Homework 3 – Mining Of Massive Datasets
Diabetes Prediction Model

For this assignment is required the dataset used is The Diabetes Health Indicators Dataset from https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

Task 1 – Creating offline.csv and online.csv and training classification model

The dataset is loaded and checked for linear dependencies between features. For that, a heatmap is used and it can be seen that the features aren't linearly correlated, so there isn't a need to do feature selection in order to remove the linearly dependent features.
Also the mean and standard deviation of the features of the dataset are checked int order to ensure that there aren't any large deviations between entries meaning there is no need to scale the data.
The dataset is then divided into two datasets with 80/20 ratio while also ensuring that the class distribution of the target variable is preserved in both datasets. The larger dataset is named offline.csv and is used to train the classification model, while the smaller dataset online.csv is used for prediction in the third task.
The offline.csv dataset is loaded and the target column "Diabetes_012" datatype is changed from float to int because classification models work with integer values only.
Next the dataset is divided into training and testing datasets, again using a 80/20 ratio while also preserving the class distribution of the target variable in both the training and testing datasets.
A Grid Search with f1-scoring is used in order to find the best hyperparametars for the three classification models, namely DecisionTreeClassifier, RandomForestClassifier and XGBClassifier.
After testing multiple values for the hyperparametars, the best ones are selected for training the actual models. Then using K-Fold Cross-Validation, the best model for this assignment is chosen.
In this case, that's the XGBClassifier which is saved locally in the file diabetes_prediction_model.pkl and used later for the third task.

Task 2 – Kafka Patient Entry Producer

The online.csv dataset is loaded and the target feature is removed, then the dataset is turned into json format that contains the feature names and the data from the dataset, which is then parsed into a dictionary.
A Kafka Producer is made that takes the feature names, an entry from the dataset and a timestamp, parses it into a json string and sends them to a local server that works from a docker container under the topic "heath_data".

Task 4 – Live Prediction

The model for predicting diabetes that was saved in task 1 is loaded and will be used for predicting new data.
The sample data send from the KafkaProducer containing entries from the online.csv dataset from task 2 is received using a KafkaConsumer that reads the data from the local server under the topic "health_data". A dataset is constructed from the json string received from the KafkaConsumer.
and the sample is then given to the model that predicts whether the patient has diabetes or not, adds the prediction to the received sample and sends it via another KafkaProducer under the new topic "health_data_predicted."

Task 5

A KafkaConsumer is reading the data received on the topic "health_data_predicted" to check if the assignment is successful.