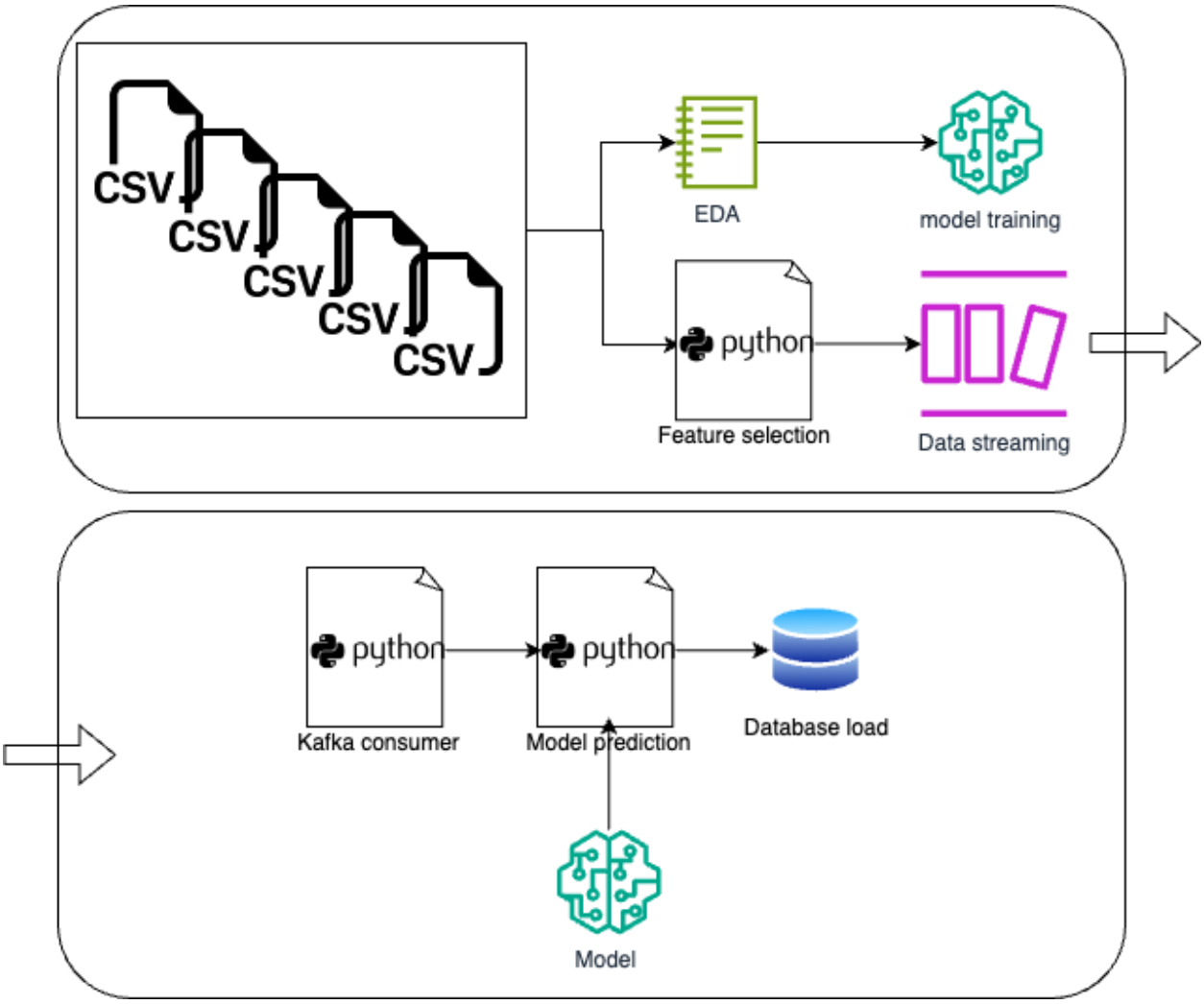
Workshop 3: Machine learning and Data streaming

Introduction

Given the 5 CSV files which have information about happiness score in different countries, train a regression machine learning model to predict the happiness score. Create the entire EDA/ETL to extract features from the files, train the model using a 70-30 data split (70% for training - 30% for testing) , stream the transformed data and then in the consumer get the data and use the trained model to predict the happiness score and store in a database the predictions with the respective inputs (features), the entire work is presented in the following figure. Finally extract a performance metric to evaluate the model using the testing data and the predicted data.



What is Expected

Is expected two train a regression model with all the EDA and feature selection, stream the data using kafka, use the trained model to predict values of the testing dataset and extract a performance metric

Data:

Five CSVs files of different years with Happiness information of different countries

Note: Data from different years may be different so keep in mind to evaluate each of the dataset and compare the features (columns).

Evidences:

EDA and model train, Model PKL file, predicted data and features in the database

Technologies

We expect you to use in this challenge:

- Python
- Jupiter Notebook
- Database (you choose)
- Kafka
- CSV files
- Scikit-learn library