Submitted From:

Name: Hamza Asif

Roll No: 23-AI-93

Document Type: Assignment Report

Subject: Machine Learning

Submitted To:

Sir Hamza Farooqi

Report of Complex Computing

Title:

Multi-Model Ensemble for Noisy Data Classification

DataSet File:

titanic.csv

1. Introduction:

Real-world datasets are often noisy. They contain missing values, outliers, and irrelevant features. These problems reduce the accuracy of machine learning models.
This project uses the **Titanic dataset**, which is a real-world dataset containing missing ages and categorical features.

The goal is to:

- Clean the data
- Train multiple models
- Compare performance
- Improve accuracy using ensemble learning

2. Data Preprocessing:

The Titanic dataset had missing values and categorical features.

2.1 Missing Values:

- The **Age** feature had missing values.
- Mean imputation was used to fill missing ages.
- Missing values in **Embarked** were filled using the mode.

This helped maintain dataset size without losing important information.

2.2 Feature Encoding:

- Gender (male/female) was converted into numeric form.
- Embarked locations were mapped to numbers.

2.3 Feature Scaling:

Standardization was applied using **StandardScaler**.
This step is important for models like **SVM**, which are sensitive to feature scale.

3. Model Development:

Three supervised learning models were trained:

3.1 Decision Tree:

- Easy to understand and interpret.
- Prone to overfitting on noisy data.
- Depth was limited to reduce overfitting.

3.2 Naïve Bayes:

- Fast and efficient.
- Assumes feature independence.
- Performs well on small and noisy datasets.

3.3 Support Vector Machine:

- Works well with high-dimensional data.
- Uses margin maximization.
- Regularization helps reduce overfitting.

4. Ensemble Learning:

4.1 Bagging:

Bagging (Bootstrap Aggregating) was applied using multiple Decision Trees.

Why Bagging helps:

- Reduces variance
- Improves stability
- Handles noisy data better

The ensemble model performed better than a single Decision Tree.

5. Performance Analysis:

Each model was evaluated using:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix

Observations:

- Naïve Bayes performed reasonably well.
- SVM showed better balance between precision and recall.
- Bagging ensemble achieved the **highest accuracy**.
- Ensemble reduced overfitting by averaging predictions.

6. Overfitting and Regularization:

Overfitting happens when a model learns noise instead of patterns.

Mitigation Techniques:

- Decision Tree depth was limited.
- SVM used regularization.
- Bagging reduced variance by combining models.

These techniques improved generalization on unseen data.

7. Conclusion:

This project demonstrated that:

- Proper preprocessing improves model performance.
- Ensemble learning performs better on noisy datasets.
- Bagging is effective for reducing overfitting.
- Using multiple models gives better insight into data behavior.

Ensemble methods are recommended for real-world noisy datasets.