BREAST CANCER DIAGNOSIS PREDICTION: MACHINE LEARNING APPROACH

Final Report

Yahan Madhuhansa

1. Introduction

Breast cancer is one of the most common cancers worldwide, and early detection significantly improves treatment outcomes. This project leverages machine learning to classify breast tumors as **malignant (cancerous)** or **benign (non-cancerous)** based on diagnostic features extracted from medical imaging. By automating this classification, I aim to assist healthcare professionals in making faster, data-driven decisions.

2. Problem Definition

Context

- Pathologists analyze tumor characteristics (e.g., size, shape, texture) to diagnose breast cancer. This process can be time-consuming and subjective.
- Misclassification errors (e.g., false negatives) can have severe consequences.

Machine Learning Problem

- **Binary Classification**: Predict diagnosis (Malignant "M" or Benign "B") using tumor features.
- **Key Challenge**: Maximize **recall for malignant cases** (minimize false negatives) while maintaining high overall accuracy.

3. Objectives

1. **Primary Goal**:
    - Develop a model to classify tumors with >95% accuracy.
2. **Secondary Goals**:
    - Identify the most predictive features
    - Compare multiple algorithms (Logistic Regression, KNN, Random forest) to find the best performer.
    - Visualize relationships

4. Data Overview

Dataset Source

- Cancer Data**: Dataset**: Contains 569 samples with 30 numeric features derived from tumor images.

Key Features

Feature Type	Examples	Description
Mean Values	radius_mean, texture_mean	Average of all cell measurements.
Standard Error	area_se, concavity_se	Variability in cell features.
Worst Values	radius_worst, concave points_worst	Largest/most abnormal observations.

Target Variable

- diagnosis:
  - **Benign (B)**: 357 samples (~63%).
  - **Malignant (M)**: 212 samples (~37%).

---

5. Step of Project

Understand data frame

- Shape, data types
- Check missing values
- Check duplicate values

EDA

- Check outliers using boxplot
- Visualize data to understand relations
- Correlation with diagnosis
- Remove unwanted columns

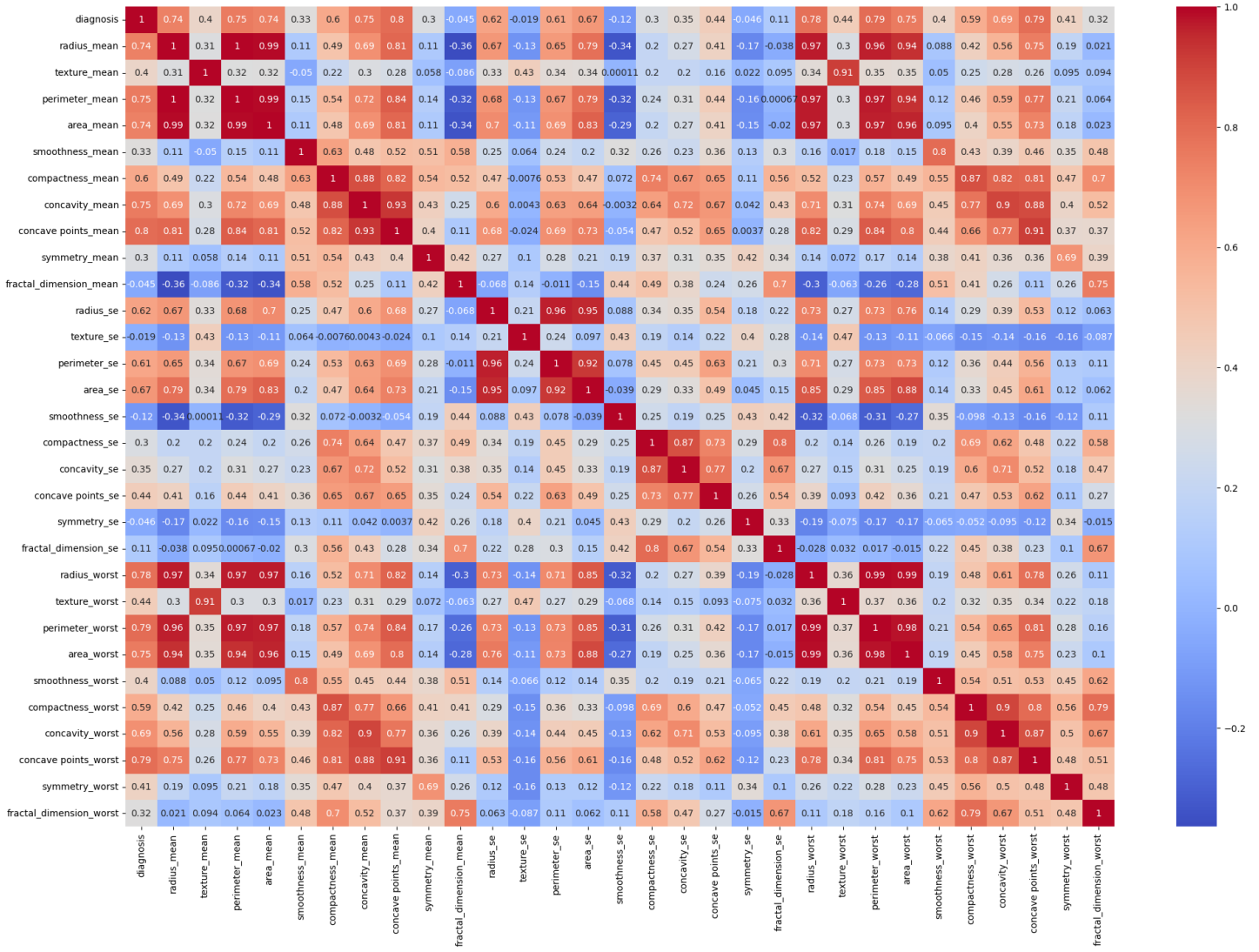
Modeling

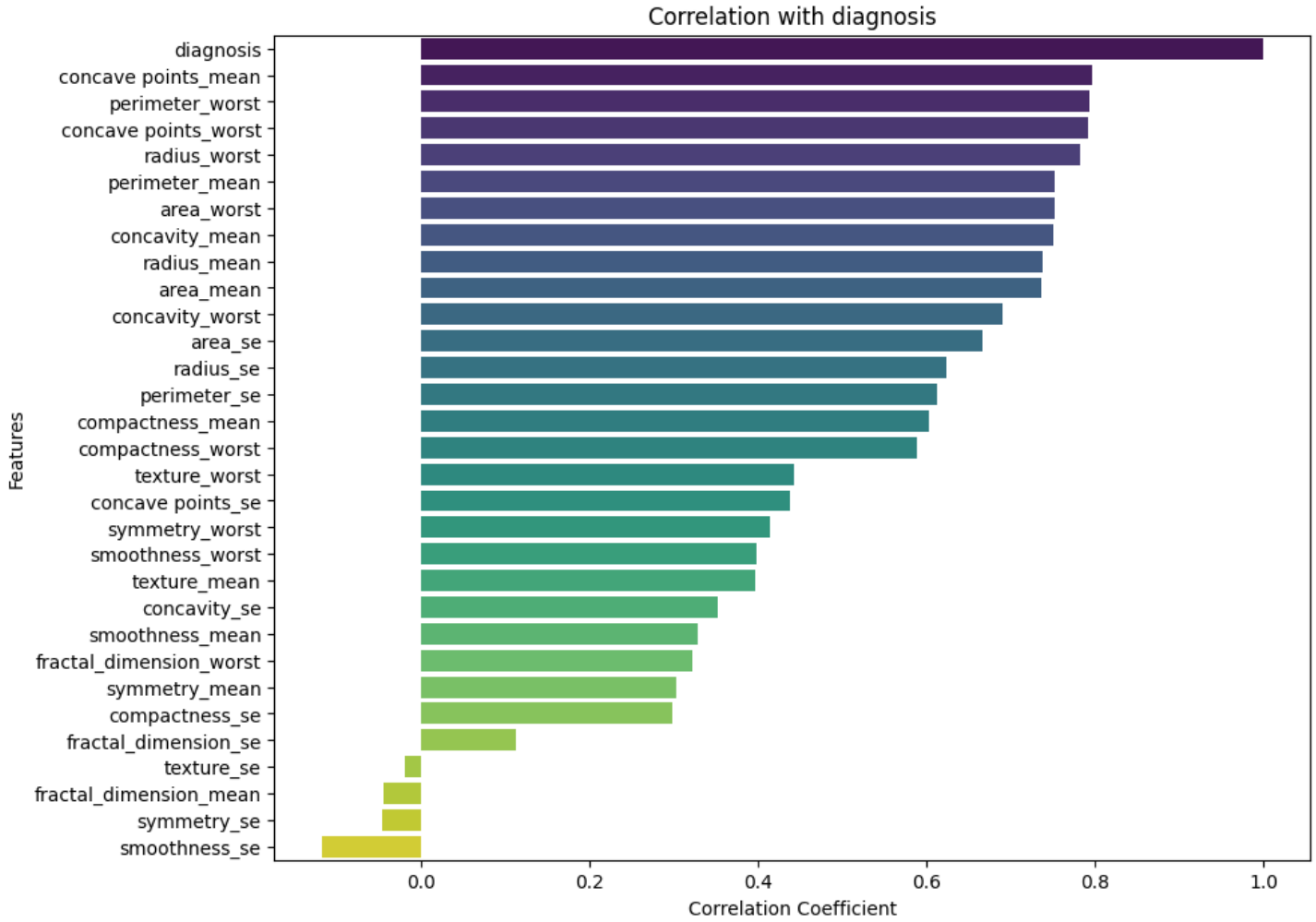
- Load and prepare Data
- Train models
  - Logistic Regression
  - KNN
  - Random forest
- Feature Importance
- Models save

6. Results & Discussion

After Understand data frame I checked **outliers**. It has too many outliers. Removing all is not a smart idea for a small data set. So I removed only extreme outliers using conditions.

After removing outliers I Checked relationships between each feature Using a **Heatmap**





And I Found that there are lot of similar features. so I can safely remove some features without reducing the model performance.

Removed features

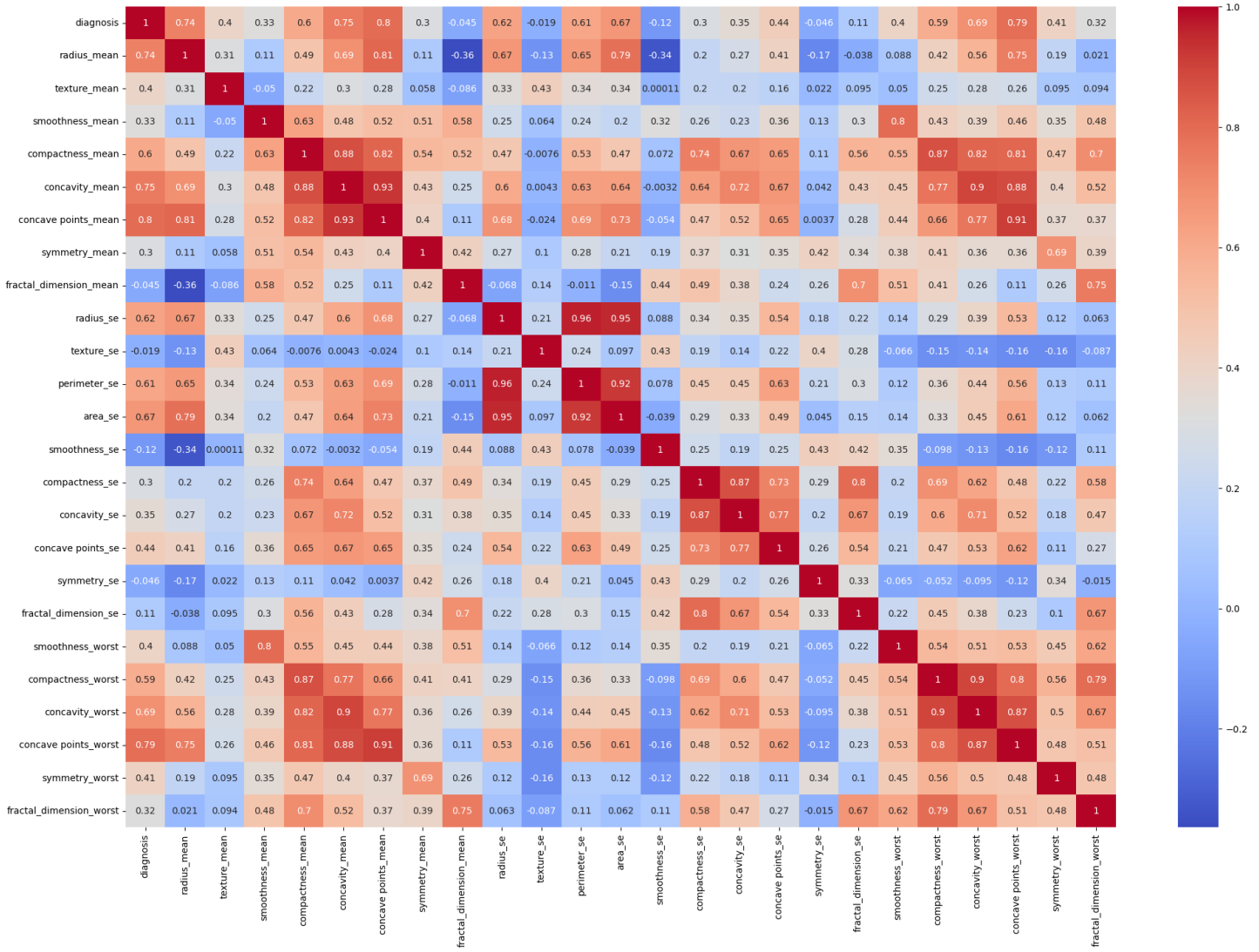
radius_mean is similar to

- perimeter_mean
- area_mean
- radius_worst
- perimeter_worst
- area_worst

texture_mean is similar to

- texture_worst

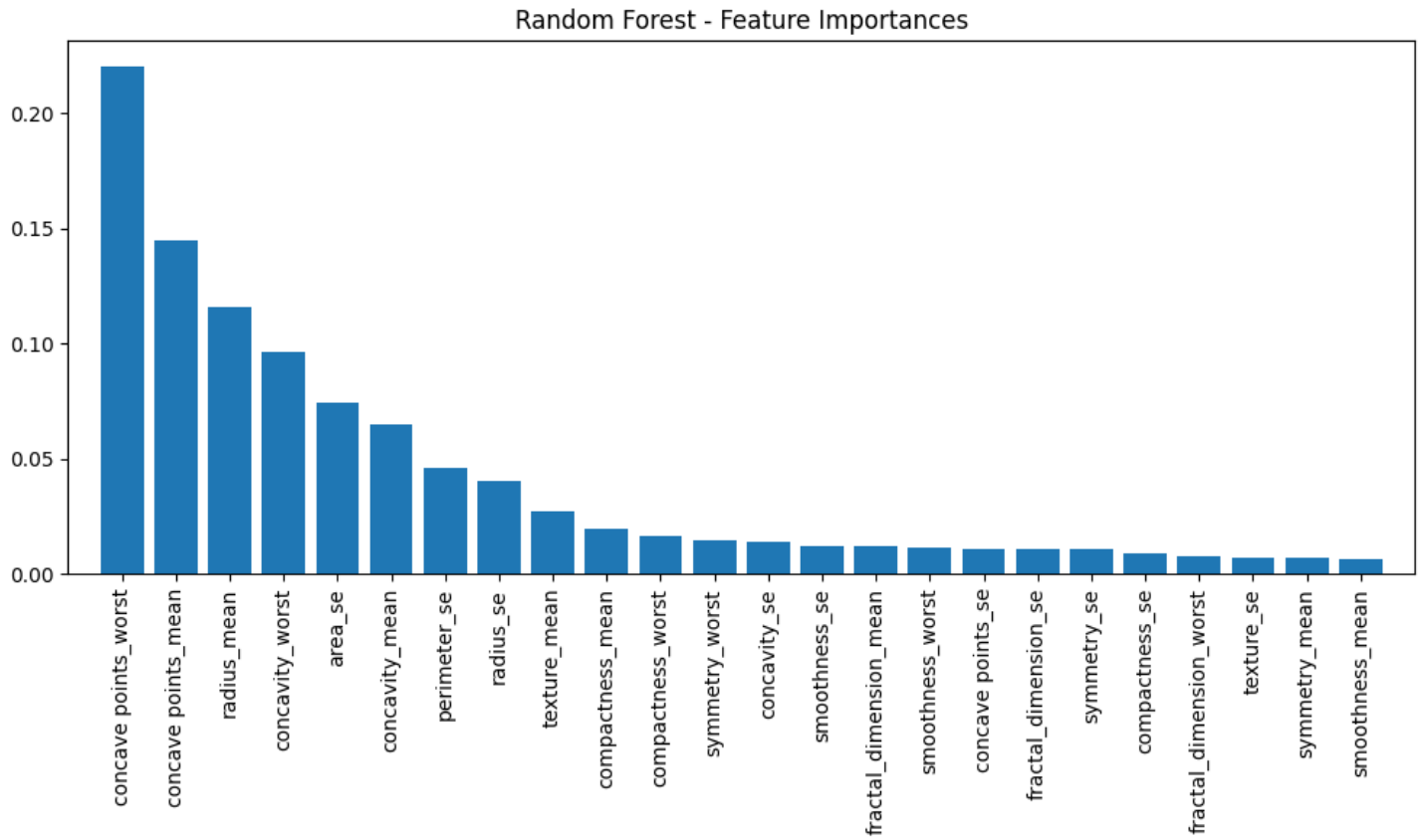
After removing features



Model performance

Model	Accuracy
Logistic Regression	0.9722222222222222
KNN	0.9907407407407407
Random Forest classifier	0.9815

Feature Importance



7. Conclusion

This project successfully built machine learning models to classify breast tumors as **malignant (M)** or **benign (B)** with high accuracy. The **KNN model performed best (99.07%)**, followed by Random Forest (98.15%) and Logistic Regression (97.22%).

Key findings:

- **Size and shape features** (like radius_mean and concave points_worst) were most important in predicting cancer.
- Removing similar features did not reduce model performance.

All models achieved over **95% accuracy**, showing that machine learning can effectively assist in breast cancer diagnosis.