machine learning techniques we will be able to predict the
- The dataset contains 32 different types of variables and all of them are numerical variables.
- We have target variable as ‘diagnosis’, which is 0 if diagnosis is ‘Benign’ and shows ‘1’ if the diagnosis is ‘Malignant’

Random Forest Classifier:

- A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
- The random forest utilizes techniques called BOOTSTRAPPING and AGGREGATING, commonly known as BAGGING.
- It is easy to use and flexible, as it handles both classification and regression problems.
- In Random forest n number of random records are taken from the data set having k number of records. Individual decision trees are constructed from each sample.
- Each decision tree will generate an output. Final output is considered based on ***Majority Voting or Averaging*** for Classification and regression respectively.


Tree 1
Tree 2
. . .
Tree n
Class 1
Class 1
Class n
Majority Voting for Classification or Averaging for Regression
Final Class

Source: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/



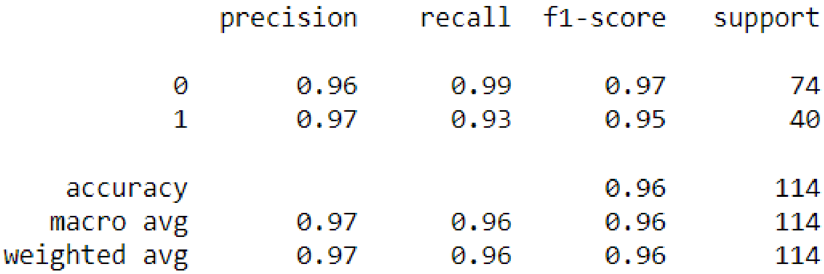
```
              precision    recall  f1-score   support

           0       0.97      0.96      0.97        74
           1       0.93      0.95      0.94        40

    accuracy                           0.96       114
   macro avg       0.95      0.95      0.95       114
weighted avg       0.96      0.96      0.96       114
```

Support Vector Classifier

- SVC, or Support Vector Classifier, is **a supervised machine learning algorithm typically used for classification tasks**. SVC works by mapping data points to a high-dimensional space and then finding the optimal hyperplane that divides the data into two classes.
- The SVM kernel is a function that takes low dimensional input space and transforms it to a higher dimensional space i.e. it converts not separable problem to separable problem. It is mostly useful in non-linear separation problem.
- We have trained the SVM model with ‘rbf’ kernel, which is non-linear kernel. Below image shows the classification report for this model



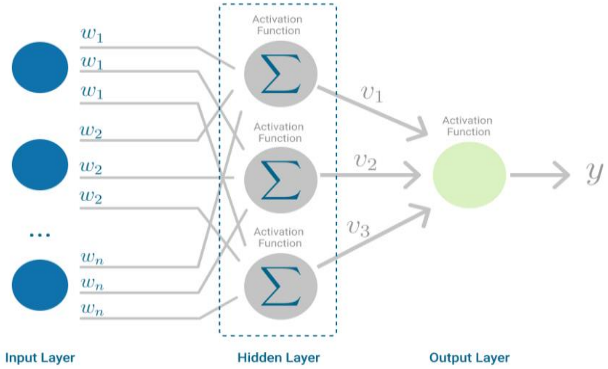
```
              precision    recall  f1-score   support

           0       0.96      0.99      0.97        74
           1       0.97      0.93      0.95        40

    accuracy                           0.96       114
   macro avg       0.97      0.96      0.96       114
weighted avg       0.97      0.96      0.96       114
```

Multi Layer Perceptron

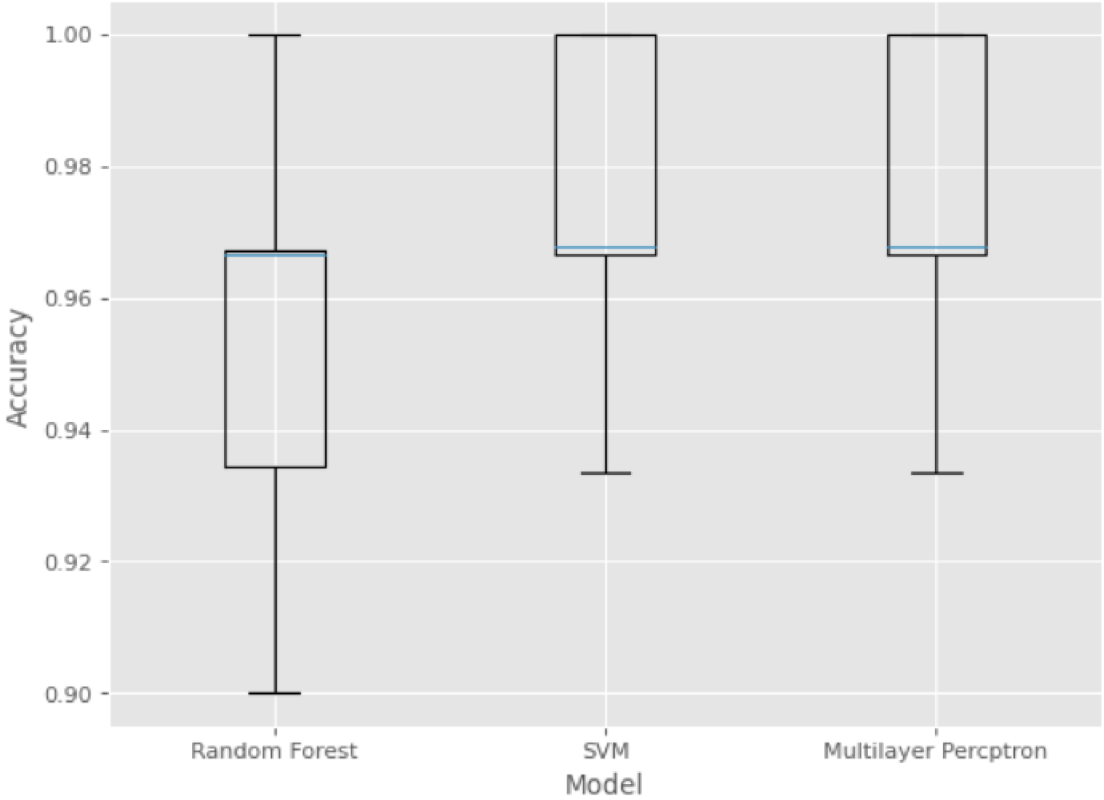
- The **Multilayer Perceptron** is a neural network where the mapping between inputs and output is non-linear.
- A Multilayer Perceptron has input and output layers, and one or more **hidden layers** with many neurons stacked together. And while in the Perceptron the neuron must have an activation function that imposes a threshold, like ReLU or sigmoid, neurons in a Multilayer Perceptron can use any arbitrary activation function.
- Multilayer Perceptron falls under the category of feedforward algorithms, because inputs are combined with the initial weights in a weighted sum and subjected to the activation function, just like in the Perceptron. But the difference is that each linear combination is propagated to the next layer.
- Each layer is *feeding* the next one with the result of their computation, their internal representation of the data. This goes all the way through the hidden layers to the output layer.
- Accuracy for MLP Classifier was 96% for our model.



Source: https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

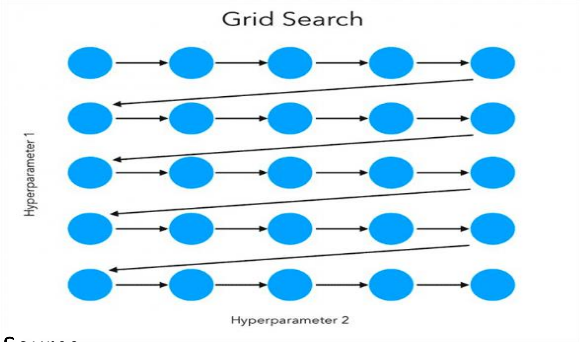
Model Evaluation without Hyperparameter Tuning



UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Hyperparameter Tuning

- GridserachCV and RandomizedsearchCV are well-known techniques for model structure selection.
- We have used GridSearchCV in our project to find the optimal hyperparameters.
- GridSearchCV is a technique of performing hyperparameter tuning to find out the best values of hyperparameter for given model.
- GridSearch involves using a different combinations of all given hyperparameters and their values and find out the performance of each combination and selects the best value for hyperparameters.
- Cross-validation is also performed while using GridSearchCV. Cross-validation is used while training the model.
- Hence, this process with cross-validation is time consuming to evaluate the best hyperparameters.



Source:
https://maelfabien.github.io/machinelearning/Explorium4/#

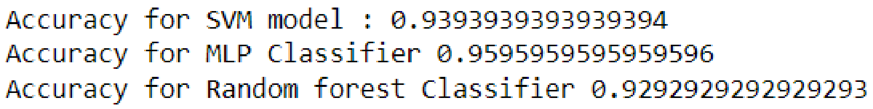
Results After Hyperparameter Tuning using GridSearchCV

We have used K-fold cross validation and GridSearchCV to find the optimal hyperparameters

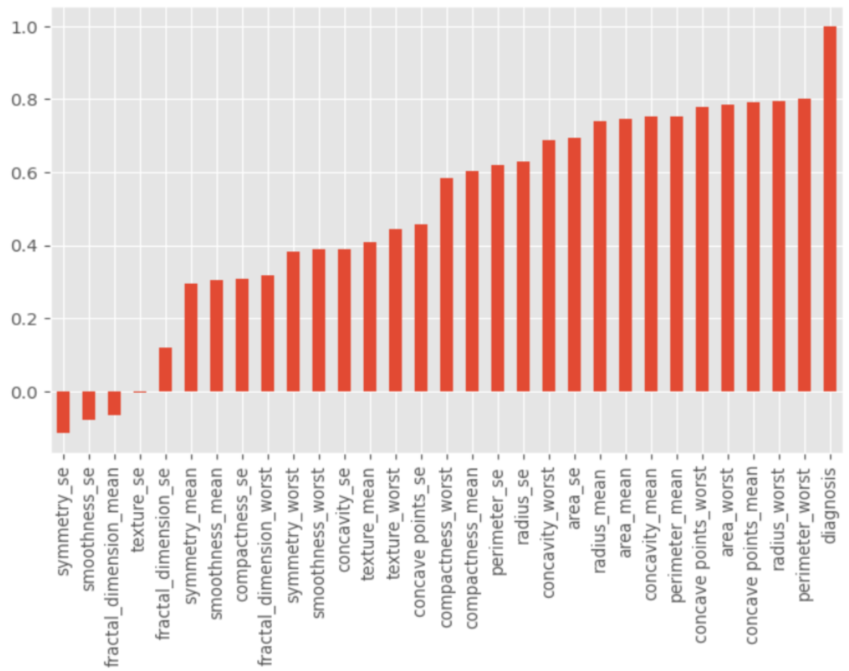
Model	Accuracy
Random Forest	0.969
Support Vector Machine	0.965
Multilayer Perceptron	0.964

Feature Selection with Pearson’s Correlation Coefficient

- The Pearson correlation method is a statistical technique used for feature selection in machine learning and data analysis.
- It measures the linear relationship between two variables, which are typically continuous numerical variables. The Pearson correlation coefficient, denoted by "r", ranges from -1 to 1, with values closer to 1 indicating a strong positive correlation, values closer to -1 indicating a strong negative correlation, and values close to 0 indicating no correlation.
- We have selected the variable with correlation coefficient greater than 5 and then trained our model.
- We have shown the results of our model on this image.

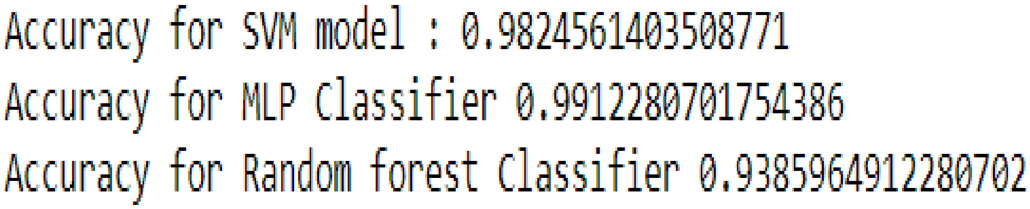


```
Accuracy for SVM model : 0.9393939393939394
Accuracy for MLP Classifier 0.9595959595959596
Accuracy for Random forest Classifier 0.9292929292929293
```



Variable Selection with Lasso:

- Lasso is a supervised algorithm wherein the process identifies the variables that are strongly associated with the response variable. This is called variable selection.
- With LassoCV method, we have selected the features accordingly and then trained the model with those features for the best model.



```
Accuracy for SVM model : 0.9824561403508771
Accuracy for MLP Classifier 0.9912280701754386
Accuracy for Random forest Classifier 0.9385964912280702
```

- After comparing this results with the one we got from Pearson’s method, we can say that LassoCV performs better and do good job in classifying the Breast cancer.