Name : Saad Bin Haseeb
Roll Number : 22F-BSAi-87

Multi-Model Ensemble for Noisy Data Classification

This report presents the design, implementation, and evaluation of a multi-model machine learning pipeline to handle real-world noisy data. The Titanic Survival dataset was used as a case study because it contains missing values, irrelevant features, and class imbalance, making it suitable for evaluating robustness and ensemble techniques.

1. Data Preprocessing

Real-world datasets often contain missing, noisy, and irrelevant data. To address these issues, several preprocessing steps were applied. Irrelevant attributes such as PassengerId, Name, Ticket, and Cabin were removed because they do not contribute meaningfully to prediction. Missing numerical values such as Age and Fare were handled using median imputation, which is robust against outliers. Categorical attributes like Embarked were imputed using the most frequent category.

Feature scaling was performed using standardization to ensure that all features contribute equally to distance-based models such as Support Vector Machines. Categorical variables were encoded numerically to make them suitable for machine learning algorithms.

2. Model Development

Three supervised learning models were trained and evaluated. A Decision Tree classifier was used due to its interpretability and ability to handle non-linear relationships. However, decision trees are prone to overfitting, so pruning was applied by limiting tree depth.

A Naïve Bayes classifier was included because of its efficiency and robustness to noisy data. Despite its assumption of feature independence, it performs well on many real-world problems.

Support Vector Machine (SVM) was used for its strong generalization ability and effectiveness in high-dimensional feature spaces. Cross-validation was applied to all models to ensure reliable performance estimation.

3. Ensemble Learning

To improve classification performance, ensemble learning techniques were applied. Bagging (Bootstrap Aggregating) was implemented using multiple decision trees. Bagging reduces variance by averaging predictions from several models trained on different subsets of data, making it particularly effective for noisy datasets.

AdaBoost was also applied as a boosting technique that focuses on correcting misclassified samples. While boosting can improve accuracy, it may amplify noise in heavily corrupted data. In this experiment, Bagging demonstrated better stability and overall performance.

4. Performance Analysis

Model performance was evaluated using accuracy, precision, recall, and F1-score. Confusion matrices were plotted to visualize classification results. Among individual models, SVM and Decision Tree showed competitive performance, while Naïve Bayes performed reasonably well despite its simplicity.

The Bagging ensemble outperformed individual classifiers by reducing overfitting and improving generalization. Regularization techniques such as tree pruning and SVM margin control effectively mitigated overfitting.

5. Conclusion

This study demonstrates that handling noisy data requires careful preprocessing, model selection, and regularization. Ensemble learning, particularly Bagging, significantly improves robustness and accuracy on real-world datasets. The results highlight the importance of combining multiple models to achieve reliable and ethical machine learning solutions.