## Artificial Intelligence

## and

## Machine Learning

Semester-IV (Batch-2022)

Case Study: adult.csv

<https://drive.google.com/file/d/1ydketq_g1cNAvKHTxnXq7Qj5O-HdJIbr/view?usp=drive_link>

A red and white sign

Description automatically generated with low confidence

**Supervised By: Submitted By:**

Faculty Name Name :Simranpreet kaur

Roll Number : 2210990854

Group : G13(A)

**Department of Computer Science and Engineering**

## Chitkara University Institute of Engineering & Technology,Chitkara University, Punjab

**DESCRIPTION**

**Libraries:**

**Pandas:** Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.

The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Pandas allows us to analyze big data and make conclusions based on statistical theories.

Pandas can clean messy data sets, and make them readable and relevant.

**Matplotlib:** matplotlib is a Python library used for creating static, interactive, and animated visualizations in a wide variety of formats. It is part of the larger Matplotlib library, which provides a comprehensive suite of tools for data visualization. Matplotlib specifically provides a collection of functions that make it easy to generate plots, charts, and other visual representations of data.

**Seaborn:** Seaborn is a data visualization library built on top of Matplotlib in Python. It provides a high-level interface for creating informative and attractive statistical graphics. Seaborn is particularly well-suited for working with complex datasets, allowing users to quickly generate visually appealing visualizations with minimal code.

**Numpy:** NumPy is a Python library that provides a multidimensional array object and various routines for processing those arrays. It is the fundamental package for scientific computing with Python, and it is widely used in many fields of science and engineering.

**Questions: Display top 10 rows**

1. display last 10 rows
2. find shape of our dataset(number of rows and number of columns)
3. Getting information about our dataset like total number rows,total number of Columns,data
4. Type of Each column and memory Requirement
5. Fetch Random Samples from DataSet(50%)
6. check NULL values in dataset
7. Perform Data Cleaning(Replace '?'

with NaN) and plot in graph with seaborn library

1. Drops all rows having missing values
2. Check for duplicate data and drop them
3. Get overall Statistics about the dataFrame
4. Bivariate Analysis on graph
5. Replace Salary values ['<=50k','>50k'] with 0 and 1
6. Which workclass getting the Highest Salary?
7. how has better Chance to get Salary>50k Male or Female
8. covert workclass Columns Datatype to category Data

**Methods:**

1. pandas.read\_csv(): **p**andas.read\_csv() is a function in the pandas library that reads a comma-separated values (csv) file into a Dataframe. It has many parameters that allow you to customize how the file is read, such as sep, header, index\_col, usecols, dtype, skiprows, nrows, na\_values, and more.
2. Data.head(): data.head(10) is a method that returns the first 10 rows of the DataFrame data. It is useful for getting a quick overview of the data or checking the format and column names. You can change the number of rows by passing a different value to the method, such as data.head(5) or data.head(20)
3. **data.tail(10) :** data.tail(10) is a method that returns the last 10 rows of the DataFrame data. It is useful for getting a quick overview of the data or checking the format and column names. You can change the number of rows by passing a different value to the method, such as data.tail(5) or data.tail(20)
4. **data.isnull().sum():** data.isnull().sum() is a method that returns the number of missing values in each column of the DataFrame data. It is useful for detecting and handling missing data in your analysis. It returns a Series with the column names and the count of null values.
5. Len(): The len() function is a built-in function in Python that returns the number of items in an object. It can be used with sequences, such as strings, lists, tuples, and ranges, or with collections, such as dictionaries, sets, and frozen sets.
6. **max():** The max() function is a built-in function in Python that returns the maximum value among the given arguments.
7. Sample() : sample() is a method provided by the random module. The sample() function is used to generate a random sample from a sequence or iterable (such as a list, tuple, or set) without replacement. This means that each element in the original sequence can only be chosen once in the resulting sample.
8. Heatmap(): In Seaborn, the heatmap() function is used to create a heatmap, which is a graphical representation of data in a matrix format. Heatmaps are particularly useful for visualizing the relationships between two categorical variables or for displaying the correlation matrix of a dataset. The colors in a heatmap represent the values of the data points in the matrix, making it easy to identify patterns and trends.
9. Info() : the **info()** method is used to display a concise summary of a DataFrame, including the data types of columns, the number of non-null values, and memory usage.
10. Isin(): isin is a method in the pandas library of Python. It checks whether each element in a DataFrame is contained in a list of values and returns a DataFrame of Boolean values of the same shape as the original DataFrame.
11. Replace(): replace() is a method that can be used to replace a part of a string with another string. It takes two or three arguments: the old substring to be replaced, the new substring to replace it with, and optionally the number of times to perform the replacement.
12. Duplicated: duplicated() is a method in the pandas library of Python. It returns a Boolean Series indicating whether each row in a DataFrame is duplicated or not.
13. Any(): any() is a built-in function in Python that returns True if any element of an iterable is truthy, and False otherwise. An iterable is an object that can be looped over, such as a list, a tuple, a set, a dictionary, or a string. A truthy value is a value that evaluates to True in a Boolean context, such as a non-zero number, a non-empty string, or a non-empty container.
14. drop\_duplicates() : drop\_duplicates, which is a method in the pandas library of Python. It removes duplicate rows from a DataFrame based on a subset of columns or all columns
15. describe() : describe() is a method in the pandas library of Python. It generates descriptive statistics for a DataFrame or a Series of numeric or object values. It summarizes the central tendency, dispersion, and shape of the data distribution, excluding NaN values. It also analyzes the unique, top, and frequency values for object data, such as strings or timestamps
16. boxplot() : boxplot() is a function in the matplotlib.pyplot module of Python. It draws a box and whisker plot for a given data set. A box and whisker plot is a graphical representation of the distribution of a numerical variable, showing the median, quartiles, and outliers of the data
17. unique() : unique() is a function that can be used to get the unique values from a list, a series, or an array in Python. It can be imported from different modules, such as pandas, numpy, or operator.
18. Value\_counts(): value\_counts() is a method that can be used to count the frequency of unique values in a pandas Series or DataFrame. It returns a Series containing the counts of each value, sorted in descending order by default.
19. Countplot() : countplot() is a function in the seaborn module of Python. It draws a bar plot that shows the counts of observations in each categorical bin. You can use various parameters to customize the appearance and behavior of the countplot, such as x, y
20. Mean() : mean() is a function that can be used to calculate the arithmetic mean of a given set of numbers in Python. The arithmetic mean is the sum of the numbers divided by the number of elements.
21. Groupby() : groupby() is used to group a pandas DataFrame or Series by one or more columns or levels and apply a function to each group.
22. Astype() : astype() is a method that can be used to change the data type of an array, a series, or a dataframe in Python. It can be imported from different modules, such as numpy, pandas, or matplotlib.pyplot