BINF 311 Data Analytics and Visualisation

Statistical Hypothesis Testing and Predicting Continuous Values

Group Project – 30 Points

Project Datasets:

This group project primarily contains three datasets namely Real Estate, Wine Quality Red and Wine Quality White with their filenames "**real-estate.csv**", "**winequality-red.csv**" and "**winequality-white.csv**" respectively.

The Real Estate dataset contains a range of information about housing such as transaction date, house age, distance to the nearest train station, convenience stores, house location and its price.

Both the Wine Quality Red and Wine Quality White datasets contain information about Wine's acidity, sugar, chloride, free sulfur dioxide, its density, its pH value, sulphates, proportion of alchohol and its quality.

Tasks to complete:

This project needs to be completed in a group of **maximum 4 students** and your individual report should include some description of what is Correlation Analysis, Hypothesis Testing and the significance of the value of P in determining the reliability of test results. Also include some description of Regression analysis for predicting continuous values.

Afterwards, use our practicals to analyse and visualise both the datasets and make a list of **python's program files** that has necessary codes to achieve correlation analysis, linear rigression, multiple regression for all the datasets.

Make a use of python’s pandas library to read/import each dataset into a dataframe, explore its data distribution and look for missing values and remove outliers if there are any. Explore the datatypes of each column using pandas different data handling methods.

Carefully explore all the datasets, identify and report **the necessity of analysis** (consider it as an open question) by showing descriptive statistics of appropriate columns and using various visualisation techniques included in the practicals to showcase data distribution and other exploratory tasks (pair plot could be one of them).

Identify any possible correlations, State appropriate hypothesis and interpret the generated value of P. Once confirmed, make use of the identified associations / correlations in your data to build regression models (single and multiple) to predict housing price and wine quality (Red and White).

Follow a chronological order of analytical tasks by creating multiple exercises and reporting your analysis with python codes and visualisation plots. Devide your report into two sections namely section A and B for **Real Estate** and **Wine Quality (Red and White)** datasets respectively.

The project files (python files with the datasets) need to be uploaded in a zipped document to the classroom and a soft copy of the report can be submitted to Turnitin system.

Both the soft and hard copies should have a name of your group, participating students and the individual hard copies need to be submitted in my office on the deadline. After the submission, the students need to come with their respective group members to demonstrate their work.

Notes:

To write a good, clear and cohesive essay, review our lecture slides, do some research and read relevant literature online.

***Your write up needs to be in your own words**, no need to write any fancy words if you cannot think of any, just explain your conceptual understanding in plain, simple English.*

*Although the project is to be conducted in groups, the report needs to be prepared individually with **at least 2000 words** and it should include relevant visualisations. The structure of the report doesn't need to be of any specific format as long as the contents are in chronological order, it is acceptable.*

*Make use of some appropriate citations using **IEEE referencing style** whenever any external content is used as a part of your report to make it more sound and concrete.*

*A guide to make references in IEEE style is attached with this assignment. A hard copy of the report needs to be submitted in my office **by the deadline** and a soft copy needs to be submitted to the Turnitin system.*

Each group needs to present their experiments in a demonstration of 5 to 10 minutes.