BA 890 Research Project Proposal Mekhal Raj Rajendran

Classifying Malignant Cancer cells from Breast Cancer data.

Initial Proposal

Breast cancer is the most common type of cancer amongst women in the world. It accounts for 25% of all cancer cases, and affected over 2.1 Million people in 2015 alone. It starts when cells in the breast begin to grow out of control. These cells usually form tumors that can be seen via X-ray or felt as lumps in the breast area.

The key challenge against its detection is how to classify tumors into malignant (cancerous) or benign(non cancerous). In this project, I am attempting to analyze the Breast Cancer Wisconsin (Diagnostic) Dataset and classify these tumors using machine learning.

I aim to produce a code notebook that will contain data cleaning, the machine learning methods used to classify the tumors and whether I was successful in predicting malignant and benign tumors. At the end of this project, I expect to classify breast cancer into malignant and benign tumors accurately.

Objective

Understand the Dataset & cleanup (if required).

Build classification models to predict whether the cancer type is Malignant.

Data

Source: Breast Cancer Wisconsin (Diagnostic) Dataset (Kaggle)

Important Features:

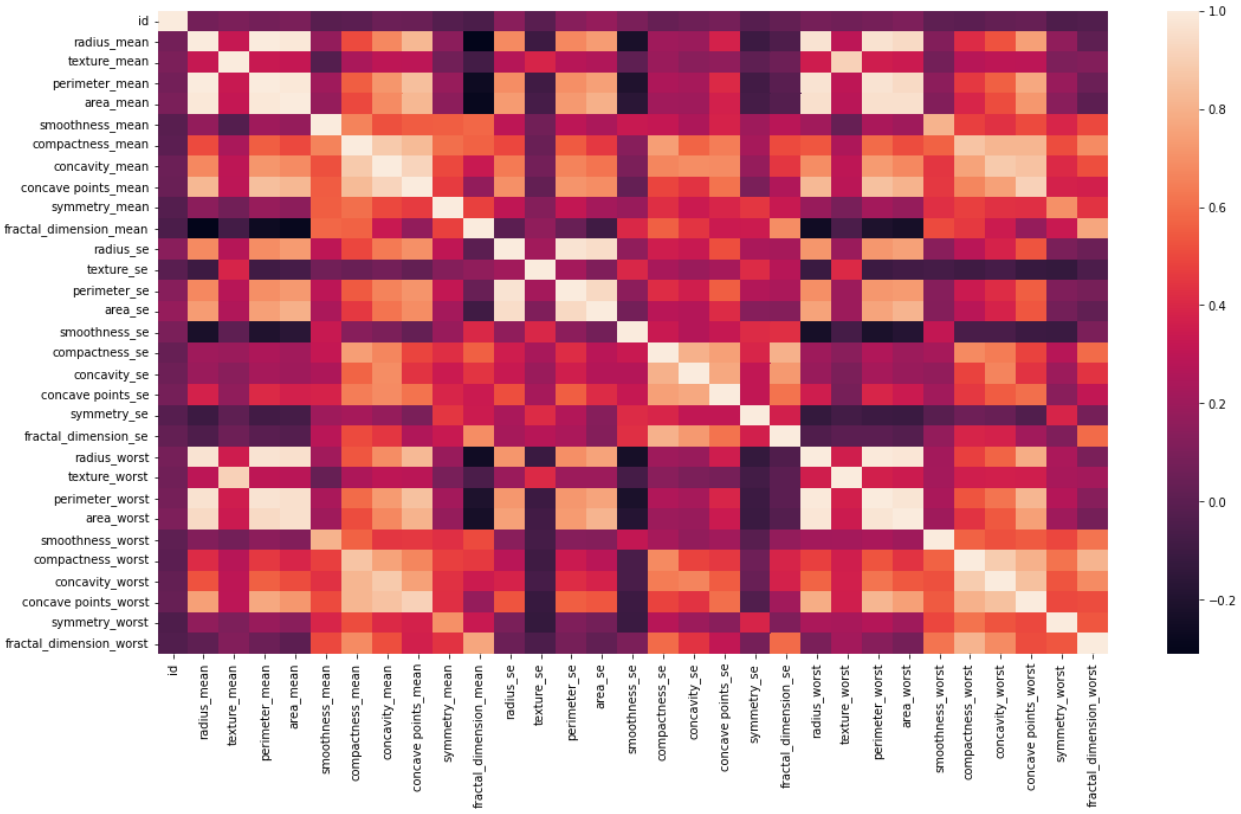
- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter2 / area−1)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension (coastline_approximation−1)

The Dataset was quite small with around 569 samples and 32 columns.

Data Cleanup

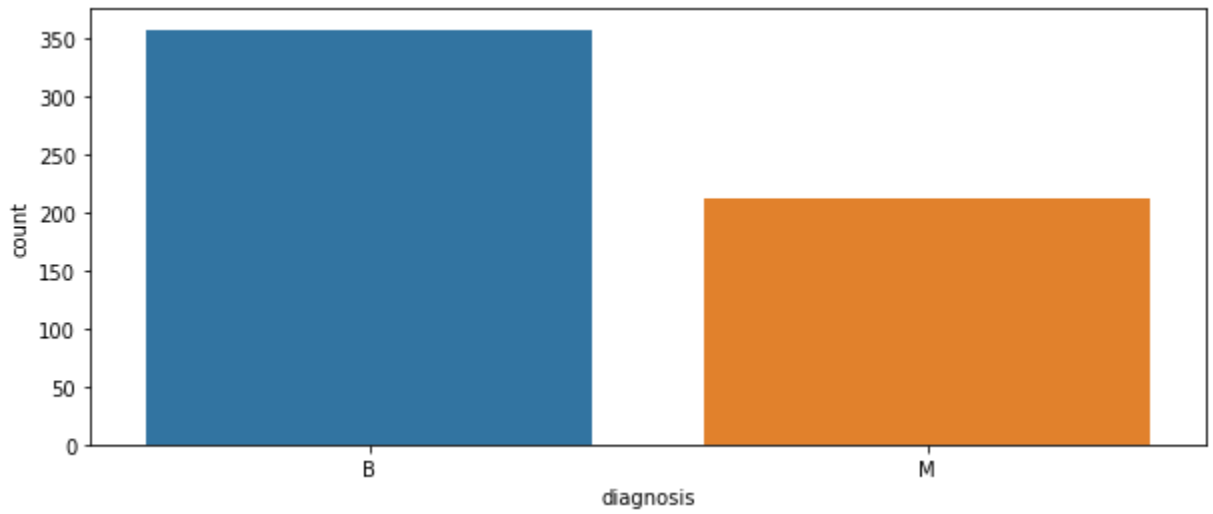
There are no null values in the dataset. Since we are trying to predict the categorical variable diagnosis, we will change its type from an object to category.

Let’s check the correlation between the variables to see which ones to drop:



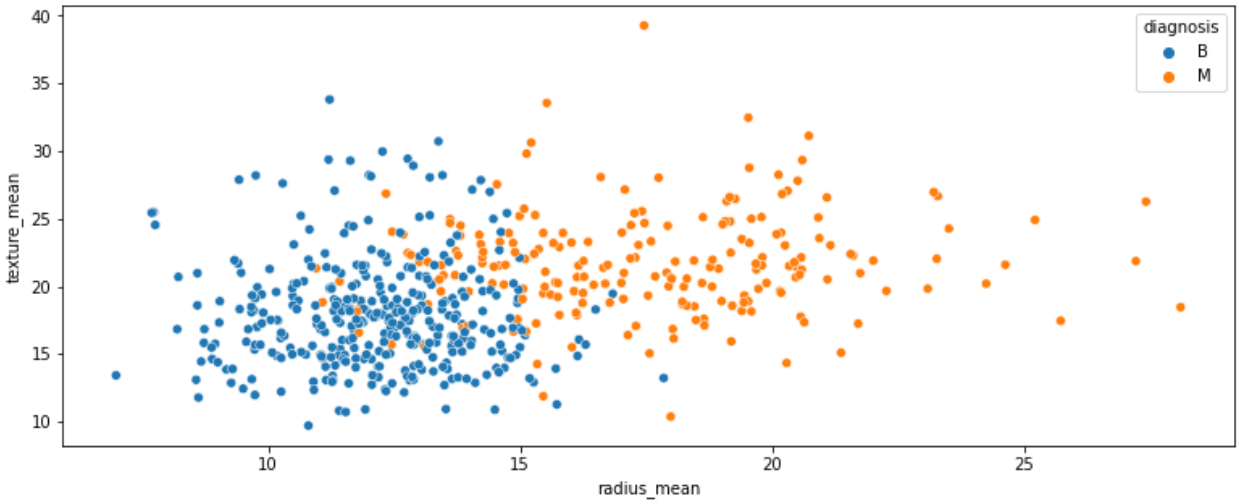
Exploratory Data Analysis

1. Let's take a look at the distribution of Benign and Malignant cases:



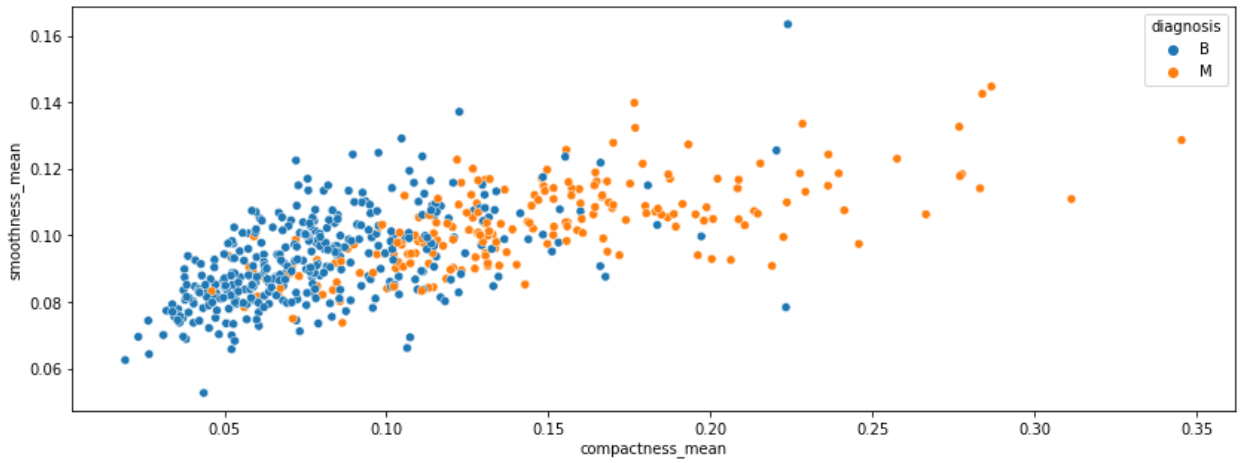
There are 357 Benign cases and 212 Malignant cases.

2. Radius_mean and texture mean



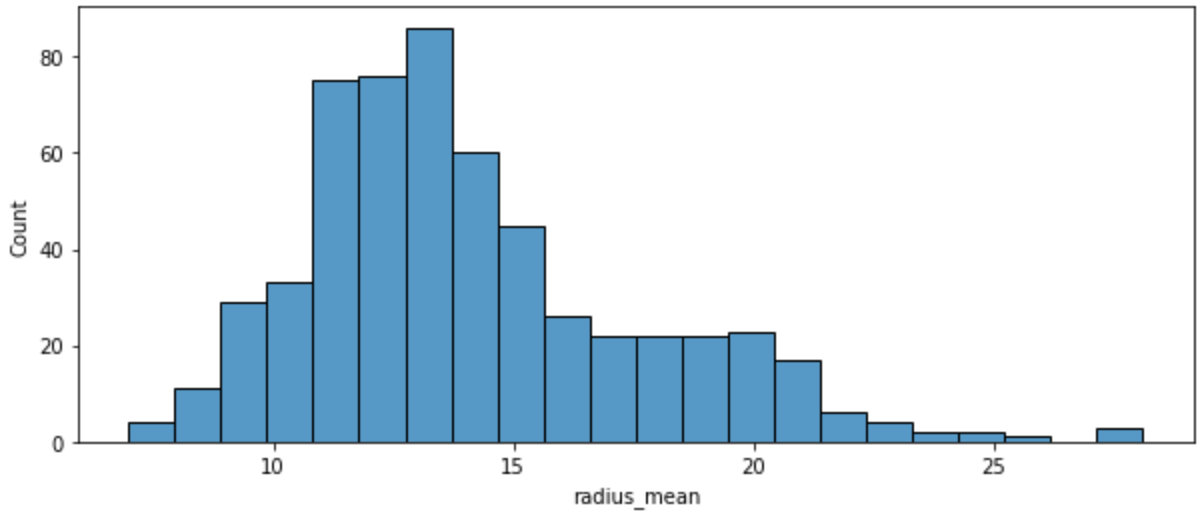
Radius_mean and texture mean for malignant tumors is higher than for benign tumors.

3. compactness_mean and smoothness_mean



compactness_mean and smoothness_mean both are higher for Benign tumors.

4. Size of the tumors:



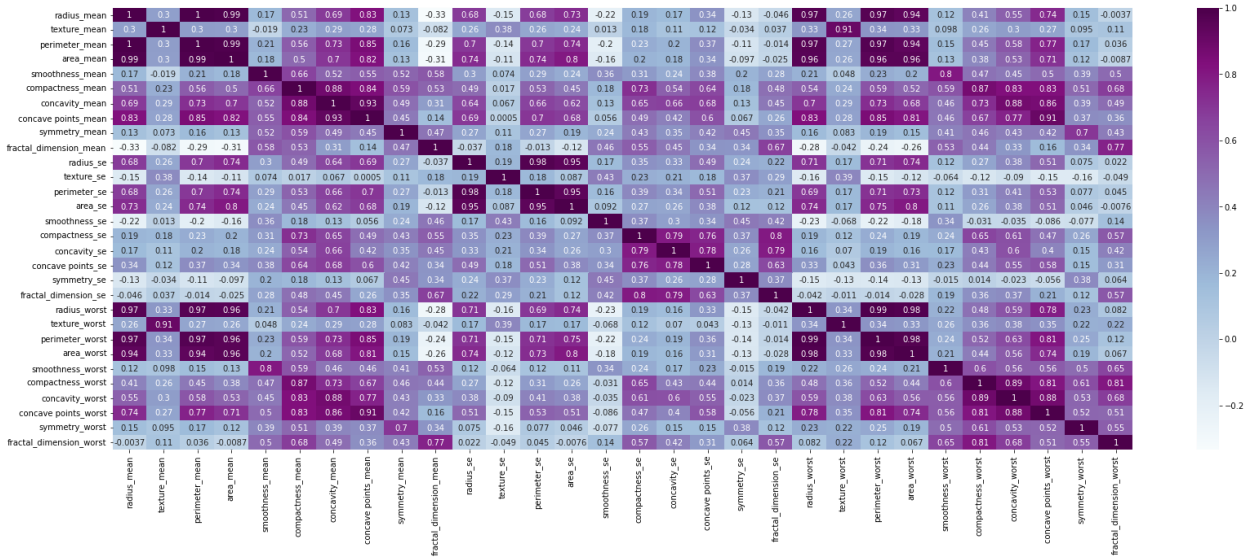
Most of the tumors lie in the size range of 12 - 15.

Creating Dummy Variables

Next, I created dummy variables for the *diagnosis* column and then dropped the *id* and *diagnosis_B* columns since we're only predicting Malignant cases.

Feature selection

I checked the correlation of different features in the dataset (the image below) and then kept only the highly correlated features in the dataset.



Dropped Features: 'area_mean', 'area_se', 'area_worst', 'compactness_worst', 'concave points_mean', 'concave points_worst', 'concavity_mean', 'concavity_worst', 'fractal_dimension_worst', 'perimeter_mean', 'perimeter_se', 'perimeter_worst', 'radius_worst', 'smoothness_worst', 'texture_worst'.

Applying Different Models

I then scaled the data, split it into training and test sets and ran different machine learning models on it.

1. Logistic Regression

Accuracy score: 0.9415204678362573

2. K-Nearest Neighbors

Accuracy score: 0.9239766081871345

3. Gaussian Naive Bayes

Accuracy score : 0.9122807017543859

4. Decision Tree

Accuracy score: 0.8830409356725146

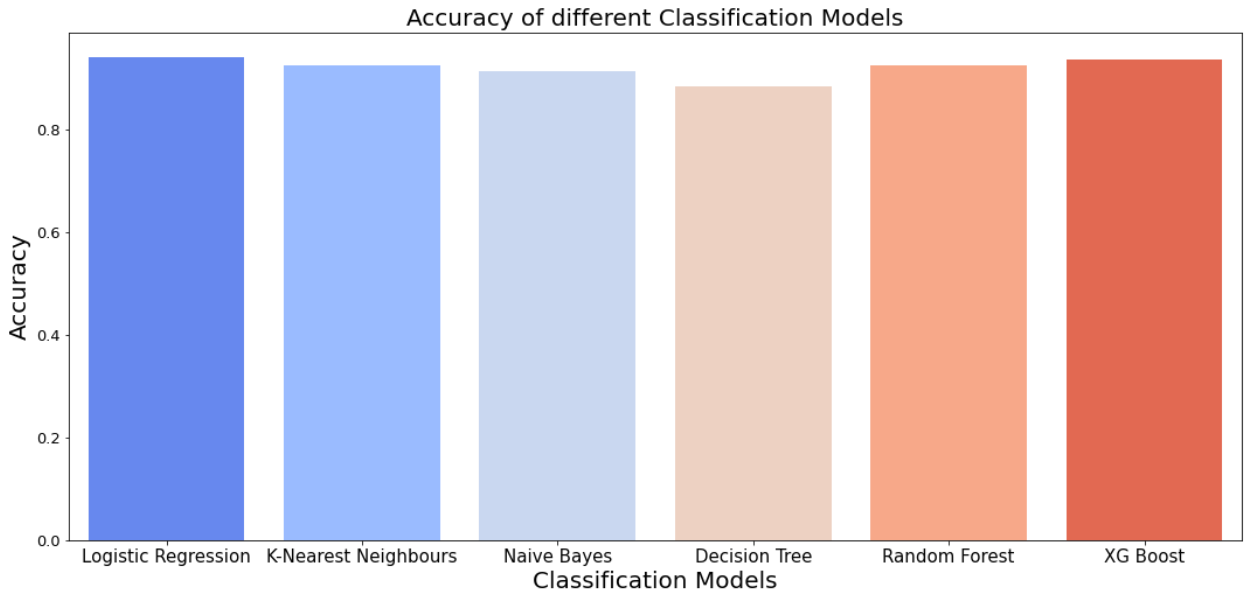
5. XGBoost Classifier

Accuracy Score: 0.935672514619883

6. Random Forest

Accuracy Score: 0.9239766081871345

Comparison of Accuracy Scores



Among all our models, Logistic regression performed the best with an accuracy of 94.1% followed by XGBoost with an accuracy of 93.5%.

Project Outcomes & Conclusions

- Visualizing the distribution of data & their relationships, helped us to get some insights on the relationship between the feature-set.
- Feature Selection was carried out and appropriate features were shortlisted.
- The Logistic Regression, Random Forest Classifier & XG-Boosting performed exceptionally well on the current dataset.
- The models were able to successfully predict Malignant cells with an accuracy 94.1%.

Limitations and Future Steps

- The Dataset was quite small totalling around 569 samples.
- Other parameters like recall can be used to measure the efficiency of the models.
- Neural Networks and related models could be used if we have access to a bigger dataset or the original scan image data.