BREAST_CANCER _CLASSIFIER

Classification Project with Multiple ML Models & Streamlit Deployment

PROJECT OVERVIEW:

Predict whether a tumor is Malignant (M) 🔴 or Benign (B) 🟢 using the Breast Cancer Wisconsin (Diagnostic) dataset.

AGENDA:

1. Dataset
2. Data Preprocessing
3. Exploratory Data Analysis (EDA)
4. Feature Selection & Explanation
5. Model Training & Evaluation
6. Streamlit App Deployment
7. Conclusion

1. DATASET :

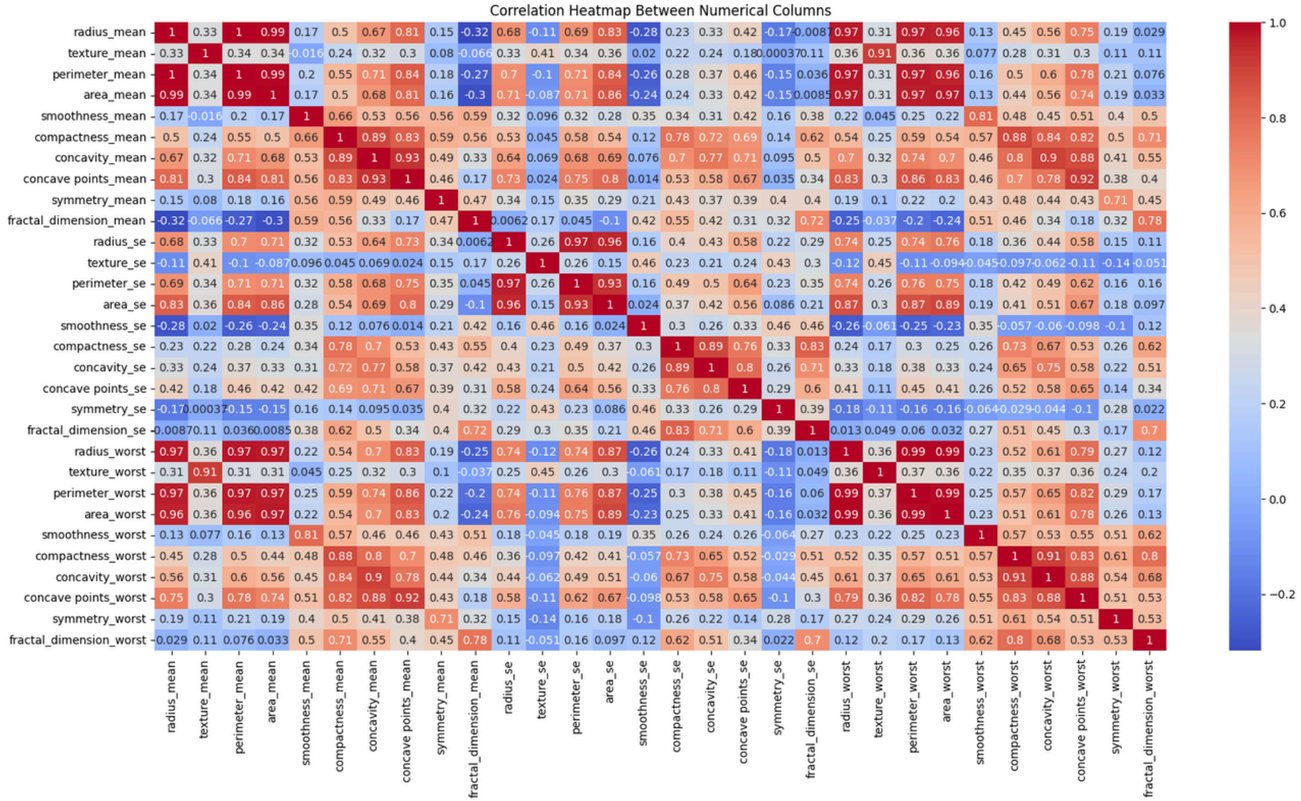
- 30 numeric features such as: radius, texture, smoothness, compactness, symmetry, fractal dimension
- Target: diagnosis (M = Malignant, B = Benign)
- Rows: 569

2. DATA PREPROCESSING:

- Removed unnecessary columns (id, Unnamed:32)
- Checked missing values & duplicates
- Handled outliers with Boxplots (replaced extreme values with boundary values)
- Normalized features using Min-Max Scaling
- Encoded target variable (Label Encoding)

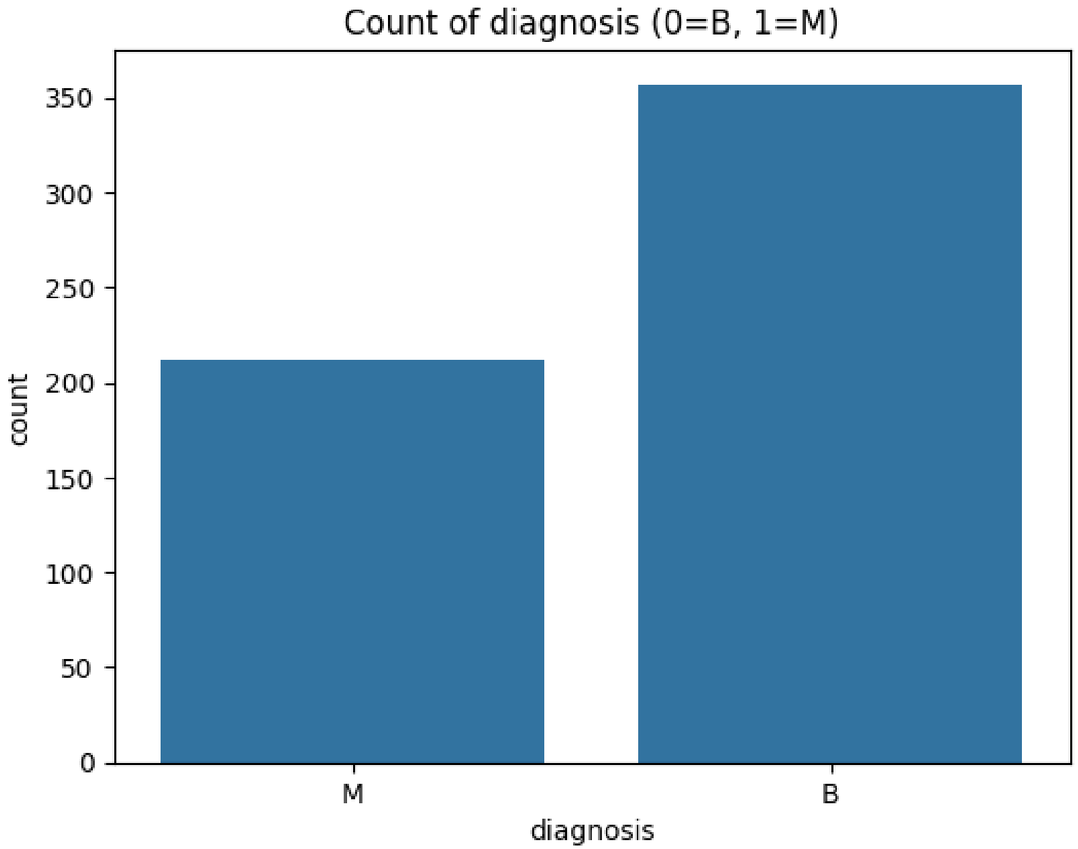
3. EXPLORATORY DATA ANALYSIS (EDA):

- Correlation Heatmap

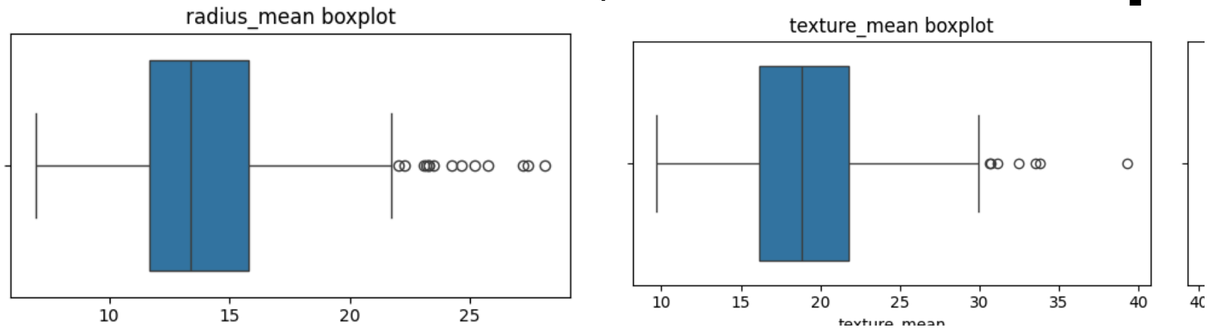


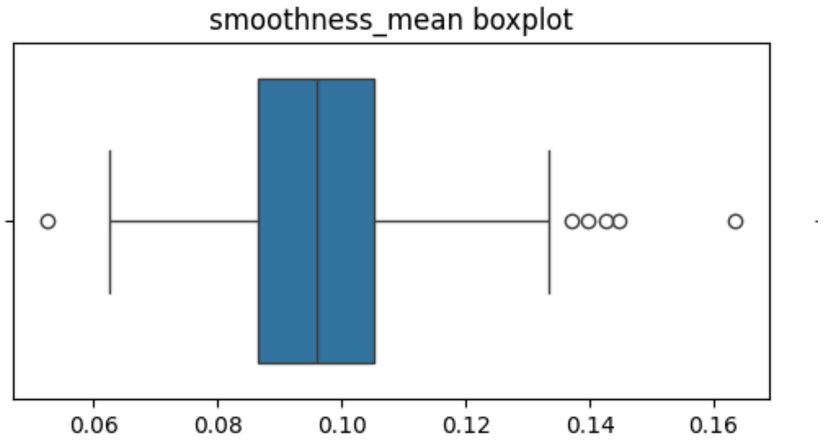
3. EXPLORATORY DATA ANALYSIS (EDA):

- Count plot for target distribution: B = 357 🟢, M = 212 🔴



- Boxplots of key features (radius_mean, texture_mean, smoothness_mean)





4. FEATURE SELECTION & EXPLANATION:

- Removed highly correlated features (≥0.8)
- Selected 13 final features for MLP model:

1. radius_mean :Average radius of the cells
2. texture_mean: Variation in texture
3. smoothness_mean: How smooth the cell edges are
4. compactness_mean: Ratio of area to perimeter²
5. symmetry_mean: Symmetry of the cells
6. fractal_dimension_mean: Complexity of cell boundary
7. radius_se: Standard error of radius
8. texture_se: Standard error of texture
9. smoothness_se: Standard error of smoothness
10. compactness_se: Standard error of compactness
11. symmetry_se: Standard error of symmetry
12. symmetry_worst: Worst (largest) symmetry
13. fractal_dimension_worst: Worst fractal dimension

5. MODEL TRAINING & EVALUATION:

- Train,Test Split: 80% training , 20% testing
  - Trained multiple classifiers: Random Forest, Decision Tree, XGBoost, Gradient Boosting, Perceptron, MLP
  - Metrics: Accuracy, Precision, Recall, F1 Score
  - Model Test Accuracy F1 Score:

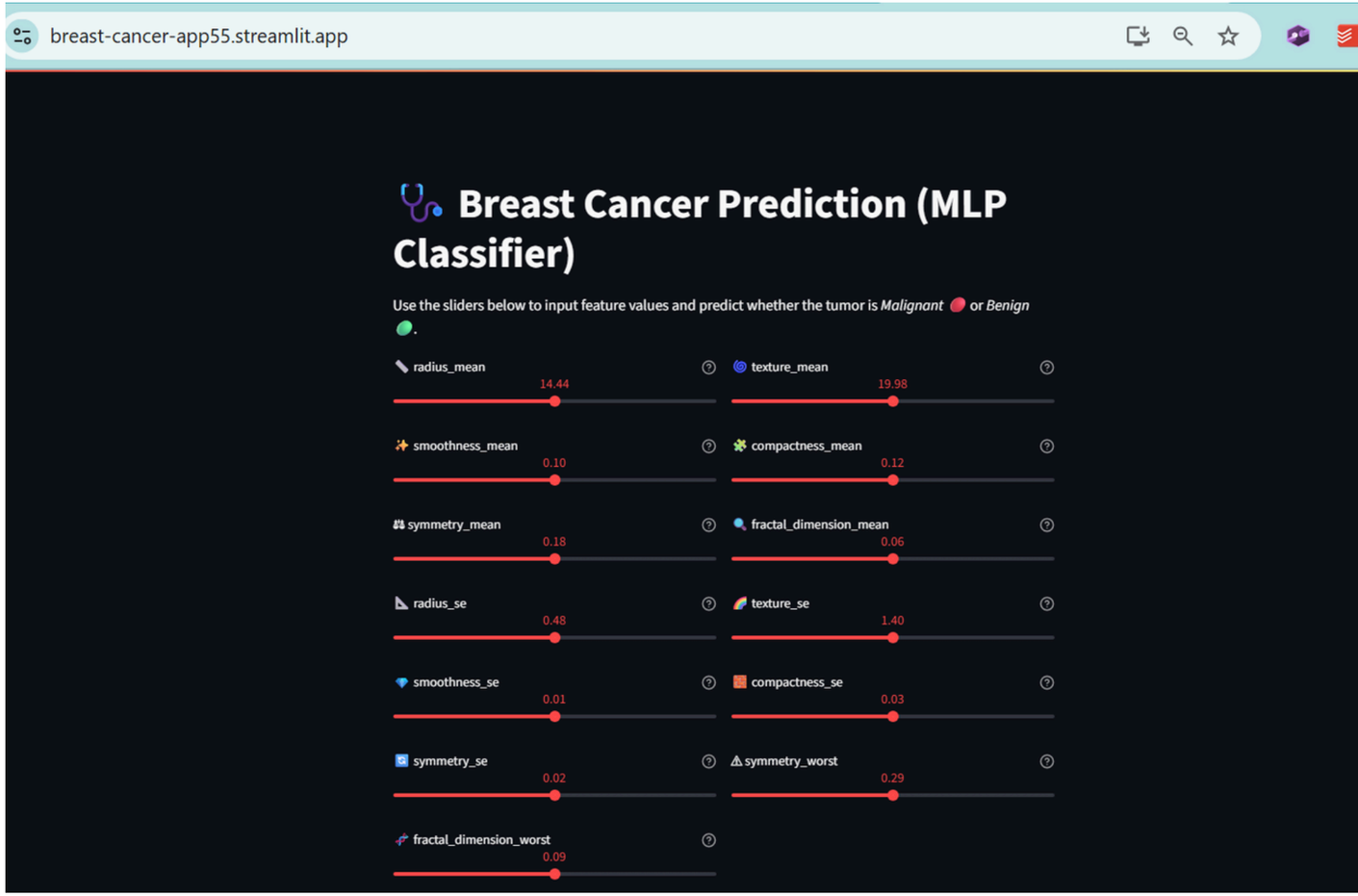
1. Random Forest 93.86% 0.916
2. Decision Tree 92.11% 0.897
3. XGBoost 95.61% 0.941
4. Gradient Boosting 94.74% 0.930
5. Perceptron 94.74% 0.929
6. MLP 96.49% 0.953

- MLP is best due to:

High Test Accuracy
Balanced Precision & Recall

6. STREAMLIT APP DEPLOYMENT:

- Interactive sliders for input
- Shows predicted tumor type 🔴🟢 and probabilities 📊
- Features have descriptive tooltips
- Streamlit link : https://breast-cancer-app55.streamlit.app/



7. CONCLUSION:

- Built robust breast cancer classifier with MLP
- Demonstrated: data preprocessing, feature selection, model comparison
- Deployed with Streamlit for easy interaction
- MLP chosen for best performance: high accuracy & F1 Score

THANK YOU