NTNU
Kunnskap for en bedre verden

Exam in IT6208
Innføring i Stordata

Contact persons during the exam:

Alexander Holt, alexander.holt@ntnu.no telephone 95 02 29 29

Xiaomeng Su, xiaomeng.su@ntnu.no telephone 48 01 81 53

Language: Please answer in English or Norwegian

Please submit a document (pdf or word) and your python code. You can put them in a zip file and upload it to Inspera.

TASK 1. CASE STUDY (50%)

The following tasks are related to case *"Dublin City Council Is Leveraging Big Data to Reduce Traffic Congestion"*. Make necessary assumptions if needed.

a)
What are the data sources used by the Dublin City Council to reduce traffic congestions? Describe the three V's of the data sources. Why is it difficult for Dublin City Council to use the traditional data processing tools and platforms?

b)
From the case, give one example of descriptive analytics, and one example of predictive analytics. What decisions are improved by these business analytics examples?

c)
In the case, it states "*Currently, the IBM team is working on ways to integrate data from rain and flood gauges into the traffic control solution- alerting controllers to potential hazards presented by extreme weather conditions, and allowing them to take timely action to reduce the impact on road users*". Now suppose that you are in charge of this project. Can you apply the CRISP model and write a brief data analytics proposal on how you plan to develop such a project?

TASK 2. DATA ANALYSIS IN PYTHON (50%)

All questions are related to the dataset satisfaction.csv, which is a dataset from an airline, containing passenger info and answers from questionnaires given to the passengers.

a)
Load the dataset and display the last 3 lines, to make sure everything is working correctly.

b)
Find the mean flight distance for all passengers.

c)
Find the mean flight distance for women under the age of 30 years.

d)
Create a line diagram for flights with a departure delay of ten minutes or more, where the departure delay in minutes is on the x-axis, and number of occurrences is on the y-axis.

e)
Create a bar chart showing the distribution of the age of the passengers.
Hint: the function size() for "grouped dataframes" can be of help here.

f)
Assume that gender, customer type, age, type of travel, and flight distance are predictors. Create a decision tree for what class (Business, Eco Plus, Eco) a passenger is likely travel on. How well does the tree predict, and how will deepening the tree affect the prediction error? You may find that the

class «Eco Plus» seems to be missing unless the tree becomes very deep. Can you think of why this is?

PS. I would advise against trying to draw very deep trees on webgraphviz.com, as it can quickly grind to a halt. In that case, it's better to simply open the dot-file in a text editor and search for the term "Plus".

g)
Are passengers traveling business class more or less likely to be satisfied than passengers traveling on Eco or Eco Plus? Create a regression tree which shows this.

h)
We assume people of the same gender and age have similar priorities when flying. If a male has been categorized as a loyal customer and have stated being satisfied with his flight, and having given the grade 4 for seat comfort, 5 when grading food or drink, 5 for gate location, 5 for wifi service, 5 for inflight entertainment, and 3 for ease of online booking. Using KNN and assuming a K of 6, how old is he likely to be?

----------The end ----------