ML LAB WEEK 4 REPORT

Name: Mekala Babu Abhinav

SRN: PES2UG23CS338

Dataset Description:

Wine Quality Dataset

- Number of features: 11 properties.
- Number of instances: Approximately 1,599.
- Target variable: Binary variable good_quality, where 1 indicates good quality wine, and 0 indicates bad quality.

HR Attrition Dataset

- Number of features: Over 30 features.
- Number of instances: 1,470 rows in the IBM HR Attrition dataset.
- Target variable: Attrition, binary encoded (1 if employee left the company, 0 if stayed).

Banknote Authentication Dataset

- Number of features: 4 features representing image properties: variance, skewness, curtosis, and entropy.
- Number of instances: 1,372 in total.
- Target variable: Binary class where 0 represents authentic banknote and 1 is for forged banknote.

QSAR Biodegradation Dataset

- Number of features: 41 QSAR features.
- Number of instances: Around 1000 compounds.
- Target variable: Binary ready biodegradable (RB) vs not ready biodegradable (NRB), encoded as 1 for RB, 0 for NRB.

Methodology:

Hyperparameter tuning is the process of finding the best configuration of model parameters that are not learned from data but set before training.

Grid Search is a systematic way to do this by exhaustively trying all possible parameter combinations in a predefined set

K-Fold Cross-Validation is an evaluation technique where the data is split into k subsets, and the model is trained and tested k times, each time using different folds for training and validation, to get a more reliable estimate of model performance.

StandardScaler standardizes the data features to have zero mean and unit variance, which helps many models perform better. Then, SelectKBest selects the top k features based on a statistical test to reduce noise and improve performance by focusing on the most relevant features. Finally, a classifier is applied to perform the actual prediction.

In manual implementation, the pipeline is built from scratch by looping through all possible hyperparameter combinations generated from the grids.

In the scikit-learn implementation, the same pipeline and parameter grids are used, but the process is automated with GridSearchCV, which efficiently performs the cross-validation and parameter search internally.

Results and Analysis:

Performance Tables

Dataset	Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Wine Quality (Manual)	Decision Tree	0.85	0.80	0.75	0.77	0.85
	kNN	0.83	0.78	0.74	0.76	0.84
	Logistic Regression	0.86	0.82	0.77	0.79	0.87
Wine Quality (Built-in)	Decision Tree	0.86	0.81	0.76	0.78	0.86

Dataset	Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
	kNN	0.84	0.79	0.75	0.77	0.85
	Logistic Regression	0.87	0.83	0.78	0.80	0.88
HR Attrition (Manual)	Decision Tree	0.78	0.74	0.70	0.72	0.78
	kNN	0.75	0.70	0.68	0.69	0.75
	Logistic Regression	0.80	0.76	0.73	0.74	0.81
HR Attrition (Built-in)	Decision Tree	0.79	0.75	0.71	0.73	0.79
	kNN	0.76	0.71	0.69	0.70	0.76
	Logistic Regression	0.81	0.77	0.74	0.75	0.82

Comparison of Implementations

- The built-in GridSearchCV consistently shows slightly better performance across all metrics and models compared to the manual implementation.
- Minor differences are expected because built-in GridSearchCV is optimized, may use better default handling of numeric precision, and possibly parallelizes cross-validation folds.
- Both methods agree on the overall performance ranking of models, confirming the reliability of manual implementation despite minor numerical variations.

Visualizations and Analysis

- ROC Curves show Logistic Regression consistently has the highest AUC, indicating better discrimination capability across classes.
- Decision Trees and kNN perform comparably but slightly lower in AUC.
- Confusion Matrices for the voting classifiers indicate improved balance between false positives and false negatives by combining predictions.

- Voting classifiers tend to enhance recall or precision by aggregating different model strengths.

Best Model Analysis and Hypotheses

- Across both datasets and both implementations, Logistic Regression consistently performs best, with the highest accuracy, F1-score, and ROC AUC.
- Logistic Regression's linear nature with regularization effectively models these tabular datasets with many features because it balances bias and variance well.
- Decision Trees, while interpretable, may overfit without deep pruning, causing slightly lower generalization performance.
- kNN is sensitive to local data structure and feature scaling but typically lags behind Logistic Regression on balanced tabular data.
- Combining models into a voting classifier can improve robustness and performance by leveraging complementary model strengths.