# Dublin Business School

# Assessment Brief

# Assessment Details

|  |  |
| --- | --- |
| Module Title: | Statistics for Data Analytics |
| Module Code: | B9DA101 |
| Module Leader: | Dr Shahram Azizi |
| Stage (if relevant): |  |
| Assessment Title: | CA two |
| Assessment Number (if relevant): |  |
| Assessment Type: |  |
| Restrictions on Time/Length : | Submission before deadline |
| Individual/Group: |  |
| Assessment Weighting: |  |
| Issue Date: |  |
| Hand In Date: |  |
| Planned Feedback Date: |  |
| Mode of Submission: | Online |

**Guideline:**

* This CA assesses students on core concept in Hypotheses tests, GLM analytics, and Bayesian analytics.
* All questions are mandatory.
* Use R/Rstudio to solve questions and perform analytics.
* Any submission after deadline will not be considered and scored.

**Question 1**

Consider a relational dataset and specify your input and output variables , then:

* 1. Train the model using 80% of this dataset and suggest an appropriate GLM to model **ouput** to **input** variables.

**(10 Marks)**

* 1. Specify the significant variables on the **output** variable at the level of 𝛼=0.05 and explore the related hypotheses test. Estimate the parameters of your model.

**(10 Marks)**

* 1. Predict the output of the test dataset using the trained model. Provide the functional form of the optimal predictive model.

**(10 Marks)**

* 1. Provide the confusion matrix and obtain the probability of correctness of predictions.

**(5 Marks)**

**(Total: 35 Marks)**

**Question 2:**

Use the trafficstop dataset in the link;

https://github.com/shazizisazi/Datasets

consider **SrchVhcl** as the categorical output variable and **DrvMale**

and **DrvRace** as the input variables. Apply Naïve Bayes classifier to predict the output variable, to do so;

a) Compute the prior probability of the output variable. (5)

b) Compute the likelihood of input variables given the value of the output variable. (10)

c) Using (a) and (b), and Bayes’ rule, find the posterior probability of the output

variable if DrvMale =1 and DrvRace =5 . (10)

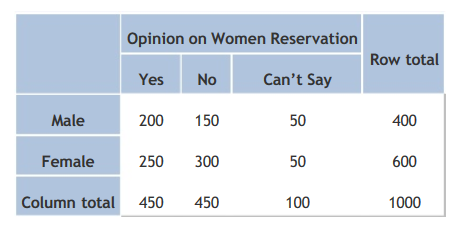
d) What is the predicted value of output variable if DrvMale =1 and DrvRace =5.

(10)

**(Total: 35 Marks)**

**Question 3**

An opinion poll surveyed a simple random sample of 1000 students. Respondents were classified by gender (male or female) and by opinion (Reservation for women, No Reservation, or No Opinion). Results are shown in the observed contingency table below.



Does the gender and opinion on women reservation are independent? Use a 0.05 level of significance. To do so,

1. State the hypotheses. **(5 Marks)**
2. Find the statistic and critical values. **(10 Marks)**
3. Explain your decision and Interpret results. **(15 Marks)**

**(Total: 30 Marks)**