Group Project 1 - IE 5640 Data Mining

Distance measurement & analysis in the Cyber-Physical lab

NOTES:

1. This is a team assignment. Refer to Canvas (under People > IE5640) to find your teammates.
2. For the report, include a cover page stating the course number, semester, team number, and the name of all team members.
3. Only one of the group members needs to submit the PDF report (on behalf of everyone in the group).
4. You may include as many headers/titles as you want, a table of content, and introduction section, and an Appendix.
5. It is fine if you need to include small portion of your code throughout the text for better understanding. However, most, if not all, of the code must be represented at the end of your report under Appendix. You still need to include at least parts of the outputs throughout the main body.
6. Quality, proper interpretation, and organization is of priority not the length of your report.

Project Description:

A robot with an ultrasonic sensor mounted on it (facing forward) started moving toward two walls in two similar scenarios, except that it moved toward a concrete wall first, and did the same traverse, under a new experiment, toward a glass wall. The start and end points in two experiments are assumed to be the same. By visualizing the data in the provided datasets, you will realize that it was initially far from walls, as it started moving and got closer to a wall, the distance values decrease. There are some spikes and anomalies in the dataset which you will handle appropriately. Note that the ultrasonic sensor should have collected 10 data points in each second, but it may have not! (due to sensor failures, programming issues, etc.)

1) **Introduction (2.5 pts.):** Write 1-2 paragraphs stating what this project is about, nature of data, etc.

2) **Data Exploration (70 pts.):** Your team is given two datasets, one for a glass wall and one for a concrete wall. You need to explore the data and:
   a) Your Python code should create three predictors (Distance, Obstacle_Type, Angle_Approach) and one output variable (Time_Collision).
      i. Distance: Sensor readings in inches ("Value" column in each Excel spreadsheet).
      ii. Obstacle_Type: Concrete, Glass. You can name the Obstacle_Type category as "Glass" for dataset 1 and "Concrete" for dataset 2.
      iii. Angle_Approach: Angle of approach which should be random values 10 ± 5 for Wall and 0 ± 5 for Glass.
      iv. Time_Collision: Time (sec) to hit an obstacle. Max speed of the robot is 24in/s. Simply calculate this target value using: Time = Distance / Velocity(of robot).
   b) Your team can trim a few seconds of last rows of data where the same distance "Value" column in the datasets is exactly repeated.
   c) After trimming, deal with seconds that do not have exactly 10 records in them as directed below:
      - You can choose to impute the median/mean or any other statistic (of your choice) of distance measurements in that particular missing time step.
      - Where your imputed records go is also optional: in the beginning of that particular second, middle, etc.
      - Whether you impute all missing records in each second next to each other or randomly spread among all other records in that second is also your choice.
   d) Report and handle missing and outlier values.
   e) Use two plots of choice that make the most sense to present the nature of the data to a reader and to determine how both scenarios were similar and/or different.
   f) Combine the two datasets for both scenarios to get around 340 to 400 rows (34 to 40 seconds).
   g) Shuffle the data (in any way you can/want) to avoid data overfitting when using k-NN and Naive Bayes.

   To better understand the above requirements, note the following:

- A picture of how the above procedures can be done is partially shown below.

	A	B	C	D	E
1	Time	Distance	Obstacle_Type	Angle_Approach	Time_Collision
2	17:18:27	470.47	Concrete	9.117203084	19.60291667
3	17:10:49	122.05	Glass	0.325557157	5.085416667
4	17:18:31	130.71	Concrete	12.56163168	5.44625
5	17:10:54	41.73	Glass	-8.122086183	1.73875
6	17:18:42	4.72	Concrete	11.86658295	0.196666667

- It is fine that if after all the above are done, still one of your datasets has more/less data points than the other.

3) **Model Training (15 pts.):** Your team needs to split the data into 60% training and 40% validation. You should:
   a) Train and validate a k-Nearest Neighbors (k-NN) model.
   b) To work with Naïve Bayes, label "Time_Collision" as Low, Medium, High based on your team`s choice of threshold. You can choose to follow this or any other three thresholds of your choice: IF(Time_Collision<2, "Close", IF(Time_Collision<6, "Mid", IF(Time_Collision >=6, "Far"))). Train and validate a Multinomial Naive Bayes (NB) model. Hint: Multinomial NB does not handle continuous values such as Distance or Angle_Approach values, it works with discrete values only!
   c) Briefly interpret the above modelling results, accuracies, and errors.

4) **New Record Regression/Classification (10 pts.):** When you are done developing k-NN and NB, consider feeding one or a couple of random datapoints, and apply both your k-NN and NB to it/them, compare and determine if the results from the same data point(s) would be similar/different across k-NN and NB, and show the output (Time_Collision) of that/those new records(s) in your report.

5) **Conclusion and Future Work (2.5 pts.)**: What would your team conclude for this project and what are the potential areas for future work?