ENGINEERING GRADUATE SALARY PREDICTION

GROUP 4:

SHRAVANI PAI
PRADEEPTHI
DURGA
KANISHK THAKUR

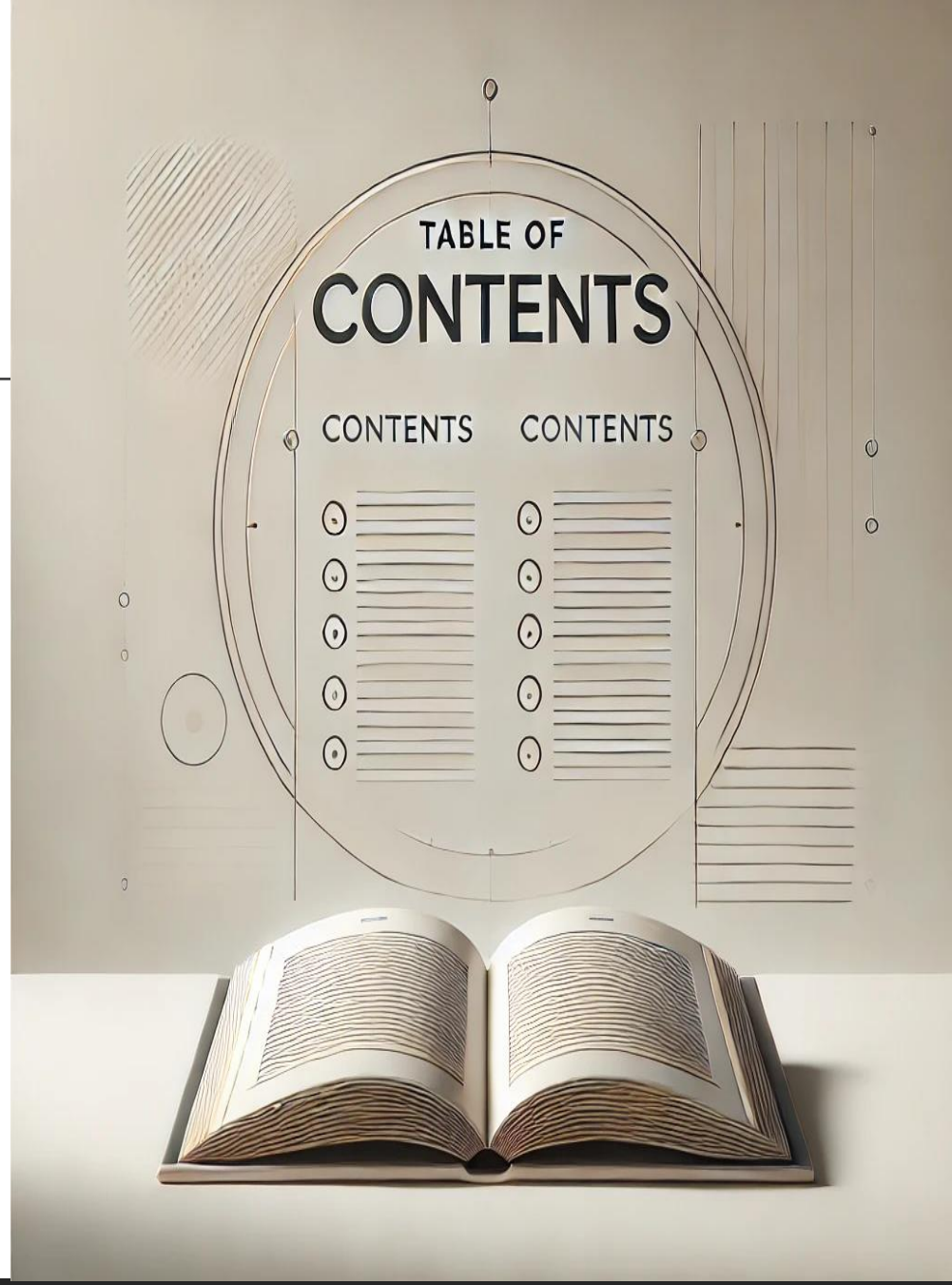
ABSTRACT

The project explores factors affecting graduate salaries, including academic performance, skills, and personality traits. Python was used for data cleaning, regression, and machine learning models like Logistic Regression, Decision Trees, and K-nearest neighbors. Charts and visualizations were used to present insights on key salary determinants.

.

CONTENTS

1. Introduction
2. Data Preprocessing
3. Exploratory Data Analysis
4. Data Modeling & Evaluation
5. Insights



INTRODUCTION

1. The Engineering Graduate Salary Dataset serves as a vital tool for understanding salary trends and factors influencing the career outcomes of graduates. Academic performance, technical skills, and personality traits are critical in shaping salary prospects.

2. Technological advancements: Leveraging machine learning and data analytics has transformed salary prediction models, enabling more accurate analysis and insights for better career planning and recruitment strategies.

DATA PREPROCESSING

DATA

- DATASET – The data set contains 34 columns and 2999 records.
- SOURCE – https://drive.google.com/drive/folders/1BdZRhiHBsYoN0zOXcr_f1cjpSZrWFGG2
- VARIABLES –

CATEGORICAL VARIABLE	CONTINUOUS VARIABLE
Gender	ID, DOB, 10^{th}, 12^{th} Percentages
10^{th}, 12^{th} Board	College tier, College GPA
Degree	Domain
Specialization	Quantitative Aptitude
College State	Personality Traits

DATA CLEANING

1.DATA CLEANING :

Replaced missing values (-1) with 0 to ensure no negative data.
Dropped irrelevant columns such as IDs and board information that aren’t needed for the analysis.

2. DATE OF BIRTH (DOB) HANDLING :

Converted the 'DOB' column to datetime format.
Calculated the age of individuals by subtracting their birthdate from today’s date and converting it to years.
Converted the 'Age' to integer format and removed the 'DOB' column.

3. GENDER ENCODING :

Encoded the 'Gender' column into numerical values:
Male = 1, Female = 0 for better processing in models.

DATA CLEANING

4. DUMMY VARIABLE CREATION :

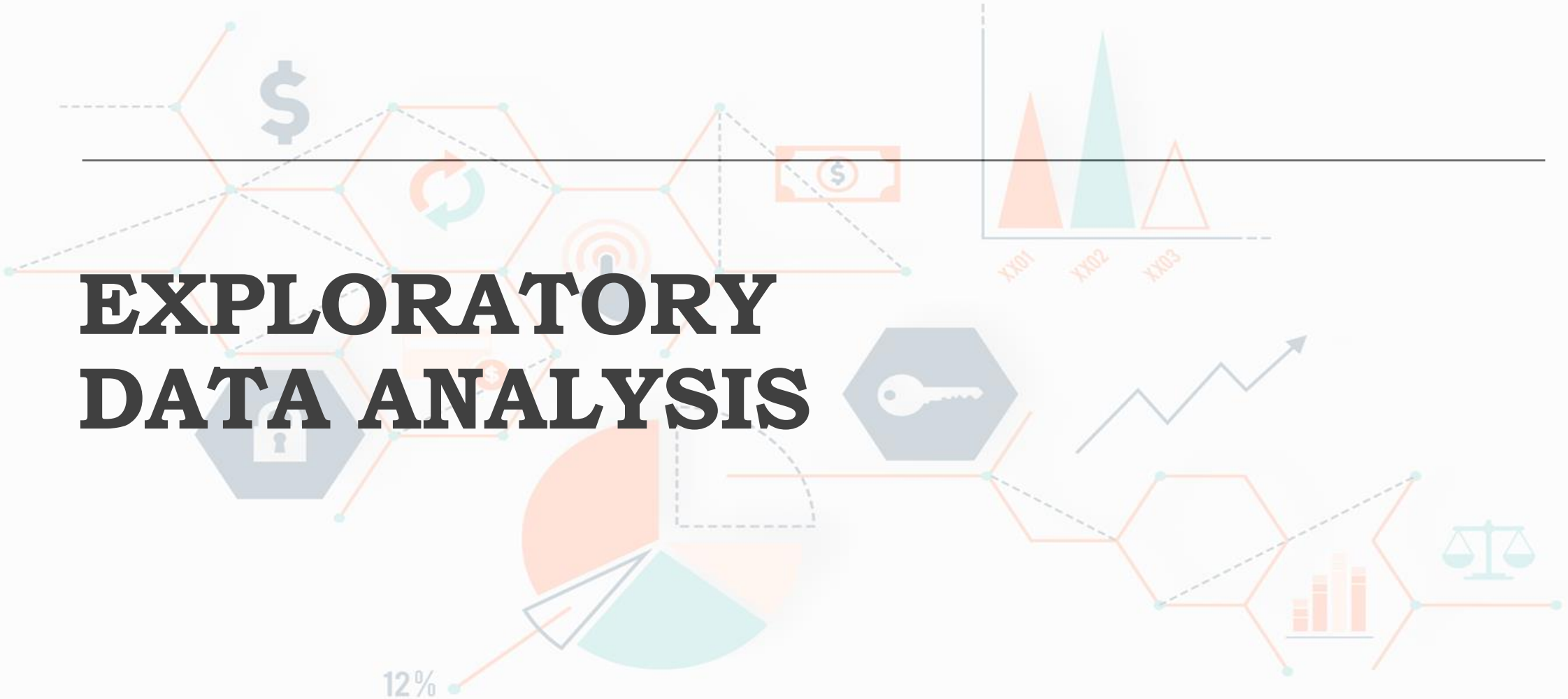
Transformed categorical columns like 'Specialization' and 'Degree' into dummy variables for easier model processing.

5. SPECIALISATION GROUPING :

Combined computer-related specializations into a single category called ‘Specialization Computer’.
Grouped electrical-related specializations into ‘Specialization Electrical’ to reduce complexity.

6. DEGREE ADJUSTMENTS :

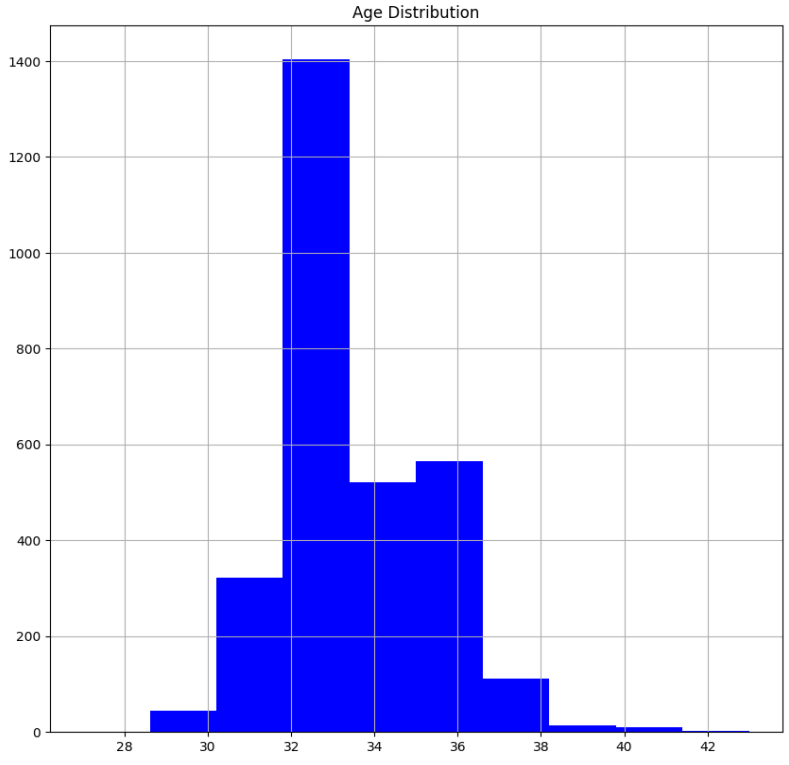
Handled specific cases where M.Tech./M.E. degrees were reclassified based on specialization groupings

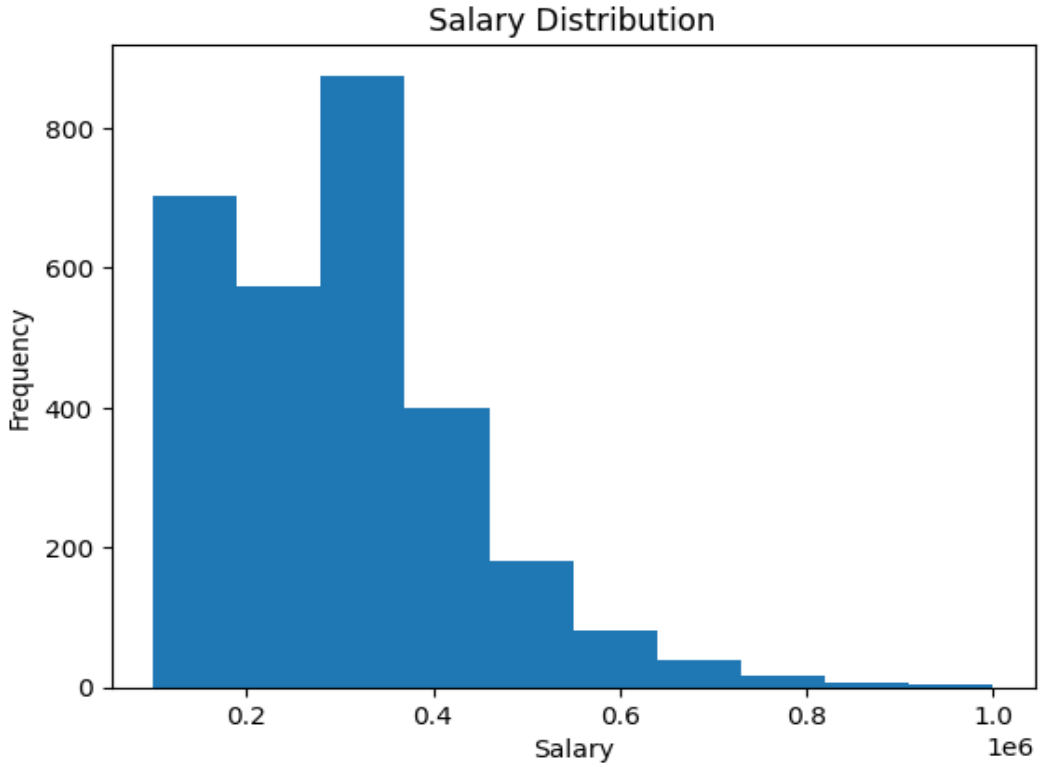


EXPLORATORY DATA ANALYSIS

UNIVARIATE DATA ANALYSIS

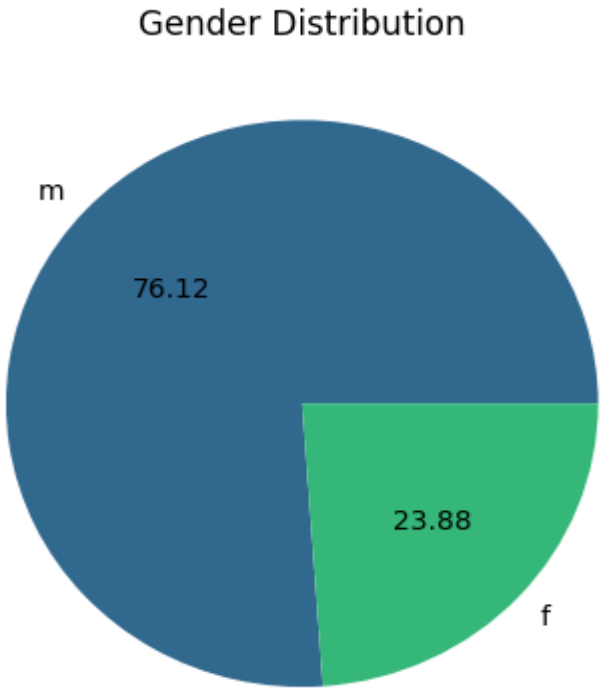
HISTOGRAM

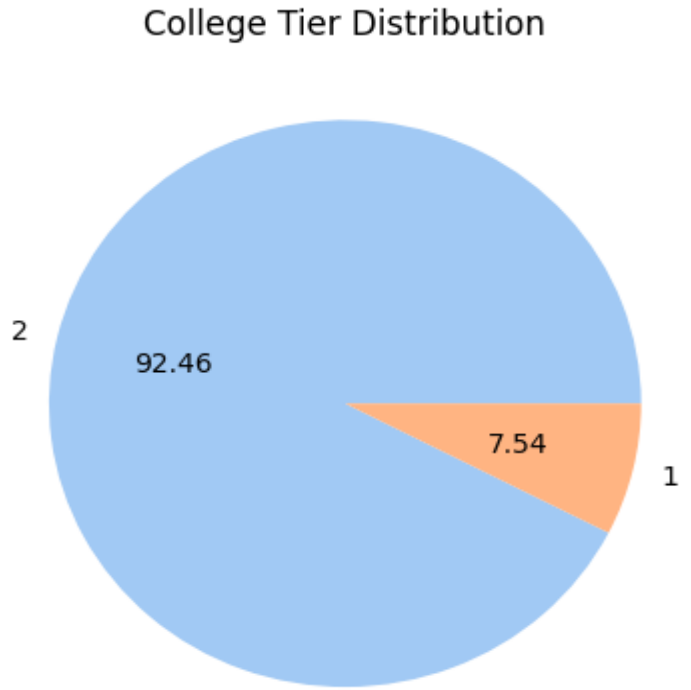




UNIVARIATE DATA ANALYSIS

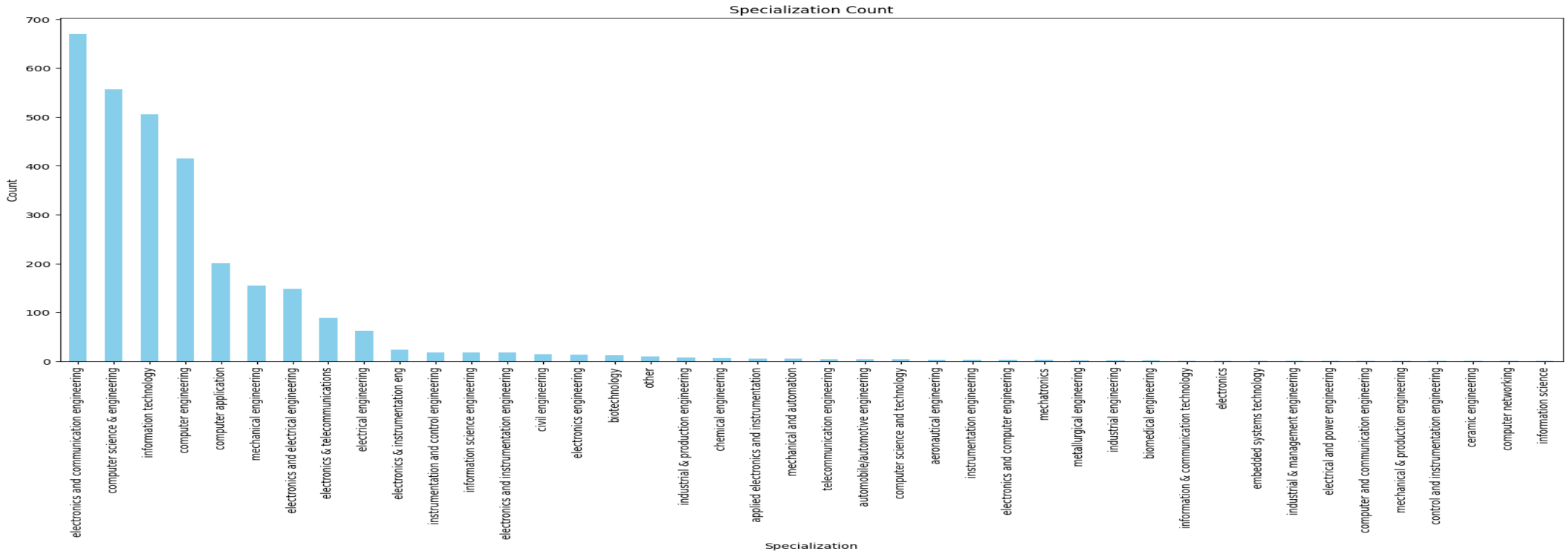
PIE – CHART





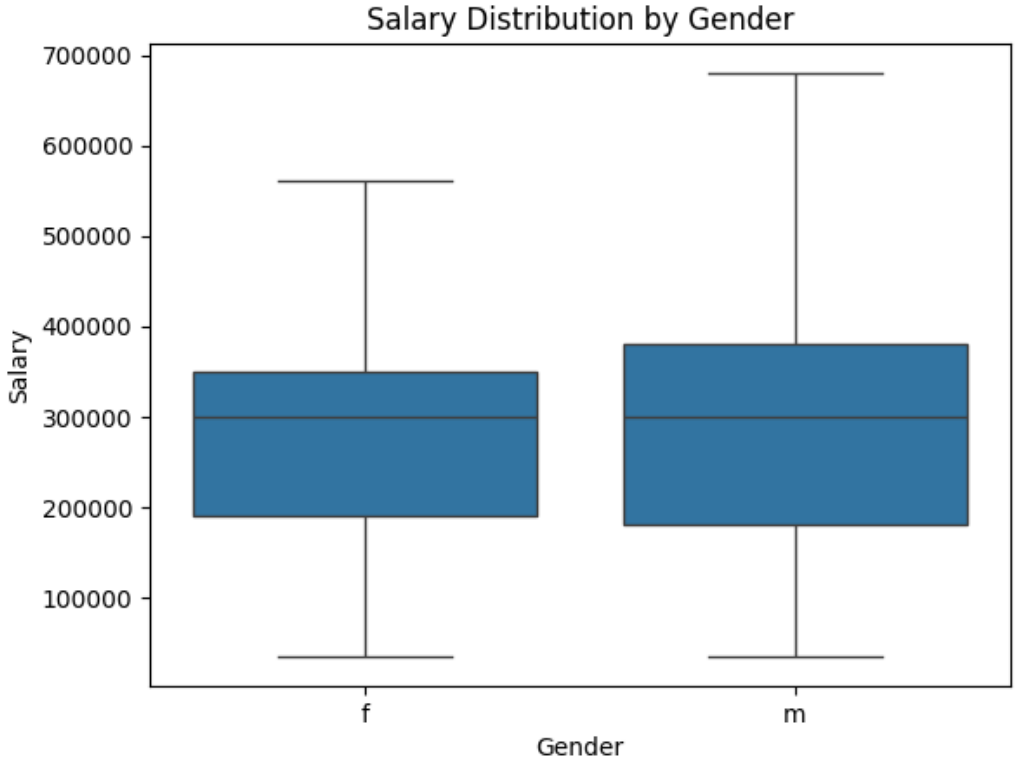
BI – VARIATE DATA ANALYSIS

BAR PLOT

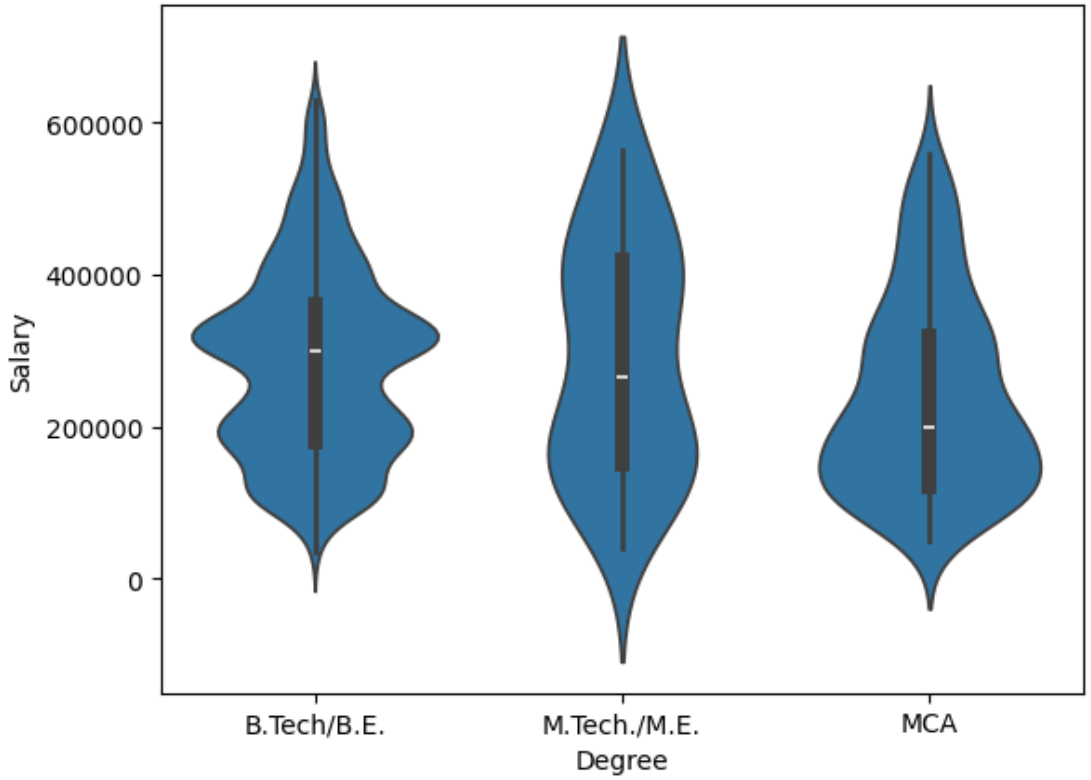


BI – VARIATE DATA ANALYSIS

BOX PLOT

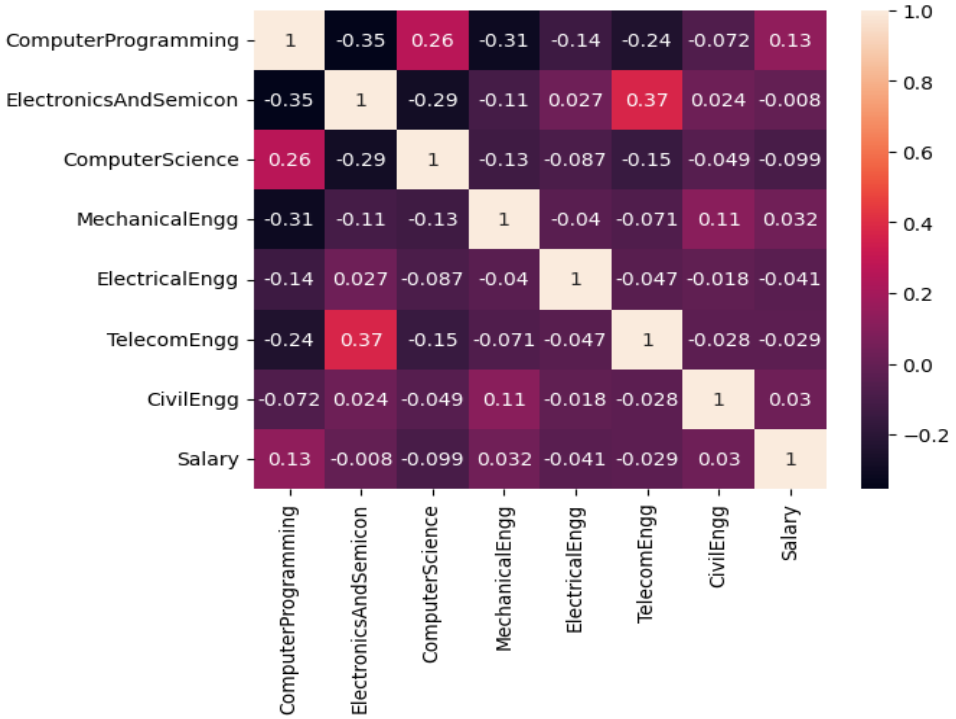


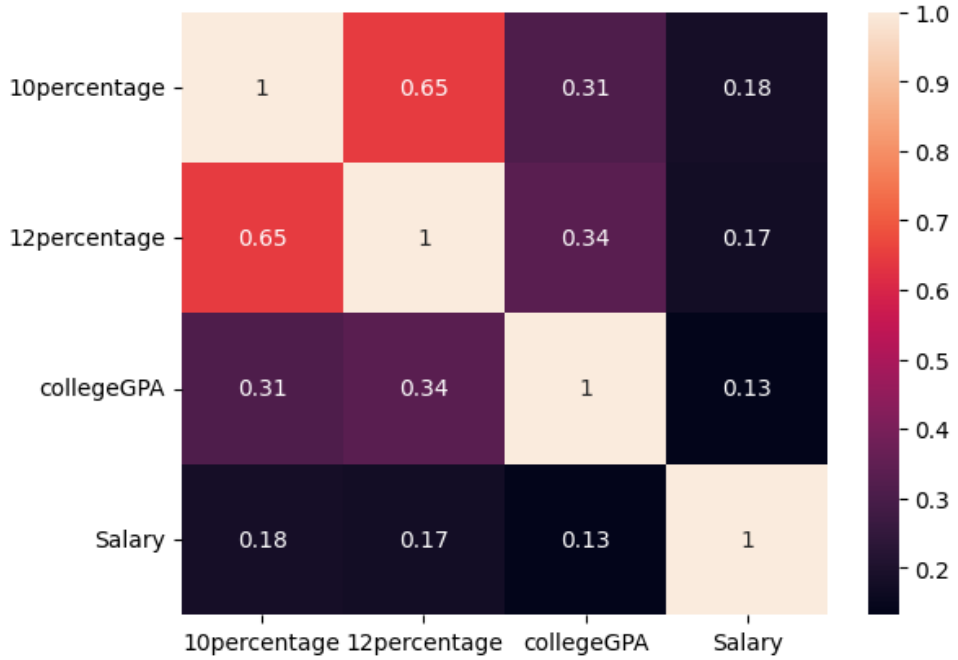
VIOLIN PLOT



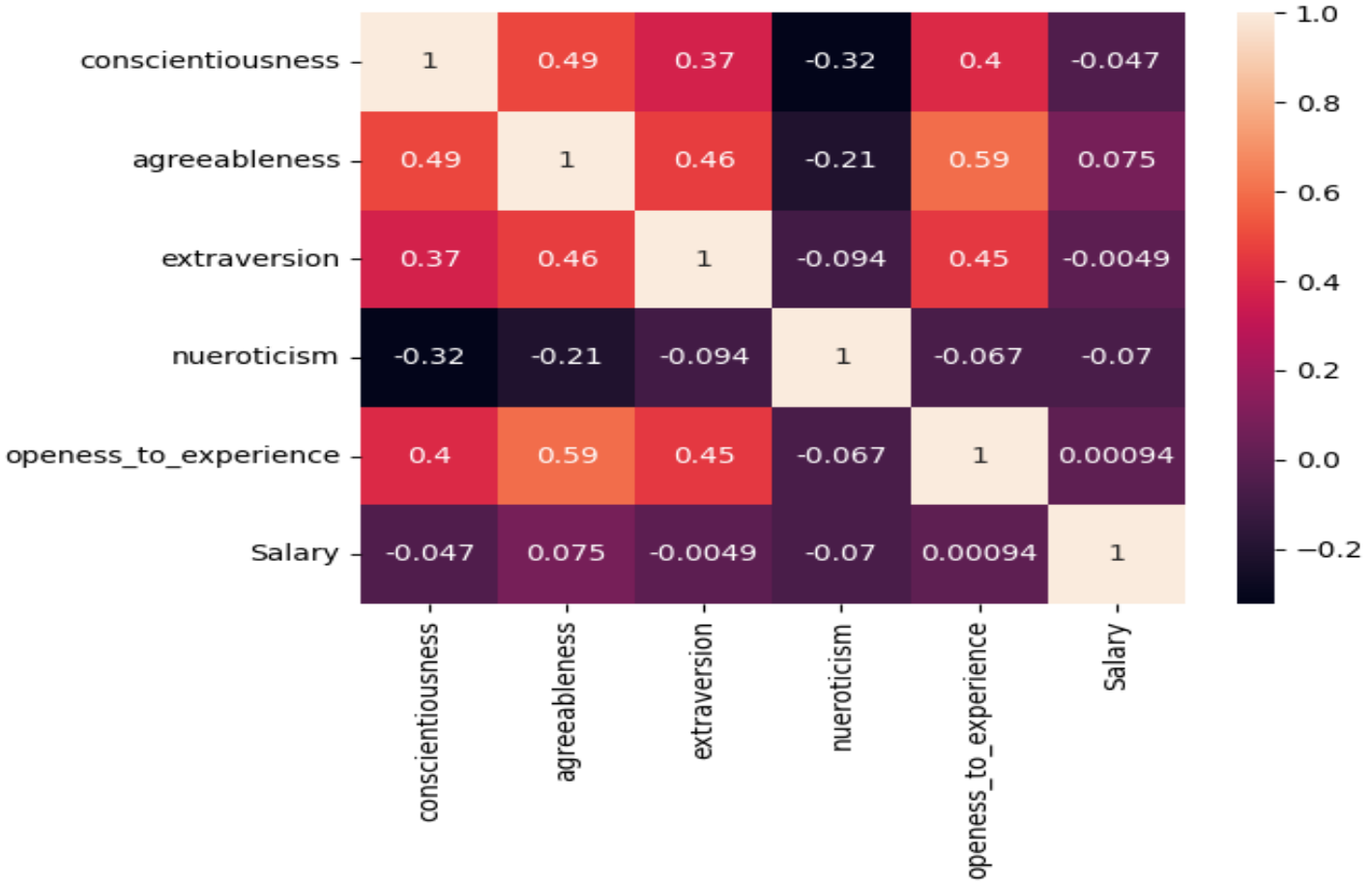
MULTIVARIATE DATA ANALYSIS

HEATMAP



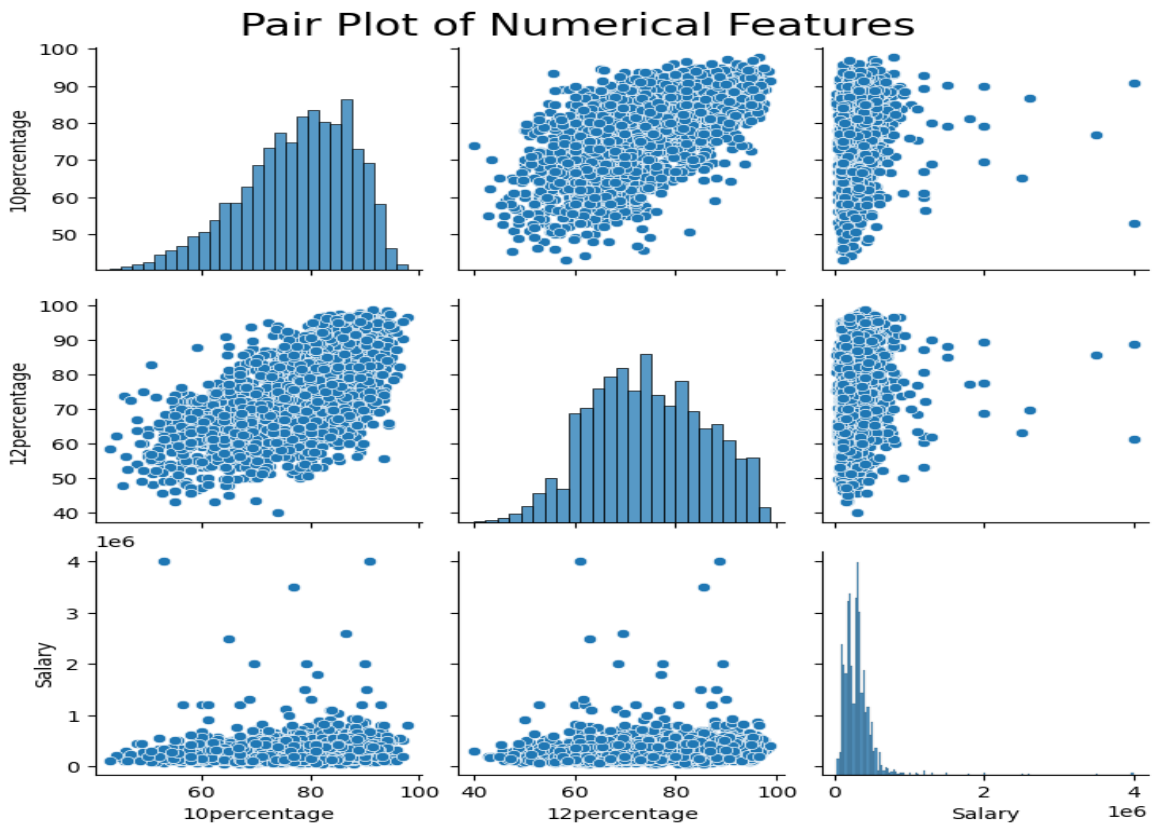


MULTIVARIATE DATA ANALYSIS



MULTIVARIATE DATA ANALYSIS

PAIR PLOT



MACHINE LEARNING ALGORITHM (CLASSIFICATION)

CLASSIFICATION MODELS

Logistic Regression:
A linear model used for binary classification tasks. It predicts the probability of a category based on input features by fitting the data to a logistic function.

K-Nearest Neighbors (KNN):
A non-parametric model that classifies new data points based on the 'k' closest data points in the feature space, using majority voting for classification.

Decision Trees:
A model that splits the data into branches based on feature values, making decisions at each node until a final classification is reached at the leaves. It captures complex relationships but is prone to overfitting.

ACCURACY COMPARISON

CLASSIFICATION MODELS	LOGISTIC REGRESSION	KNN CLASSIFICATION	DECISION-TREE CLASSIFIER
TRAIN-TEST SPLIT	ACCURACY	ACCURACY	ACCURACY
80-20	70.62%	69.49%	66.29%
70-30	70.60%	69.34%	67.58%
90-10	71.05%	69.17%	63.53%
85-15	69.84%	70.10%	66.83%
75-25	71.04%	69.53%	66.51%

ALGORITHM COMPARISON

ALGORITHM	ACCURACY
LOGISTIC REGRESSION	71.05%
K – NEAREST NEIGHBOUR	70.10%
DECISION TREE	67.58%

INSIGHTS

Data Cleaning: Addressed missing values and encoded categorical data for accurate analysis.

EDA: Uncovered key factors like academic performance and skills impacting salary outcomes.

Models Evaluated: Logistic Regression, KNN, and Decision Trees

Best Model: Logistic Regression, with the highest accuracy of 71.05% (90-10 split).

KNN Performance: Slightly lower accuracy than Logistic Regression.

Decision Tree: Consistently the lowest accuracy, with 63.53% at the 90-10 split.

Conclusion: Logistic Regression is the most reliable model across train-test splits.

THANK YOU

- SHRAVANI PAI
- PRADEEPTHI
- DURGA
- KANISHK THAKUR

COLAB NOTEBOOK - https://colab.research.google.com/drive/1qtnUH8isiuqd6BUM7MpzJcz7OGIuDD02?usp=drive_link#scrollTo=3SRDtf11z8I8