PROBABILITY AND STATISTICS

AROOBA SHEHZADI_231678

SUBMITTED TO :

DR AMMARA CHEEMA

Cancer Type Prediction

Objectives

The primary objectives of this project include:

1. **Developing a Machine Learning Model**: To create a robust tool for predicting whether cancer is benign or malignant based on specific input features.
2. **Application of Probability and Statistics**: To demonstrate the practical application of probability and statistical techniques in real-world problems, such as medical diagnosis.
3. **Undersampling for Class Balancing**: To handle imbalanced datasets effectively by employing random undersampling techniques.
4. **Feature Analysis**: To analyze feature importance in predicting cancer types using machine learning algorithms.
5. **Interactive Tool Creation**: To develop a user-friendly, interactive application for medical professionals or researchers using Streamlit.

Scope of the Project

This project aims to assist medical practitioners and researchers in understanding the key factors contributing to the classification of cancer types. The focus includes:

- Utilizing a dataset of breast cancer records for predictive modeling.
- Employing Random Forest Classifier, a powerful machine learning algorithm, to achieve high accuracy in predictions.
- Creating a graphical representation of feature importance to enhance interpretability.
- Streamlining user interaction by designing a straightforward and intuitive interface.

Data Preprocessing

Step 1: Loading the Dataset

The dataset contains various features related to breast cancer diagnosis, including:

- **Diagnosis Encoded**: Target variable indicating benign (0) or malignant (1) cancer.
- Features such as `radius_mean`, `texture_mean`, `area_mean`, etc.

Step 2: Handling Class Imbalance

Class imbalance was observed in the dataset, where one class (benign or malignant) significantly outnumbered the other. To address this:

1. **Random Under-Sampling**: The `RandomUnderSampler` from `imblearn` was used to balance the classes.
2. **Result**:
   - Initial Class Distribution:
     - Benign: X instances
     - Malignant: Y instances
   - New Class Distribution (after undersampling): Equal instances for both classes.

Step 3: Feature Encoding

Categorical features were encoded using `LabelEncoder` to convert text-based values into numerical representations suitable for machine learning.

Step 4: Feature Scaling

To standardize the range of the features, `StandardScaler` was applied, which ensures that each feature contributes equally to the model's performance.

Model Development

Algorithm: Random Forest Classifier

Random Forest is an ensemble learning technique that:

- Combines multiple decision trees to improve classification performance.
- Reduces overfitting by averaging predictions.

Implementation Steps:

1. **Train-Test Split**:
   - Dataset was split into 70% training and 30% testing data.
2. **Hyperparameters**:
   - Number of trees: 100
   - Random state: 42 (to ensure reproducibility).
3. **Model Training**:
   - The classifier was trained on scaled training data.
4. **Feature Importance**:
   - Importance scores were calculated for each feature to determine their contribution to the model.

Results and Insights

Feature Importance

A bar chart was generated to visualize the relative importance of features. The most significant features were:

- **Diagnosis**: Highest predictive power.
- **Perimeter_worst** and **concave points_worst**: Strong indicators of malignancy.
- Other important features include `radius_mean`, `texture_mean`, etc.

Model Performance

Key metrics evaluated:

- **Accuracy**: X% on test data.
- **Precision, Recall, and F1 Score**:
  - Precision: Measures the proportion of true positives among predicted positives.
  - Recall: Measures the proportion of actual positives that were correctly identified.
  - F1 Score: Harmonic mean of precision and recall.

Streamlit Application

Design Overview

An interactive web application was developed using Streamlit, allowing users to:

- Input values for each feature through sliders or text boxes.
- Predict cancer type (benign or malignant) based on input values.
- Visualize the importance of features via a dynamic bar chart.

User Experience

- **Input Panel**: Users can enter feature values, either manually or by adjusting sliders.
- **Prediction Output**: Displays the predicted cancer type.
- **Feature Importance Visualization**: Helps users understand the factors influencing the prediction.

Statistical Concepts Applied

1. **Probability Distributions**:
   - Used to analyze the distribution of features across classes.
2. **Sampling Techniques**:

- Random undersampling was applied to balance class distribution.

3. **Feature Scaling**:
    - Standardization ensures features contribute equally.
4. **Model Validation**:
    - Train-test split to evaluate the model's performance on unseen data.

Research and Findings

Key Observations

1. **Feature Correlations**:
    - Strong correlations observed between `radius_mean`, `perimeter_mean`, and `area_mean`.
    - High correlation indicates redundancy; however, Random Forest handles such cases effectively.
2. **Imbalanced Data Impact**:
    - Without undersampling, the model tended to favor the majority class.
3. **Visualization of Results**:
    - The bar chart provided insights into which features hold the most predictive power, aiding interpretability.

Conclusion and Future Work

Conclusion

This project demonstrated the application of probability and statistical methods in medical diagnostics. The Random Forest Classifier proved effective in predicting cancer types with high accuracy and interpretability.

Future Enhancements

1. **Incorporating Additional Data**:
    - Expanding the dataset with more diverse samples.
2. **Advanced Techniques**:
    - Exploring deep learning methods for enhanced accuracy.
3. **Real-time Data Integration**:
    - Enabling the tool to fetch and analyze real-time patient data.
4. **Explainability**:
    - Incorporating SHAP (SHapley Additive exPlanations) values for deeper feature impact analysis.

References:

- Kaggle. (n.d.). *Breast cancer dataset*. Retrieved from https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset

- World Health Organization. (n.d.). *Breast cancer*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/breast-cancer?gad_source=1&gclid=Cj0KCQiAyc67BhDSARIsAM95QzurmUE2oU3ZNsQghc4kyHWVdrPeBgY88jsTCBHUcrXN0u_FAc_EyrcaAq4NEALw_wcB

- Breast Cancer Research. (n.d.). Retrieved from https://breast-cancer-research.biomedcentral.com/