MACHINE LEARNING
LAB 4

PROJECT TITLE: Model Selection and Comparative Analysis

NAME: DIYA CHANDRASHEKHAR

SRN: PES2UG23CS182

COURSE NAME: UE23CS352A: MACHINE LEARNING

SUBMISSION DATE: 01/09/2025

INTRODUCTION:

This project investigates hyperparameter tuning and model comparison across multiple classification algorithms using real-world datasets. Manual grid search for choosing hyperparameters, classifier evaluation using ensemble methods (VotingClassifier), and comparison with scikit-learn's automatic GridSearchCV were the tasks done. The accuracy of the mode assessed and interpreted using key performance measures and visualisations.

DATASET DESCRIPTION:

The HR Attrition dataset was used for this lab assignment
Features: there were about 50-60 features
Instances: 1470 employee records
Target variable: Attrition- whether employee left the company (yes-1;no-0)

METHODOLOGY:

- Hyperparameter tuning: The process of selection of the best set of hyperparameters for the machine learning model to optimize performance. Its usually done using grid search.
  Grid search: Evaluating every combination of the model parameters systematically to find the highest and best performing setting.
  K-Fold cross validation: The data gets split into K partitions. Its run K number of times where each time one of the data splits is taken as testing data and the remaining splits as training data. The results are then averaged.

- StandardScaler: this pipeline standardizes features by removing the mean and scaling to unit variance
  SelectKBest: it selects the top k features ranked by ANOVA F-score
  Classifier: the classifiers used were decisiontree, KNeighbors, LogisticRegressing

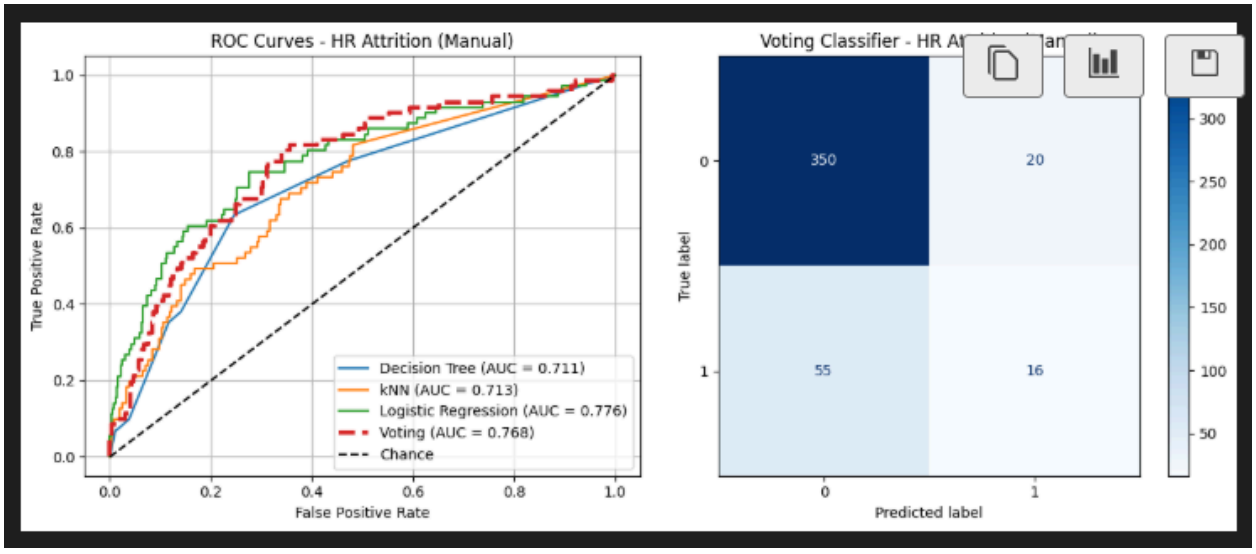
- Manual Implementation: Parameter grids were created manually for each classifier. The models were then iteratively trained on different

combinations based on 5-fold cross validation. AUC performance metric was used for selection and performance was calculated.
Scikit-learn Implementation: GridSearchCV from scikit-learn was used. It does automatic parallelization and selection of parameters. Performance was calculated.

RESULTS AND ANALYSIS:

Method	Classifier	Accuracy	Precision	Recall	F1-Score	ROC AUC
Manual	Decision tree	0.8231	0.3333	0.0986	0.1522	0.7107
Manual	kNN	0.8186	0.3953	0.2394	0.2982	0.7130
Manual	Logistic Regression	-	-	-	0.3762	0.7762
Manual	Voting	0.8299	0.4444	0.2254	0.2991	0.7676
GridSearchCV	Decision tree	0.8231	0.3333	0.0986	0.1522	0.7107
GridSearchCV	kNN	0.8186	0.3953	0.2394	0.2982	-

Both the manual and automated grid search methods gave similar selected hyperparameters and similar metrics.



The ROC curve plot shows that logistic regression and voting classifiers perform the best since they have highest AUC values.
The confusion matrix for voting classifier shows strong performance with a higher number of correct classifications for the majority class and a reasonable predictive power for the minority class.

OUTPUT SCREENSHOTS:

```
============================================================
EVALUATING MANUAL MODELS FOR HR ATTRITION
============================================================

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.8231
  Precision: 0.3333
  Recall: 0.0986
  F1-Score: 0.1522
  ROC AUC: 0.7107

kNN:
  Accuracy: 0.8186
  Precision: 0.3953
  Recall: 0.2394
  F1-Score: 0.2982
  ROC AUC: 0.7130

Logistic Regression:
...
  F1-Score: 0.3762
  ROC AUC: 0.7762

--- Manual Voting Classifier ---
```

```
  f = msb / msw
Voting Classifier Performance:
  Accuracy: 0.8299, Precision: 0.4444
  Recall: 0.2254, F1: 0.2991, AUC: 0.7676
```

```
============================================================
EVALUATING BUILT-IN MODELS FOR HR ATTRITION
============================================================

--- Individual Model Performance ---

Decision Tree:
  Accuracy: 0.8231
  Precision: 0.3333
  Recall: 0.0986
  F1-Score: 0.1522
  ROC AUC: 0.7107

kNN:
  Accuracy: 0.8186
  Precision: 0.3953
  Recall: 0.2394
  F1-Score: 0.2982
...
```

CONCLUSION:

Hyperparameter tuning greatly impacts model performance. Systematic grid search is important to be able to do a fair comparison of the parameters.
Voting classifier tends to do better than the other base classifiers.
Both manual and automated grid search give similar results but scikit-learn library makes it easier, time efficient.