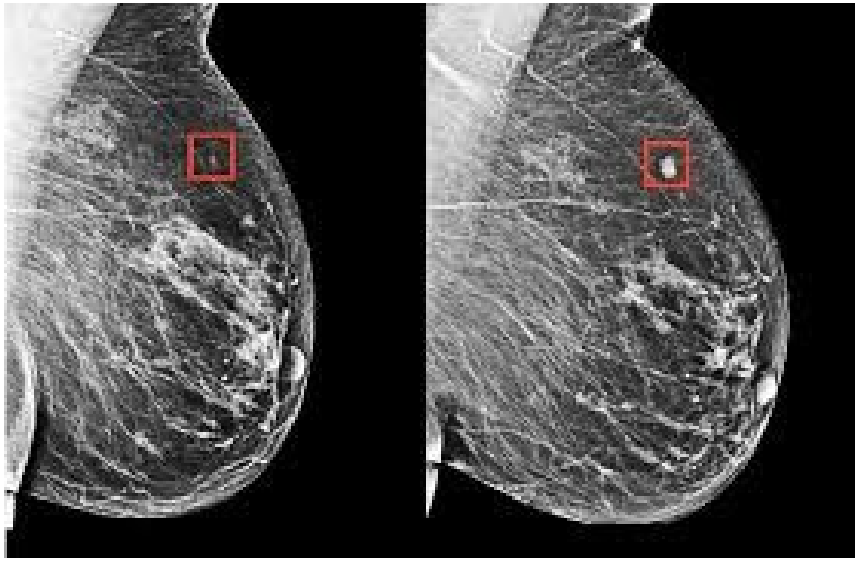
Capstone 3 Project: Breast Cancer Prediction

Dev Joshi[1]
(1) Spring Board Bootcamp

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 malignant tumor to aid in clinical diagnosis.



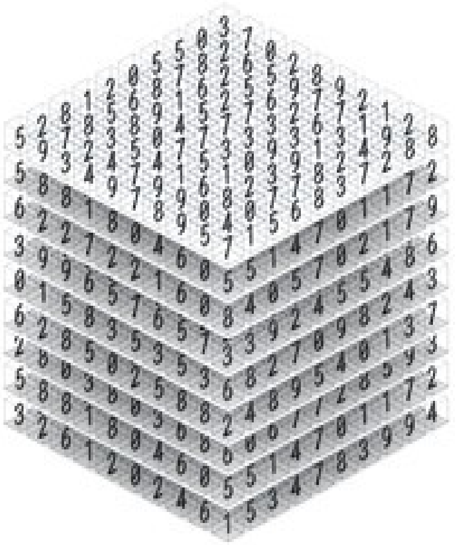
The dataset source is :

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+Diagnostic

- The project goal is to deploy machine learning algorithms to accurately distinguish between a benign and malignant tumor.

Dataset

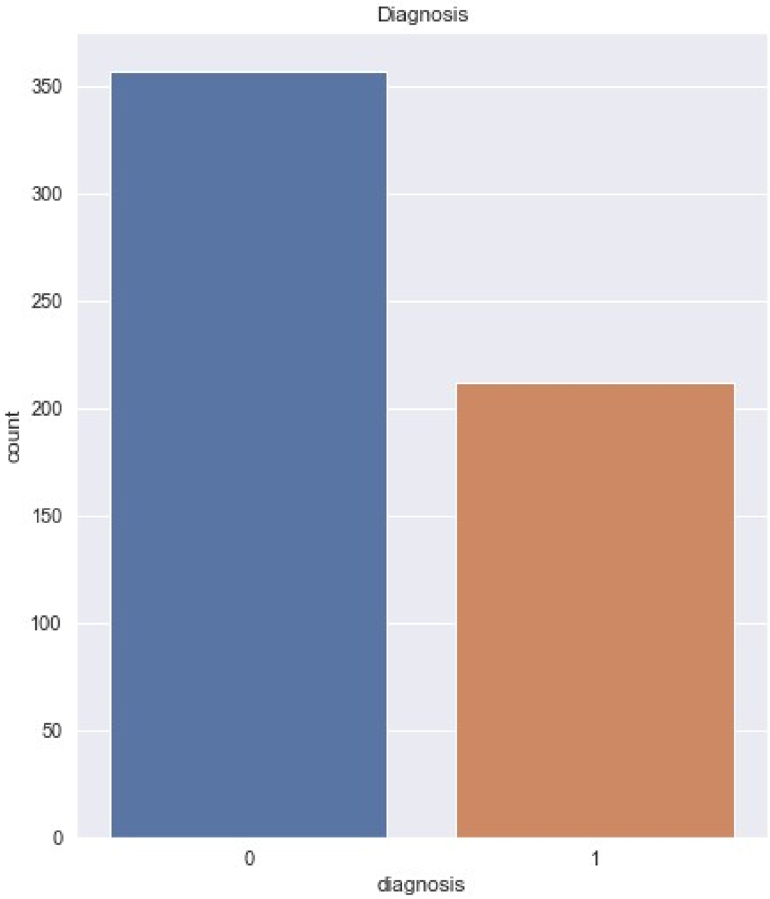
- The dataset consists of 569 rows and 33 columns. We removed two columns – ‘id’ and ‘Unnamed: 32’ as they weren’t essential for the analysis.
- We explored the data and looked at various columns of the dataframe. We made the target variable (column) ‘diagnosis’ as numeric by assigning ‘1’ as Malignant and ‘0’ as benign.

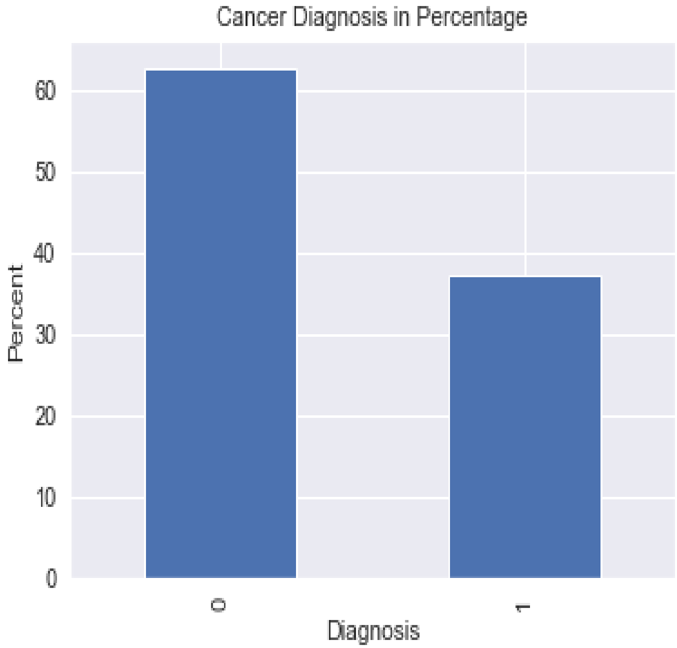


Columns in the dataset

- 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
- 10fractal_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
- 30fractal_dimension_worst

Diagnosis

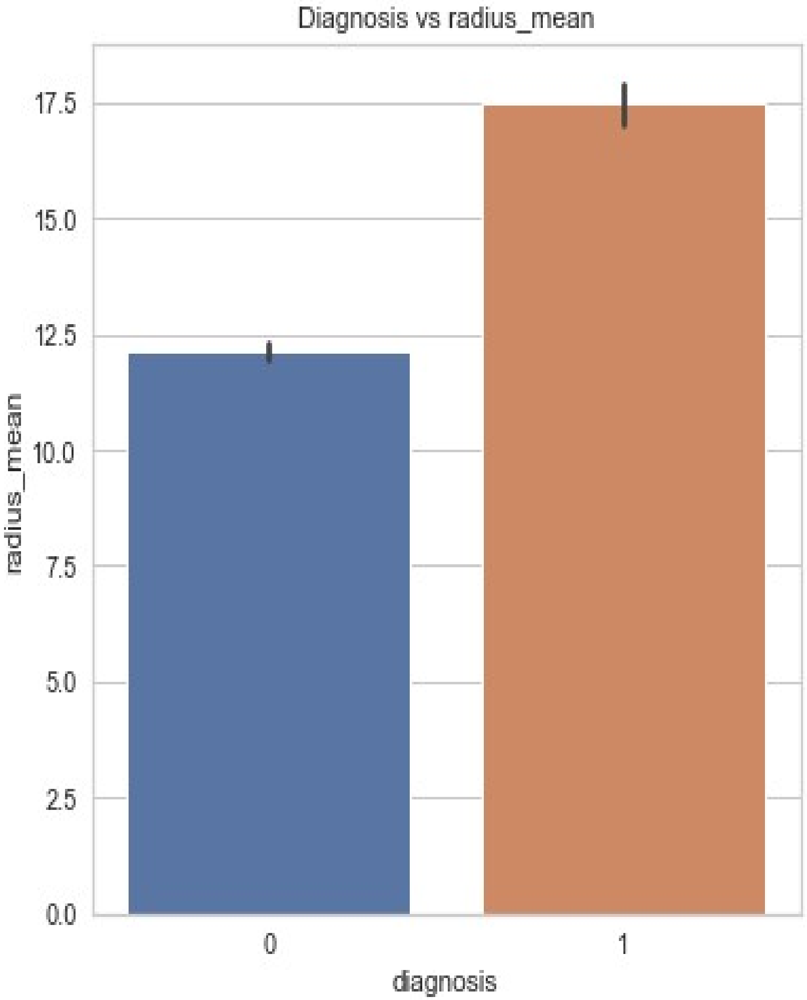




- We plotted the diagnosis column in the dataset to observe that the number of the benign cases are 357 and the number of malignant cases are 212 (image in the left).

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

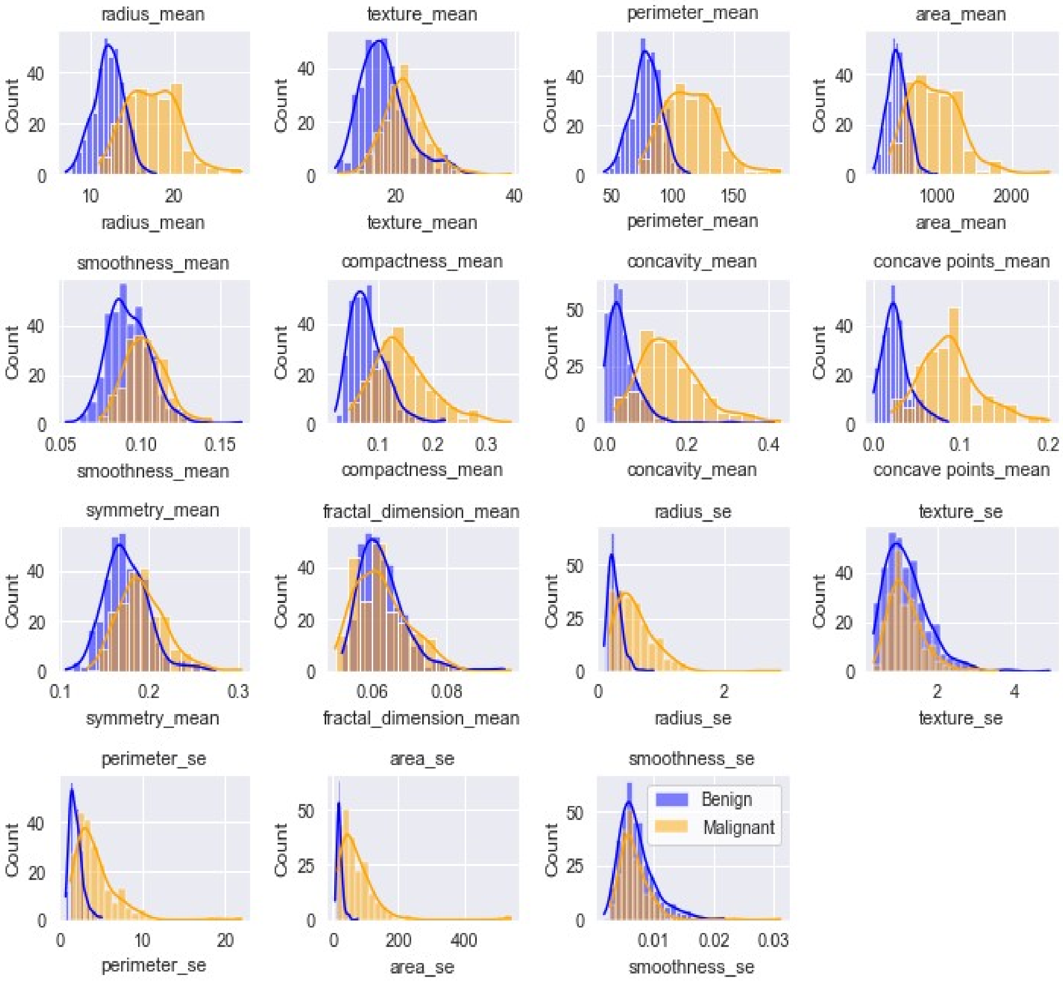
Diagnosis



- 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 cells as seen in the adjacent image .

Features

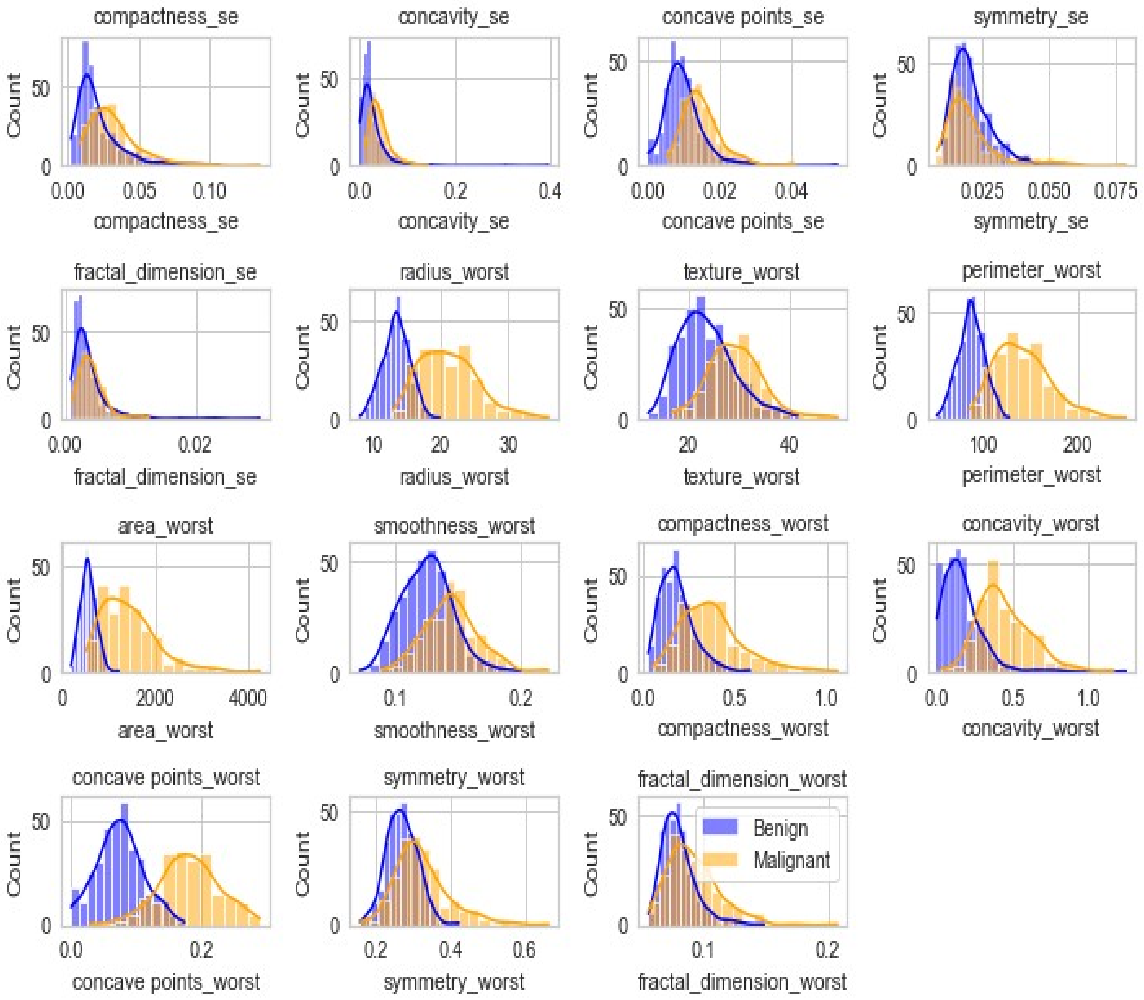
- Similarly, the other features in the data can also be compared as seen in the adjacent image.
- We see that the mean radius is greater for the malign cells as compared to the benign cells (1st subplot).
- For most of the features (except few – namely – symmetry_mean, fractal_dimension_mean, texture_se,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



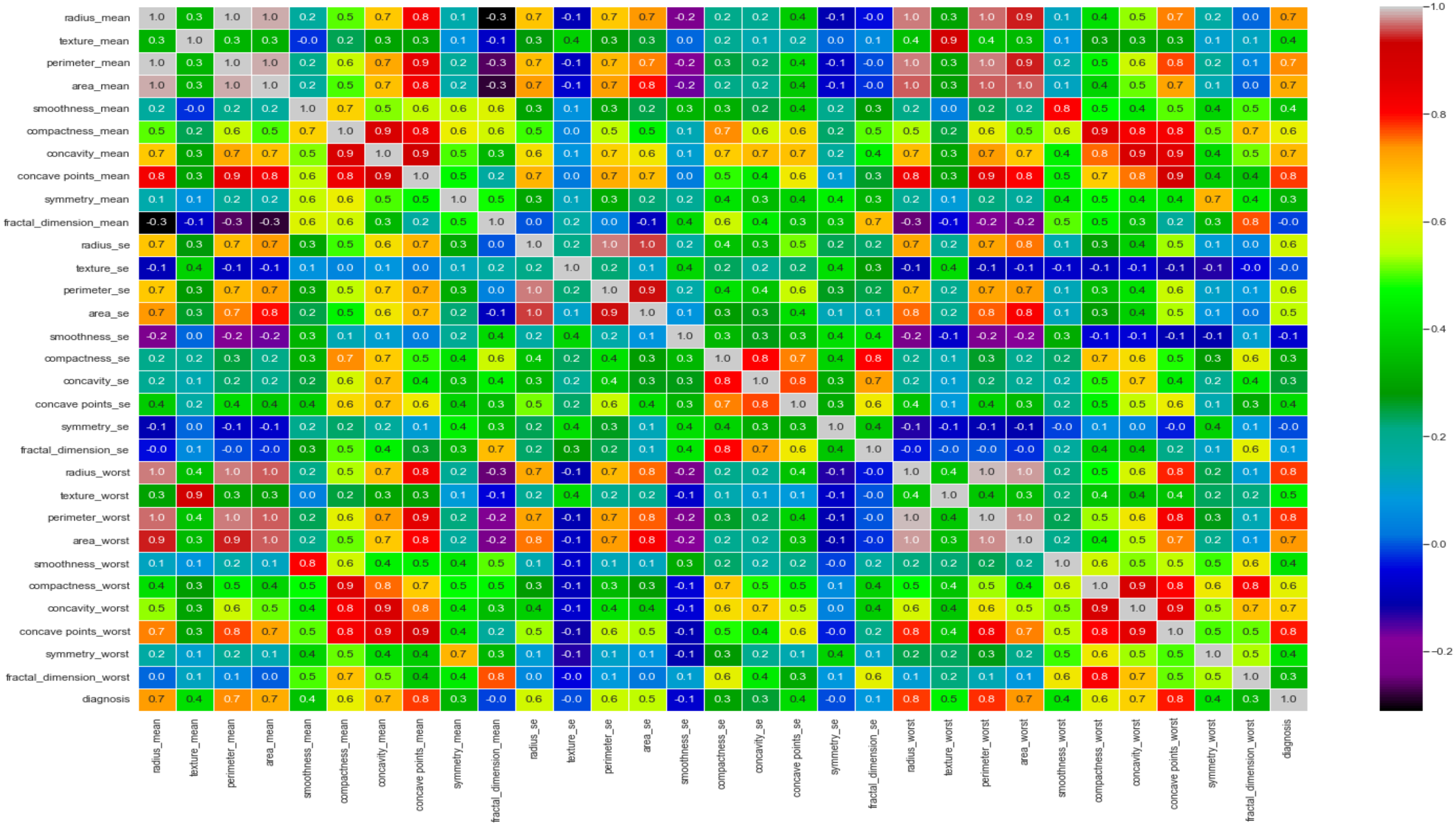
Features

- 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.

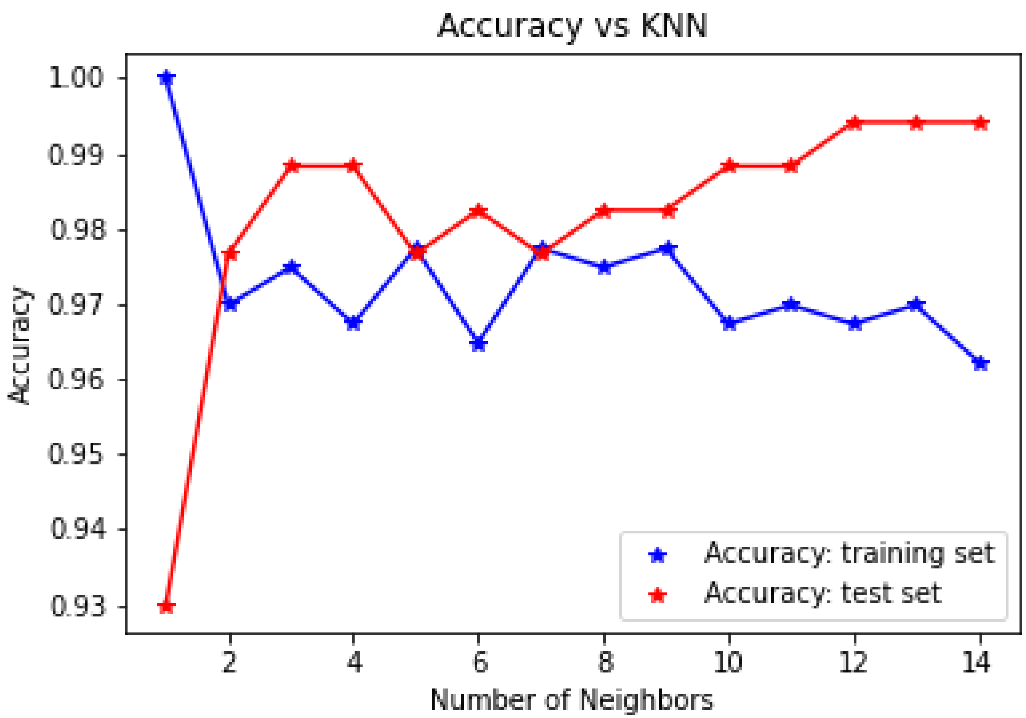


Correlation Map



- We checked the correlation map with respect to the diagnosis.
- As we can see in the adjacent image , it has a low correlation - but positive - with most of the features. It has negative correlation with few of the features.

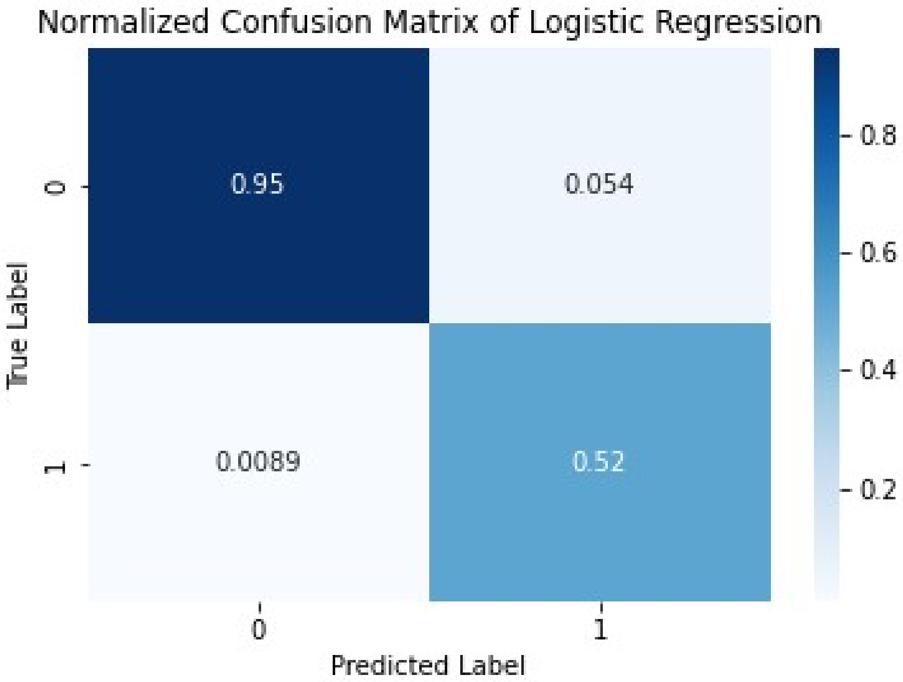
Training and Modeling



- We 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 above.

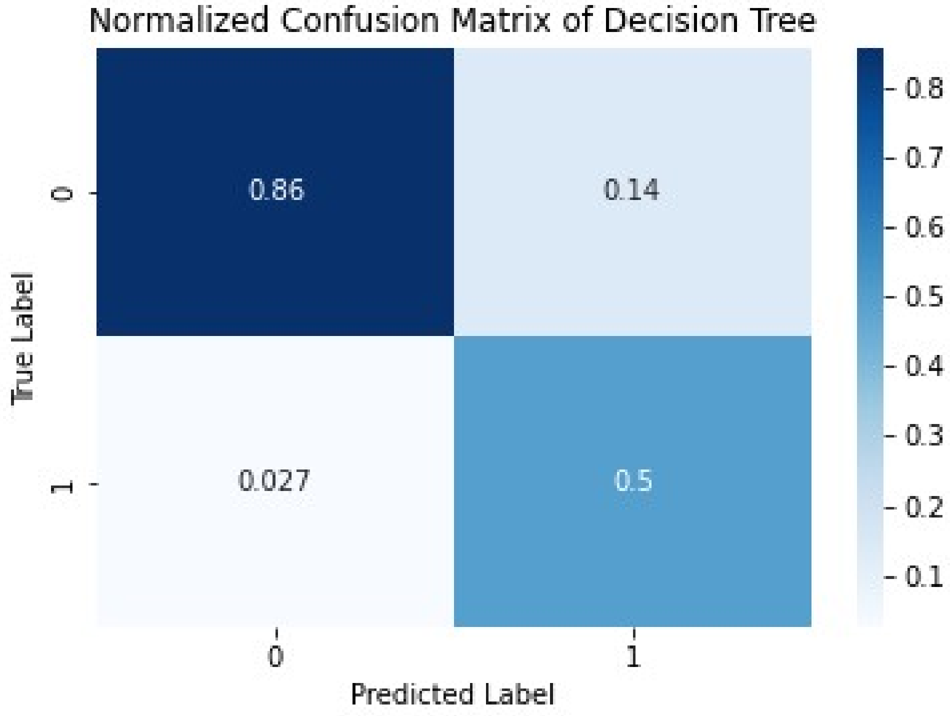
Confusion Matrix



- 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 0.92.

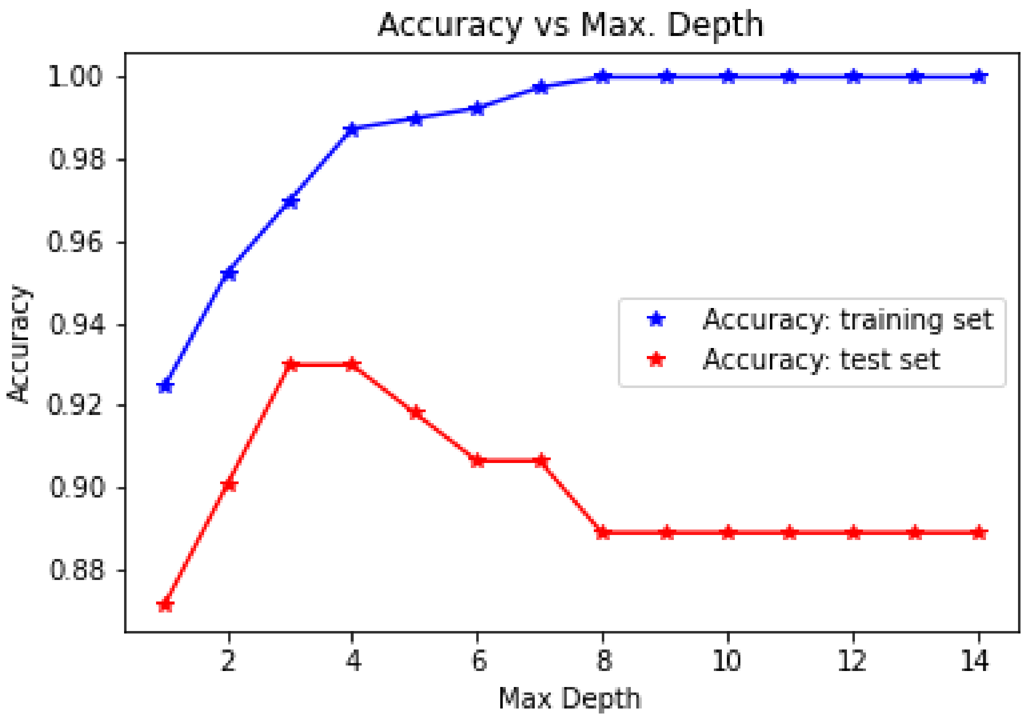
Confusion Matrix



- We applied Decision tree modeling to build confusion matrix for this model. We fit the model on the training set of the data and employed the model to make prediction using the test data. We used this predicted data and true test data for the target variable to build the confusion matrix.

- The accuracy score from this confusion matrix is 0.91.

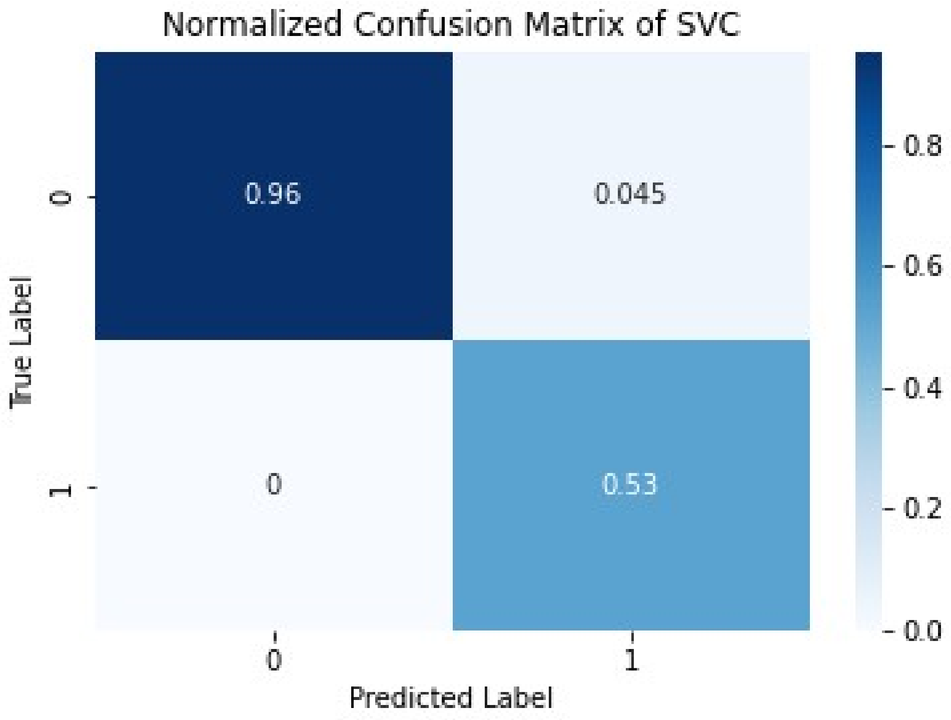
Decision Tree : Max. Depth



- We checked 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.

Confusion Matrix



- We employed the support vector machine model to build the confusion matrix.
- The accuracy score from this confusion matrix is 0.99.

Performance Metrics:Table

Model	Precision	Recall	F-score	Accuracy	Revenue
Decision Tree	0.96	0.78	0.86	0.91	+$232,000.00
Logistic Regression	0.82	1.00	0.90	0.92	+$253,000.00
Support Vector Classifier	0.98	1.00	0.99	0.99	+$255,000.00
‘Naive’	0.75	0.86	0.80	0.82	+$ 185,000.00

- The SVC model has the highest accuracy.
- The cost-analysis shows that SVC model would make the highest revenue gain as compared to a naive model.

Conclusions

- We built a machine learning-based classifier that if a predicts if a breast tumor is benign or malign to aid in clinical diagnosis, based on the information provided.
- While building this cancer cell predictor, we learned about common preprocessing steps such as label encoding, and handling outliers.
- We implemented three different machine learning models, and evaluated the performance using the accuracy score.
- Based on the accuracy score, we found the Support Vector Machine model to be most accurate.
- We have used python's machine learning libraries to implement machine learning algorithms. In the future, we can investigate to estimate the tangible benefits of the predictions of these machine learning models.

Thank You !