Amrita School of Computing

Dept of CSE

19CSE304 Foundations of Datascience

* Final Assignment(Srl No 6): Data Manipulation, Introduction to Time Series Data, Classifier and Regression.
* Marks: 05 marks out of 10 marks assigned for the EvaluationLab4.
* Date of issue: 25 Nov 2022
* Individual work: Due date for submission: 01 Jan 2023

------------------------------------------------------------------------------------------------------------------------------

**A. Data Manipulation-Pandas**

Q1. Demonstrate the following pre-processing activities: Indexing, missing value manipulation, concat, append, merge and join, aggregation, grouping, transform, apply, and pivot tables.

Reference: Ch3 Data Manipulation with Pandas (pp: 97 to 190) “Python Data Science Handbook” by Jake VanderPlas

*Dataset: US states data pp155(Students 1 to 25);*

*Birthrate data pp177(Studentss26 to 49);*

*Recipe data pp184 (50 and beyond)*

B**. Time Series**

Q2. Demonstrate the following functions:

* Index by time
* Resample, shifting and windowing
* Time-shifts (shift() and tshift()
* Rolling windows
* Visualize the data

*Dataset- Seattle Bicycle Counts -Ref “Python Data Science Handbook” by JakeVanderPlas.*

Reference: Ch3 Data Manipulation with Pandas( Pp188 to 202)

Q3. Demonstrate the following pre-processing activities:

* Indexing, Selection, Subsetting (pp94)
* Date Ranges, Frequencies, and Shifting (Leading and Lagging) Data
* Time Zone Localization and Conversion
* Periods and Period Arithmetic
* Period Frequency Conversion
* Resampling and Frequency Conversion
* Downsampling and Upsampling
* Moving Window Functions
* Exponentially Weighted Functions

*Dataset: stock\_px\_2.csv or any suitable data -Ref “Python for Data Analysis” by Wes McKinney*

Reference: Ch3 Data Manipulation with Pandas( Pp317 to 362)

**C. Classifier**

(Submit Q4 and any one from Q5, and Q6 i.e., (i) to include KNN classifier and any 2 other classifier models; and (ii) Carry out validation.

*Dataset: ckd.csv, wine.csv, breast-cancer.csv - Ref Text Book Adhikari.*

Q4-Consider the Chronic Kidney Disease (CKD) data. Prepare a classifier model as stated above. *Given that a new patient Alice has Haemoglobin****0*** *and Glucose levels* ***1.1*** *standard deviations above average, find if Alice is likely to have CKD.*

Q5-Prepare a classifier model for wine dataset – wine.csv.

Q6-Prepare a classifier model for breast-cancer.csv dataset.

**D. Regression**

Predicting a quantitative variable –Regression

Q7. Predict Home prices employing multiple regression. Compare performance with a KNN Regressor model used for prediction. Draw a residual plot for these predictions and offer your comments.

*Dataset: house.csv -Ref Text Book Adhikari.*

Q8. Implement a logistic regression model – employ scikit-learn library functions.

*Dataset: Train and Test from Titanic – Ref “Python for Data Analysis” by Wes McKinney*

==================================================================================