ACCELERATED KOMPUTING(AnK)

1) COVID 19 DATASET

As a Supervised Machine Learning beginner's practice project, I cleaned and analyzed the data from the file I downloaded from ourworldindata org using ***pandas*** framework.

Using ***pd.read_csv,*** and pasting the path of the file, it was stored in a variable and top five and end five data was shown using ***.head()*** and ***.tail()*** feature. The ***.shape*** attribute returned a tuple showing the number of rows and columns. Using ***.value_counts()*** a column was selected and number of occurrences of each country was shown. ***.isna().any()*** created a new dataframe with Boolean values checked and displayed any missing values in particular column. ***.isna().sum()*** created a Boolean dataframe and gave the sum for each True value treated a 1. The covid cases of only Nepal was filtered by creating a new dataframe and head/tail used to display top and bottom five datas.

For visualization, seaborn and matplotlib was imported where Seaborn is a statistical data visualization library based on Matplotlib. ***.set*** and ***.lineplot*** was used with appropriate figure size to create a time series plot

2) POKEMON DATASET

Similar trend was followed while dealing with the dataset of pokemons. But other than previously used attributes, ***.min(), .max(), .mean(), .median***() was used in columns with numerical data to give the minimum, maximum, mean and median values of the column respectively. Column rename was also done using ***.rename(columns={'original name' : 'new name'}, inplace =True)*** where the old name was mapped to new name.

To visualize the distribution of HP of grass type pokemons, a histogram was plotted. ***.histplot*** was used and a ***Kernel Density Estimate*** was also added . Another bar plot/count plot was plotted to show how many were legendary grass type pokemons. ***.countplot*** was used for this purpose. Similar plot was made for various columns too. A scatter plot was also generated using ***.scatterplot*** and it was found that as the special attack increased, special defense also increased and vice-versa.

3) CREDIT CARD FRAUD DETECTION

The dataset was loaded into the pandas dataframe and the first five rows was displayed. Then, the date_time column was converted to datetime data type using ***to_datetime()*** function and hour, day and month was extracted and made new columns using ***.dt*** accessor. The irrelevant features and columns were dropped from the dataset using ***drop*** . The categorical columns in the dataset was handled and ***LabelEncoder()*** was used to encode these columns into numerical format by running a loop which iterated through each column name. ***.fit_transform(_)*** was used to fit the encoder to the data in current column, then, it transforms the data in that column,

replacing the original categorical values with the assigned numerical labels. A function called haversine is defined that took four arguments:latitude and longitude of first and second points; ***def haversine(lat1, lon1, lat2, lon2) .. np.array([geodesic((a, b), (c, d)).km for a,b,c,d in zip(lat1, lon1, lat2, lon2)])*** calculated the geodesic distance for each pair of coordinates using ***geopy.distance*** library. Before calculating the distance, ***.dropna*** was used to handle any missing values in specified columns as geodesic function cannot handle missing coordinate values.

Now, the data is prepared for machine learning model by separating the features from the target variable. A list is created containing the name of columns that is used as input features for the model to learn patterns and make predictions. Then, a bar plot is generated to show the distribution of target variable without SMOTE. The imbalance in dataset's target variable was visually observed. Now, Synthetic Minority Over-sampling Technique is applied to deal with imbalanced datasets by setting random state. The countplot was then visualized and minority class was oversampled to balance the dataset. Now, the data was split into training and testing sets by using ***train_test_split*** function using 20% of data for testing and 80% for training. It returned four variables: features for training and testing set and target for training and testing set. LightGBM classifier model was trained by using ***lgb.LGBMClassifier(_)*** to initialize with hyperparameters such as gradient boosting decision tree type, maximum leaves in one tree, step size shrinking during boosting, number of boosting rounds, depth of tree, etc. ***lgb_model.fit(_)*** trained the LightGBM model using training data to learn from patterns to make predictions on unseen data. ***Lgb_model.predict(_)*** was used to make predictions on the test dataset which can be compared with actual target values to evaluate model's performance. The classification report was generated which contained precision, recall, f1 score and support for each fraud and non fraud case. Also ROC AUC score was calculated to be high which indicated good performance. ***lgb.plot_importance(_)*** was used to understand which features the model considered most important when making predictions about fraudulent transactions visually. The False Positive Rate, True Positive Rate and threshold was calculated to plot the Receiver Operating Characteristic and to calculate Area Under the Curve.

Finally, the ROC Curve and AUC was visually displayed . The ROC AUC score was 0.99 which showed the model's excellent power to distinguish fraudulent and non fraudulent transactions. The curve was also close to the top left corner which confirmed strong performance. The lgbmodel and labelencoders was saved using joblib library for future deployment.

4) CODE AND FILE

https://colab.research.google.com/drive/1pIkhAEPDyl52WOiQCs4Aa9-I52Hj3NXB?usp=sharing

https://drive.google.com/file/d/1jfE2enSMZyKHIWxeLjqiSJsoG8iZoESb/view?usp=sharing

5) SIC

I realized that I had failed to explain about my Samsung capstone project as it had been long time after its completion and I had not looked into it since then. You can view my documented project by clicking the link below:

https://drive.google.com/drive/folders/1TF-wTkk5QAqRo9qK8HxHlIN-uwJt-MSB?usp=sharing

Made By:

- Prayash Phuyal (078BME031.prayash@pcampus.edu.np)