Leticia Genao

May 9, 2024

ANA680 Week 1 Homework

Breast Cancer Diagnosis: A Multi-Model Machine Learning Approach

**Introduction**

This report presents a comparison of various classification models to predict breast cancer using a dataset obtained from the UCI Machine Learning Repository. The primary objective is to evaluate and compare the performance of several machine learning models, including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Naïve Bayes, Decision Tree, Random Forest, and XGBoost, using default parameters.

**Dataset Description**

The dataset used in this study is the Breast Cancer Wisconsin (Original) Dataset from the UCI Machine Learning Repository. It comprises 683 instances, each with 10 features, and a class label that indicates whether the breast cancer is benign (label = 2) or malignant (label = 4).

**Methodology**

**Data Preprocessing**

The class labels were converted from 2 (benign) and 4 (malignant) to 0 and 1 respectively, to facilitate the use of binary classification techniques. No additional cleaning or feature engineering was performed to maintain the integrity of the original data as per the assignment instructions.

**Model Building**

Each model was implemented using Python with the Scikit-learn library. The dataset was randomly split into training (75%) and testing (25%) sets to evaluate the models' performances.

**Model Evaluation**

The models were evaluated based on their accuracy and the confusion matrix. These metrics help in understanding the models' capabilities in correctly predicting the class labels.

**Results**

The following tables and figures summarize the performance of each model:

|  |  |  |
| --- | --- | --- |
| **Logistic Regression** | | |
| **Accuracy: 95.32%** | | |
|  | **Predicted: Benign** | **Predicted: Malignant** |
| **Actual: Benign** | TN: 102 | FP: 1 |
| **Actual: Malignant** | FN: 7 | TP: 61 |

|  |  |  |
| --- | --- | --- |
| **KNN (k=5)** | | |
| **Accuracy: 94.73%** | | |
|  | **Predicted: Benign** | **Predicted: Malignant** |
| **Actual: Benign** | TN: 102 | FP: 1 |
| **Actual: Malignant** | FN: 8 | TP: 6 |

|  |  |  |
| --- | --- | --- |
| **SVM (Linear)** | | |
| **Accuracy: 95.32%** | | |
|  | **Predicted: Benign** | **Predicted: Malignant** |
| **Actual: Benign** | TN: 102 | FP: 1 |
| **Actual: Malignant** | FN: 7 | TP: 61 |

|  |  |  |
| --- | --- | --- |
| **SVM (RBF)** | | |
| **Accuracy: 94.73%** | | |
|  | **Predicted: Benign** | **Predicted: Malignant** |
| **Actual: Benign** | TN: 101 | FP: 2 |
| **Actual: Malignant** | FN: 7 | TP: 61 |

|  |  |  |
| --- | --- | --- |
| **Naïve Bayes** | | |
| **Accuracy: 96.49%** | | |
|  | **Predicted: Benign** | **Predicted: Malignant** |
| **Actual: Benign** | TN: 100 | FP: 3 |
| **Actual: Malignant** | FN: 3 | TP: 65 |

|  |  |  |
| --- | --- | --- |
| **Decision Tree** | | |
| **Accuracy: 94.15%** | | |
|  | **Predicted: Benign** | **Predicted: Malignant** |
| **Actual: Benign** | TN: 101 | FP: 2 |
| **Actual: Malignant** | FN: 8 | TP: 60 |

|  |  |  |
| --- | --- | --- |
| **Random Forest** | | |
| **Accuracy: 93.56%** | | |
|  | **Predicted: Benign** | **Predicted: Malignant** |
| **Actual: Benign** | TN: 102 | FP: 1 |
| **Actual: Malignant** | FN: 10 | TP: 58 |

|  |  |  |
| --- | --- | --- |
| **XGBOOST** | | |
| **Accuracy: 95.32%** | | |
|  | **Predicted: Benign** | **Predicted: Malignant** |
| **Actual: Benign** | TN: 102 | FP: 1 |
| **Actual: Malignant** | FN: 7 | TP: 61 |

**True Positives (TP)**: The number of correctly predicted malignant cases.

**False Positives (FP)**: The number of benign cases incorrectly classified as malignant.

**True Negatives (TN)**: The number of correctly predicted benign cases.

**False Negatives (FN)**: The number of malignant cases incorrectly classified as benign.

**Discussion**

The models' performances differed, with Random Forest exhibiting the lowest accuracy at 93.56% and Naïve Bayes the greatest at 96.49%. The reasons for this variation in performance may be traced back to a number of inherent aspects of how each model handles the properties of the data and how sophisticated their algorithms are.

* **Naïve Bayes** performed well, probably because of its efficiency in binary classification tasks and its assumption of feature independence, which seemed to work well with this dataset despite the real interdependencies.
* **XGBoost**, **SVM (Linear)**, and **Logistic Regression** all reported accuracy rates of 95.32%. The robustness of these models in drawing linear decision boundaries makes them appropriate for the dataset's apparent linear separability.
* Given that **SVM (RBF)** and **KNN (k=5)** both have accuracy rates of 94.73%, it is possible that noise and outliers had a minor impact due to their reliance on radial basis functions and neighbor closeness, which might cause overfitting.
* The results from **Decision Tree** and **Random Forest** were not very impressive, which could be attributed to their distinct overfitting and underfitting tendencies. Without adjustment, the decision tree's simplicity and the forest's randomness might not be able to adequately capture the dataset's complexity or patterns without overfitting to noise.

**Error Analysis**

The models' sensitivity and specificity are demonstrated by the false positives and false negatives that were noted for each model. Despite their high accuracy, models such as Naïve Bayes exhibited a propensity towards increased false positives, suggesting that they may have been overly sensitive in identifying instances as malignant.

**Conclusion**

This comprehensive comparison reveals which models perform better both qualitatively and statistically by examining how they manage prediction errors. For this dataset, XGBoost and Logistic Regression demonstrated a reasonable trade-off between sensitivity and specificity, making them appropriate options. The clinical cost of false positives vs false negatives, however, may influence the model selection. To further improve model accuracy and lessen potential biases, future research may investigate feature selection strategies, ensemble methods, and parameter tuning.

**References**

* Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [<https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)>Irvine, CA: University of California, School of Information and Computer Science.

**Appendix**

* GitHub Repository: [Breast\_Cancer\_Classification](http://github.com/leticiagenao/breast_cancer_classification)