Submitted By: Pratik B Kanase
Kanasepratik2@gmail.com
+91-8087475700

Assignment: Prediction of benign or malignant cancer tumors

Q1. Please show all your work in the Python Jupyter notebook.

Detailed Python jupyter notebook is attached with the mail.

Q2. Using data visualization tools, please explain how we can understand the data structure

- With the Data visualization approach, we can apply a variety of techniques on the dataset which include an understanding of data, maximize insight into the dataset, identifying anomalies, identifying outliers, understanding behavior and distribution of data

- With the help of different charts, diagrams, plots we can understand data in a very well manner also it will help to extract meaningful information from the data that can be applicable and beneficial for businesses. Visualization provides a unique perspective on the dataset. You can visualize data in lots of different ways.

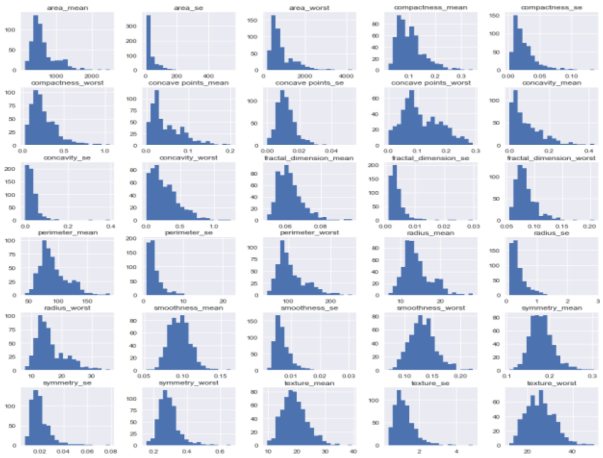
- The main goal is to filter large datasets into visual graphics to allow for an easy understanding of complex relationships within the data. Revealing previously unnoticed key points about the data sources to help decision-makers compose data analysis reports.

Following is the approach I have used for the visualization and analysis of the data.

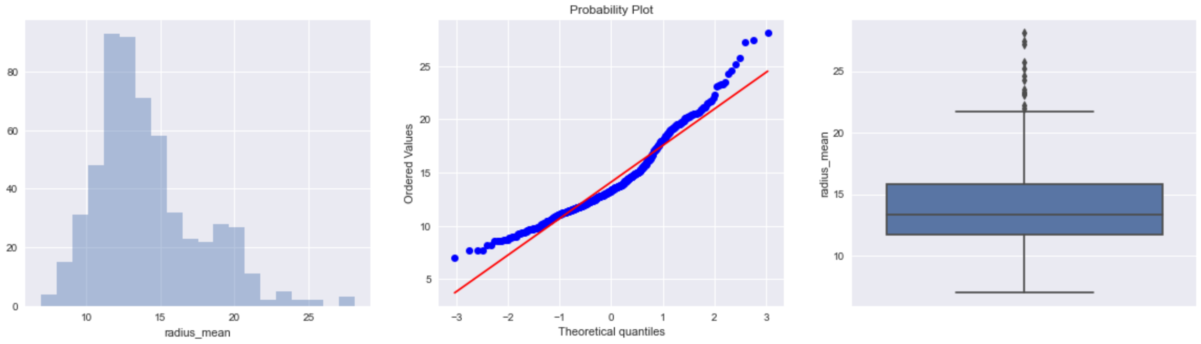
- **Univariate Analysis of Data:**

In univariate analysis, if data is present in the continuous format then we can analyze it using histogram, frequency plot, and density plots.

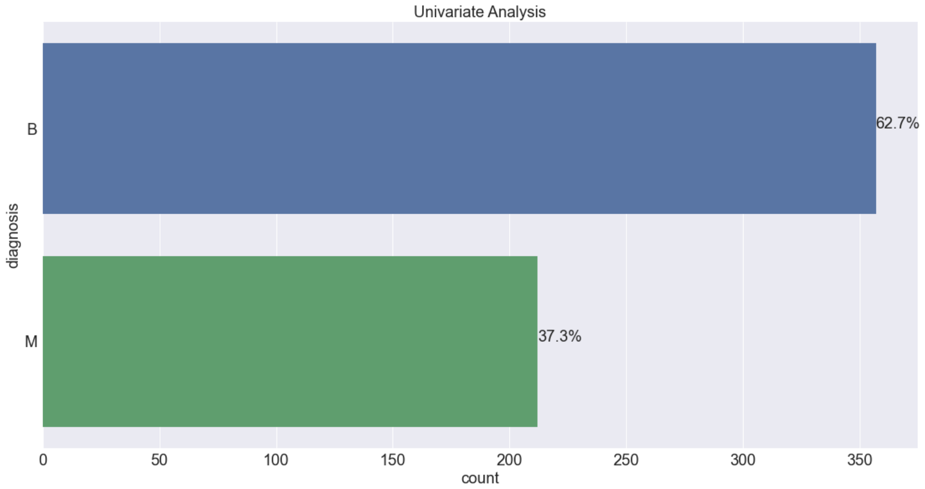
Following Fig Shows the univariate analysis with the histogram of all independent features.



Following fig shows the univariate analysis with the help of qq plot, box plot and histogram



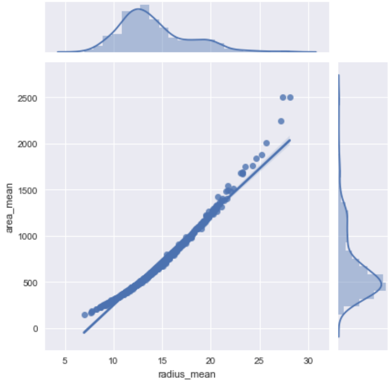
Following fig shows the univariate analysis of categorical dependent feature



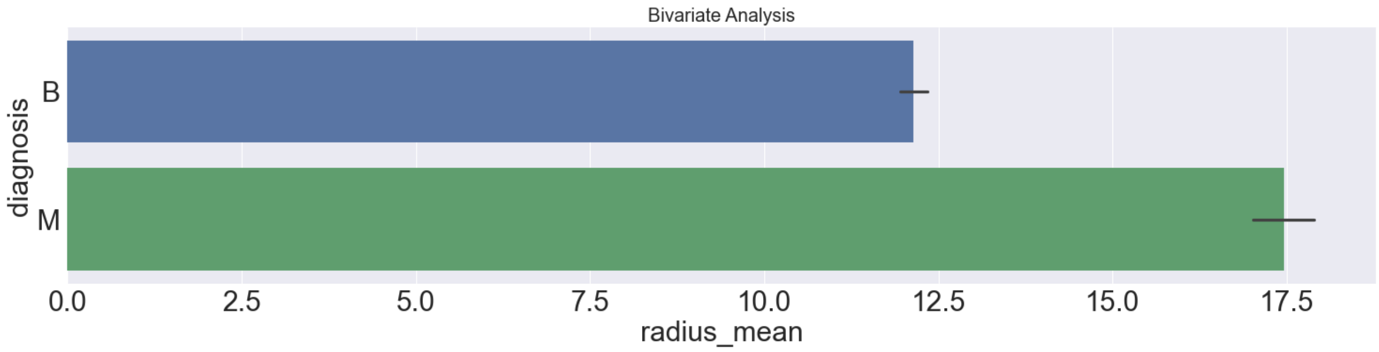
Whereas if the data is present in the discrete format then we can analyze it using barplot and counterplot.

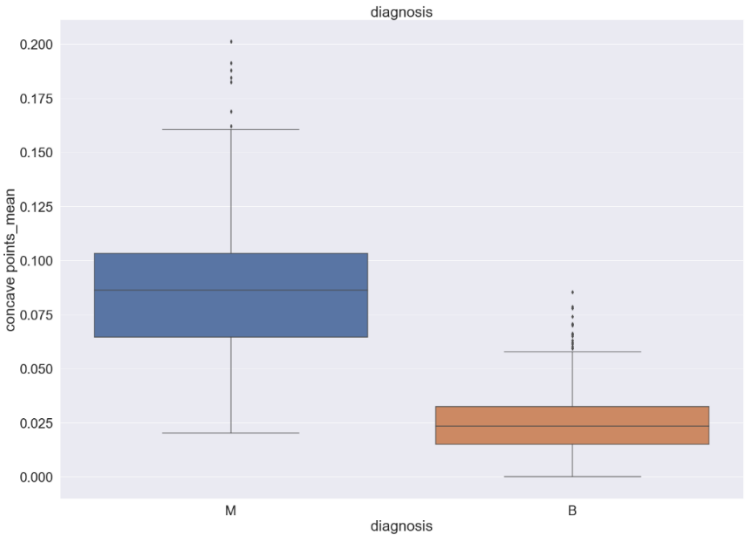
- **Bivariate Analysis of Data:**

If both the input features are continuous in nature then we can use a scatter plot to analyze the data.



Bivariate analysis with the help of barplot and boxplot.





And if one of the features is continuous and another one is discrete then we can use a bar plot, box plot, or violin plot accordingly.

- **Multivariate Analysis**

In multiverse analysis, we can use a heatmap or contour plot to understand the significance of all multiple e features present in the dataset.

Q3. Please explain if dimensionality reduction is required/possible or not. How did you check?

- The main aim of using the dimensionality reduction technique, in this case, was to avoid the Curse of Dimensionality. As the number of features increases, the model becomes more complex. The more the number of features, the more the chances of over fitting. Hence to avoid such types of problems associated with the model I have used the PCA as a dimensionality reduction technique.

- Another aim to use the principal component analysis for this data is to reduce the complexity of the model by expecting only useful information from the data. With the help of PCA we can transfer the data from a higher dimension to a lower dimension thus it will help to reduce complexity and over fitting of data.

- From the analysis of the given data, it is observed that there was multicollinearity present inside the dataset. Hence PCA is the main methodology to take care of multicollinearity by removing redundant features.

- It’s not feasible to analyze each and every variable at a microscopic level. It might take much more time to perform any meaningful analysis. Hence dimensionality reduction with PCA is a better way to deal with high-dimensional data so that we can quickly extract patterns and insights from it. Some algorithms do not perform well when we have large dimensions. So reducing these dimensions needs to happen for the algorithm to be useful

.

Problem with Many Input Variables:

- The performance of machine learning algorithms can degrade with too many input variables.

- We can consider the columns of data representing dimensions on an n-dimensional feature space and the rows of data as points in that space. This is a useful geometric interpretation of a dataset.

- Having a large number of dimensions in the feature space can mean that the volume of that space is very large, and in turn, the points that we have in that space (rows of data) often represent a small and non-representative sample.

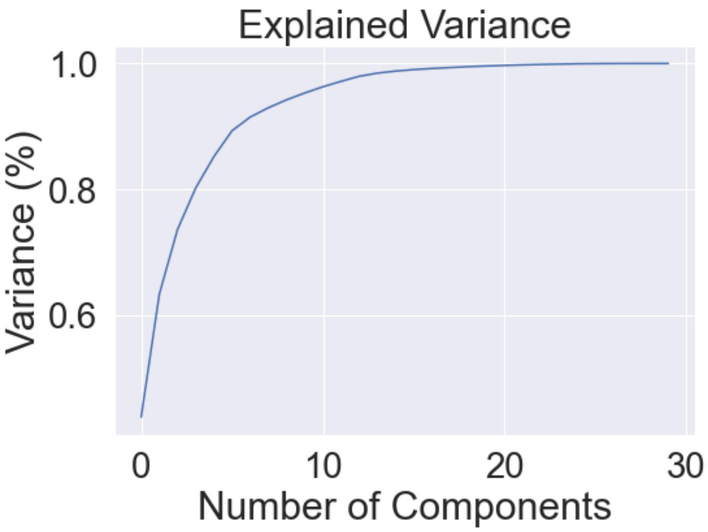
- This can dramatically impact the performance of machine learning algorithms fit on data with many input features, generally referred to as the “curse of dimensionality.

Q.4 Please explain eigen-vectors and eigen-values and their importance.

- Principal components can be geometrically seen as the directions of high-dimensional data that capture the maximum amount of variance and project it onto a smaller dimensional subspace while keeping most of the information.

- We need to make your data values between 0 to 1 because PCA will get affected by the scale of features. Hence with the help of a standard scalar, we need to standardize our data.
- After Standardization, we create a covariance matrix, this covariance matrix will tell that how much two features are related, and also this will help us to get eigenvalues and eigenvectors.

- Now we have eigenvalues and eigenvectors. So we can sort eigenvectors on the basis of eigenvalues to get top rated eigenvectors. The following graph shows the relationship between variance and the number of components.



Explained Variance Ratio:

```
array([0.48764841, 0.16814246, 0.086523  , 0.0629324 , 0.05309537,
       0.03971485, 0.02384921, 0.01701543, 0.01176614, 0.01005909,
       0.0087148 , 0.00702621, 0.00646907, 0.00459661, 0.00264262])
```

- from the graph, we can see that nearly 15 eigenvectors having good eigenvalues. So out of 30 eigenvectors, we will take only 15 eigenvectors for further analysis, because rest features having 0 eigenvalues, it means their presence will not affect the output of the model.so the shape of the eigenvectors will reduce to (15, 30) from (30,30). We will use 15, 30 eigenvectors to create our new Dataset.

- Hence we the help of PCA the objective of dimensionality reduction is achieved.

Q.5 Which classification methods are you using? How do you decide among different methods?

I have used the following classification algorithms for prediction:

1. Random Forest without hyperparameter tunning
2. Random forest with hyperparameter tunning
3. Random forest with recursive feature elimination
4. Support vector classifier
5. Random forest with principal component analysis
6. Logistic regression with principal component analysis

On the basis of interpretation of the following parameters I chose the best-suited algorithm:

- Accuracy and/or Interpretability of the output
- Speed or Training time
- The accuracy of the model.
- The interpretability of the model.
- The complexity of the model.
- The scalability of the model.
- How long does it take to build, train, and test the model
- Roc_auc_score, F1 Score, and Precision
- Confusion matrix analysis

Depending upon the analysis and comparison of the above parameters with respect to different algorithms I have chosen the best performing algorithm

---

Q.6 Please provide a confusion matrix and explain how it can help us to check the reliability of the result.

The following are the detailed analysis of the classification evaluation matrix for different algorithms.

Model	Confusion Matrix	Train Score	Test Score	Precision	Recall	F1 Score	roc auc score
Random Forest without RFE and Without Hyperparameter Tunning	[[100 8] [5 103]]	1.0	0.939	0.9259	0.95238	0.93896	0.93981
Random Forest with Feature selection and With Hyperparameter Tunning	**[[99 9] [2 106]]**	**1.0**	**0.949**	**0.9166**	**0.98019**	**0.94736**	**0.94907**
Support Vector Classifier with Optimum Feature selection	[[101 7] [5 103]]	1.0	0.944	0.9351	0.95283	0.94392	0.94444
Random Forest on PCA Data	[[95 13] [15 93]]	0.997	0.870	0.8796	0.86363	0.871559	0.87037
Logistic Regression on PCA Data	[[101 7] [8 100]]	0.997	0.930	0.9351	0.92660	0.93087	0.93055

- From the above analysis we can interpret that random forest with hyper parameter tuning gives us the best auc score of 0.94 and best precision and recall values comparing with other algorithms. Hence the random forest with the hyper parameter tuning is best suited model for this data.

- Calculating a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making.

- The main problem with classification accuracy is that it hides the detail you need to better understand the performance of your classification model.

- Precision is a useful metric in cases where False Positive is a higher concern than False Negatives. The recall is a useful metric in cases where False Negative trumps False Positive. The recall is important in medical cases where it doesn’t matter whether we raise a false alarm but the actual positive cases should not go undetected

- F1-score is a harmonic mean of Precision and Recall, and so it gives a combined idea about these two metrics. It is maximum when Precision is equal to Recall.

Detailed Description of Test Results:

1. **Model: Random Forest without Feature selection and Without Hyper parameter Tuning**

- Confusion Matrix:
  - [[103   5]
  - [4 104]]

- Training Score:
  - 1.0

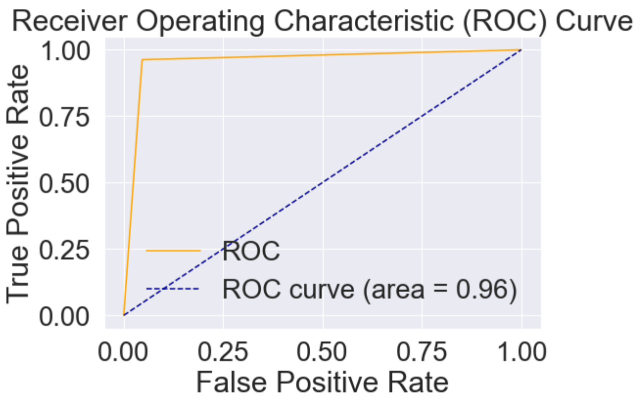
- Test Score/Accuracy Score:
  - 0.9583333333333334

- Precision:
  - 0.9537037037037037

- Recall: 0.9626168224299065

- F1_Score:
  - 0.958139534883721

- roc_auc_score:
  - 0.9583333333333333



2. Model: Random Forest with Feature selection and Without Hyperparameter Tunning

- Confusion Matrix:
  - [[100   8]
  - [5 103]]

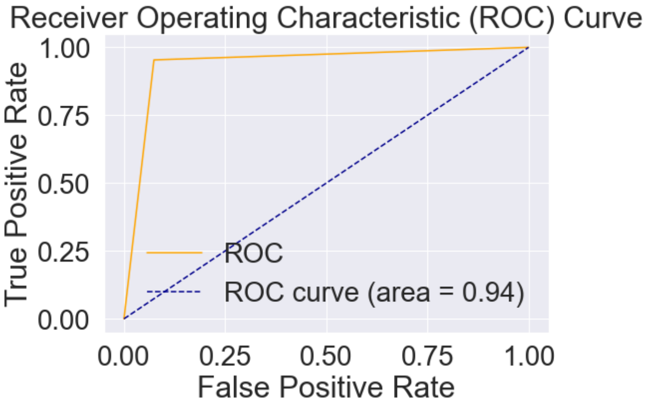
- Training Score:
  - 1.0

- Test/Accuracy Score:
  - 0.9398148148148148

- Precision:
  - 0.9259259259259259
-
- Recall: 0.9523809523809523

- F1_Score:
  - 0.9389671361502347

- roc_auc_score:
  - 0.9398148148148148



3. Model : Random Forest with Feature selection and With Hyperparameter Tunning

- Confusion Matrix:
  - [[99   9]
  - [2 106]]

- Training Score:
  - 1.0

- Test/Accuracy Score:
  - 0.9490740740740741

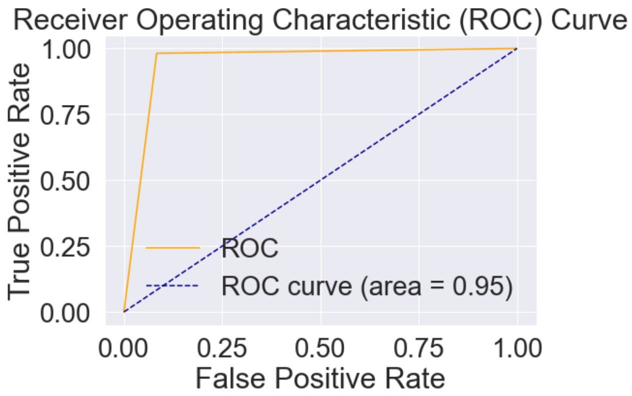
- Precision:

    - 0.9166666666666666

- Recall: 0.9801980198019802

- F1_Score:
    - 0.9473684210526315

- roc_auc_score:
    - 0.9490740740740741



4. Model: Support Vector Classifier with Optimum Feature selection

- Confusion Matrix:
    - [[101   7]
    - [5 103]]

- Training Score:
    - 1.0

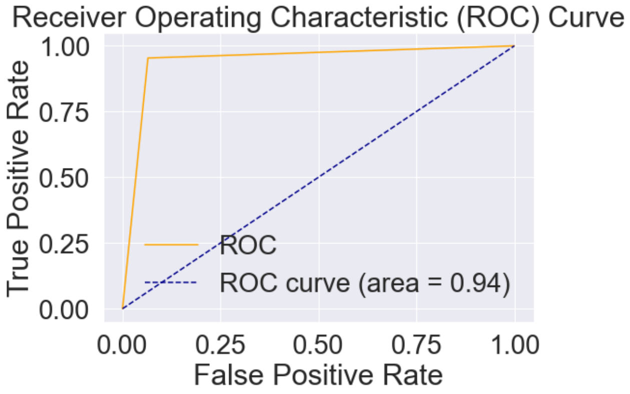
- Test/Accuracy Score:
    - 0.9444444444444444

- Precision:
    - 0.9351851851851852

- Recall: 0.9528301886792453

- F1_Score:
    - 0.9439252336448598

- roc_auc_score:
    - 0.9444444444444445

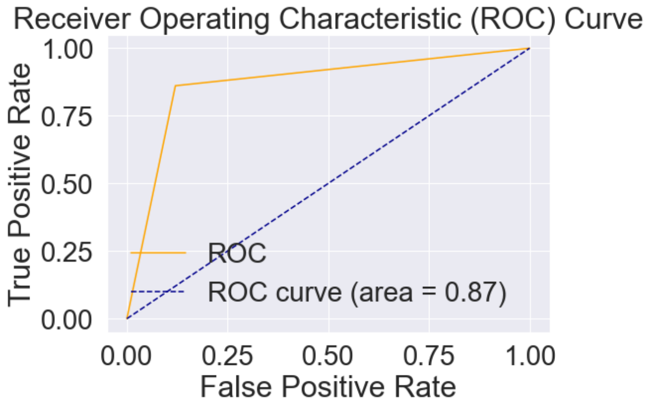


5. **Model: Random Forest on PCA Data**

- Confusion Matrix:
  - [[95 13]
  - [15 93]]

- Training Score:
  - 0.9979919678714859

- Test/Accuracy Score:
  - 0.8703703703703703



- Precision:
  - 0.8796296296296297

- Recall: 0.8636363636363636

- F1_Score:
  - 0.8715596330275229

- roc_auc_score:
  - 0.8703703703703705

6. **Model: Logistic Regression on PCA Data**

- Confusion Matrix:
  - [[101   7]
  - [8 100]]

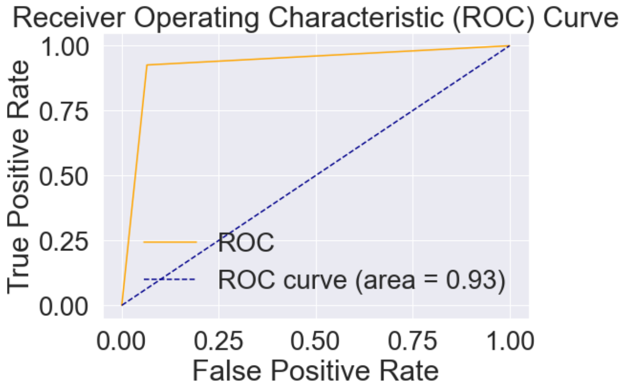
- Training Score:
  - 0.9979919678714859

- Test/Accuracy Score:
  - 0.9305555555555556

- Precision:
  - 0.9351851851851852

- Recall: 0.926605504587156

- F1_Score:
  - 0.9308755760368663
- roc_auc_score:
  - 0.9305555555555557



Q.7 Please provide the learning curve and explain how it can help us in determining whether the model is being over-fit or under-fit.

- The learning curve visualizes the effect of the number of observations on the performance metrics. Learning curves can be used as the diagnostic tool to analyze and interpret the model performance.

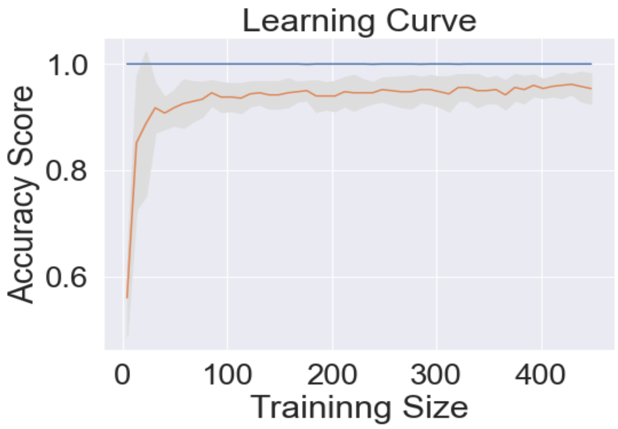
- The learning curve is the graph which is drawn between performance metric and number of observations. On the learning curve, we can observe and analyze the performance metric of the training set and performance metric on the test set with the number of observations. With the help of this learning curve, we can analyze how our data and its parameters are affecting the performance of the model.

- As we are increasing the number of data points the testing accuracy is increasing, it is increasing after some point after this it started to saturate. From the learning curve we can analyze that if we increase the size of the dataset, that will not guarantee to increase in the model performance with better accuracy.

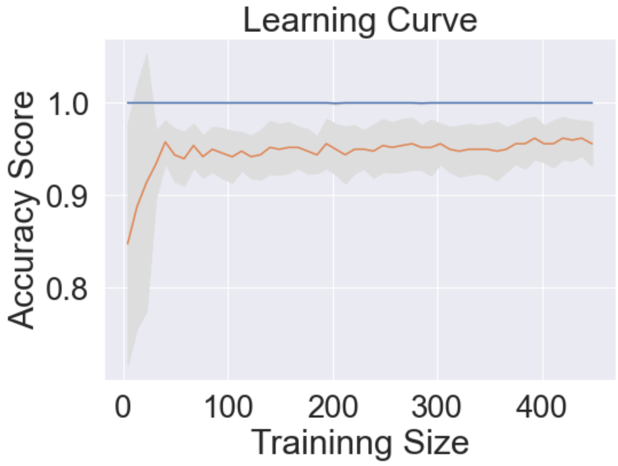
- The model with low bias and high variance is not a desirable condition for a better machine learning approach. For the stability of the model, this condition should be avoided. In this case, the total error is high. Hence we need to select some between the value of bias and variance where the total error is reduced and model complexity is optimum

The following are the learning curves and their analysis from the given dataset.

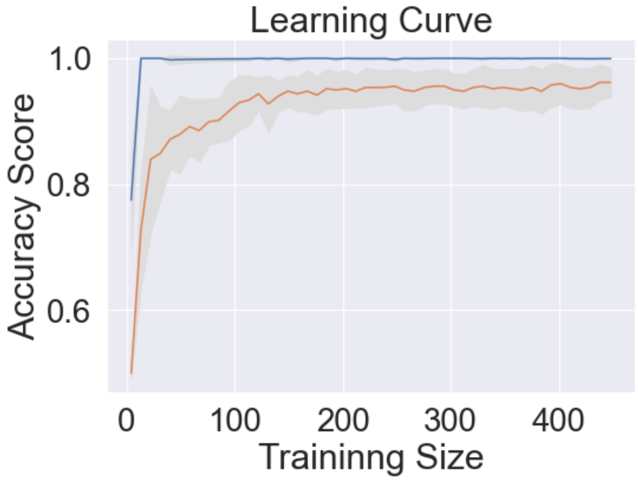
1.Model: Random Forest without Feature selection and Without Hyperparameter Tunning



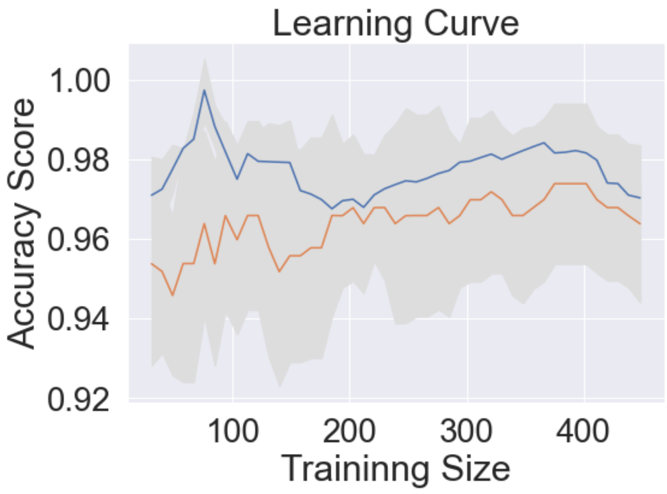
2. Model: Random Forest with Feature selection and With Hyperparameter Tunning



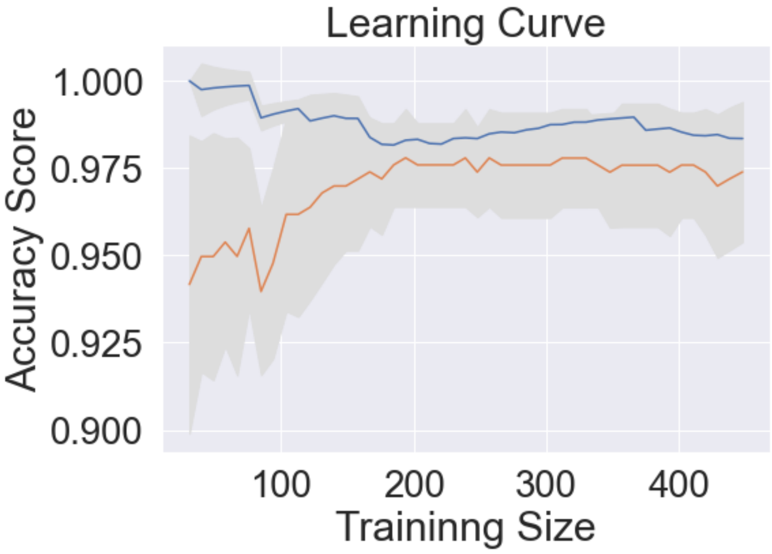
3. Model: Support Vector Classifier with Optimum Feature selection



4. Model: Random Forest on PCA Data



5.Model: Logistic Regression on PCA Data



Q.8 When do you consider adding the “regularization parameter” to the model? And how it will help to improve the model performance?

- This is a form of regression, that constrains/ regularizes or shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model, to avoid the risk of overfitting.

- When there is a high difference between train accuracy and the test accuracy then we can say that the model is overfitted to the train data. We need to add the stabilizing parameter to the model to make it stable. Hence by using the regularizing methods (L1 regularization or L2 regularization) we can make the model more stable. Hence we can tackle the problem of overfitting with the help of regularizing parameters.Hence by adding the regularizing parameters leads to an increase in the overall model performance.

Q.9 Please briefly explain how reinforcement-learning can be utilized in fraud detection models

- The major disadvantage of the traditional process is the occurrence of false positives. This means completely normal customers just looking to make a purchase will go away from your business. A false positive not only affects the sale in the process but also lifetime value generated from the customer. Thus manual reviews based on rules should be the last line of defense in the fraud detection strategy.

- Reinforcement learning differs from supervised learning in not needing labeled input/output pairs to be presented, and in not needing sub-optimal actions to be explicitly corrected. Instead, the focus is on finding a balance between exploration (of uncharted territory) and exploitation

- In fraud detection models we can apply the idea of reward-based hypothesis to predict the output based on reinforcement learning. Reinforcement learning is the study of decision making over time with consequences. With the aim of the agent to maximize the expected reward, we can apply reinforcement learning methodology to solve the problems related to the fraud detection

---

Q.10 Please describe when to use logistic sigmoid, tanh, and Fourier as a basis function.

An activation function is a very important feature of an artificial neural network, they basically decide whether the neuron should be activated or not.In artificial neural networks, the activation function defines the output of that node given an input or set of inputs.

We can use the Sigmoid Logistic function for the following conditions:

- Sigmoid is very popular in classification problems
- The output of the sigmoid function always ranges between 0 and 1 Sigmoid is S-shaped, ‘monotonic’ & ‘differential’ function
- A derivative of the sigmoid function (f’(x)) will lie between 0 and 0.25.
- A derivative of the sigmoid function is not “monotonic”.

- The sigmoid function is not “zero-centric”. This makes the gradient updates go too far in different directions. 0 < output < 1, and it makes optimization h

We can use the Tanh function for the following conditions:

- Tanh is the modified version of sigmoid function. Hence have similar properties of sigmoid function.
- Output is zero “centric
- Optimization is easier
- The function and its derivative both are monotonic
- Derivative /Differential of the Tanh function (f’(x)) will lies between 0 and 1.
- “Tanh is preferred over the sigmoid function since it is zero centered and the gradients are not restricted to move in a certain direction”
- Derivative of Tanh function suffers “Vanishing gradient and exploding gradient problem”.
- Slow convergence- as its computationally heavy.(Reason use of exponential math function)