University Of Arizona
College of Information Science

Analysis of Different Classification Model

Viren Sasalu, Bhaskar Bijoy Kashyap, Gaurangi Raul, Tushar Vimal
Team: Visual Voyagers
Date: 12-11-2024

Introduction

Breast cancer is a critical health issue where early detection improves outcomes. Using the Wisconsin Breast Cancer dataset, this project trains and evaluates three machine learning models: Logistic Regression (LG), Support Vector Machine (SVM), and Naive Bayes (NB). These models are assessed using key metrics like accuracy, precision, recall, and AUC (Area Under the Curve). Visualizations, including ROC curves and confusion matrices, compare their performance. The aim is to identify the most effective model for breast cancer diagnosis and demonstrate how machine learning can enable accurate, automated decision-making in healthcare.

Description of Data

The Wisconsin Breast Cancer dataset Which is a widely used benchmark dataset containing diagnostic features derived from fine needle aspiration biopsies of breast masses, used to classify tumors as benign or malignant.

Models Explored

- Logistic Regression (LR):
- Support Vector Machine (SVM)
- Naive Bayes (NB)

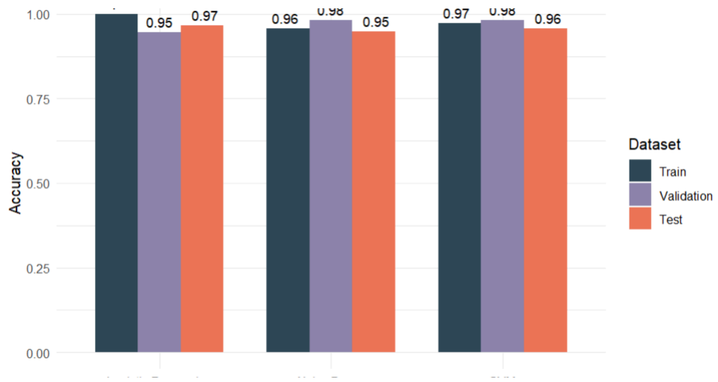

Fig 1 : Accuracy across different models

This chart compares the accuracy scores of Logistic Regression, Naive Bayes, and SVM models across training, validation, and test datasets. All three models demonstrate high and comparable accuracy, with slight variations, suggesting strong and consistent performance across datasets.

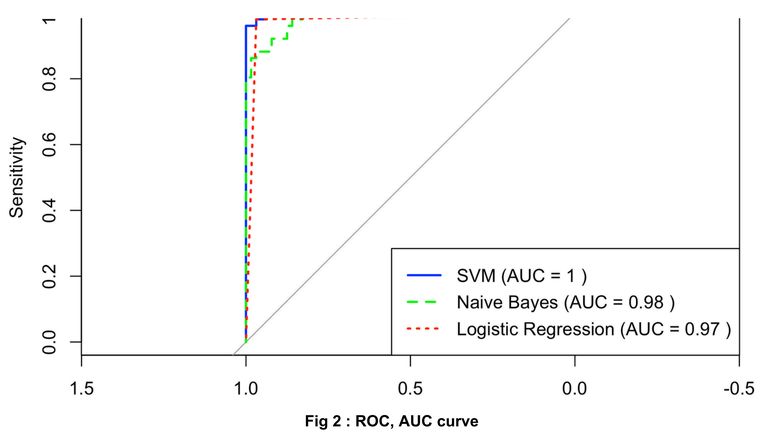

Fig 2 : ROC, AUC curve

This ROC curve comparison illustrates the performance of SVM, Naive Bayes, and Logistic Regression models, showing their respective AUC values. The SVM model achieves the highest AUC of 1.0, followed by Naive Bayes (0.98) and Logistic Regression (0.97).

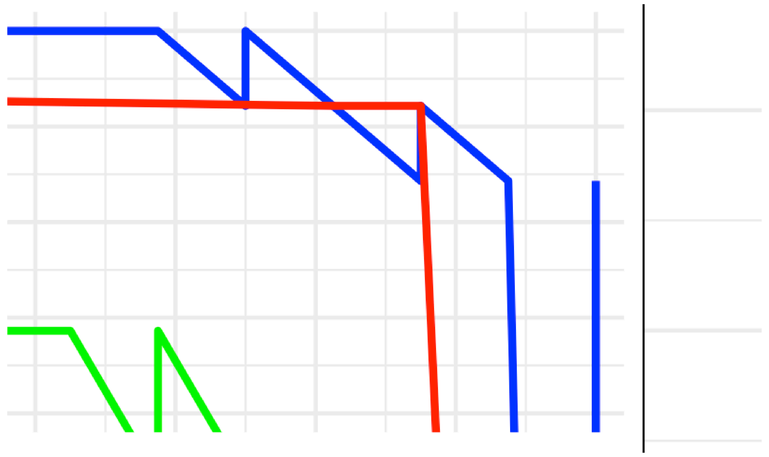

Fig 3 : Zoom in of figure 2

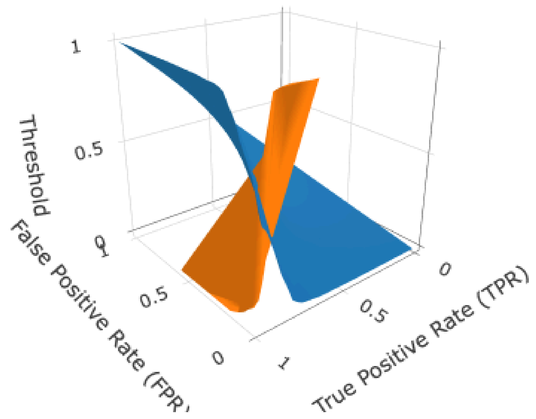

Fig 4 : ROC 3D plane

This 3D ROC surface plot compares the relationship between the True Positive Rate, False Positive Rate, and classification thresholds for two models. The intersecting surfaces suggest differing trade-offs between TPR and FPR at varying thresholds, providing a visual insight into model performance across threshold levels.

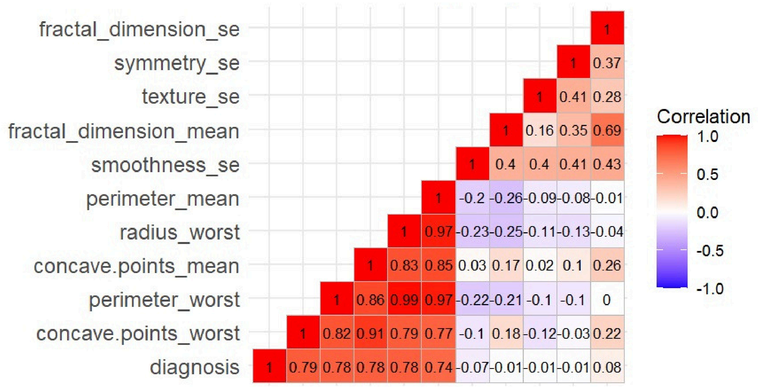

Fig 5 : Heatmap of correlation matrix

This heatmap visualizes the correlation matrix between various independent and dependent variables, with color intensity indicating the strength and direction of the correlation. Strong positive correlations (red) are observed between attributes like radius_mean, perimeter_mean, and area_mean, while weaker or negative correlations (blue) are evident for some attributes like fractal_dimension_mean.

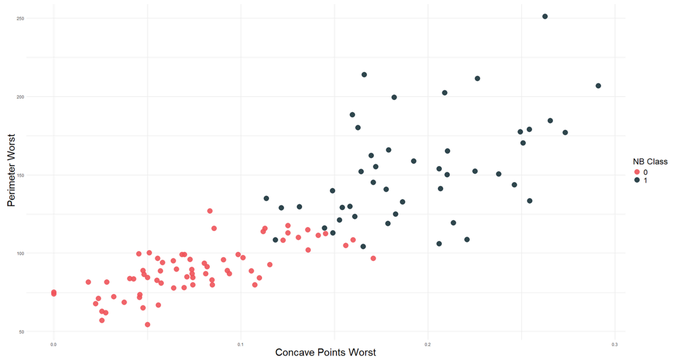

Fig 6 : Classification boundary by Naive Bayes using two of the best features

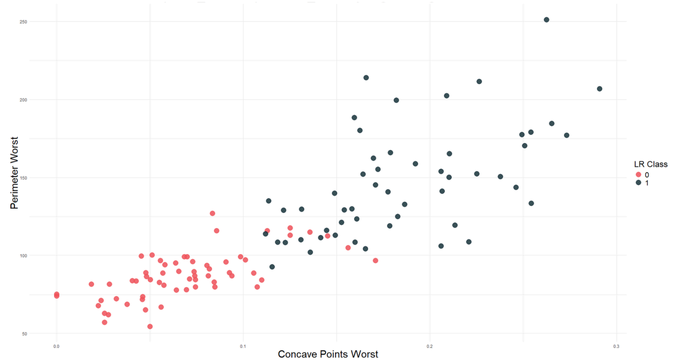

Fig 7 : Classification boundary by Logistic regression using two of the best features

These scatter plots compare the classification results of Logistic Regression (LR) and Naive Bayes models based on two highly correlated features, concave.points_worst and perimeter_worst, with points color-coded by predicted class. Both models show a clear separation of classes, but the Logistic Regression model appears to create more distinct class boundaries, particularly for higher values of the features.

Conclusion

Logistic Regression performed the best as it assumes linear separability between the target and features, which aligns well with the dataset's characteristics. while Naive Bayes performed the worst due to its assumption of feature independence, which was violated as many predictors were highly correlated. SVM is highly sensitive to hyper-parameter if the data is linearly separable might not offer significant advantages over logistic regression. Hence, model selection is very important.

Future Work

- Optimize the parameters of each model through hyperparameter tuning and analyze the resulting improvements.
- Investigate additional models like Decision Trees, KNNs, and compare their performance with the current ones.
- Study and visualize alternative model evaluation metrics for deeper insights.

References

- Dataset Source:

UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Dataset. Link

- Machine Learning Algorithms:

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.

- R Libraries Used:

Chang, W., et al. (2023). plotly: Create Interactive Web Graphics via 'plotly.js'. R package version 4.10.1.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

- Correlation Analysis:

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia & Analgesia, 126(5), 1763–1768.