Machine Learning
Assignment 3 - Supervised Learning (Classification)

Vikas Thammanna Gowda

Due: 04/27/2025

Instructions

Collaboration Policy

This is an individual assignment. You may discuss concepts, problem formulations, and approaches to solving the problems, but you must write your own code and explanations.

- You are **not allowed** to share solutions, source code, or exact approaches.
- Any external sources (books, online resources, discussions, etc.) you refer to must be cited in your write-up.
- You do not need to cite course lecture notes, textbooks, or materials provided as part of the course.

Assignment Structure

Your submission consists of two parts:

Coding Component (Submit as `<your_name>_PA03.ipynb`)

- Use Markdown cells in Jupyter Notebook to add each question before solving it.
- Write clean, readable, and well-commented code.
- Define functions for repetitive tasks instead of redundant code.
- Ensure all visualizations are clear, properly labeled, and provide meaningful insights.
- Use at least two different types of visualizations for each data exploration question (e.g., histogram and box plot).

Report Write-up (Submit as `<your_name>_PA03.pdf`)

- Add each question to your write-up before answering them.
- The report should mirror the coding component and provide interpretations of results.
- Use Times New Roman, size 14 for questions, size 12 for answers.
- Ensure the document is justified and structured.
- Include properly labeled figures and tables, centered with captions.
- All the plots must be complete, be of the same size, and be centered with a figure number and a figure name.
- Clearly explain decisions regarding missing data handling, feature selection, scaling, and outlier removal.

Dataset Description

You will use the dataset `Clean_Used_Car_Sales_PA03.csv` for this assignment. This dataset contains cleaned and preprocessed used car sales data for supervised learning tasks. You will build classification models on this data.

Assignment Questions: Coding vs. Write-Up

Each question involves both a coding component (implementation) and a write-up component (interpretation).

1. Decision Tree & Random Forest Classification

Coding:

- Use 25% of the data as your test set.
- Build the following classifiers to classify the fuel type.
  - Decision Tree Classifier
  - Random Forest Classifier
- Use the default arguments for both models.

Write-up:

- Create and present the Confusion Matrix (showing actual vs. predicted).
- Report the Accuracy, Precision, F1-score, and Recall.
- Which model performs better and why?

2. Varying Test Set Sizes for Fuel Classification

Coding:

- Repeat 1 with test sizes of 20%, 30%, and 50% instead of 25%.

Write-up:

- Compare and comment on how the split size affects performance metrics.

3. Decision Tree: Varying Max Depth

Coding:

- Use 20% of the data as your test set.
- Train a Decision Tree with three different maximum depths: 5, 8, and 11.

Write-up:

- Present and compare the Confusion Matrix for each max depth.
- Report the Accuracy, Precision, F1-score, and Recall.
- Which model (i.e., which max depth) performs best and why?

4. Decision Tree: Varying Minimum Leaf Samples

Coding:

- Use 20% of the data as your test set.
- Train a Decision Tree with minimum leaf sample values of 20, 40, 80, and 100.

Write-up:

- Present and compare the Confusion Matrix for each minimum leaf sample.
- Report the Accuracy, Precision, F1-score, and Recall.
- Which setting performs best and why?

Grading Rubric (100 Points)

Category	**Criteria**	**Points**
Code Quality	Code is well-structured, commented, and readable.	10
Tables Quality	Clear, labeled, and informative.	10
Classification models	Correct application of regression models and PCA.	50
Discussion of results	Quality of explanations, rationale, and justifications.	30
Total	**Final Score**	**100**

Table 1: Grading Rubric

Final Submission Checklist

- Jupyter Notebook (`.ipynb`)
- PDF Write-up (`.pdf`)
- All tables included
- Formatted report with proper justifications
- Sources cited where applicable