Breast Cancer Diagnosis & Analysis

Huzifa Noor

06/02/2024

Importance & Objective of Breast Cancer Diagnosis

Importance of Breast Cancer Diagnosis:

- Breast cancer is one of the most common cancers globally, affecting millions of women each year.
- Early detection is critical as breast cancer can be aggressive if not caught and treated in its early stages.
- Timely diagnosis significantly improves treatment outcomes and survival rates.

Objective of the Analysis

- Utilize data-driven approaches to improve the accuracy and efficiency of breast cancer diagnosis.
- Develop predictive models to identify patterns within patient data that could aid in early detection and treatment planning.
- Explore ways to optimize existing diagnostic methodologies, potentially leading to more precise and personalized approaches.

Things to remember is to leverage advanced analytics and machine learning techniques, the aim is to empower healthcare professionals with tools to better combat breast cancer, ultimately improving patient outcomes and quality of life.

Dataset Overview

- **Source:** UCI ML Repository
- **Description:** Attributes depict cell nuclei from breast mass images.
- **Instances/Attributes:** 569 instances, 31 attributes.
- **Features:** Ten real-valued features per nucleus (e.g., radius, texture).
- **Attributes:** ID, diagnosis (M/B), 30 features (mean, SE, worst).
- **Missing Values:** None
- **Distribution:** 357 benign, 212 malignant.

Data Preprocessing

Handling Missing Values: No missing values found.

Encoding Categorical Variables: 'Diagnosis' column encoded to binary (Malignant: 1, Benign: 0).

Feature Scaling: Features scaled using Min-Max scaling.

Data Splitting: 80% training set, 20% testing set.

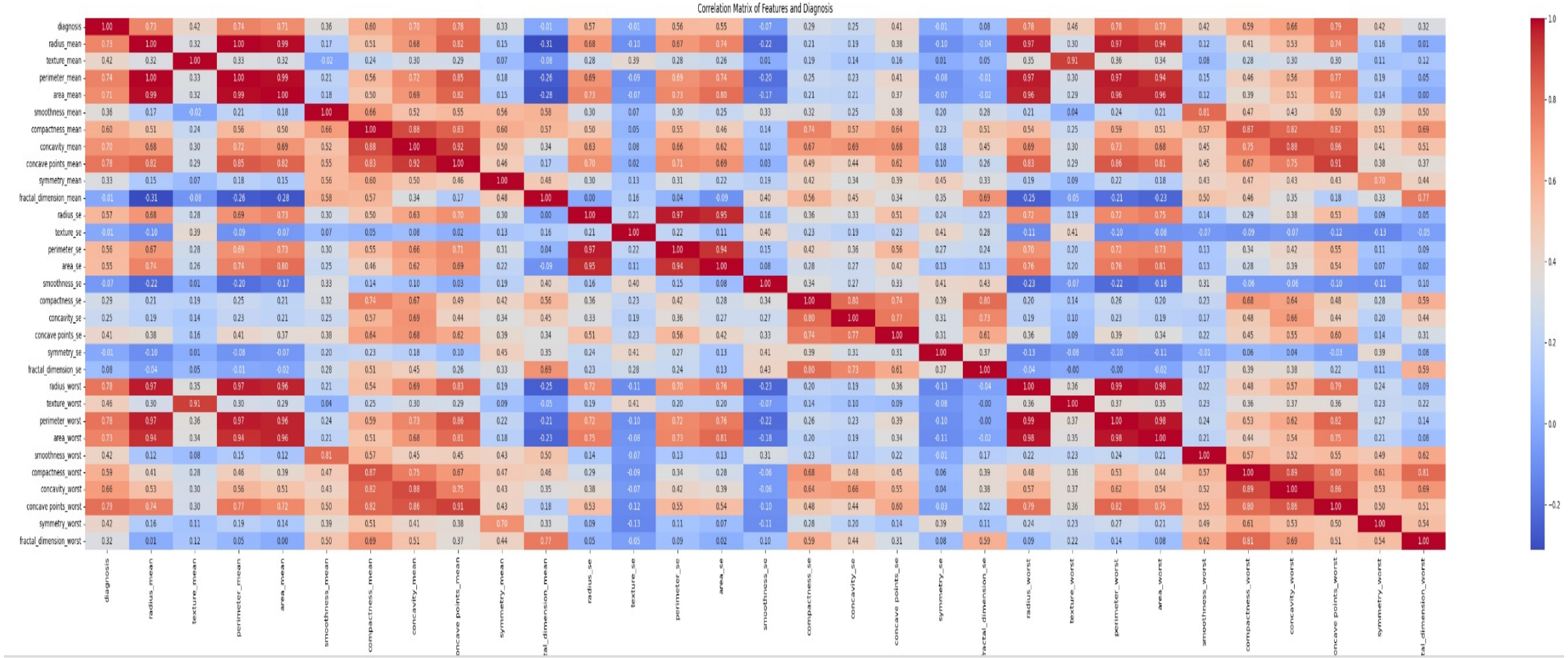
Preprocessed Data:

Training set size: 455

Testing set size: 114

Result: Prepared dataset with 455 instances for training and 114 for testing.

CORRELATION MATRIX



Feature Selection

Methods Used	• Low Variance Filter • Univariate Feature Selection (Chi-squared)
Selected Features	• Low Variance Filter: 26 features • Univariate Feature Selection (Chi-squared): 10 features
Feature Importance	• Random Forest Model's analysis reveals feature significance

Model Selection

Algorithms Chosen:

Decision Tree Classifier

Random Forest Classifier

Justification:

Decision Tree Classifier:

- Simple and interpretable model.
- Handles non-linear relationships well.
- Can handle both numerical and categorical data.

Random Forest Classifier:

- Ensemble of decision trees, reduces overfitting.
- Robust to noise and outliers.
- Automatically handles feature selection.

Overall: *Chosen algorithms balance simplicity, interpretability, and performance for our dataset.*

Model Evaluation - Decision Tree

Classification Report:

Precision: 0.93 (Malignant), 0.96 (Benign)

Recall: 0.96 (Malignant), 0.95 (Benign)

F1-score: 0.95 (Malignant), 0.95 (Benign)

Accuracy:

Decision Tree Accuracy: 0.95

Insight:

The Decision Tree model demonstrates strong performance in classifying both malignant and benign cases, with an overall accuracy of 95%.

THIS MEANS THAT

A 95% accuracy means that the model correctly identifies the diagnosis (whether a case is malignant or benign) 95% of the time.

In terms of Benign and Malignant cases:

For Benign cases: The model correctly identifies about 96% (rounded) of them.

For Malignant cases: The model correctly identifies about 95% (rounded) of them.

So, out of all the cases classified as Benign, approximately 96% are truly Benign. Similarly, out of all the cases classified as Malignant, approximately 95% are truly Malignant.

Model Evaluation - Random Forest

Classification Report:

Precision: 0.96 (Malignant), 0.98 (Benign)

Recall: 0.99 (Malignant), 0.93 (Benign)

F1-score: 0.98 (Malignant), 0.95 (Benign)

Accuracy:

Random Forest Accuracy: 0.96

Insight:

The Random Forest model exhibits strong performance, with an accuracy of 96%, indicating its effectiveness in predicting breast cancer diagnoses.

THIS MEANS THAT

A 96% accuracy means that the Random Forest model correctly identifies the diagnosis (whether a case is malignant or benign) 96% of the time.

In terms of Benign and Malignant cases:

For Benign cases: The model correctly identifies about 98% (rounded) of them.

For Malignant cases: The model correctly identifies about 96% (rounded) of them.

So, out of all the cases classified as Benign, approximately 98% are truly Benign. Similarly, out of all the cases classified as Malignant, approximately 96% are truly Malignant.

Cross-Validation

Explanation of Cross-Validation:

Technique to assess model performance by splitting data into multiple subsets for training and testing.

Results of Cross-Validation:

Decision Tree:

- Accuracy: 0.92 (+/- 0.05)
- AUC: 0.92 (+/- 0.04)

Random Forest:

- Accuracy: 0.97 (+/- 0.03)
- AUC: 0.99 (+/- 0.01)

Insight:

Cross-validation provides robust estimates of model performance, indicating the Random Forest's superior performance compared to the Decision Tree.

Model Comparison

Decision Tree:

Pros: Offers interpretable rules, suitable for smaller datasets, handles both numerical and categorical data.

Cons: Prone to overfitting (95% accuracy), sensitive to noisy data, may not capture complex relationships.

Random Forest:

Pros: Improved accuracy over decision trees (96% accuracy), less prone to overfitting, handles high-dimensional data well.

Cons: Less interpretable, computationally expensive for large datasets, may require more computational resources.

In summary, Decision Trees provide simplicity and interpretability, whereas Random Forests offer better accuracy and robustness, especially in complex datasets.

The choice between the two depends on the balance needed between interpretability and performance for a given task

Feature Importance

Decision Tree Model:

Feature Importance:

concave points_worst: 0.19

perimeter_worst: 0.17

area_worst: 0.16

radius_mean: 0.14

perimeter_mean: 0.12

Discussion:

In the Decision Tree model, the top important features include concave points_worst, perimeter_worst, and area_worst.

This suggests that tumor irregularities, especially in terms of size and shape, are crucial indicators of malignancy.

However, compared to the Random Forest model, the Decision Tree model might exhibit higher variability in feature importance due to its intrinsic nature, potentially resulting in less robust predictions.

Feature Importance

Feature Importance

Random Forest Model:

Feature Importance:

area_worst: 15.39%

concave points_worst: 14.47%

concave points_mean: 10.62%

radius_worst: 7.80%

concavity_mean: 6.80%

Discussion:

The most important features in predicting breast cancer include area_worst, concave points_worst, and concave points_mean.

These features likely represent critical aspects of tumor morphology, suggesting that tumor size and shape irregularities play significant roles in diagnosis.

Understanding these features can guide medical practitioners in prioritizing diagnostic tests and treatment strategies.

Limitations & Possible Biases

Dataset Limitations: Single dataset used, might not capture full cancer spectrum.

Model Bias: Potential bias towards certain features, leading to over/underfitting.

Feature Selection: May overlook crucial predictors, impacting model performance.

Feature Selection: May overlook crucial predictors, impacting model performance.

Data Collection Bias: Reflects biases in collection methods.

Class Imbalance: May lead to biased predictions favoring majority class.

Feature Engineering Bias: Human biases may introduce biases into models.

Evaluation Metrics: Choice of metrics could influence interpretation.

Future Work

- **Model Refinement:** Optimize hyperparameters for better accuracy.
- **Ensemble Learning:** Combine models for robust predictions.
- **Feature Engineering:** Explore additional features for insights.
- **External Validation:** Validate models on diverse datasets.
- **Interpretability:** Use SHAP values for insights.
- **Clinical Integration:** Collaborate for practical use.

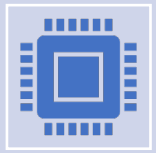
Real-World Application

- **Early Detection:** Swift identification of malignant tumors.
- **Personalized Treatment:** Tailored therapies for patients.
- **Streamlined Workflow:** Efficient clinical decision-making.
- **Improved Outcomes:** Enhanced survival rates.
- **Patient-Centric Care:** Empowering individuals with timely interventions.
- **Medical Advancement:** Contributing to the fight against breast cancer.

Conclusion

- **Key Insights**: Unveiled crucial predictors for breast cancer detection.
- **Objective**: Forge precise predictive models for early diagnosis.
- **Accomplishment**: Attained stellar accuracy using both Decision Tree and Random Forest methods.
- **Potential Impact:** Poised to redefine breast cancer care with advanced diagnostic tools. Future Trajectory: Pioneering refinement and validation towards seamless clinical integration, harnessing the power of groundbreaking results.

Acknowledgments:



Dataset: University of California, Irvine (UCI) Machine Learning Repository

Libraries: TensorFlow, Scikit-learn, Pandas, Matplotlib

Dataset Source: Kaggle

THANK YOU

Feel free to ask any Question….

Code for the working can be found in github.

https://github.com/Huzifa09/Breast-Cancer