BREAST CANCER DETECTION USING MACHINE LEARNING

COURSE:
DATA ANALYSIS WITH PYTHON

DONE BY:
AMARACHI FLORENCE ONYEDINMA-N

DATE:
07/08/2025

PROJECT OVERVIEW

This report presents a simplified explanation of the Breast Cancer Detection Model, developed under the Zion Tech Hub Capstone Program. The goal was to build a tool that helps predict whether a tumor is benign (non-cancerous) or malignant (cancerous) based on patient diagnostic data. Using machine learning, the model supports early detection, which is key to improving health outcomes.

STATEMENT OF THE PROBLEM

Breast cancer remains one of the leading causes of death among women worldwide. Early detection significantly increases the chances of successful treatment, yet many patients receive diagnoses at advanced stages due to delays in screening or limited access to diagnostic tools. Traditional methods like biopsies and manual scans, while effective, are time-consuming, costly, and prone to human error.

There is a growing need for automated, accurate, and accessible diagnostic support systems that can analyze medical data and flag high-risk cases early. However, integrating machine learning into healthcare comes with challenges, including handling imbalanced datasets, choosing the right model, and ensuring predictions are interpretable and trustworthy for medical professionals.

This project aims to bridge that gap by developing a machine learning model that assists in detecting whether a breast tumor is malignant or benign, using real-world diagnostic data. It seeks to complement, not replace, clinical judgment, while reducing the burden on healthcare systems and improving patient outcomes.

OBJECTIVE

Develop a machine learning model to predict breast cancer malignancy based on diagnostic features, with potential to assist in early detection and treatment planning.

POTENTIAL BENEFITS

- Enable earlier detection of malignant breast tumors, improving patient outcomes
- Reduce unnecessary biopsies for benign cases
- Assist radiologists in diagnostic decision-making
- Improve hospital resource allocation by identifying high-risk cases
- Provide insights into key factors contributing to malignancy

METHODOLOGY

I used the **"Wisconsin Breast Cancer Dataset,"** which contains measurements from breast tumor cells (e.g., size, shape, texture). Each record is labeled as either: **Benign (B): Non-cancerous** and **Malignant (M): Cancerous.**

1. Data Preprocessing

- Handled missing values and cleaned the dataset.
- Standardized column names and removed duplicates.
- All features were numeric; no categorical encoding needed.
- Dataset was split into training and testing sets (80/20).
- Explored feature relationships and distributions.

2. Model Selection

Five classification models were evaluated:

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)

Each model was evaluated using accuracy, precision, recall, and F1-score.

3. Model Training and Choosing the Best Model

- All models were trained on the training set.
- Random Forest had the best performance across all evaluation metrics, achieving over 95% accuracy in predicting breast cancer diagnosis.

4. Fine-Tuning the Model:

- Hyperparameter tuning was applied on the Random Forest model using GridSearchCV.

5. Model Evaluation

- Metrics included:
  - Accuracy: 0.97
  - Precision: 1.00
  - Recall: 0.93
  - F1-Score: 0.96
- Confusion Matrix showed 3 false negatives and 0 false positives.
- Cross-validation confirmed model reliability.

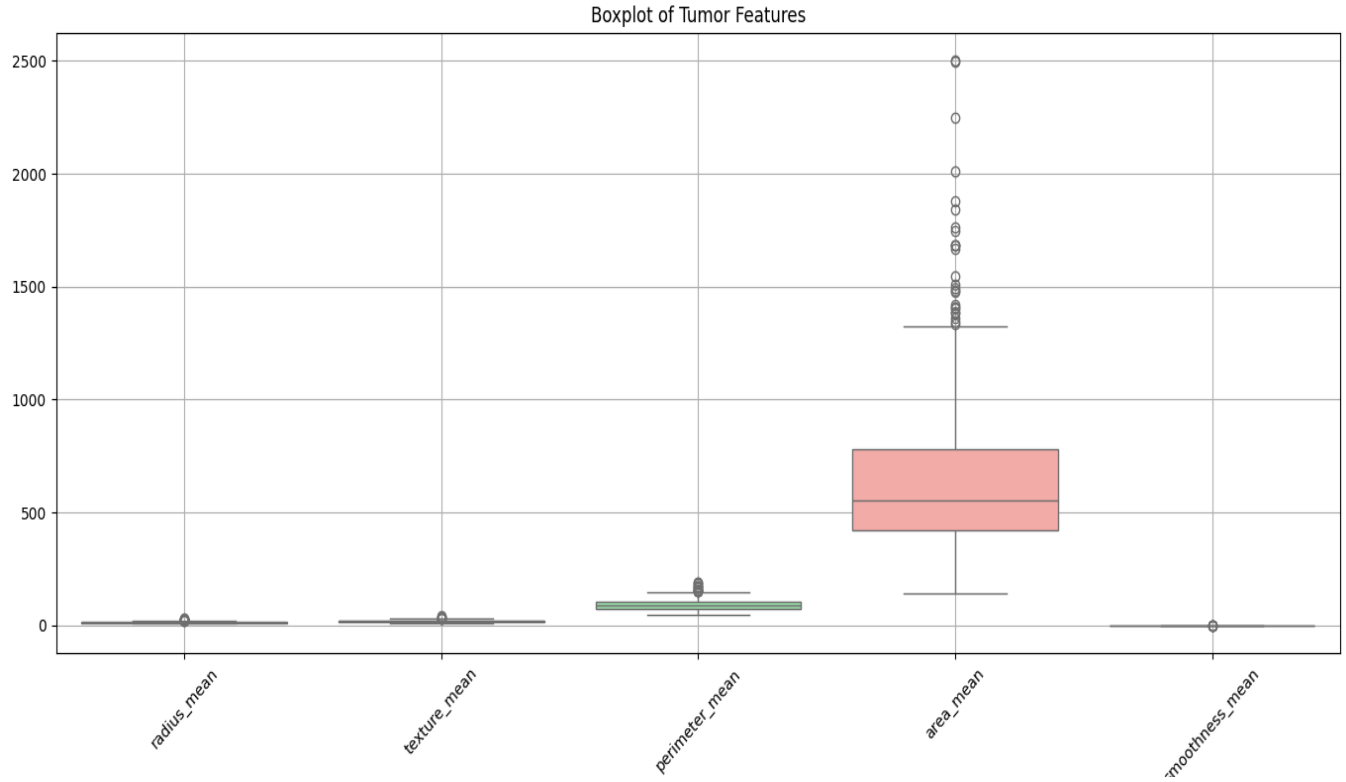
6. Model Deployment and Building a Streamlit App

- The best model (Random Forest) was saved using joblib.
- An interactive web-based diagnostic tool using Streamlit was developed.
- The app includes explanations, feature charts, and diagnosis results.
- Users can input tumor measurements manually or use sample test values.
- A prediction is displayed with confidence percentage.
- A bar chart visually compares the top 10 most significant input values.
- Automatic interpretation with findings, recommendation, and conclusion is displayed.

CHART ANALYSIS & VISUAL INTERPRETATIONS

1. Boxplot of Tumor Features

This chart illustrates the distribution of various tumor characteristics, helping us understand their spread and identify potential outliers.



Insights:

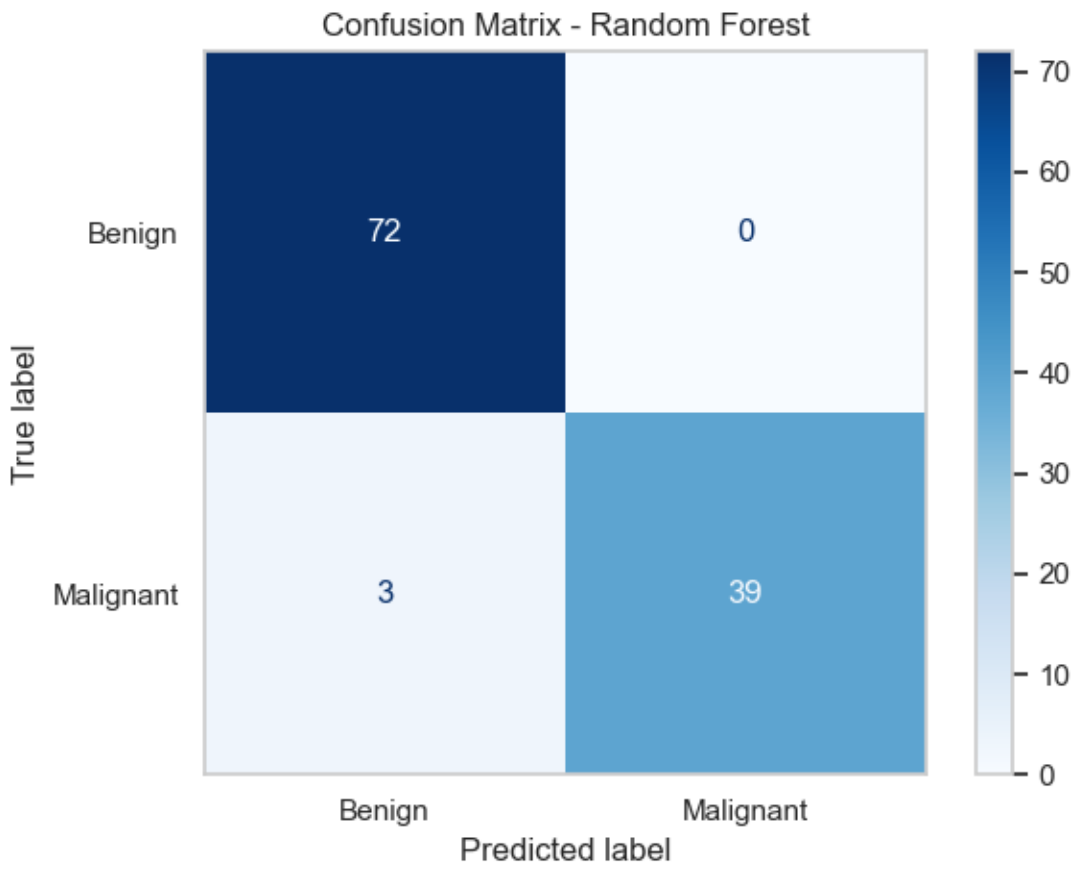
- There is a significant variation in features like **area_mean,** with some tumors exhibiting exceptionally high values.
- Presence of outliers suggests specific tumors might have distinct characteristics that warrant further investigation.

Recommendation: Focus on those outlier tumors for additional study, as they could provide insights into aggressive forms of the disease or unique patterns in tumor behavior.

Conclusion: This visualization helps us appreciate the complexity and variability in tumor characteristics, guiding us toward areas that may require more research and attention.

2. Confusion Matrix - Random Forest

This matrix displays the performance of our Random Forest model in classifying tumors as benign or malignant.



Insights:

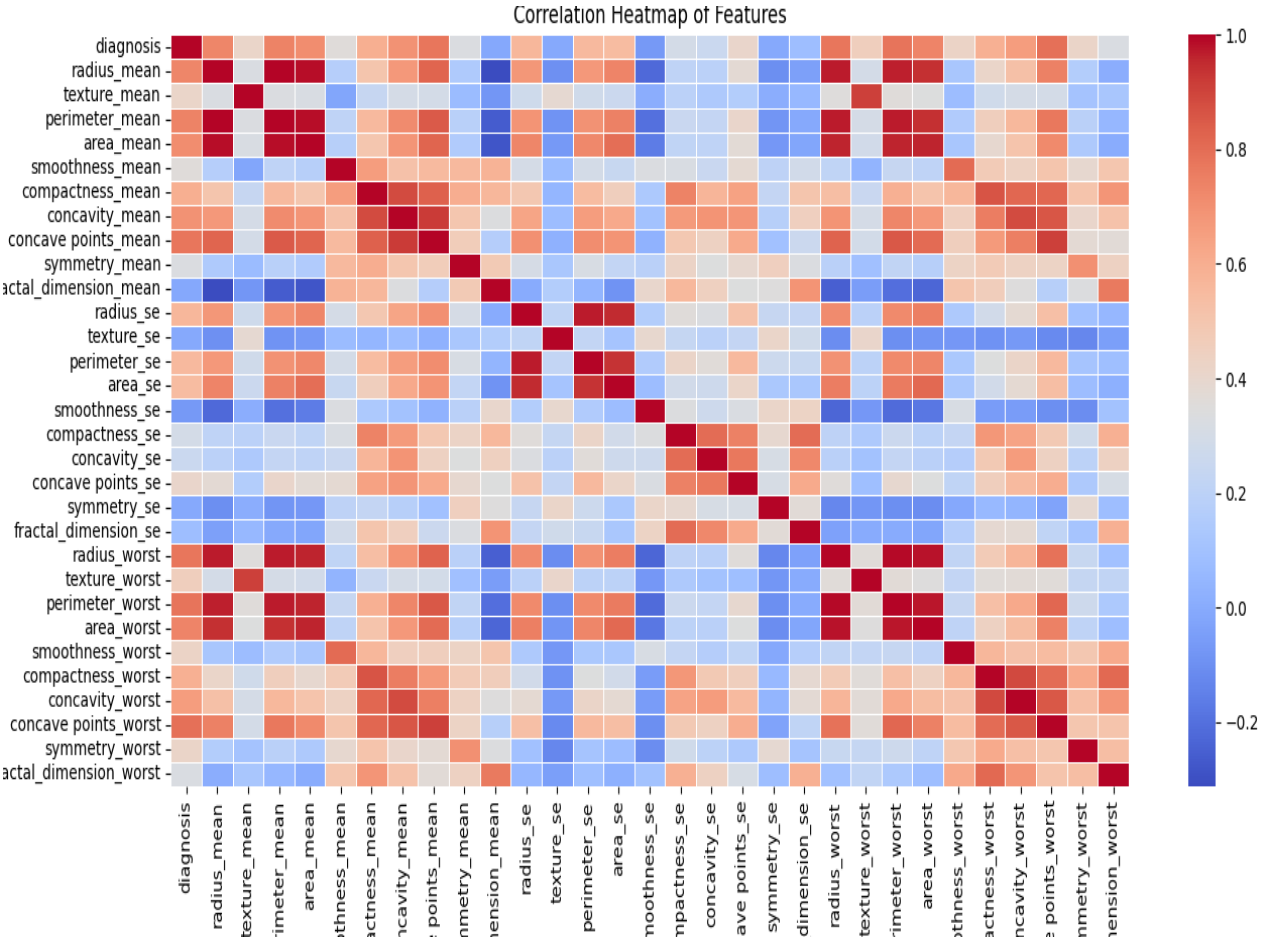
- The model accurately identified 72 benign tumors but misclassified 3 malignant tumors as benign.
- It successfully identified 39 malignant tumors, showcasing a strong detection capability.

Recommendation: Since there are a few false negatives, we should consider refining the model further to minimize these misclassifications, ensuring patient safety.

Conclusion: The model is generally effective, but we need to enhance its sensitivity to ensure it catches every malignant case.

3. Correlation Heatmap of Features

The aim of this chart is to show how different tumor features correlate with each other, revealing potential relationships.



Insights:

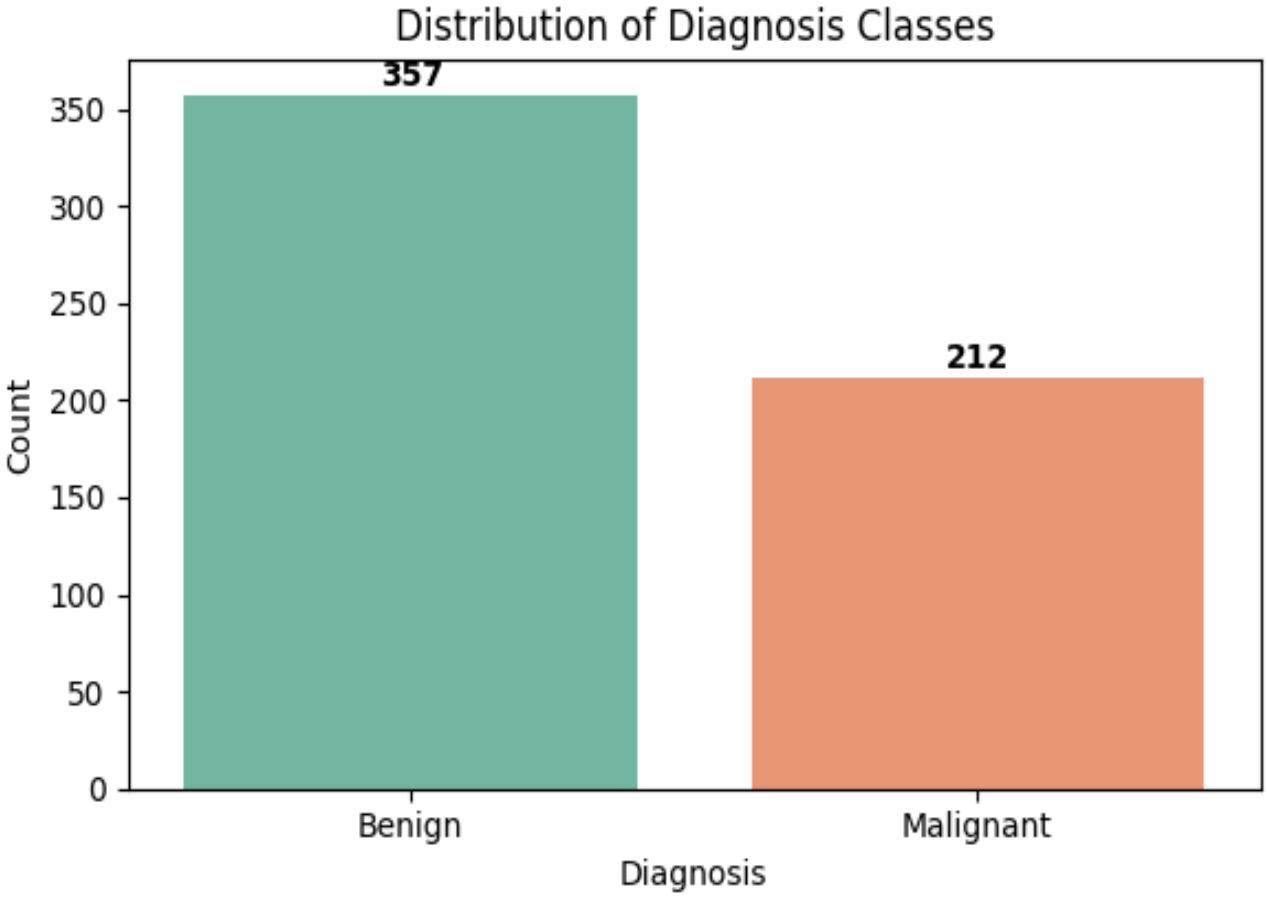
- Strong correlations (close to 1 or -1) are visible between certain features, like radius_mean and perimeter_mean, suggesting that as one increases, the other tends to increase or decrease correspondingly.
- Understanding these relationships can help refine the model and improve predictions.

Recommendation: Leverage these correlations to streamline our feature selection process, focusing on the most impactful variables for model training.

Conclusion: This heatmap provides valuable insights into feature interdependencies, paving the way for more accurate and efficient modeling.

4. Distribution of Diagnosis Classes

This bar chart illustrates the distribution of tumor diagnoses, showing the balance between benign and malignant cases.



Insights:

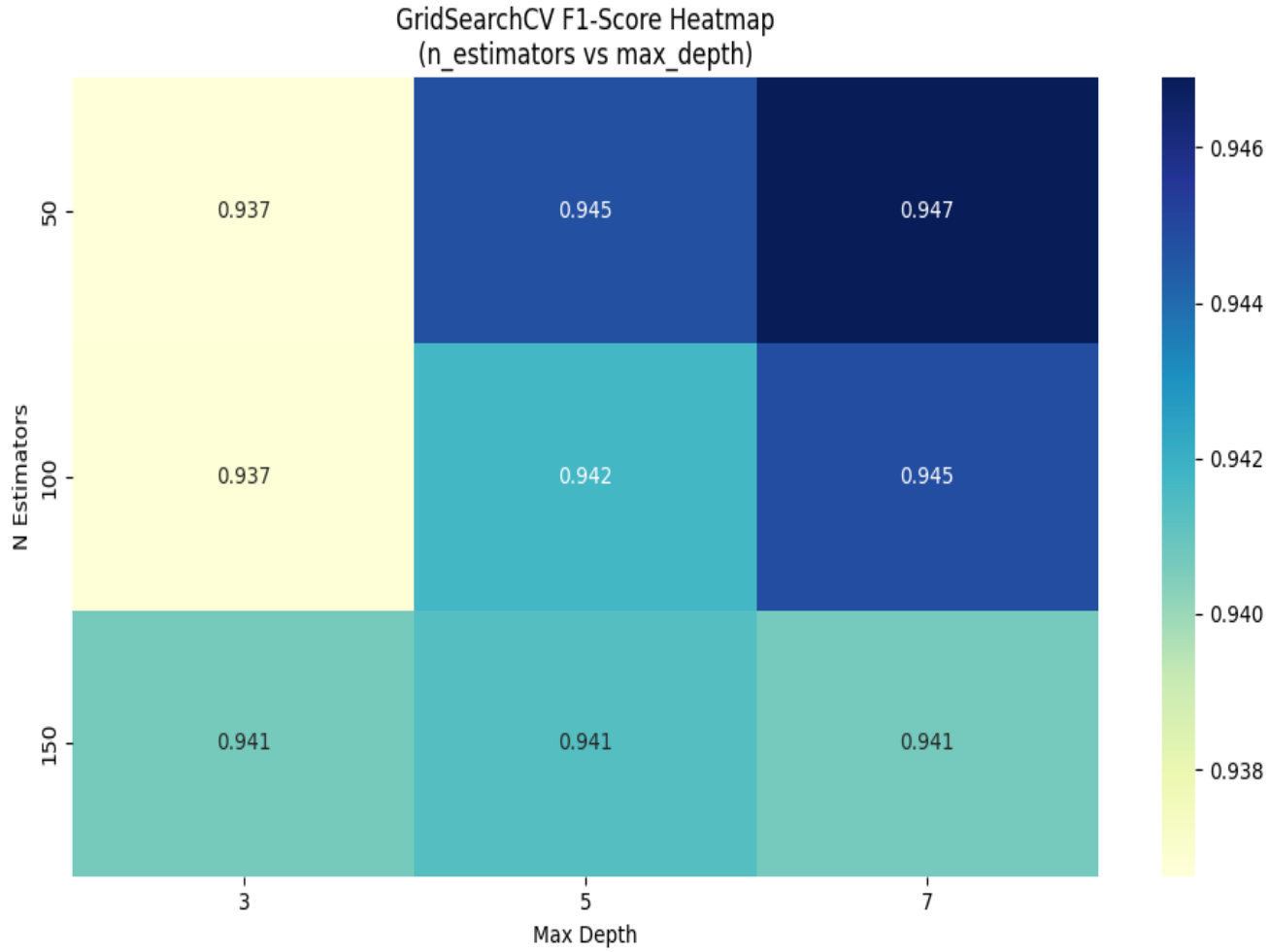
- There are 357 benign cases compared to 212 malignant cases, indicating a higher prevalence of benign tumors in the dataset.
- This imbalance may affect model training and performance.

Recommendation: Consider techniques to address this imbalance, such as data augmentation or adjusting classification thresholds to improve detection of malignant cases.

Conclusion: Understanding the distribution of diagnoses is crucial for training effective models that accurately reflect real-world scenarios.

5. GridSearchCV F1-Score Heatmap

Heatmap displays the F1-scores of different model configurations based on varying numbers of estimators and maximum depths.



Insights:

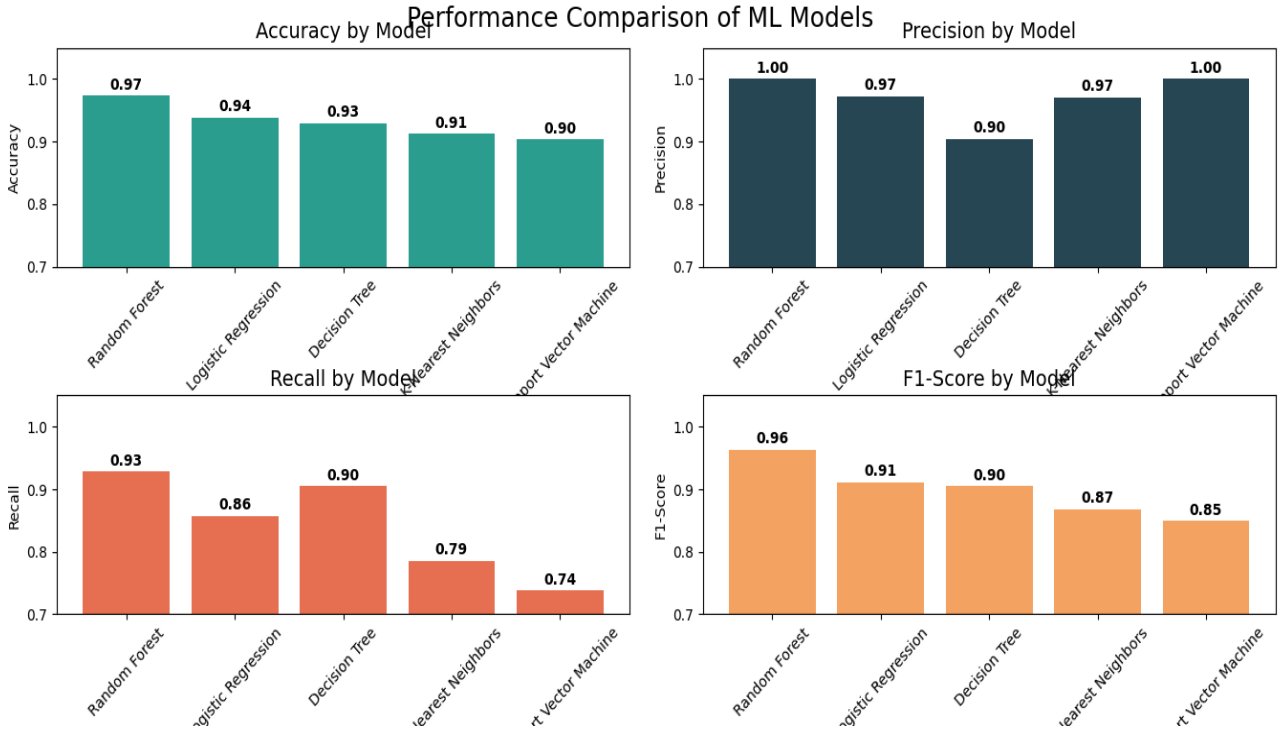
- Higher F1-scores are observed with specific combinations of parameters, indicating optimal settings for our model.
- The best-performing configurations can significantly enhance our predictive accuracy.

Recommendation: Utilize the identified optimal parameters in our model to achieve the best possible performance.

Conclusion: This analysis empowers us to make data-driven decisions in tuning our model for improved outcomes.

6. Performance Comparison of ML Models

This series of charts compares various machine learning models based on accuracy, precision, recall, and F1-score.



Insights:

- The Random Forest model exhibits the highest accuracy and precision, making it a robust choice for our classification task.
- Other models, while decent, do not surpass Random Forest in overall performance.

Recommendation: Continue using the Random Forest model as our primary classification tool, while also exploring potential enhancements from other models.

Conclusion: This comparison reinforces our confidence in the Random Forest model's capabilities, ensuring we are well-equipped to make informed decisions moving forward.

STREAMLIT APP USER INTERFACE

The **Breast Cancer Detection App** provides an interactive platform for users to enter tumor measurements and receive predictions. Key features include:

- **User Guidance:** Clear instructions on input values and their significance.
- **Example Data:** Users can utilize pre-filled example measurements for testing.
- **Visual Feedback:** Users receive a graphical representation of the top influencing tumor features in the prediction.

Findings and Recommendations: After predictions, the app highlights the key features impacting the diagnosis, encouraging informed discussions between patients and healthcare providers.

PROJECT SUMMARY

In this capstone project, I built a simple but powerful tool using machine learning to help detect breast cancer. The goal was to predict whether a tumor is **benign (not cancer)** or **malignant (cancer)** by analyzing patient test results.

I started by carefully examining the raw data, how tumor size, shape, and texture vary, and spotting patterns, unusual values, and imbalances. This helped us clean the data and prepare it properly.

After testing several machine learning models, it was discovered that the Random Forest model gave the most accurate and reliable results. It was fine-tuned to make it even better, and then a simple web app was built where predictions can be made easily by entering tumor details.

The model was tested thoroughly and reached **97% accuracy**. Most importantly, it rarely missed cancer cases, which is crucial in healthcare. Every step, from data understanding to final predictions, was guided by visuals that made the findings easier to explain and decisions easier to make.

KEY RECOMMENDATIONS

- **Handle Imbalances Carefully:** Because there were more non-cancer cases in the data, we should keep using techniques that help the model treat all patients fairly and not miss cancer cases.
- **Focus on What Matters:** Some features in the data are more useful than others. Removing repetitive or unhelpful information keeps the model faster and clearer.
- **Monitor Missed Cancer Cases:** We had a few instances where cancer wasn't detected (false negatives). These should be reviewed and reduced as much as possible.
- **Keep Updating the Model:** Just like medical knowledge evolves, so should this model. It should be retrained often with new data to stay reliable.
- **Support, Not Replace Doctors:** This model is a decision support tool. It helps doctors, but it doesn't replace professional medical judgment or testing.

CONCLUSION

This project shows how technology and healthcare can work hand-in-hand. By using machine learning, I built a practical and easy-to-use tool that can assist in early detection of breast cancer.

The model performs very well, is backed by strong data insights, and has been made accessible through a simple web application.

Ultimately, this project is not just about building a smart tool; it's about empowering people and healthcare providers with something that could make early diagnosis faster, easier, and more accurate.

APPENDIX

Tools Used: Python, Pandas, Scikit-learn, Streamlit, Plotly
Dataset: Breast Cancer Wisconsin (Diagnostic) Dataset
Author: Amarachi Florence Onyedinma-Nwamaghiro
Platform: Zion Tech Hub.
Course: Data analysis with python