MET CS 677 Term Project

100 points

1. Assignment Description

Term project is an important integrated part of the course.

Select a medium size data set from the available public data sets for analysis.

Follow general design and analysis process in your project.

- Describe a research scenario and state your research question.
- Describe the data
- Data visualization and plots to show different aspects of the dataset
- Preprocessing of the data
- Analysis the data using as many algorithms that are covered in the course, and as they are applicable. At least two algorithms. You are welcome to add algorithms that we have not covered in the class as well for extra points.
- Conclusion

Term Project Submissions

At the end of the term project your final report should include

- **A document capturing details of the term project (4 to 10 pages)**
- **All your Python code**
- **Presentation slides and 10 min presentation in the class**
- **Copy of this document with the boxes filled up**

2. Research Scenario Description (no more than 200 words)

Describe your research scenario in no more than 200 words. This is a general description of the use case. Follow class examples, first describe the overall scenario and then specify a specific research question.

An Auto Insurance Company's data has been taken into consideration for the following research scenario. Here we analyze the data on various data fields such as credit score, size of family, vehicle type and visualize the patterns in the process. This research can answer the questions like which customer is more likely to claim an insurance in the future and which customer would not be able to claim the insurance given the credit score. It will help the insurance company to get an insight of their customer's insurance needs and they can further modify their insurance policy such that they both profit from it. Further the research can also provide an insight on which vehicle type the insurance was claimed the most. Finally, the research scenario will use data science and machine learning models to classify the data and come to a certain conclusion which can be used by the company to forecast the insurance claims of their customer and further build a strategy for the same.

3. Describe the data set (no more than 200 words)

Briefly, describe the data set, including each data field. If possible provide a Link to the main data set source.

The data taken here is medium size data of an Auto Insurance company. It has fields like –

a) Credit score- It is an important data field as insurance company maps the customer with their credit score availability.

b) Vehicle type- This data field gives the customer's car model such as Car, Van, Utility, Truck.

c) Size of Family- It gives the customer's family members size which will be used by the auto insurance company.

d) Age- Simple data field which will map the customer to their age column. It can be anywhere from 18 to 60 years old.

e) Gender- Simple data field which will map the customer to their gender category.

There are more data fields like engine HP etc in the data set which will give more insight on customer's details.

3. Research Question (no more than 100 words)

Describe briefly in one or two sentences the main research question.

The research question for the research scenario are-

1) Which machine learning models turns out to be most accurate for the given data analysis?
2) Which customers are more likely to claim an insurance in the following years?

4. State Your Conclusion (no more than 100 words)

State the conclusion so that a none-data scientist can understand.

After using different machine learning models like KNN, Random Forest, Decision Tree, Naïve Bayes, Logistic Regression and Support Vector Machine we found KNN has best accuracy rate with 73.12% for our given data. Therefore, using K nearest Neighbor algorithm the company can predict which customer is more likely to claim the insurance.

Grading will be done based on

1. **Originality of selected data set and data analysis approach**
2. **Data Preparation, cleanup, preprocessing and data presentation**
3. **Data analysis**
4. **Quality of your code, output results, and conclusion**
5. **Quality of 10 min class presentation**