Artificial Intelligence and Machine Learning

Project Report Semester-IV (Batch-2022)

Case Study: Adult Dataset

[https://drive.google.com/file/d/1VC3dEPrL30siFOCbTn8UhhSVu1ew6t5I/view?usp=sharing](https://drive.google.com/file/d/1VC3dEPrL30siFOCbTn8UhhSVu1ew6t5I/view?usp=sharing" \t "https://www.smartwebsolutions.org/text-to-url/_blank)

**Supervised By: Submitted By:**

Rajeev Bhardwaj Tanishka Kautish

2210990894

G-13

**Department of Computer Science and Engineering**

**Chitkara University Institute of Engineering & Technology,**

**Chitkara University, Punjab**

**Description about Case Study: -**

1. Display Top 10 Rows:

How can we view the top 10 rows of our dataset?

1. Display Last 10 Rows:

How can we check the last 10 rows of our dataset?

1. Find the Shape of Our Dataset (Number of Rows and Number of Columns):

Where can we find the shape of our dataset, i.e., the number of rows and columns?

1. Getting Information About Our Dataset:

How do we obtain information about our dataset, including the total number of rows, total number of columns, data type of each column, and memory requirements?

1. Fetch Random Samples from Dataset (50%):

How do we randomly sample 50% of the data from our dataset?

1. Check Null Values in Dataset:

How do we identify and handle null values in our dataset?

1. Perform Data Cleaning (Replace '?' with NaN) and Plot in Graph with Seaborn Library:

How can we replace occurrences of '?' with NaN in our dataset as part of data cleaning, and then plot the results using the Seaborn library?

1. Drop All Rows Having Missing Values:

How do we remove all rows from our dataset that contain missing values?

1. Check for Duplicate Data and Drop Them:

How can we identify and drop duplicate rows in our dataset?

1. Get Overall Statistics About the Data Frame:

How do we obtain overall statistical information about the entire data frame?

1. Bivariate Analysis on Graph:

How can we perform bivariate analysis and visualize relationships between two variables using graphs?

1. Replace Salary Values ['<=50k', '>50k'] with 0 and 1:

How do we convert salary values '<=50k' and '>50k' to numerical values 0 and 1 in our dataset?

1. Which work class is Getting the Highest Salary?

How can we determine which work class has the highest average salary in our dataset?

1. Who has a Better Chance to Get Salary >50k, Male or Female?:

How do we analyze and compare the chances of getting a salary >50k between male and female individuals?

1. Convert work class Column Datatype to Category Data:

How can we convert the data type of the ‘Work Class’ column to the category data type?

# Library: -.

1. Pandas
2. Seaborn
3. Matplotlib
4. Numpy

**Methods: -**

1. **read\_csv():**

Description: Reads a CSV file and converts it into a data frame.

1. **tail():**

Description: Displays the last few rows of the data frame.

1. **head():**

Description: Displays the first few rows of the data frame.

1. **shape():**

Description: Returns the shape (number of rows, number of columns) of the data frame.

1. **info():**

Description: Provides basic information about the data frame, such as column types and missing values.

1. **sample():** Description: Check Random Samples from Dataset
2. **isnull():**

Description: Returns True/False for each value in the data frame, indicating whether the value is missing (NaN) or not.

1. **duplicated():**

Description: Check for duplicate data in the dataset

1. **drop\_duplicates():**

Description: Delete the duplicated data

1. **dropna():**

Description: used to remove missing (NaN) values from a Data Frame or Series

1. **describe():**

Description: Checks overall statistics of the dataset

1. **sns.heatmap():**

Description: Plots rectangular data as a color-encoded matrix, where values are represented by colors.

1. **sns.boxplot():**

Description: Draws a box plot to show the distribution of numerical data across different levels of one or more categorical variables.

1. **max():**

Description: Returns the maximum value in a column of the data frame.

1. **min():**

Description: Returns the minimum value in a column of the data frame.

1. **mean():**

Description: Calculates the mean (average) value of a column in the data frame.

1. **len():**

Description: Returns the number of rows in the data frame

1. **group\_by():**

Description: a method for grouping data by one or more columns, enabling the application of aggregate functions to each group independently.

1. **astype():**

Description: a method for explicitly converting the data type of a pandas object (e.g., DataFrame column) to a specified data type.

1. **sort\_values():**

Description: Helps arrange the data in either ascending or descending order based on the values in the specified columns

1. **to\_numeric:**

Description: Used to convert the values of a Series to numeric format.

1. **value\_counts():**

Description: Counts the unique values in a specific column of the data frame.

1. **replace():**

Description: Used to replace values in a Data Frame or Series.