BREAST CANCER DETECTION USING MACHINE LEARNING

COURSE:
DATA ANALYSIS WITH PYTHON

DONE BY:
AMARACHI FLORENCE ONYEDINMA-N

DATE:
25/08/2025

PROJECT OVERVIEW

In this project, I built a machine learning model that can help predict whether a breast tumor is benign (non-cancerous) or malignant (cancerous) based on diagnostic data. This work was part of my capstone project under the Zion Tech Hub program.

The main idea behind this project was to support early detection, because in healthcare, early detection often makes all the difference. Using data science, I created a tool that can help simplify diagnosis, assist doctors in decision-making, and show how machine learning can actually be applied to solve real-world problems.

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

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

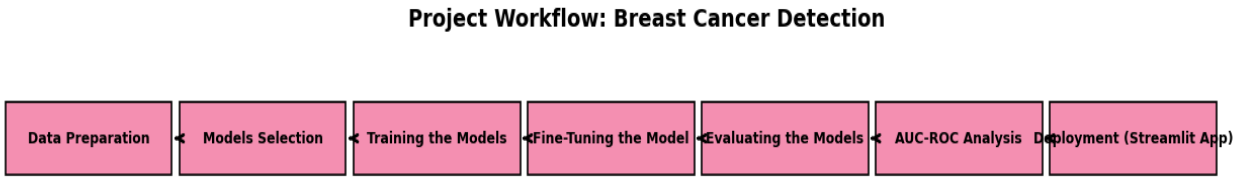
POTENTIAL BENEFITS

- 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

PROJECT WORKFLOW

The workflow of this project can be summarized as follows:



Data Preparation → Models Selection → Training the Models → Fine-Tuning the Model → Evaluating the Model → AUC-ROC Analysis → Deployment (Streamlit App)

1. Data Preparation

The Wisconsin Breast Cancer Dataset was used, which contains medical measurements of breast tumors such as their size, shape, and texture.

- First, the dataset was cleaned by removing duplicate records to avoid repeated information.
- Then, it was divided into two groups: one for *teaching the model* (80%) and one for *testing it* (20%).
- Since all the values were already numbers, no extra conversion was needed.
- Data exploration was done to understand patterns, unusual values, and relationships between features.

2. Models Selection

To find the best method, five different machine learning approaches were tested:

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

3. Training the Models

Each model was trained using the teaching set and then tested with the test set. Their results were compared using key measures of success.

The **Random Forest model** performed the best, achieving more than **95% accuracy**.

4. Fine-Tuning the Model

To make the Random Forest even stronger, its internal settings (like how many "trees" it uses and how deep they go were adjusted). This process, called ***hyperparameter tuning,*** helped maximize performance.

5. Evaluating the Model

The final version of Random Forest achieved:

- **97% Accuracy** (overall correctness)
- **Precision** (no false alarms for cancer)
- **0.93 Recall** (it caught most cancer cases)
- **0.96 F1-Score** (a balance between precision and recall)

The confusion matrix showed that only **3 cancer cases were missed**, which is very important because in medicine, missing a diagnosis can be dangerous.

6. AUC-ROC Analysis

Since there were more non-cancer cases than cancer cases, accuracy alone wasn't enough. A special test called the **AUC-ROC curve** was used. The model scored **0.99 (out of 1)**, which means it is extremely reliable at telling apart cancerous and non-cancerous tumors across different scenarios.

7. Deployment (Streamlit App)

Finally, an **interactive web app** was created so that anyone can use the model. With this app, users can:

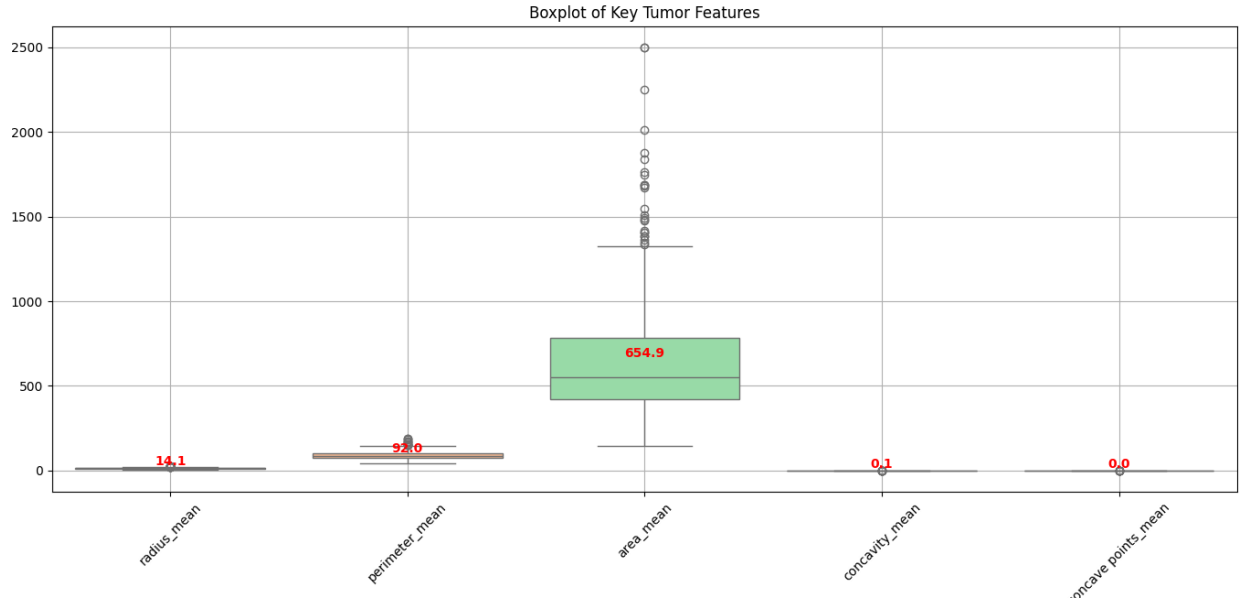
- Enter tumor measurements manually.
- Instantly see whether the tumor is predicted as benign or malignant, along with a confidence score.
- View charts showing which tumor features had the most influence on the prediction.

For efficiency, the app focuses only on the **top 10 most important features** found by Random Forest. This makes predictions faster without reducing accuracy.

CHART ANALYSIS & INTERPRETATIONS

1. **Boxplot of Tumor Features**

Objective: The goal of this chart was to check how tumor features like radius, perimeter, and area are distributed and to spot any outliers.



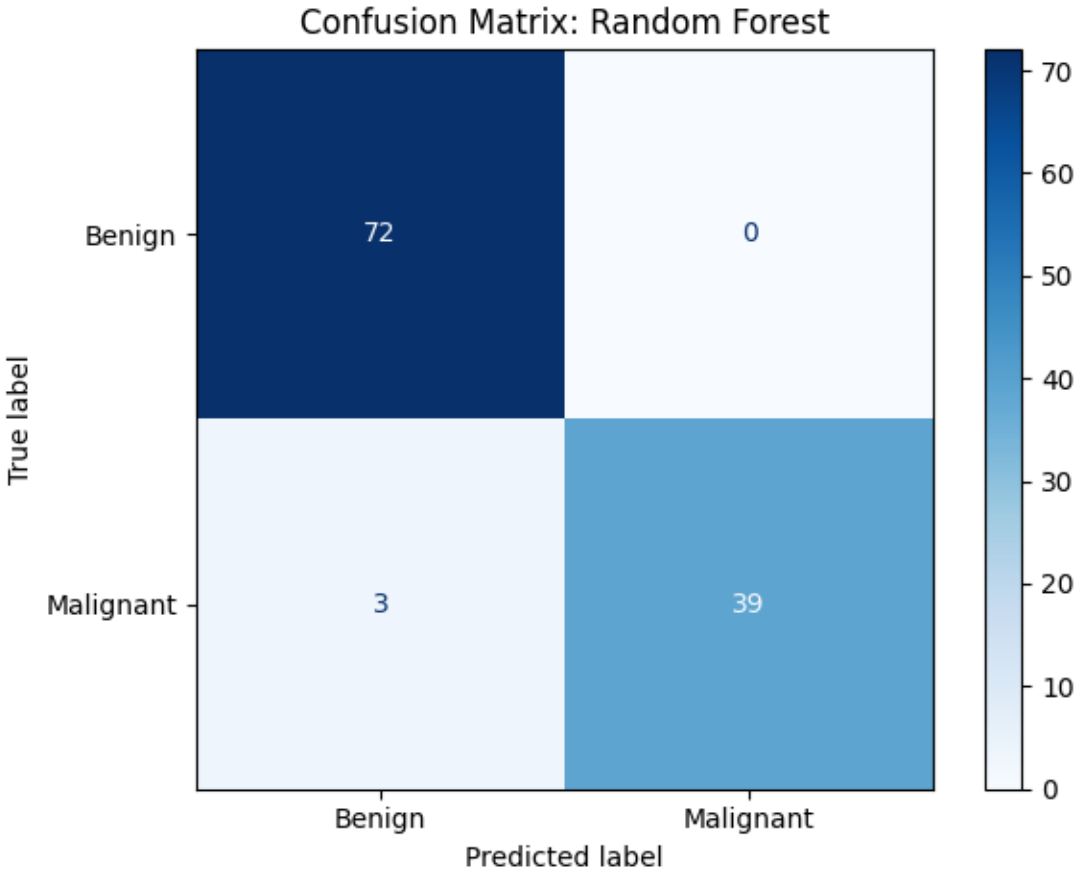
Insights: From the chart, it was noticed that *area_mean* has some very large outliers, showing that some tumors are unusually big compared to the rest.

***Recommendation*:** Those outliers should be looked at more closely, they could represent aggressive tumors or special cases worth deeper study.

Conclusion: This chart helped to understand the wide variability in tumor sizes and why some cases may stand out more than others.

2. Confusion Matrix – Random Forest

Objective: This chart was used to evaluate how well the Random Forest model classified tumors.



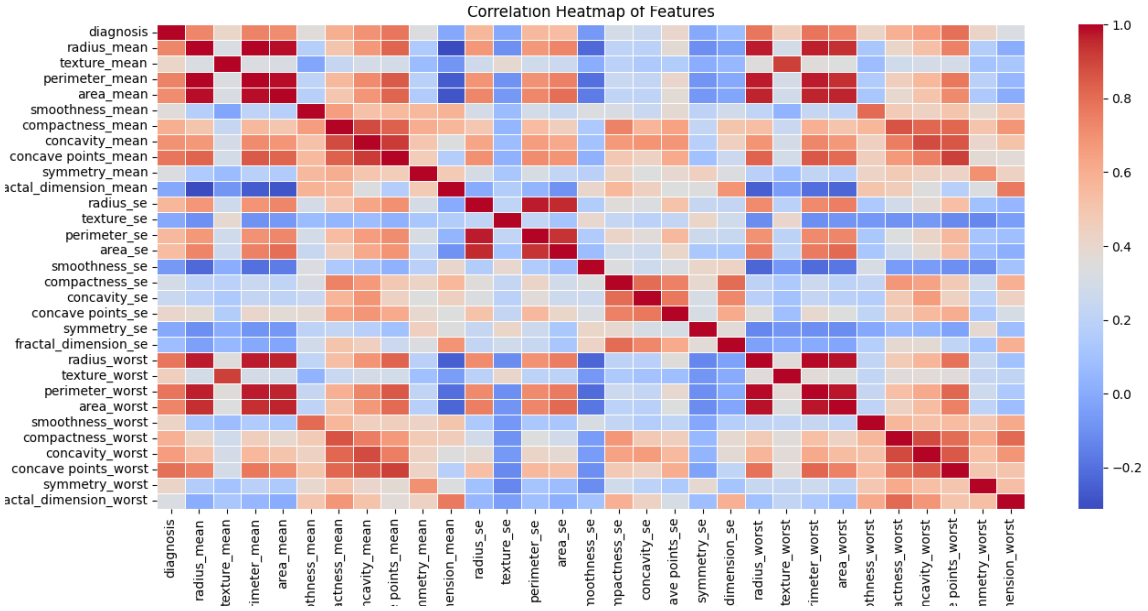
Insights: The model did really well. It correctly predicted **72 benign cases** and **39 malignant cases**, but it did miss **3 malignant cases** (false negatives).

Recommendation: Even though performance is strong, the model need to be refined further to reduce false negatives since missing a cancer case is very costly in real life.

Conclusion: The confusion matrix shows Random Forest is reliable, but there's room to make it even more sensitive.

3. Correlation Heatmap of Features

Objective: This chart was used to see which tumor features were related to each other.



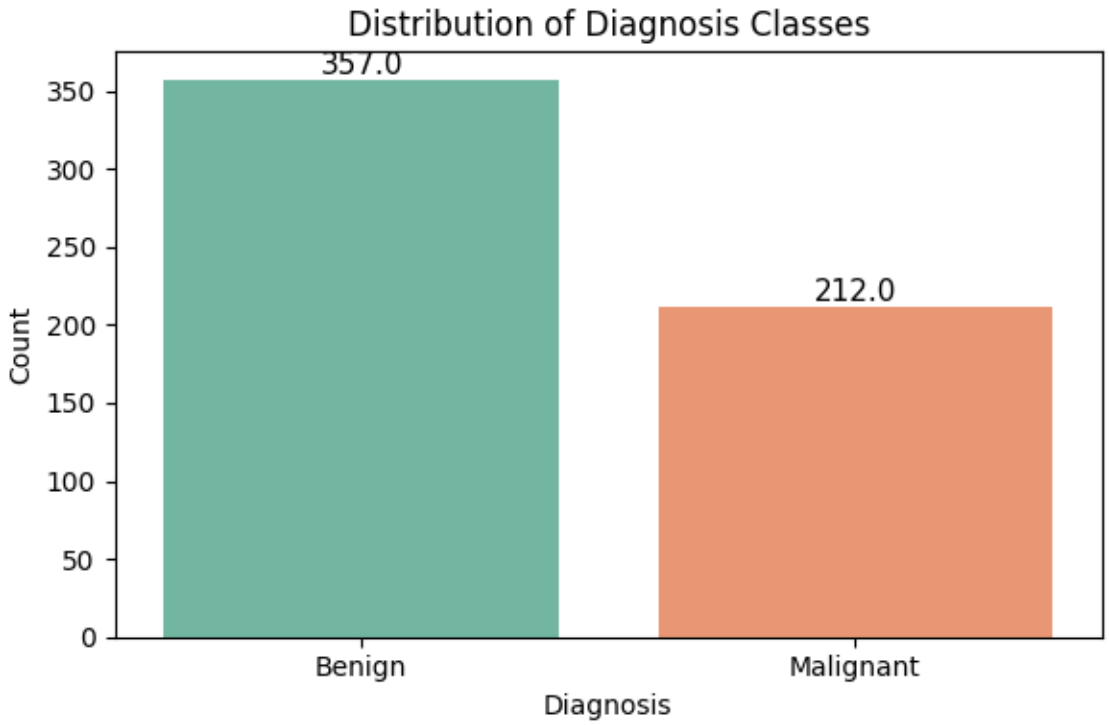
Insights: Features like *radius_mean, perimeter_mean,* and *area_mean* are very strongly correlated. This tells that they capture similar information. On the other hand, some features like *symmetry_worst* had negative correlations with others.

Recommendation: For better efficiency, redundant features can be removed then focus on the most impactful ones.

Conclusion: The heatmap was useful in guiding which features are most important and how to simplify the dataset without losing valuable information.

4. Distribution of Diagnosis Classes

Objective: This chart helped us to understand how balanced or imbalanced the dataset is.



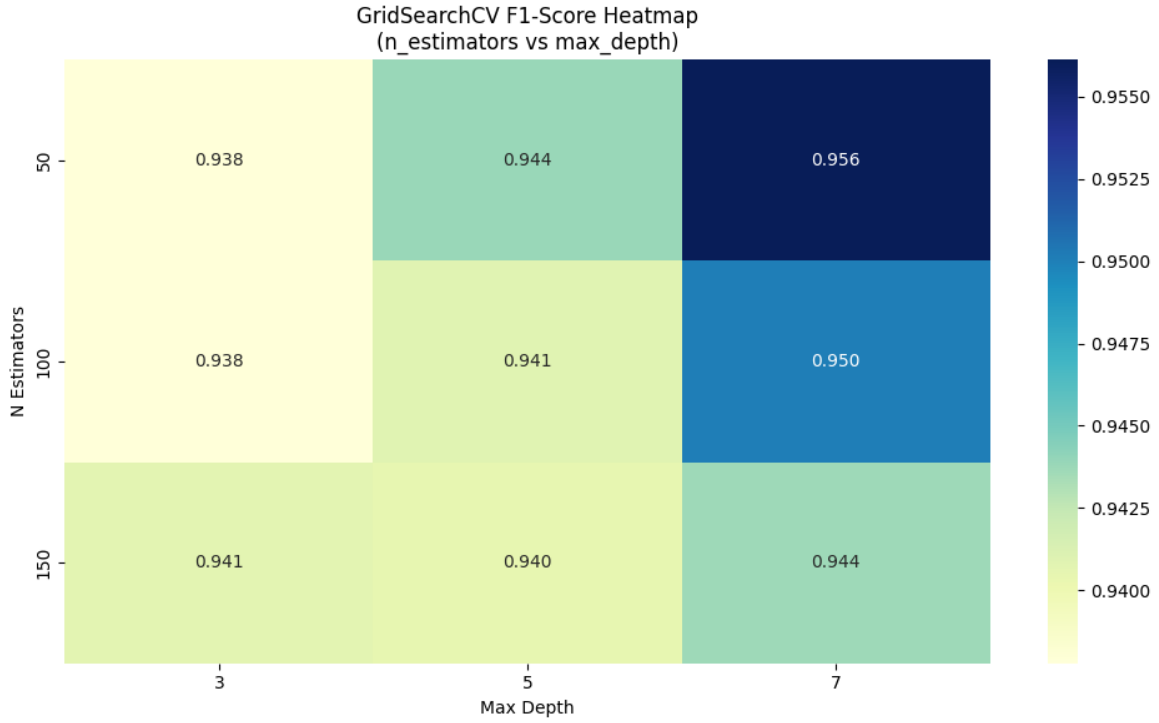
Insights: The dataset has more **benign cases (357)** compared to **malignant cases (212)**. This imbalance could affect how models learn.

Recommendation: To balance this, oversampling malignant cases or adjusting class weights can be done so that the model doesn't lean too heavily towards benign.

Conclusion: Understanding this imbalance was key in making sure the model treats both classes fairly.

5. GridSearchCV F1-Score Heatmap

Objective: This chart was used to visualize the effect of different Random Forest parameters on performance.



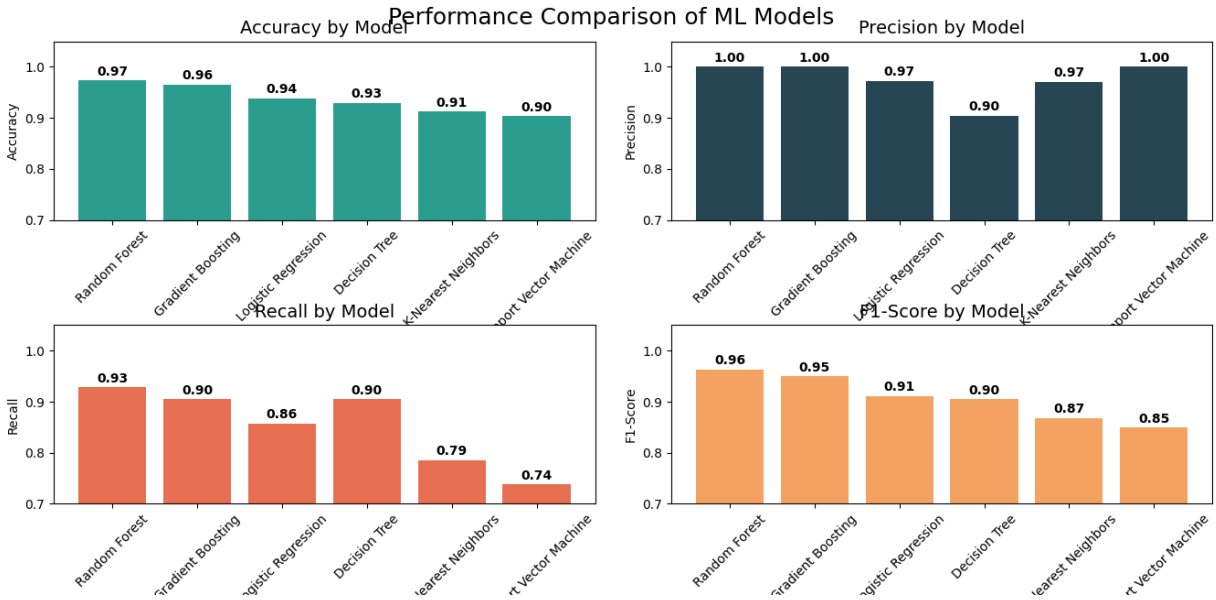
***Insights*:** It was found that the best results came with **150 estimators and max depth of 3**. That gave the highest F1 score.

Recommendation: These optimal settings should be used in the final model, and it can also be experimented with more configurations later.

Conclusion: GridSearchCV made it easier to fine-tune the model and squeeze out the best performance.

6. Performance Comparison of ML Models

Objective: Here, all the models, Random Forest, Logistic Regression, SVM, KNN, and Decision Tree were compared, across different metrics.



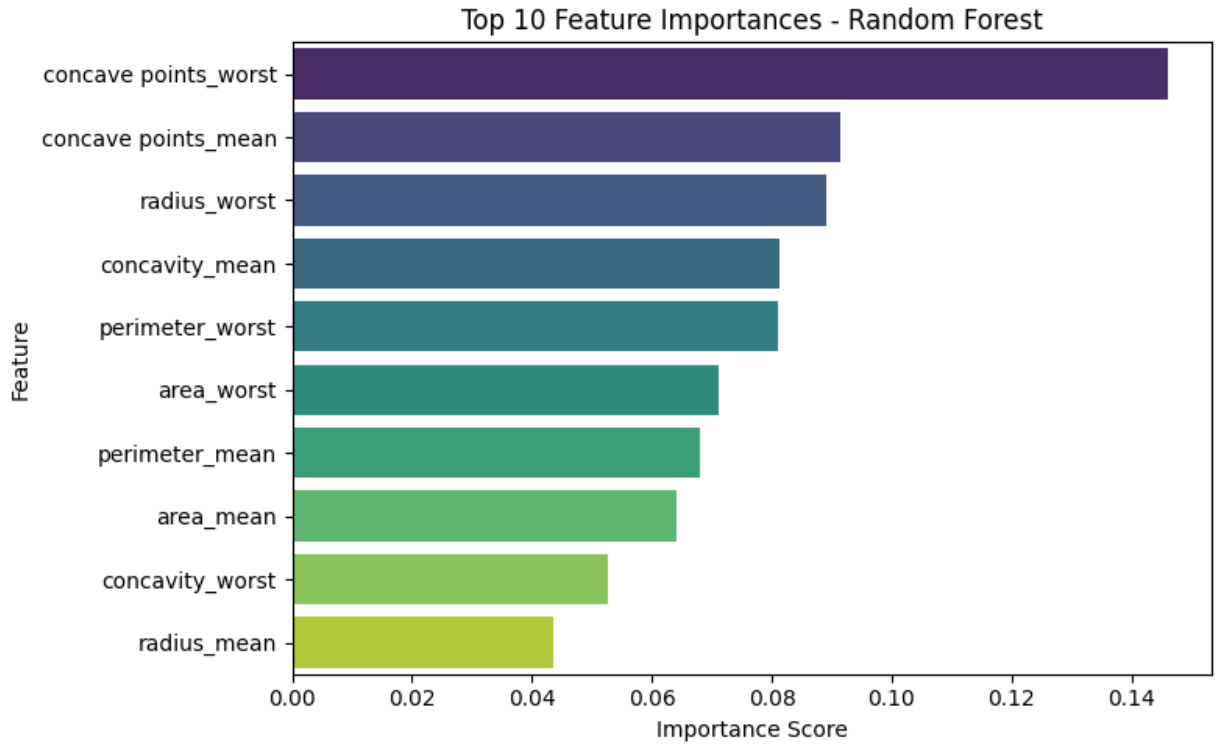
***Insights*:** Random Forest consistently came out on top with the best accuracy, precision, and recall.

Recommendation: Stick with Random Forest for production use, but keep exploring other models like Gradient Boosting for potential improvements.

Conclusion: This comparison confirmed that Random Forest is the most dependable choice for this dataset.

Top 10 Feature Importances – Random Forest

Objective: To identify which features contribute the most to predictions.



Insights:

- The Random Forest model highlighted **concave points_worst** and **concave points_mean** as the strongest predictors of malignancy.
- Tumor size and shape features such as **radius_worst, area_worst, and perimeter_worst** also ranked highly.
- This aligns with medical understanding, where irregular shapes and larger tumors often indicate cancer.

Recommendation: Diagnostic tools should prioritize these top features since they carry the most weight in decision-making.

Conclusion: This confirmed that tumor concavity and size-related features are critical in breast cancer detection.

STREAMLIT APP USER INTERFACE

The Breast Cancer Detection App provides an interactive platform for users to enter tumor measurements and receive predictions.

Key Features:

- **Branding & Awareness:** The app integrates a **Breast Cancer Awareness logo** and clear instructions for usability.
- **User Guidance:** Each input field has tooltips, ensuring non-technical users can understand the significance of features.
- **Example Data:** Pre-filled values are available for testing, based on dataset averages.
- **Visual Feedback:** After prediction, a bar chart highlights the **entered feature values** and the **top influencing features**.

Real-World Use Case:
While educational, this app demonstrates how AI-powered systems can support clinicians by providing **quick, reliable, and interpretable predictions**, reducing manual workload and assisting in early detection.

PROJECT SUMMARY

In this capstone project, a simple but powerful tool was built 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.

RECOMMENDATIONS:

From the insights gained in this project, here are some clear next steps:

1. **Handle data imbalance carefully:** Since benign cases are more than malignant ones, balancing techniques (like oversampling or class weights) should be applied to ensure the model treats both classes fairly.

2. **Focus on the most impactful features:** Some features overlap in the information they provide. Removing redundant ones will make the model simpler, faster, and just as accurate.

3. **Reduce false negatives:** While the model performed very well, those 3 missed cancer cases are critical in real life. More tuning and exploring models like Gradient Boosting could help reduce them further.

4. **Keep the model updated:** As new medical data becomes available, retraining the model will help keep predictions relevant and trustworthy.

5. **Support tool, not replacement:** This model is best seen as a decision support system to complement doctors, not replace their expertise.

LIMITATIONS & FUTURE WORK

- **Dataset Size:** The Wisconsin dataset is relatively small and may not represent global populations.
- **Class Imbalance:** Benign cases outnumber malignant ones, potentially biasing predictions.
- **Explainability:** Random Forest is powerful but not easily interpretable. More interpretable tools (e.g., **SHAP, LIME**) should be added for clinicians.
- **External Validation:** The model has not been tested with hospital datasets or imaging data.

FUTURE WORK:

- Explore larger, more diverse datasets for better generalization.
- Integrate ensemble methods such as Gradient Boosting or XGBoost.
- Include AUC-ROC and calibration plots in the evaluation for more reliability.
- Combine clinical features with imaging data for hybrid diagnostic tools.

CONCLUSIONS:

This project was not just about building a machine learning model, it was about discovering how **data science and healthcare can work together**.

By cleaning the dataset, exploring it visually, and testing different models, a tool that reached **97% accuracy** was developed and proved reliable for breast cancer prediction.

Each visualization told part of the story: correlations revealed hidden patterns, the confusion matrix showed model strengths and weaknesses, and model comparisons confirmed the best approach.

Beyond the numbers, this project highlighted the importance of balancing technical accuracy with real-world impact. It shows that when machine learning is applied thoughtfully, it can support medical professionals, improve decision-making, and potentially save lives.

REFERENCES

UCI Machine Learning Repository, Breast Cancer Wisconsin (Diagnostic) Dataset.
Scikit-learn Documentation: https://scikit-learn.org/stable/
Streamlit Documentation: https://docs.streamlit.io/
Plotly Express Documentation: https://plotly.com/python/plotly-express/
Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.

ZION TECH HUB