INTEL UNNATI INDUSTRIAL TRAINING

PROJECT REPORT

ON

PROBLEM STATEMENT

"Knowledge Representation and Insights Generation from Structured Dataset"

Institution:

Baba Farid College of Engineering and Technology.

Submitted By: Team Learners.

Team Members:

- Harmandeep Kaur
- Harmandeep Singh
- Jashandeep Singh
- Vansh
- Vikramjeet Singh

Table Of Content:

- Introduction
- Acknowledgement
- Data Description
- Methodology
- Result
- Conclusion

ACKNOWLEDGEMENT:-

I would like to extend my sincere gratitude to Intel Corporation for providing me with the opportunity to participate in the Intel Industrial Program and work on Problem Statement 12. This experience has been instrumental in enhancing my knowledge and skills in data analysis and machine learning.

First and foremost, I am deeply thankful to my mentor, **"Er Sumeet Bharti",** for their invaluable guidance, constant support, and insightful feedback throughout the project. Their expertise and patience were crucial in overcoming challenges and achieving the project objectives.

I would also like to express my appreciation to the entire Intel Industrial Program team for their comprehensive support and for creating an environment conducive to learning and growth. The resources and workshops provided were immensely beneficial.

Sincerely,

Team Learners

Introduction:

In today's data-driven world, effectively utilizing structured datasets is crucial for gaining a competitive edge. **Knowledge Representation** involves organizing raw data into a structured form that highlights relationships and meanings, making it easier to work with. **Insights Generation** is the process of analysing this organized data to uncover patterns and trends that can inform decision-making.

Objectives:->

Enhance Data Accessibility and Usability:

- Organize and structure data for easy access and use, enabling informed decision-making.

Enable Effective Knowledge Representation:

- Develop models that accurately represent data relationships and meanings.

Uncover Hidden Patterns and Insights:

- Use analytical techniques to identify patterns, trends, and correlations within the data.

Enhance Data Visualization and Communication:

- Create visualizations that effectively communicate insights to stakeholders.

Data Description:

Source of dataset:" UCI Machine Learning Repository"

About Dataset:

The Adult dataset, also known as the "Census Income" dataset, is a widely used benchmark dataset in the field of machine learning and data analysis. It is derived from the 1994 U.S. Census database and is primarily used for tasks such as classification, regression, and data preprocessing exercises.

The Adult dataset contains demographic and employment information for approximately 48,842 individuals. The primary objective when working with this dataset is to predict whether an individual's income exceeds $50,000 per year based on their demographic and employment attributes.

Features of dataset:

1. Age: Continuous. The age of the individual.

2. Workclass: Categorical. The classification of the individual's employment (e.g., Private, Self-emp-not-inc, Local-gov, etc.).

3. fnlwgt: Continuous. The final weight or the number of people the census believes the entry represents.

4. Education: Categorical. The highest level of education achieved (e.g., Bachelors, HS-grad, 11th, etc.).

5. Education-Num: Continuous. The numeric representation of the education level.

6. Marital Status: Categorical. The marital status of the individual (e.g., Married-civ-spouse, Divorced, Never-married, etc.).

7. Occupation: Categorical. The type of work performed by the individual (e.g., Tech-support, Craft-repair, Other-service, etc.).

8. Relationship: Categorical. The individual's relationship status (e.g., Wife, Own-child, Husband, etc.).

9. Race: Categorical. The race of the individual (e.g., White, Asian-Pac-Islander, Amer-Indian-Eskimo, etc.).

10. Sex: Categorical. The gender of the individual (Male or Female).

Handling Missing Values: The dataset contains some missing values, represented by the '?' symbol, particularly in the workclass, occupation, and native-country attributes. These need to be handled appropriately, either by removing records with missing values or by imputing them.

Encoding Categorical Variables: Many of the attributes are categorical and need to be encoded numerically for use in machine learning algorithms. Techniques such as one-hot encoding or label encoding are commonly applied.

Feature Scaling: Continuous variables like age, fnlwgt, education-num, capital-gain, capital-loss, and hours-per-week may require scaling to ensure they contribute equally to the model's performance.

For numerical values: fill with mean of column values.

For categorical values: fill with mode of column values.

Methodology:

Data Collection and Preparation

- **Data Source Identification**: Identify the sources of structured data, such as databases, CSV files, APIs, etc..
- **Data Cleaning**: Handle missing values, remove duplicates, correct errors, and standardize formats.
- **Data Transformation**: Normalize or standardize data, create new features, and perform aggregations if necessary.

Exploratory Data Analysis (EDA):

Unique Values and Dataset Information: The dataset was examined for unique values, general information, and statistical summary.

Visualization: Various visualizations were used, including histograms, correlation heatmaps, distribution plots, and boxplots for each feature to understand the data better.

Knowledge Representation:

Data Visualization: The data was visualized using histograms, correlation heatmaps, and bar plots to represent the distribution and relationships among different features.

Preprocessing:

Label Encoding: Converting Categorical values into numerical values using LabelEncoder.

Data Normalization: The features were normalized using Min-Max Scaler to remove the median and scale according to the Inter-Quartile Range.

Outlier Detection: Detect outlier is important part for ensuring model accuarcy .For Outlier Detection we use two methods:**Z-Score and IQR method** . But in our case we **use Z-Score Method** because Using IQR Method it reduced our dataset in quite large amount.

Pattern Identification:

Model Training: A Random Forest Classifier was used to train the model on the dataset.

Model Evaluation: The model's performance was evaluated using accuracy score, classification report, confusion matrix, and cross validation scores.

Insights Generation:

Feature Importance: The importance of each feature was determined using the feature importances and attribute of the Random Forest model and visualized using bar plots

Results: Knowledge Representation and Insights Generation from Adult Dataset

Data Overview

- **Dataset**: 48,842 instances, 14 attributes (age, workclass, education, marital status, occupation, etc.)
- **Target Variable**: Income (`<=50K` or `>50K`).

Data Preprocessing

- **Missing Values**: Imputed using mode.
- **Encoding**: One-hot encoding for categorical variables.
- **Scaling**: Standardized numerical features.

Exploratory Data Analysis (EDA)

- **Age Distribution**: Majority are 20-50 years old.
- **Education Levels**: High school and some college are most common.
- **Occupations**: Common roles include "Exec-managerial" and "Prof-specialty".
- **Income Insights**:
  - **Age**: Higher income more common in older age groups.
  - **Education**: Higher education correlates with higher income.
  - **Hours per Week**: Higher income linked to more hours worked.

Knowledge Representation:

In this we plot two types of graphs:

- Frequency Graph
- Distributed Graph

For Numerical Values we plot **Histogram** .

For Categorical Values we plot **Bar graph**

For Boolean values we plot **Pie Chart**

We Draw Distributed graph considering target column.

If target column is Numerical we draw **for Numerical values Scatterplot and for categorical we draw Stripplot**

If target column is Categorical we draw **for Numerical values Strippplpot and for categorical we draw Countplot**

If target column is Boolean we draw **for Numerical values Stripplot and for categorical we draw Countplot.**

Insights Generation

- **Patterns**: Identified demographic factors influencing income.
- **Predictive Modeling**: Developed models (e.g., logistic regression, decision trees) to predict income levels, achieving high accuracy.
- **Visualization**: Created dashboards showing key insights (e.g., income distribution by age and education).

Conclusion:

This comprehensive study, conducted as part of the Intel Training Program Problem Statement 12, focused on knowledge representation and insight generation using the Adult dataset. Sourced from the UCI Machine Learning Repository, the Adult dataset includes demographic and employment-related information for approximately 48,842 individuals. The primary objective was to predict whether an individual's income exceeds $50,000 annually and to generate actionable insights from the analysis.

Throughout this project, a meticulous and systematic approach was adopted, encompassing data preprocessing, exploratory data analysis (EDA), feature engineering, model training, evaluation, and insight generation. The results obtained from each phase have provided valuable findings and significant recommendations.

Recommendations:

1. Policy Implications: Insights derived from the analysis can inform policy decisions aimed at addressing income inequality. For instance, enhancing educational opportunities and vocational training programs could significantly impact income distribution.

2. **Targeted Interventions:** The findings suggest targeted interventions for specific demographic groups that are more likely to earn less than $50,000 annually, such as those with lower education levels or certain occupational roles.

3. Further Research: The study opens avenues for further research into the causal relationships between demographic factors and income. Investigating these relationships could yield deeper insights and more effective strategies to mitigate income disparities.