Final Project

Dr. Amir Rahnama

March 14, 2025

1 Project Overview

In this project, you will work with a real-world dataset containing numerical, categorical, and textual data. The goal is to apply advanced data preprocessing techniques, perform feature engineering, and build an effective machine learning model to classify or predict a target variable. The dataset will contain noise, missing values, and unstructured text, making it a challenging yet valuable experience.

2 Dataset Description

The dataset consists of:

- **Customer ID**: Unique identifier for each customer.
- **Name**: Textual feature representing customer names.
- **Age**: Numerical continuous data.
- **Annual Income**: Numerical feature with outliers and missing values.
- **Education Level**: Categorical ordinal feature ("High School", "Bachelor", "Master", "PhD").
- **Occupation**: Categorical feature representing job types.
- **City**: Categorical nominal feature representing location.
- **Customer Reviews**: Textual feedback given by customers.
- **Purchase Frequency**: Number of purchases made in the last year.
- **Product Category**: The type of product purchased by the customer.
- **Target Variable**: Customer satisfaction level ("Low", "Medium", "High").

3 Project Tasks

Students should complete the following tasks:

3.1 Data Exploration and Cleaning

- Load the dataset and explore its structure.
- Handle missing values using appropriate techniques for numerical, categorical, and text features.
- Identify and remove or treat outliers using statistical methods.

3.2 Feature Engineering

- Normalize and standardize numerical features.
- Encode categorical variables using both one-hot encoding and label encoding where necessary.
- Extract meaningful insights from textual data using NLP techniques (e.g., TF-IDF, word embeddings).

3.3 Dimensionality Reduction and Clustering

- Apply PCA or another dimensionality reduction technique to reduce dataset complexity.
- Perform K-Means or hierarchical clustering to segment customers into groups.

3.4 Machine Learning Model Development

- Split the dataset into training and testing sets.
- Train at least three different classification models (e.g., Decision Tree, SVM, Random Forest, XGBoost) to predict customer satisfaction levels.
- Evaluate the models using accuracy, precision, recall, and F1-score.
- Tune hyperparameters using GridSearchCV or RandomizedSearchCV.

3.5 Model Interpretation and Visualization

- Visualize feature importance for the best-performing model.
- Create customer segments and analyze key factors influencing satisfaction.
- Present findings using plots, tables, and a final conclusion.

4 Project Submission Requirements

Students should submit the following:

- A well-documented Python code implementation.
- Data visualizations supporting key observations.
- Any additional insights or conclusions drawn from their analysis.
- Final work should be submitted in pdf file format ONLY.