Dev Joshi

Final Report : Benign or Malignant Cancer Prediction

Executive Summary:

The analysis of the various features of digitized imaging of fine needle aspirate (FNA) of a breast tumor cell mass to distinguish between a benign and a malignant tumor to aid in clinical diagnosis is challenging. We deploy machine learning models to utilize the attributes of the breast tumor cell to distinguish a benign and a malignant cell, resulting in a ~ 75 % increase in revenue over and above a naive model doing the same job.

Problem Statement

The dataset consisting of digitized imaging of fine needle aspirate (FNA) of a breast tumor cell mass can be used to distinguish between a benign and a malignant tumor to aid in clinical diagnosis. Each cell nucleus has ten real-valued features which describe characteristics of the cell nuclei present in the image. The project goal is to deploy machine learning algorithms to accurately distinguish between a benign and a malignant tumor to aid in clinical diagnosis. The original dataset is available at the https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+Diagnostic .

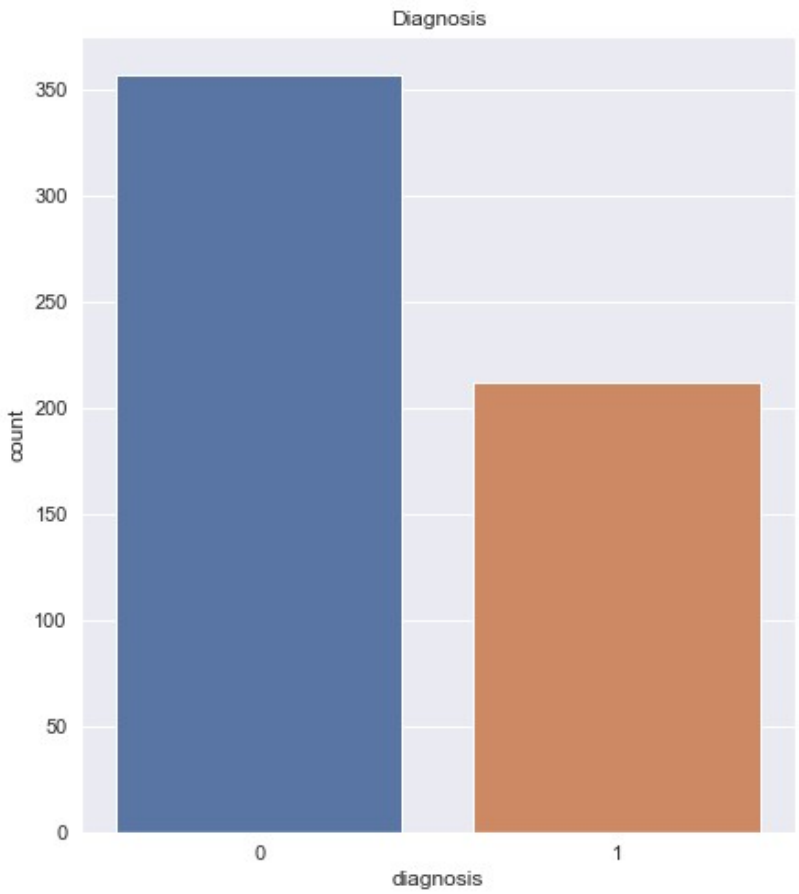
Data Wrangling: We began by downloading the dataset from the above link. The dataset consisted of 569 rows and 33 columns. We removed two columns – 'id' and 'Unnamed: 32' as they weren't essential for the analysis. The various columns of the dataset are :

0 diagnosis
1 radius_mean
2 texture_mean
3 perimeter_mean
4 area_mean
5 smoothness_mean
6 compactness_mean
7 concavity_mean
8 concave points_mean
9 symmetry_mean
10 fractal_dimension_mean
11 radius_se
12 texture_se
13 perimeter_se

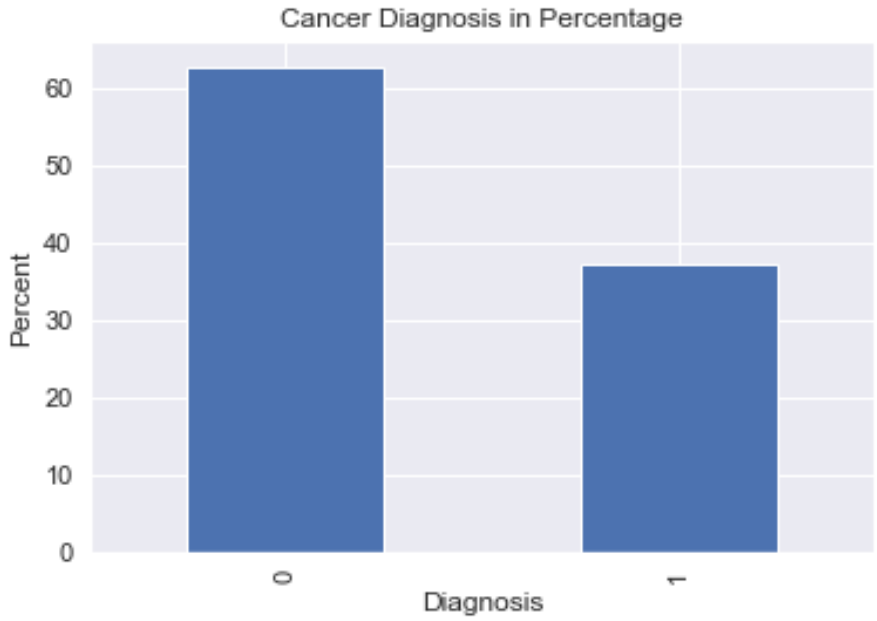
14 area_se
15 smoothness_se
16 compactness_se
17 concavity_se
18 concave points_se
19 symmetry_se
20 fractal_dimension_se
21 radius_worst
22 texture_worst
23 perimeter_worst
24 area_worst
25 smoothness_worst
26 compactness_worst
27 concavity_worst
28 concave points_worst
29 symmetry_worst
30 fractal_dimension_worst

We explored the data and looked at various columns of the dataframe. We made the target variable (column) ‘diagnosis ‘ as numeric by assigning ‘1’ as Malignant and ‘0’ as benign.

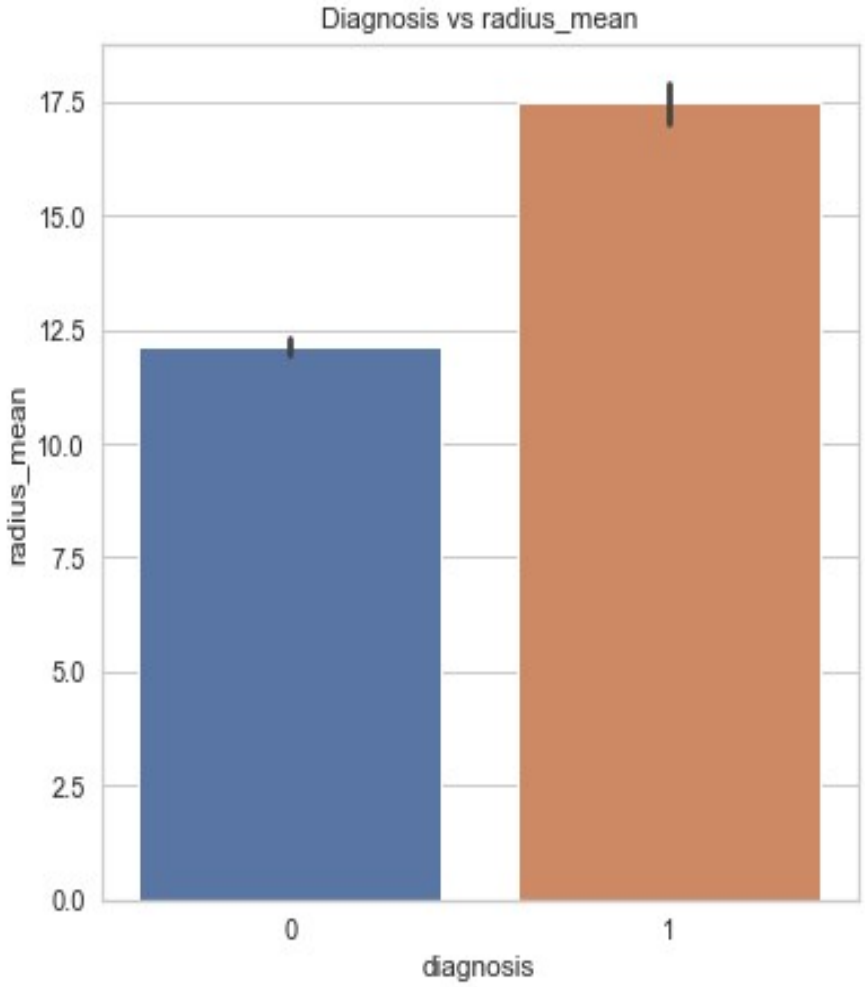
We plotted this column to observe that the number of the benign cases are 357 and the number of malignant cases are 212.



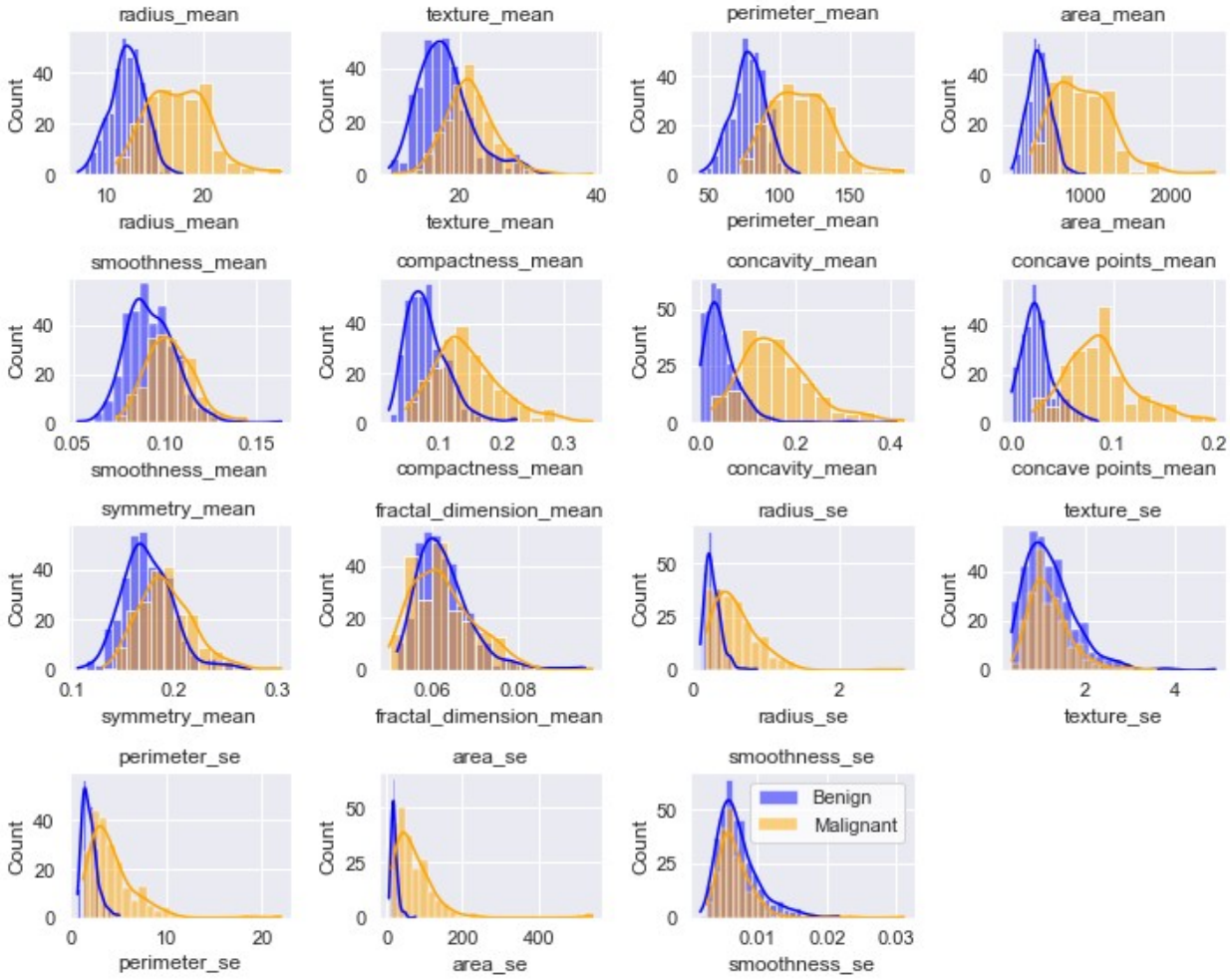
The same plot in terms of percentage yielded 62.7 % and 37.3 % for 0 (benign) and 1(malignant) respectively as seen in the following image.



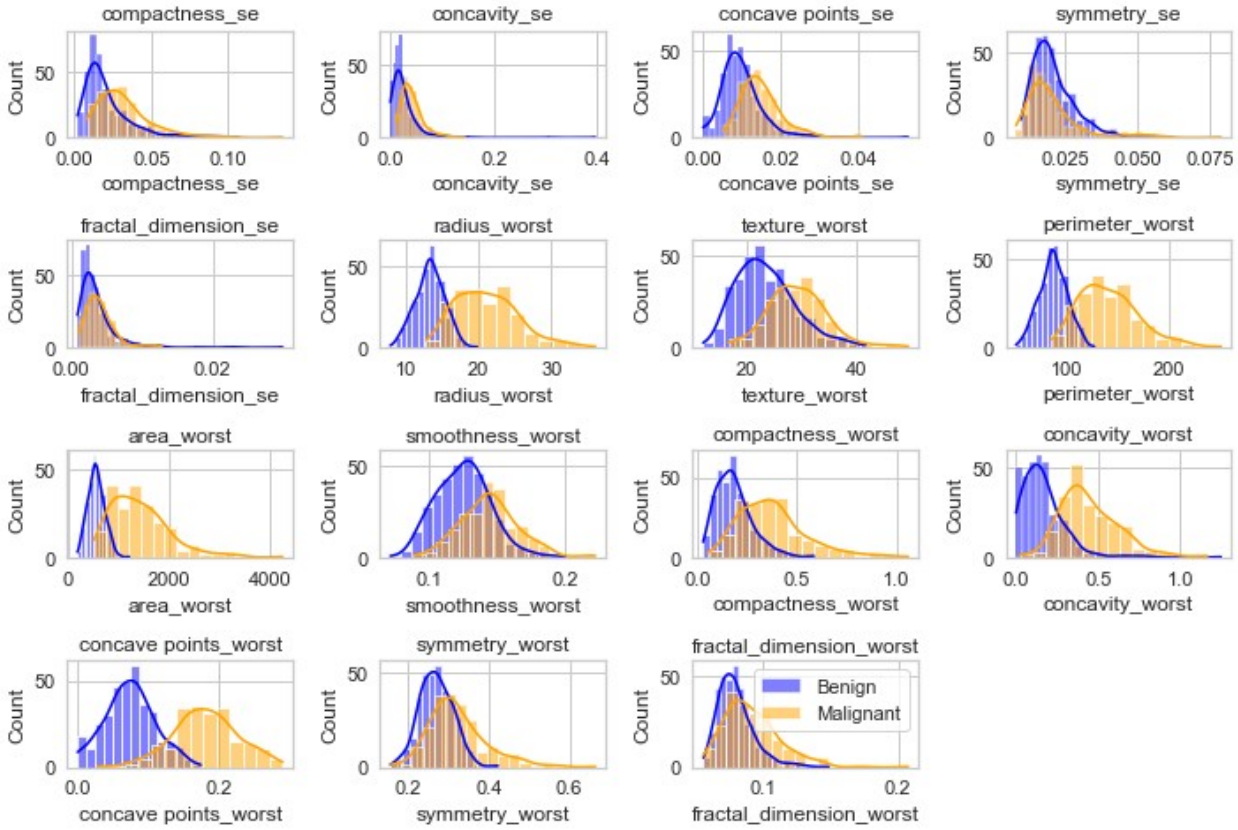
Exploratory Data Analysis: After initial data wrangling, we sought to explore the various features of the data. We looked at the various tumor features in relation to the diagnosis. For example, we see that the malign cells have greater mean radius than benign cell as seen in the adjacent image .



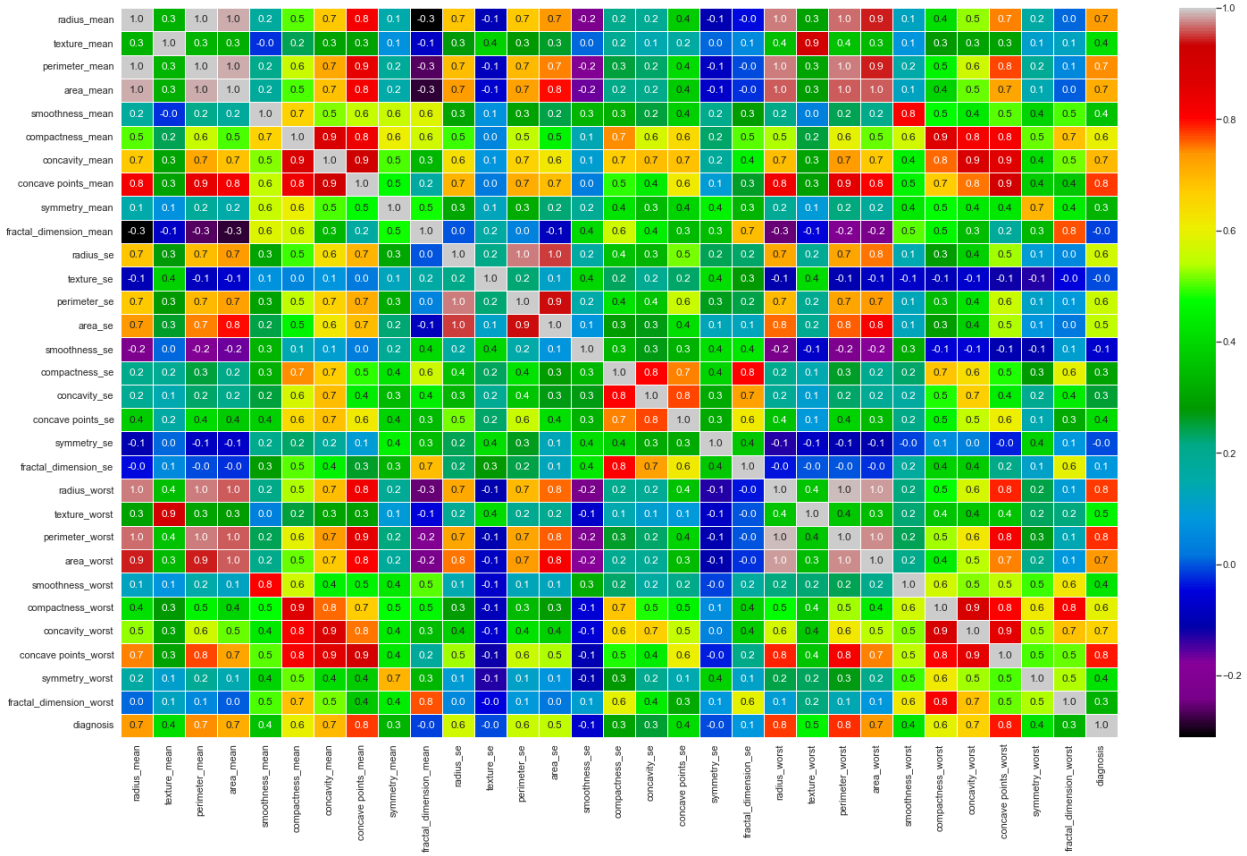
Similarly, the other features in the data can also be compared as seen in the image below. From this image also, we see that the mean radius is greater for the malign cells as compared to the benign cells (1st subplot). For most of the features (except few – namely – symmetry_mean, fractal_dimension_mean, texture_se, smoothness_se), the feature corresponding to the malign cells has higher value as compared to that of the benign cells.



We plotted similar image for the remaining features as well. Here also, we see the malign cells have higher value, although lesser count, than the benign cells for most of the features.


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
Count
Benign
Malignant

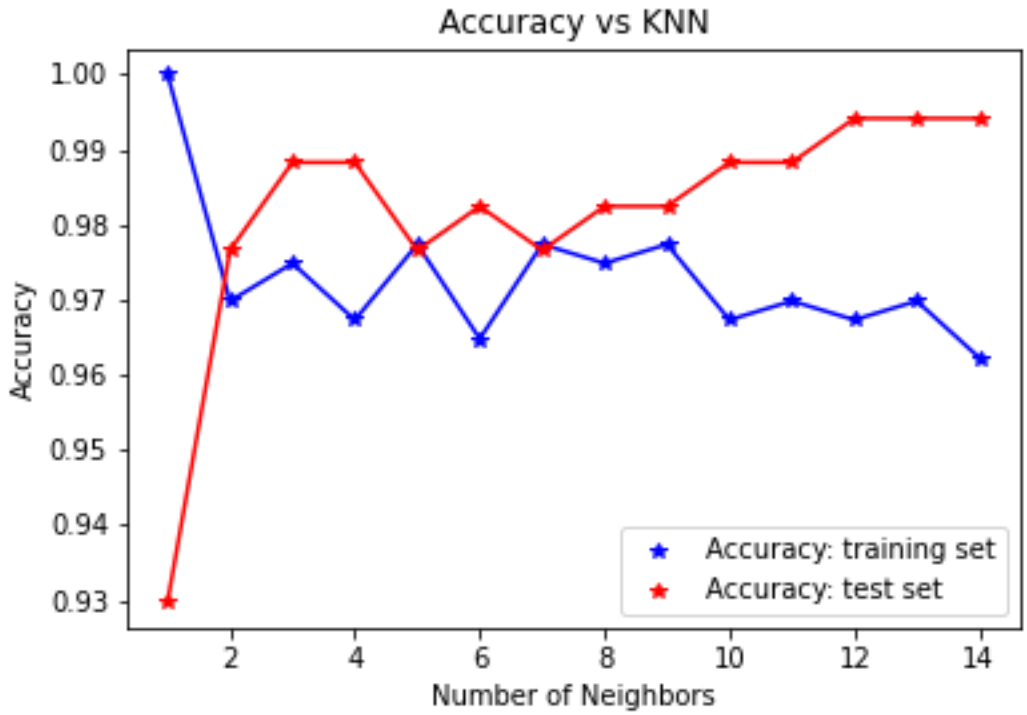
We also checked the correlation map with respect to the diagnosis feature. As we can see in the image below, it has a low correlation - but positive - with most of the features. It has negative correlation with few of the features. Lastly, we saved the data for further analysis in the subsequent units.



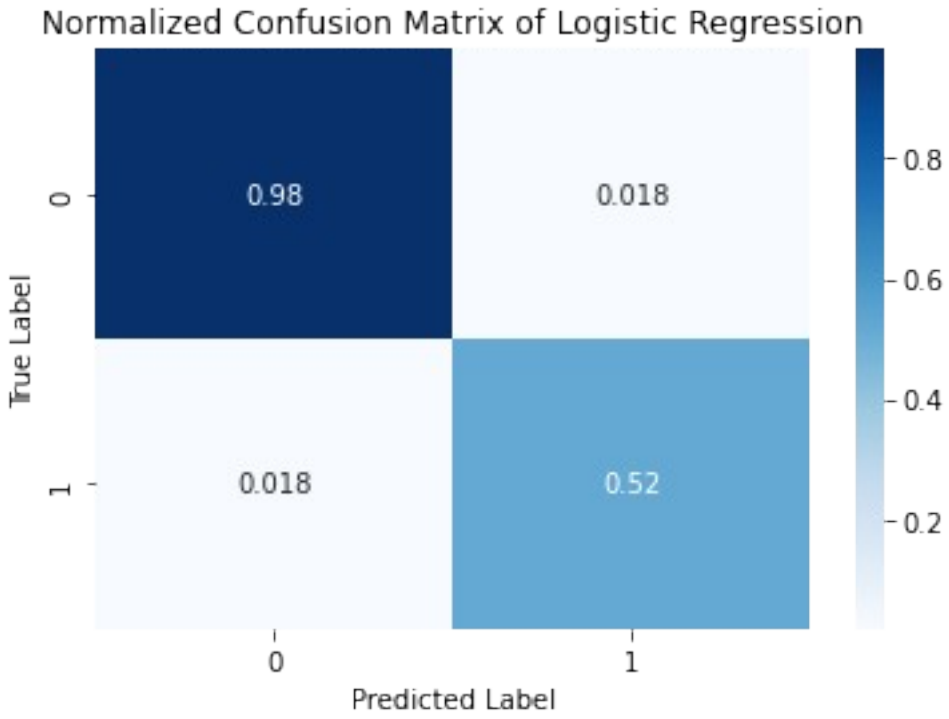
Preprocessing, Training and Modeling:

We downloaded the saved data from the previous section. We then divided the data into those that will be used to train the model and those that will be used to predict the approval : 70 % for training and 30 % for testing. We also applied standard scaling for X_train and X_test data so that no feature (with larger values) dominates over others (with smaller values).

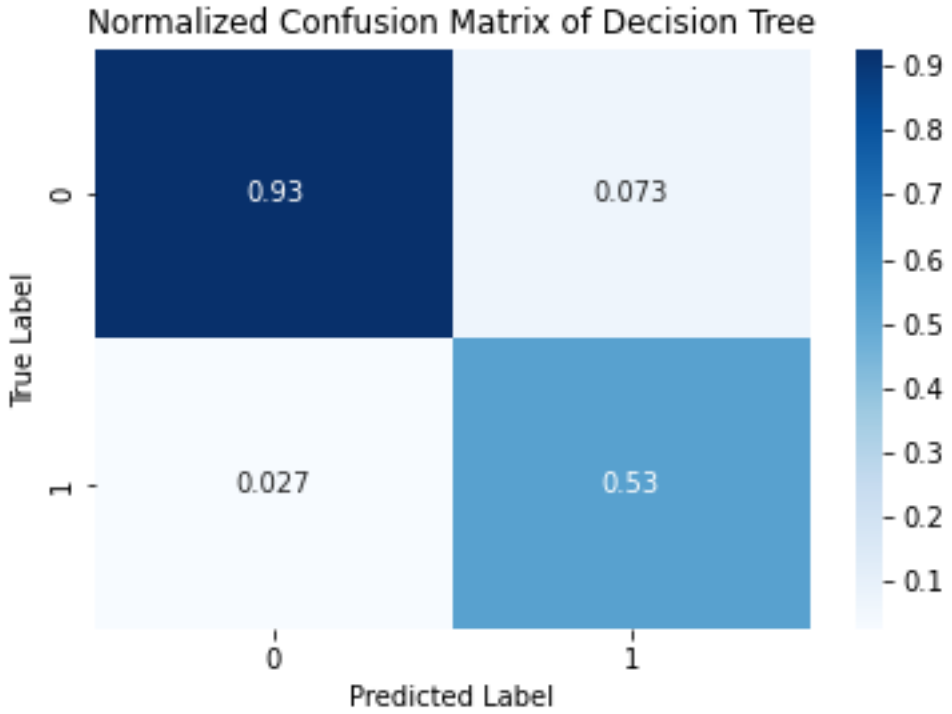
We applied the K-nearest neighbor model in the splitted training data and calculated the accuracy for both training and testing data by varying the number of neighbors from 1 to 15. We find the maximum accuracy occurs for both the training and the test set when the number of neighbors is 5 as seen in the image below.



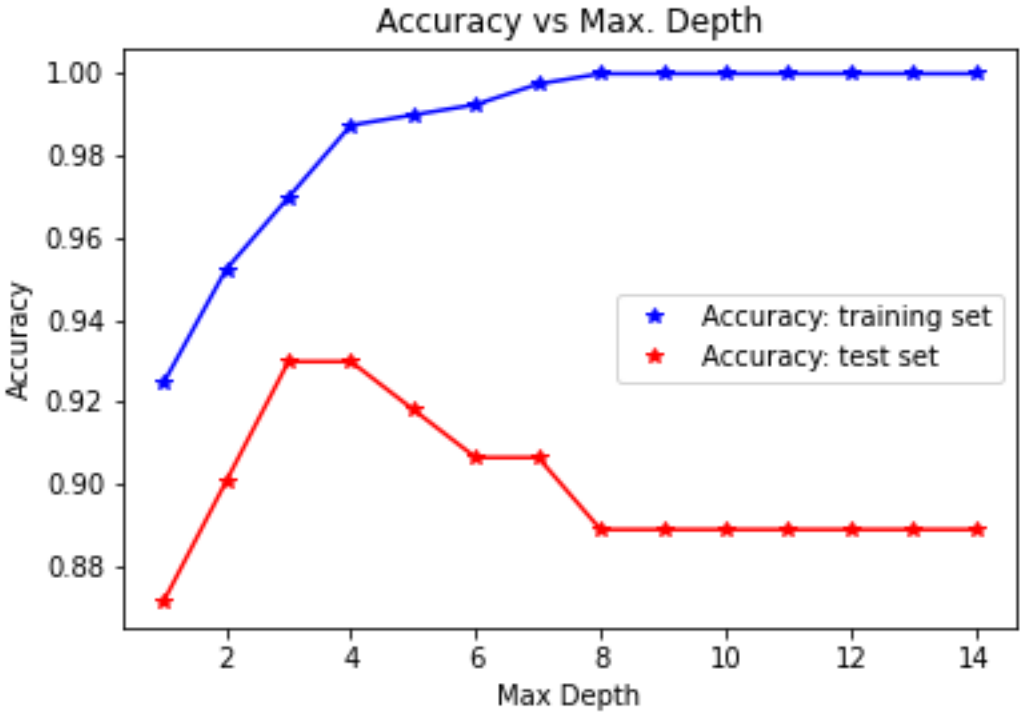
We then applied Logistic Regression model and plotted the confusion matrix. The confusion matrix summarizes the performance of a machine learning model on a set of test data. The plot displays True Negatives, False Positives (upper row) and false negatives , True Positives (lower row). To recall, 1 is malign cell and 0 is benign cell. The accuracy score from this confusion matrix is 0.96. There are other metrics such as Precision , Recall and F1-score which could be calculated from the confusion matrix.



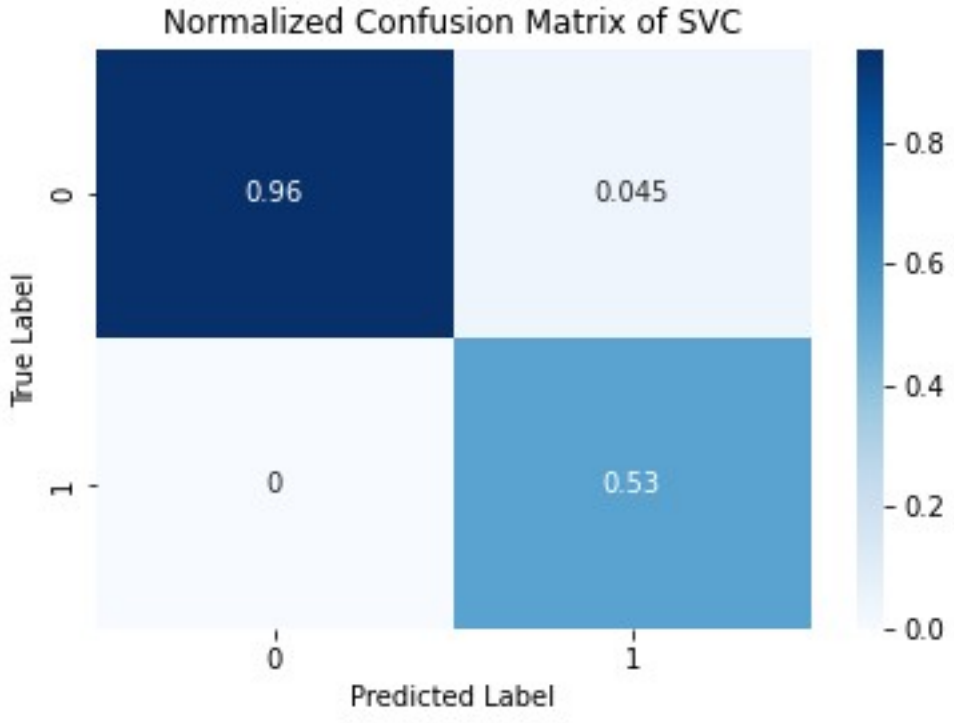
We then moved to Decision tree modeling and repeated the same process as above to plot confusion matrix for this model. We fit the model on the training set of the data and use the model to make prediction using the test data. We use this predicted data and true test data for the target variable to build the confusion matrix. The accuracy for this model was found to be ~ 0.89.



Also, we made a plot to check the variation of the accuracy with respect to the maximum depth used in the Decision Tree model. We found the accuracy for both the training and the test sets to be highest when the maximum depth is 4.



Finally, we repeated the process of plotting the confusion matrix with the support vector machine model. The confusion matrix of this model is included below. We get the accuracy of this model to be 0.97.



Based on the accuracy, the support vector machine model performed best in our exercise.

We can look at other metrics as well such as Precision , Recall and F1-score to make a detailed comparison between the models. In that case we ought to compare between precision and recall. If the precision of a model is high, it suggests that the classifier is returning accurate results (high true positives), but at the cost of missing a number of actual positives (low recall). In another case, if the recall is high, it means the classifier is returning most of the positive results, but with a number of false positives. The trade-off between precision and recall will also depend upon the threshold chosen for the classification in the model. The precision-recall trade-off is a very challenging problem for data scientists to solve when working with imbalanced data and in some cases, precision should be prioritized while in other cases recall should be prioritized, there is no universal right or wrong answer. This can be future work in this investigation.

Model	Precision	Recall	F-score	Accuracy
Decision Tree	0.96	0.78	0.86	0.91
Logistic Regression	0.82	1.00	0.90	0.92
SVC	0.98	1.00	0.99	0.99
'Naive'	0.75	0.86	0.80	0.82

COST ANALYSIS

Finally, we performed an analysis to understand what benefits, say , a diagnosing clinic may achieve by employing the above-mentioned models in comparison to without using these. For comparison, we chose a 'naive model' based upon which the clinic would decide the malignancy of the tumor cell. We assumed a cell is malignant if the mean radius is greater than the mean of the mean radius of all the cells in the data. This helped us to build a confusion matrix. The confusion matrix has four categories as mentioned earlier : True positive (Correct Alarm), True Negative, False Positive (False alarm), False Negative . We assigned monetary value to each of these categories : cost of False Positive as - $ 1,000, revenue of True-positive as + $5,000 and opportunity cost of False Negative - $ 5000. A false positive (say, cost to identify it as a false alarm) of identification of a malignant cell is assigned as - $1000. The opportunity cost of a False negative identification (a potentially malignant cell predicted as benign and hence, loss to the

clinic) is - $ 5000. Based upon these assumptions, the total revenue gain with the naive model was +$185,000. The total revenue gain with the Logistic Regression model was $ 253,000. Also, with the Decision Tree model, the total revenue gain was $232,000. With the Support Vector Model, the total revenue gain was $255,000. The threshold of the models could be further changed to see how the cost analysis 'improves' upon the naive model.