BREAST CANCER ANALYSIS

Team - *Data Alchemists* | Advanced Data Science | INFO 7390 | Spring 2024

Objective

Breast cancer, one of the most common cancers affecting women worldwide, poses significant health, emotional, and economic challenges. Early detection is crucial in improving the prognosis and survival rates of breast cancer patients. With advancements in data science and machine learning, there is a promising opportunity to leverage these technologies to enhance diagnostic accuracies and tailor treatment strategies effectively. This project utilizes supervised machine learning algorithms to analyze the breast cancer dataset from Kaggle, aiming to develop predictive models that accurately classify and predict the severity of breast cancer tumors. By doing so, we aspire to contribute to the early detection and personalized treatment of breast cancer, ultimately aiming to reduce its impact on patients and healthcare systems globally.

Target Variable

Targeting 'Diagnosis' as the variable is justified due to its substantial healthcare impact, clinical relevance, preventive potential, quality data, ML challenge, and public health value. The accurate classification of breast cancer cases is crucial for effective patient management and treatment planning. This focus not only addresses a critical public health issue but also leverages high-quality dataset features to challenge and advance machine learning methodologies in medical diagnostics.

Machine Learning Algorithms

Machine learning algorithms play a crucial role in heart disease prediction by leveraging patterns and insights from medical data to provide accurate assessments of an individual's risk. The algorithms that we have employed for our project is as follows:

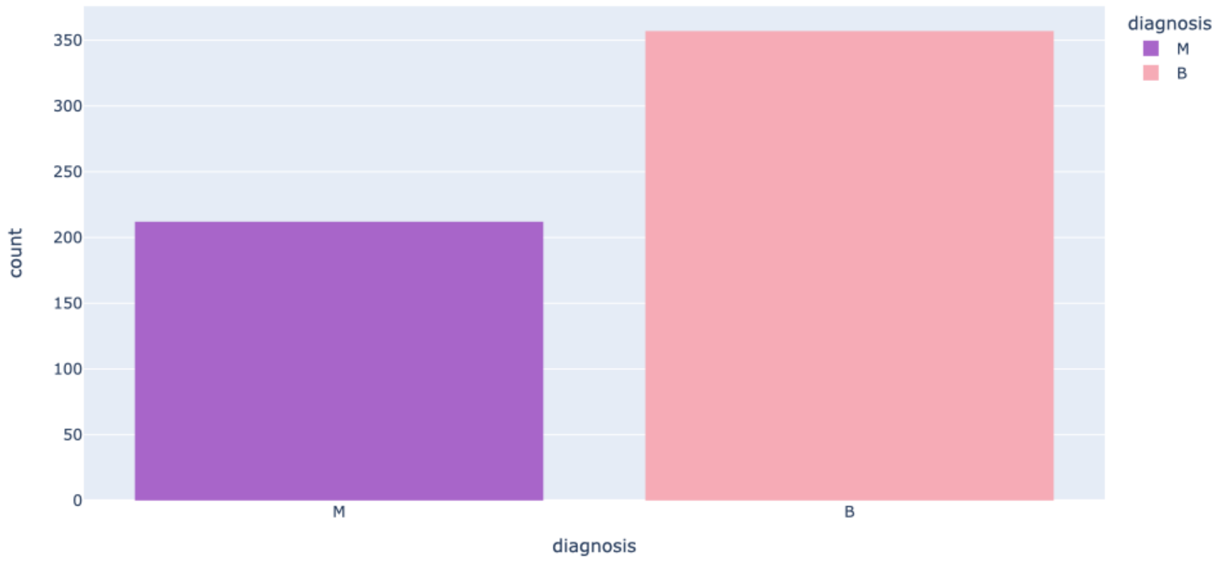
- **Decision Tree**
- **Random Forest**
- **Gaussian Naïve Biased**
- **Logistic Regression**
- **K-Nearest Neighbors**
- **XG Boost**

Data Set Description

Breast cancer is a prevalent cancer affecting a large number of women globally, and early diagnosis significantly improves treatment outcomes. This dataset contains several features derived from digitized images of fine needle aspirate (FNA) of breast mass. The features encode characteristics of the cell nuclei present in the images and are used to predict whether a tumor is malignant (M) or benign (B).
Here are the key features included in the dataset:

- ID: Patient identification number
- The diagnosis of breast tissues (M = malignant, B = benign)
- Radius Mean: Mean of distances from center to points on the perimeter
- Texture Mean: Standard deviation of gray-scale values
- Perimeter Mean: Mean size of the core tumor
- Area Mean: Area of the tumor
- Smoothness Mean: Local variation in radius lengths.
- Compactness Mean: Perimeter^2 / area - 1.0.
- Concavity Mean: Severity of concave portions of the contour
- Concave Points Mean: Number of concave portions of the contour
- Symmetry Mean
- Fractal Dimension Mean: "coastline approximation" – 1

BREAST CANCER DIAGNOSIS DISTRIBUTION - VISUALIZATION



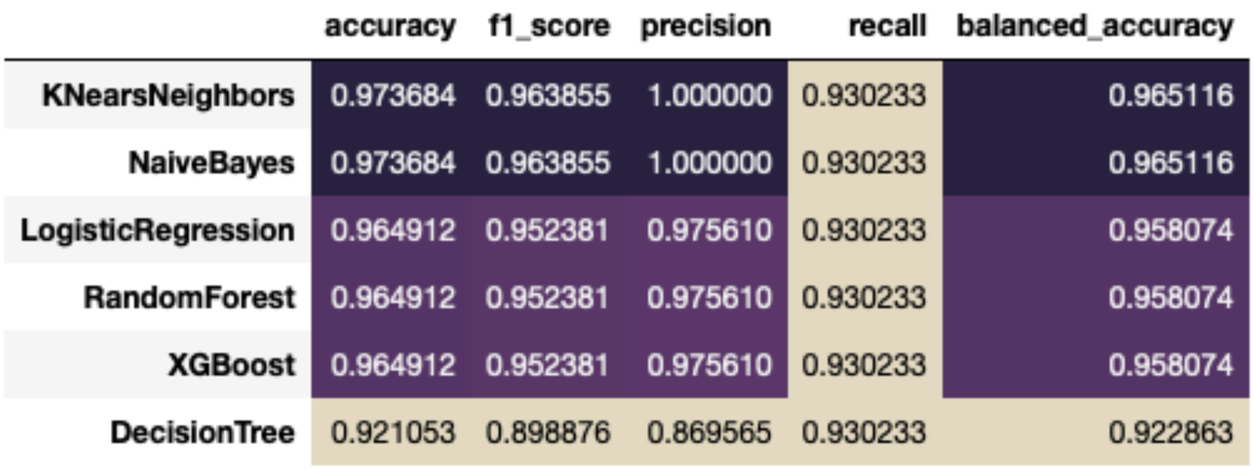
Additional Analytics & Conclusion

From the employed list of algorithms, *Random Forest* and additionally *Decision Tree* showed the top performance with the following performance metrics,

K Nearest Neighbors– Accuracy: 97.36% | f1 Score: 96.38%

Naive Bayes – Accuracy: 97.36% | f1 Score: 96.38%

Logistic Regression – Accuracy: 96.49% | f1 Score: 95.28%



	accuracy	f1_score	precision	recall	balanced_accuracy
KNearsNeighbors	0.973684	0.963855	1.000000	0.930233	0.965116
NaiveBayes	0.973684	0.963855	1.000000	0.930233	0.965116
LogisticRegression	0.964912	0.952381	0.975610	0.930233	0.958074
RandomForest	0.964912	0.952381	0.975610	0.930233	0.958074
XGBoost	0.964912	0.952381	0.975610	0.930233	0.958074
DecisionTree	0.921053	0.898876	0.869565	0.930233	0.922863

Scope of Future Work

Although the Logistic Regression, Naïve Bayes, K- Nearest Neighbors exhibited the top performance yet, we believe that by integrating advanced imaging analytics, such as mammographic texture analysis and automated tumor segmentation, we could enhance the predictions for obtaining higher accuracies. This would further aid in decisiveness and precision in diagnosing breast cancer at earlier stages, ultimately contributing to improved treatment outcomes and saving lives on humanitarian grounds.

Conclusion

- Our best performing models on the Breast cancer dataset are: Naive Bayes and K-Nearest Neighbors (KNN).
- From the employed list of algorithms, Naive Bayes and K-Nearest Neighbors (KNN) showed the top performance with the following metrics:
    - **Naive Baye**s – Balanced Accuracy: 96.51% | f1 Score: 96.38%
    - **K-Nearest Neighbors (KNN)** – Balanced Accuracy: 96.51% | f1 Score: 96.38%
- Naive Bayes and K-Nearest Neighbors (KNN) algorithms perform similarly and achieve the highest precision scores.
- Decision Tree is the only algorithm that performs better without feature selection**.**