Comsats University Islamabad

Assignment 2

Name:	Muhammad Aqib
Registration Number:	FA24-RAI-011
Submitted To:	Dr. Muhammad Imran
Date:	29-Nov-2024

Report on Classification Models for Breast Cancer Diagnosis

Objective:
The goal of this project was to look for a research paper of W category, which includes code, dataset and EDA. I chose a research paper which was centered around **Breast Cancer Tumor detection** and classification of tumors whether the tumor is malignant or benign. I compared and evaluated various machine learning models for distinguishing between malignant and benign cases in a breast cancer dataset. The dataset included features extracted from tumor measurements and was used to train and test multiple classifiers. I performed various Eda steps and then trained several machine learning models mentioned in the paper and set their accuracies side by side with the paper stated accuracy there were slight deficiencies in my accuracies due to lack of prevailing knowledge. The model accuracies are stated in the table below.

Reason for behind my model performance:

My model under performed a little when set side by side with the models in the papers here is what I think could be the potential reason for this.
No space for Improvement: the accuracy of the SVM, LR, KNN, and EC mentioned in the paper was already too high for me to surpass it.

Dataset Customized to each Model: The author of the paper had customized the dataset for each model specifically that's why the each model performed well on the dataset. Whereas I did generalize Eda steps not specific customization for each model.

Steps Taken:

1. **Exploration and Preprocessing:**
   - Loaded and analyzed the dataset for basic understanding.
   - Verified and handled missing or irrelevant data (if any).
   - Scaled the features using standardization to optimize model performance.
2. **Model Implementation:**
   - Implemented the following classifiers using a pipeline to ensure modularity and efficiency:
     - **Support Vector Machine (SVM):** Achieved high accuracy by optimizing hyperparameters such as C and kernel type.

- **Logistic Regression:** Used to establish a baseline, showing reasonable performance due to its simplicity.
- **K-Nearest Neighbors (KNN):** Tuned the number of neighbors to maximize accuracy.
- **Extreme Gradient Boosting (XGBoost):** Achieved the highest accuracy by leveraging its ensemble nature and ability to handle feature importance.

3. **Evaluation Metrics:**
   - Accuracy was used as the primary metric to evaluate the classifiers.
   - Additional metrics like confusion matrix, precision, and recall could be explored for deeper insights.

Results Summary:

Model	**Accuracy**
Support Vector Machine	**98%**
Logistic Regression	**97%**
K-Nearest Neighbors	**94%**
Extreme Gradient Boosting	**96%**

GitHub Repository Link:
https://github.com/Aqibkhan037/Artificial-intelligence.git

Conclusion:
The project demonstrates how different machine learning techniques can be applied to the same dataset and yield varying results. XGBoost emerged as the most effective model in terms of accuracy, reflecting its robustness in handling structured datasets. The modular pipeline approach used ensures scalability and adaptability to other datasets or problems.

Further exploration can include feature engineering, hyperparameter tuning, and testing with additional datasets to enhance model performance.