Data Mining Final Practical File

24/48029

Vivaan Singh Adhikari November 24, 2025

Assignment Overview

Document Contents:

Assignment Details:

- **Course:** Data Mining DSE
- **Instructor:** Prof. Archana Gahalaut
- **Hardware:** No specifications
- **Software:** Python, Pandas, Typst(documentation)

Introduction

This assignment entails my solutions to the question assigned as per the course's guidelines. All the final files are available on https://github.com/user7537/coursework/

Code

QApply simple K-means algorithm for clustering any dataset. Compare the performance of clusters by varying the algorithm parameters. For a given set of parameters, plot a line graph depicting MSE obtained after each iteration.

```
import pandas as pd

# Load dataset 1
df = pd.read_csv("japanese_credit_screening/crx.data", header=None,
na_values="?")

# Assign column names (A1…A16) if you want; or keep numeric indexing.
col_names = [f"A{i}" for i in range(1,17)]
df.columns = col_names

# Missing values
df = df.fillna(df.median(numeric_only=True))
df = df.fillna(df.mode().iloc[0])

# Remove outliers on numeric columns
num_cols = df.select_dtypes(include="number").columns
for col in num_cols:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    df = df[(df[col] >= Q1 - 1.5*IQR) & (df[col] <= Q3 + 1.5*IQR)]

# Domain specific: for attribute A1 values must be ‘b’ or ‘a’ only
df = df[df["A1"].isin(["b","a"])]

df.to_csv("q1_cleaned_ds1.csv", index=False)
print("Dataset1 cleaned and saved to q1_cleaned_ds1.csv")
```