**PRACTO**

**Objective:**

Machine learning model that can predict the consultation fee for any doctors using their key features.

**Teach stack: -** Python (Beautiful soup, Selenium, NumPy, Pandas, Matplotlib, Seaborn)

Excel

PowerPoint

**Steps: -**

1. Imported libraries and module.
2. Web Scraping and storing data into a dataframe

🡪 Created an empty dataframe(df)

🡪 Created list of cities whose data I have to web scrape.

🡪 Created a variable ‘Speciality’ to which we will assign the name of speciality, this we have to

change each time for different speciality.

🡪 After this for loop will execute which will iterate the list containing the name of cities and

web scraping will start.

🡪 Here I have used chrome driver to automatically navigate to web pages.

🡪 After this a variable ‘url’ is created to which we will assign the link of webpage from where

scraping has to be done.

Here I have used the concept ‘Formatted string’ of string concatenation, here the name of

city will concatenate with the url.

🡪 To collect the data the very first step is to scroll down the web page till end and then get the

HTML of that web page.

🡪 After scrolling, first the link of each doctor profile is collected and then by using that link it

will open that link and the data of each doctor will scraped.

🡪 After collecting all the data, all data is stored in dataframe ‘df’.

🡪 Similarly, same steps to follow to scrape data of other specialities.

1. After web scraping I have concat all the csv file into one.
2. The next step is of data cleaning as there are some noises as well as data type issue in the dataframe.
3. After all the above steps now our dataframe is ready for analysis.
4. Now I have performed EDA (Exploratory Data Analysis) to know more about data and find some useful insights.
5. Now next step is of Data preprocessing, this is necessary to build the machine learning model.

🡪 Dropping unnecessary columns which are not necessary to build the model.

🡪 Encoding (to convert categorical column into numerical column) for this I have used:

* Label encoding
* One hot encoding

🡪 Extracting independent and dependent variables.

🡪 Splitting data into training and testing,

🡪 Standardization for feature scaling.

1. Now I have built a Linear regression model (because I have to predict salary which is continuous value).

🡪 On evaluating the model, the metrics that I get are:

MSE = 93471.58

RMSE = 305.73

R-squared = 0.5

1. Now to further improve the I have PCA (Principal Component Analysis) and again built Linear Regression model.

🡪 On evaluating the model, the metrics that I get are:

MSE = 93876.21934263923

RMSE = 306.39226384267477

R-squared = 0.4979855513863929

Since the feature are less in dataframe so there is not significant change on MSE, RMSE, and R-squared values.

1. After this I have created a Dashboard using Excel showing various insights.
2. At last, I have created a PowerPoint presentation.