Capstone Project 1

Predictive Modeling for Breast Cancer Using KNN, Naive Bayes, Decision Tree, Random Forest

Name: CHEVALA AKSHAY RAJ

Batch: DST 20923

Start Date: SEPTEMBER 15 2023

1) problem statement:

The problem at hand is to develop predictive models for breast cancer diagnosis using machine learning techniques. The dataset contains features extracted from digitized images of fine needle aspirates (FNA) of breast masses, and the goal is to predict whether a mass is benign or malignant based on these features. The importance of this problem lies in providing a reliable and automated tool for early detection of breast cancer, which can significantly improve patient outcomes through timely intervention.

2) List the techniques/approaches that are used:

- **k-Nearest Neighbours (k-NN):** A simple and intuitive algorithm that classifies a data point based on the majority class of its k-nearest neighbours in the feature space.
- **Decision Tree:** A tree-like model that makes decisions based on splitting features at different nodes, providing interpretability and visualization of the decision-making process.
- **Random Forest:** An ensemble of decision trees that aggregates their predictions, often leading to improved accuracy and generalization.
- **Bernoulli Naive Bayes:** Based on Bayes' theorem, this probabilistic algorithm is suitable for binary classification tasks and assumes independence between features.

3) Why using these techniques:

- **Diversity of Approaches:** By employing a variety of algorithms, we ensure a comprehensive exploration of different model architectures and their suitability for the given problem.
- **Comparative Analysis:** Using multiple techniques allows us to compare their performance, understand their strengths and weaknesses, and select the most suitable model for the specific task.

4) Steps used or followed:

1. **Data Loading and Exploration:** Loaded the dataset and performed exploratory data analysis (EDA) to understand the distribution of features and the target variable.
2. **Data Preprocessing:** Handled missing values, dropped unnecessary columns, and converted categorical variables into numerical format.

3. **Feature Standardization:** Standardized features to ensure uniform scales, preventing any particular feature from dominating the model.
4. **Model Training:** Employed k-NN, Decision Tree, Random Forest, and Bernoulli Naive Bayes for training predictive models.
5. **Model Evaluation:** Assessed model performance using accuracy, confusion matrix, and classification reports.
6. **Visualization:** Created visualizations such as confusion matrices and decision tree diagrams for better interpretation.

5) Results :

- **k-NN:** Accuracy - 96.49%
- **Decision Tree:** Accuracy - 95.61%
- **Random Forest:** Accuracy - 99.12%
- **Bernoulli Naive Bayes:** Accuracy - 96.49%

6) Qualities:

- **Accuracy:** Achieved high accuracy across various models, indicating their effectiveness in predicting breast cancer diagnoses.
- **Interpretability:** Decision Tree and Random Forest models provide insights into feature importance and decision pathways.
- **Robustness:** Models demonstrated consistent performance on the test set, suggesting their generalization ability.

7) Suggestions:

- **Optimization:** Further hyperparameter tuning may enhance the performance of the models.
- **Ensemble Methods:** Consider exploring ensemble methods to combine the strengths of different models.
- **Domain Expertise:** Collaborate with medical experts to incorporate domain knowledge and potentially improve model interpretability.
- **Continuous Monitoring:** Regularly update and reevaluate models with new data to ensure their relevance and accuracy over time.
- **Ethical Considerations:** Prioritize ethical considerations when deploying models in a medical context, ensuring patient privacy and informed consent.