Streaming using Apache Spark Structured Streaming and Apache Kafka

A large air quality dataset of India is streamed using apache kafka and apache spark structured streaming. ML library is used for applying machine learning on the streamed data. The code is entirely written in python language. The air quality dataset is taken from kaggle <https://www.kaggle.com/sohelranaccselab/air-quality-data-in-india-2015-2020?select=city_hour.csv>. The ultimate aim of this project is to use apache kafka to fetch the data and feed to spark engiene and use machine learning to predict the quality of air. Initially the dataset is cleaned and the final column is converted into 2 categories( good and satisfactory are converted to good. Severe, poor and very poor are converted to poor). Now this dataset is given to kafka producer. Kafka producer converts csv file to json format and stores the produced data in kafka topic. Now the spark(kafka consumer) creates a session and reads the stream from the kafka topic and converts to json object type. Now, ml library is used to apply machine learning on the streamed batches. Logistic regression is used to classify the air quality and finally the stream batches along with the prediction are displayed in the console.