Multivariate time series analysis

BITS F464 Machine Learning

Assignment 2

Submitted By:

Group 8

Aditya Agrawal : 2020B5A42010P

Niranjan Chaudhari: 2020B5A30929P

Rahul James: 2020A2PS1334P

Submitted to:

Prof. Navneet Goyal,

CSIS Department,BITS Pilani

Introduction

We used the dataset of Air Quality Index with 15 columns, including labels such as Date, Time, 10 columns for various gas concentrations, and 3 columns for temperature, relative humidity, and absolute humidity. This dataset is available in the CSV file named 'AirQualityUCI.csv'.

Initially, the dataset needed cleaning due to random values of "-200" present in various columns. We cleaned the dataset by replacing these "-200" values with the mean of the respective columns. The cleaned dataset is stored in the file named 'Cleaned_AirQualityUCI.csv'.

For solving Q2, we created a new column called 'Feels Like'. This column contained labels 'HOT' and 'COLD' based on the temperature values. For temperatures greater than 25, we labeled it as 'HOT', and for temperatures less than 25, it was labeled as 'COLD'.

The dataset with the 'Feels Like' column is stored in the CSV named 'Categorical_AirQualityUCI.csv'. We removed the 'Feels Like' column, which contained the labels, to implement clustering (solving Q3).

Results

Q1)
RMSE-Root Mean Square Error
MAE-Mean Absolute Error

Models	RMSE	MAE
Logistic Regression	2.217	1.647
Random Forest	0.233	0.13
Extra Random Trees	0.283	0.16
AdaBoost	2.057	1.611

Q2)

Models	Precision	Recall	F1-Score
Naive Bayes	0.91	0.886	0.891
KNN	0.895	0.897	0.895
SVM	0.862	0.865	0.855
Decision Trees	1	1	1

Q3)
SSE-Sum of Squared Errors

Models	SSE
K-means	68041.845
EM-Clustering	2.175
K-medoids	0

Citations

https://www.kaggle.com/datasets/aayushkandpal/air-quality-time-series-data-uci

https://scikit-learn.org/