IBM Machine Learning

Breast Cancer Prediction

Executive Summary

This report presents an analysis of the Diagnostic Wisconsin Breast Cancer Database using machine learning techniques. The objective of this project is to develop a predictive model that accurately classifies breast cancer cases as either malignant or benign based on clinical and pathological features. The analysis demonstrates the potential to improve early detection, treatment planning, and resource allocation in breast cancer management. Seven machine learning models were trained, evaluated, and compared based on performance metrics. Support Vector Classifier with radial basis function kernel achieves highest accuracy (0.991) and provides valuable insights into the factors driving breast cancer classification.

Introduction

Breast cancer is a significant public health concern, and accurate classification of cancer cases is critical for effective diagnosis and treatment. This analysis aims to develop a machine learning model that can assist in predicting the malignancy or benignancy of breast cancer based on relevant features. The insights gained from this analysis can have a direct impact on improving patient outcomes, treatment decisions, and resource allocation in the healthcare sector.

Methodology

The dataset used for this analysis is the Diagnostic Wisconsin Breast Cancer Database, which comprises clinical and pathological features such as 'id', 'diagnosis', 'radius_mean', 'texture_mean', 'perimeter_mean','area_mean', 'smoothness_mean', 'compactness_mean', 'concavity_mean','concave points_mean', 'symmetry_mean', 'fractal_dimension_mean', 'radius_se', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se','compactness_se', 'concavity_se', 'concave points_se', 'symmetry_se','fractal_dimension_se', 'radius_worst', 'texture_worst','perimeter_worst', 'area_worst', 'smoothness_worst', 'compactness_worst', 'concavity_worst', 'concave points_worst', 'symmetry_worst', 'fractal_dimension_worst'. The data preprocessing steps included renaming columns, normalizing numeric variables, encoding categorical variables, and splitting the data into training and testing sets using stratified strategy. Seven different machine learning algorithms, namely Logistic Regression, Random Forest, and Support Vector Machines, K Neighbors Classifier, Gaussian Naive Bayes Classifier, Decision Tree Classifier and Stochastic Gradient Descent Classifier were trained and evaluated for their performance.

Dataset Description

Features of the dataset are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.
Number of instances in the dataset is 569. Number of attributes are 32 (ID, diagnosis, 30 real-valued input features)

Attribute information

1. ID number
2. Diagnosis (M = malignant, B = benign)
3. Ten real-valued features are computed for each cell nucleus:
   - Radius (mean of distances from center to points on the perimeter)
   - Texture (standard deviation of gray-scale values)
   - Perimeter
   - Area
   - Smoothness (local variation in radius lengths)
   - Compactness (perimeter^2 / area - 1.0)
   - Concavity (severity of concave portions of the contour)
   - Concave points (number of concave portions of the contour)
   - Symmetry
   - Fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is

Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. All feature values are recorded with four significant digits. There are no missing attribute values in the dataset.

Data Exploration and Preprocessing

Exploratory data analysis revealed imbalance of the output variable as seen in fig1, hence, usage of stratified train-test split to maintain the ratio. Exploratory data analysis also revealed correlations between radius, perimeter and area seem to be high as shown in fig2, indicating their importance in the classification task. Numerical variables were normalized using Min Max Scaler, categorical variables were encoded using label encoding to facilitate model training.

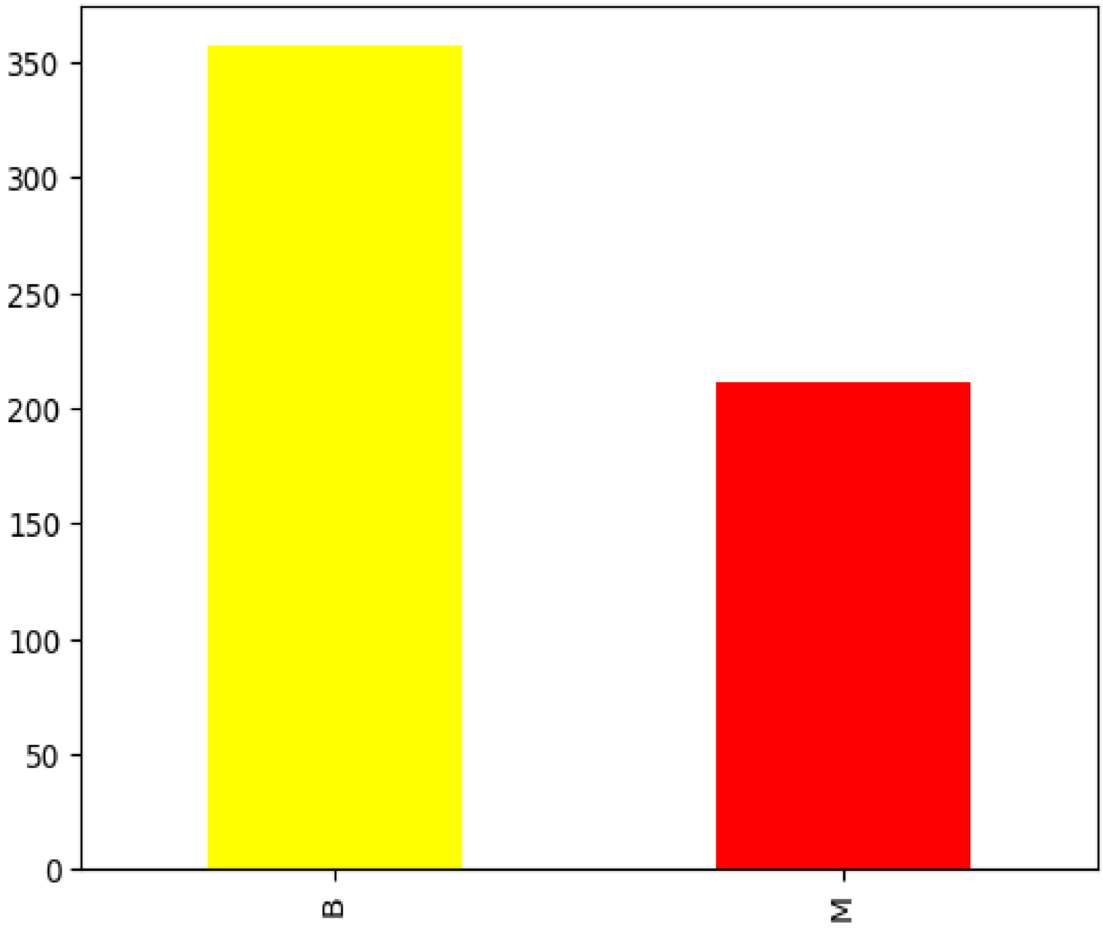

Fig.1

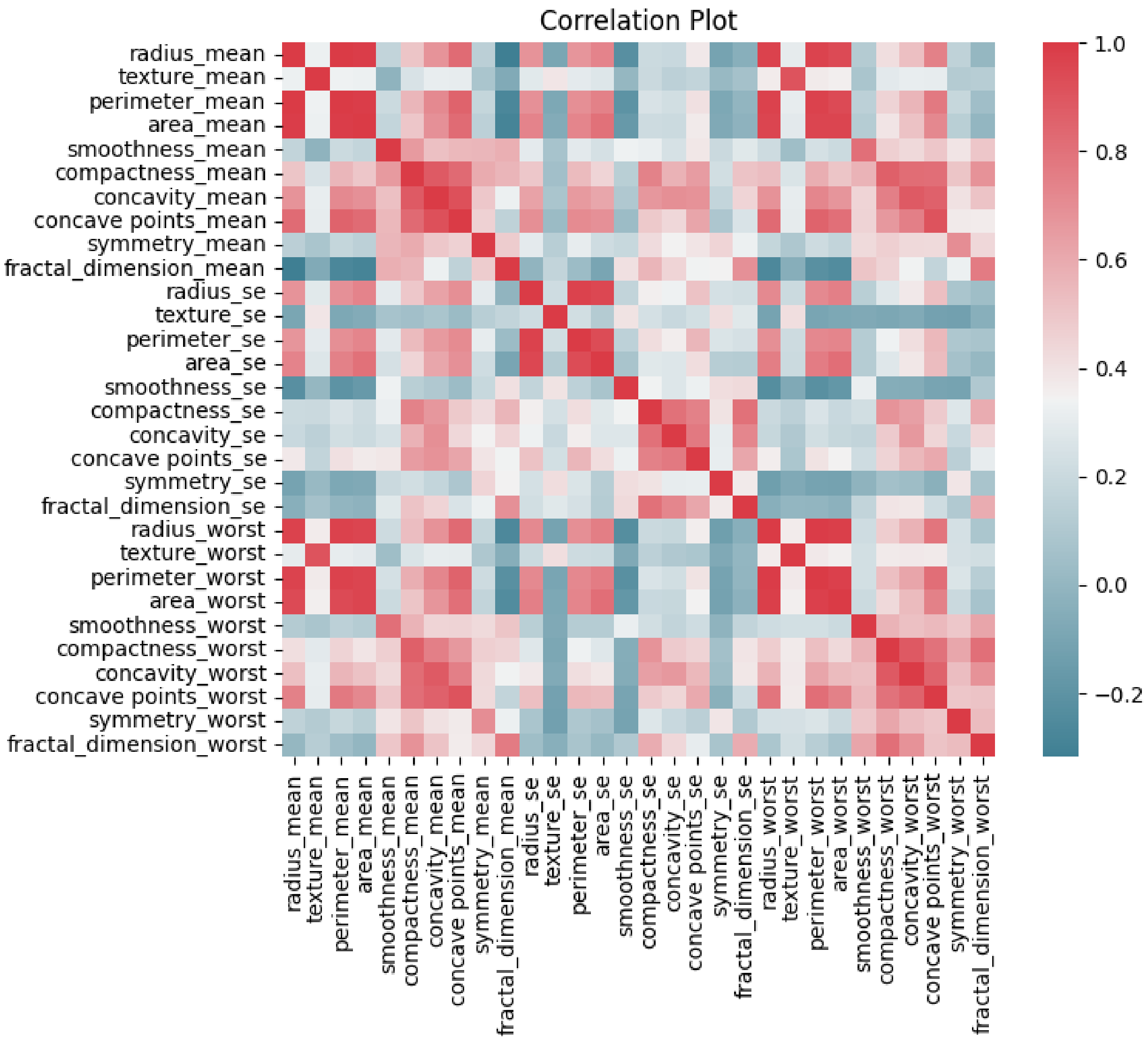

Fig. 2

Model Development and Evaluation

Seven different machine learning algorithms, namely Logistic Regression, Random Forest, and Support Vector Machines, K Neighbors Classifier, Gaussian Naive Bayes Classifier, Decision Tree Classifier and Stochastic Gradient Descent Classifier were trained on the preprocessed dataset. Model hyperparameters were tuned using Grid search CV to optimize performance. The models were evaluated based on accuracy. And the best performing model was Support Vector Classifier which was used for training with the best parameters achieved using Grid search CV. Results of Grid search CV are listed below:

- LogisticRegression()
    - best parameters are: {'C': 100.0, 'penalty': 'l2'}
    - accuracy: 0.9757004830917875
- RandomForestClassifier()
    - best parameters are: {'criterion': 'entropy', 'max_depth': 5, 'max_features': 'log2', 'n_estimators': 500}
    - accuracy: 0.9668115942028985
- SVC()
    - best parameters are: {'C': 10, 'gamma': 1, 'kernel': 'rbf'}
    - accuracy: 0.9779710144927536
- KNeighborsClassifier()
    - best parameters are: {'n_jobs': -1, 'n_neighbors': 13, 'weights': 'distance'}
    - accuracy: 0.9714009661835747
- GaussianNB()
    - best parameters are: {'var_smoothing': 0.001}
    - accuracy: 0.9337198067632851
- DecisionTreeClassifier()
    - best parameters are: {'ccp_alpha': 0.01, 'criterion': 'entropy', 'max_depth': 9, 'max_features': 'sqrt'}
    - accuracy: 0.9537681159420289
- SGDClassifier()
    - best parameters are: {'alpha': 0.001, 'loss': 'log', 'n_jobs': -1, 'penalty': 'l2'}
    - accuracy: 0.9779227053140097

Results

The evaluation of the models showed that all three achieved high accuracy on the test set. However, the Support Vector Machines (SVM) model outperformed the other models, demonstrating superior performance across multiple metrics. The SVM model achieved an accuracy of 0.991, precision of 1.0, sensitivity/recall of 0.976, specificity of 1.0, and an F1-score of 0.988. These metrics indicate the model's ability to correctly classify both malignant and benign breast cancer cases with high accuracy and reliability. Fig 3 shows the confusion matrix.

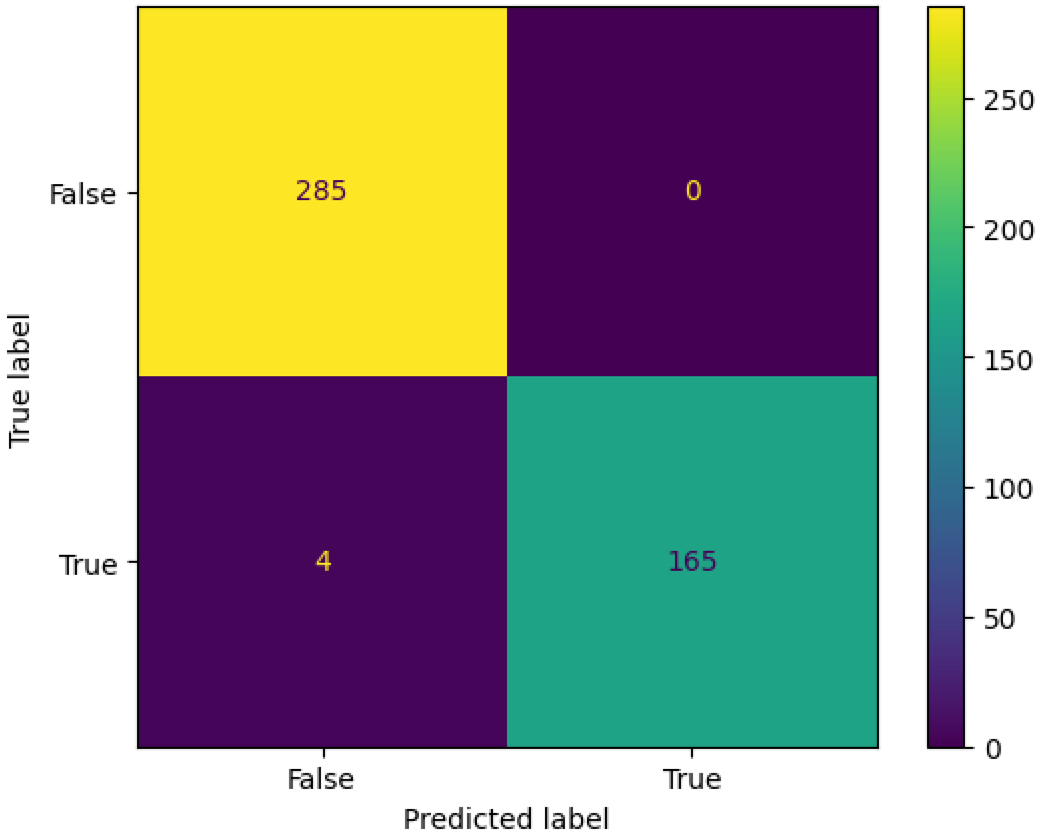

Fig. 3

Discussion and Recommendations:

Based on the evaluation results, the Support Vector Machines (SVM) model emerges as the top-performing model for breast cancer classification. It exhibits exceptional performance across various metrics, including accuracy, precision, sensitivity/recall, specificity, and F1-score. The SVM model's excellent performance makes it an ideal choice for accurate breast cancer classification, offering high confidence in predicting the malignancy or benignancy of breast cancer cases.

Conclusion

This analysis successfully developed a machine learning model for breast cancer classification based on clinical and pathological features. The Support Vector Machines (SVM) model demonstrated superior performance, achieving high accuracy and precision while maintaining high sensitivity and specificity. The model's strong performance provides valuable insights for early detection, treatment planning, and resource allocation in breast cancer management. Further research and analysis can focus on refining the SVM model or exploring ensemble techniques to enhance the predictive power and generalizability of the model.

References

- http://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
- http://www.cs.wisc.edu/~olvi/uwmp/mpml.html
- http://www.cs.wisc.edu/~olvi/uwmp/cancer.html
- K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992
- K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34