NAME: TUNUGUNTLA GOURI NITEESHA

ROLLNO:22481A42B4

BRANCH: CSE(AI&ML)

SECTION: ’B’

ACADEMIC YEAR: 2025-2026

CLASS: 3RD YEAR

**Title: Predictive Job Preference Analysis with Data Science**

For this project, a dataset related to job seekers' behaviour was selected. The dataset was collected from a open-data platform Kaggle, containing details such as education level, experience, training hours, and job-seeking status.

<https://1drv.ms/x/c/28882fa0eab2662e/EdYzRI21K_VCtIfdIVc1XRUBhTL3RX3BcsBkOKbmhOurFw?e=eeKePU>

And I initialized path through goggle colab notebook.

MY CODE: <https://colab.research.google.com/drive/1SAsd6SWCWJ-roihXPw-UUUfHfLhUjLww?usp=sharing>

**Abstraction:**

The task at hand is the analysis of a dataset based on data science for coming up with valuable insights. The goal here is to clean and preprocess data, conduct exploratory data analysis (EDA), and implement machine learning models to be used in classification and clustering. Inferences from the analysis can be utilized in decision making when making decisions based on data and also when enhancing prediction modeling for use. The study involves strict sequential processes of data preprocessing, visualization of data, feature extraction, and model testing towards providing reliable output.

**Existing System:**

The existing data analysis approach is the traditional manual inspection using spreadsheets and elementary statistical packages that are non-scalable and non-automatic. The processes are unsuitable for big data and don't use the latest machine learning. Missing values, pattern discovery, and prediction decision-making are poorly catered for by the existing system.

**Proposed System:**

To rectify the shortcomings of the existing system, the proposed system includes:

Automated Data Cleaning: Missing values, outliers, and standardization handling.

Exploratory Data Analysis (EDA): Frequency polygons, histograms, and boxplots visualization techniques to identify data distribution.

Feature Engineering: Label encoding for encoding categorical variables.

Machine Learning Models: Classification and clustering models for predictive modeling.

Evaluation & Insights: Accuracy metrics to measure the model's performance and drawing conclusions based on the result.

**Techniques Used:**

Data Preprocessing: Handling missing values, outlier detection, feature scaling.

EDA & Visualization: Histograms, boxplots, frequency polygons, heatmaps of correlation.

Feature Engineering: Label encoded categorical features to make them suitable for machine learning.

Machine Learning Models: Applied classification and clustering models.

Evaluation Metrics: Used accuracy, precision, recall, and other metrics for performance.

**Requirements Software Requirements:**

Software requirements:

Python (with libraries: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn)

Google Colab

Hardware Requirements:

Minimum 8GB RAM (for smooth computation)

Multi-core processor (for execution speed)

Enough storage for processing datasets

**Dataset Gathering & Preprocessing:**

Dataset Origin: Retrieved from online databases (Kaggle, etc.).

Preprocessing Steps:

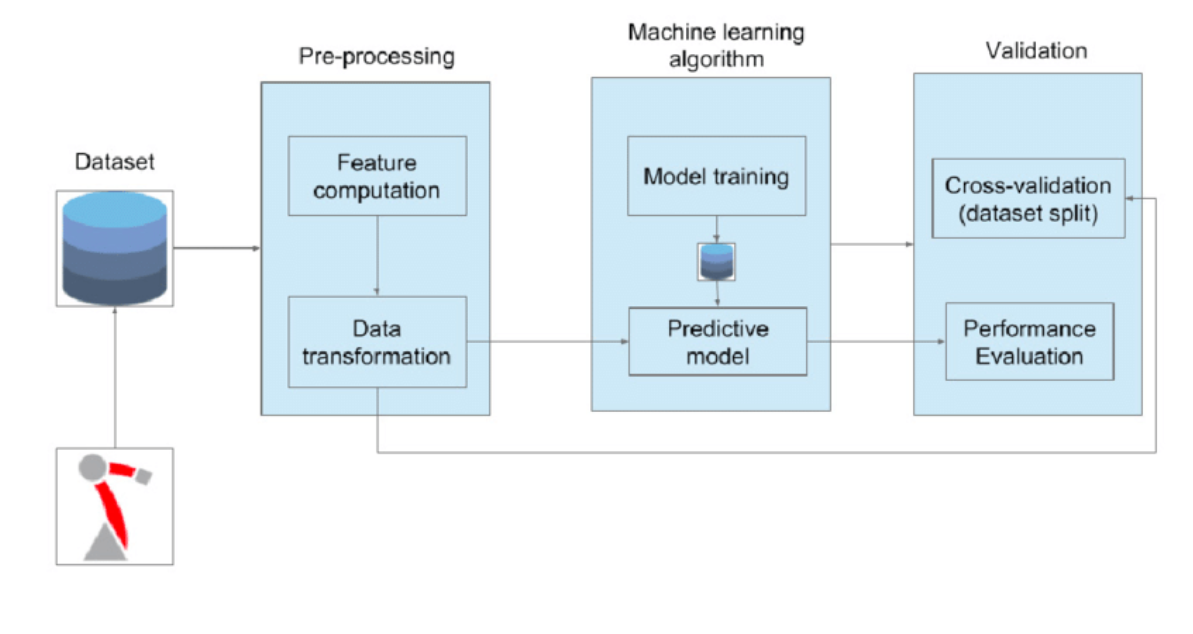
Scanned for missing values and dealt with accordingly.

Dropped outliers by using IQR (Interquartile Range) approach.

Performed label encoding for category attributes (i.e., level of education, studied university).

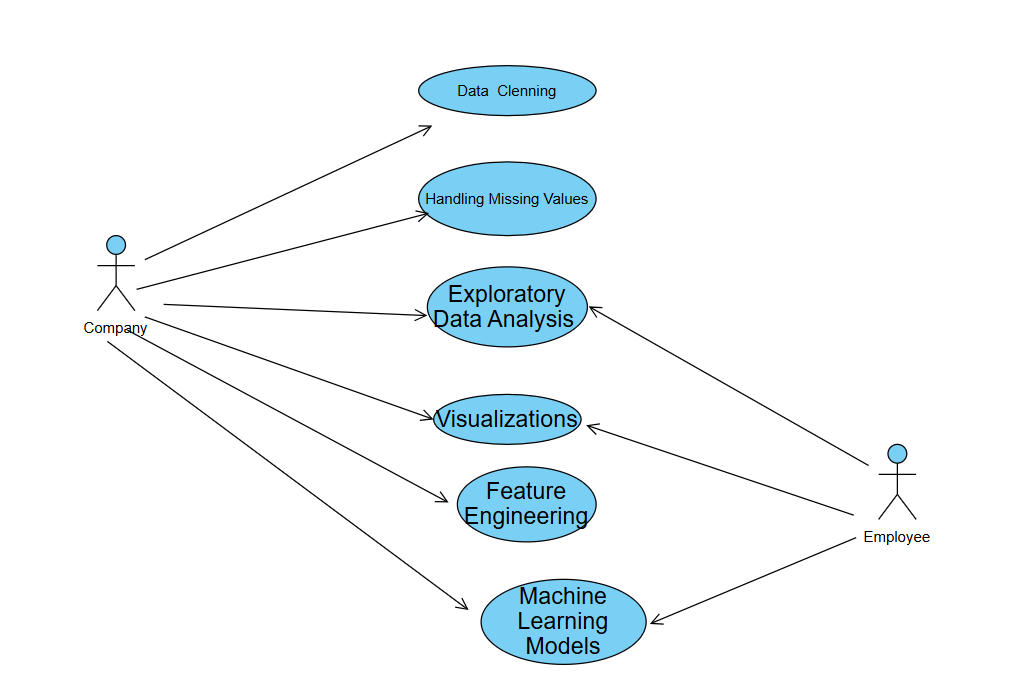
Scaled numerical attributes where necessary.

**Architecture Diagram:**



**UML Diagram (Use Case):**

A UML use case diagram illustrates the company-employee interaction:



**Steps Performed in the Project:**

Data Cleaning: Treated missing values, identified outliers, normalized the data.

Exploratory Data Analysis: Visualized distributions via histograms, boxplots, and frequency polygons.

Feature Engineering: Converted categorical features to numerical features using LabelEncoder.

Data Visualization: Used heatmaps, scatter plots, and bar plots.

Machine Learning Implementation:

Classification Model: Performed ML algorithm to classify the data.

Clustering Model: Implemented clustering techniques to group similar data points.

Evaluation: Measured model performance on accuracy and other parameters.

**Key Findings:**

Recorded findings and scope for improvement.

The dataset had strong variations in education levels and job-seeking behaviour.

EDA and visualizations exposed insightful patterns and relationships between variables.

Increased training hours tended to raise the probability of job-seeking success.

There was significant influence of level of education on the employment situation, with high levels having associated them with improving work opportunities.

**Challenges & Improvements Challenges Faced:**

Handling missing values without skewing the data.

Outlier detection and removal without data integrity loss.

Effective encoding of categorical variables.

Selecting an appropriate machine learning model for classification and clustering.

Hyperparameter tuning for model accuracy enhancement.

Providing good data balancing to avoid biased predictions.

Reducing computation time for big datasets.

Feature importance interpretation and feature selection.

Defending against class imbalances in classification problems.

Improving model generalization to prevent overfitting.

**Changes Made:**

Applied label encoding on categorical variables.

Enhanced data preprocessing methods to preserve key information.

Applied outlier removal methods for robust model training.

Applied applied feature scaling to maintain model training consistency.

**Conclusion:**

The project achieved doing data preprocessing, exploratory data analysis, and applying machine learning to a dataset. Classifying and clustering the models illustrated the way data is structured and it behaves. Intensifying challenges as the process progressed permitted intensified understanding of prime improvements for use in subsequent projects. Overall, the project illustrates why data science procedures matter when getting decision-making concepts from information examination.