Swinburne University of Technology Sarawak Campus
Faculty of Engineering, Computing and Science

COS10022 Introduction to Data Science

Assignment 2 - *Semester 2, 2020*

Assessment Title	:	Building a Naïve Bayes model to predict the survival of Titanic passengers
Assessment Weighting	:	30%
Due Date	:	**Monday, 23 November 2020** at 11.59 pm (Malaysian local time)

Assessable Item

Data science solution involving data preparation, model building, and model evaluation tasks. This assignment is to be completed in the same group as Assignment 1. Refer to the Unit Outline for the late submission penalty and group work policy.

Purpose of Assignment

This assignment aims at evaluating students' achievement of the following unit learning outcomes:

4. **Analyse business and organisational problems and formulate them into data science tasks.**
5. **Evaluate suitable techniques and tools for specific data science tasks.**
6. **Develop analytics plan for a given business case study.**
7. **Perform data analytics tasks using Microsoft Excel.**

In this assignment, your group is tasked with building and evaluating Naïve Bayes model based on the given Titanic disaster datasets. The analytic goal is to accurately predict the survival of individual Titanic passenger. You are required to produce your solution using Microsoft Excel.

Submission Requirements

Create a folder and name it with your group number (e.g. **group1**). In this folder, create the following subfolders:

- *task1*
  This subfolder should contain 2 files: **test.xlsx** and **titanic_full_survivor_list.xls**
- *task2*
  This subfolder should contain only 1 file: **train.xlsx**
- *task3*
  This subfolder should contain only 1 file: **test-deployed.xlsx**
- *task4*
  This subfolder should contain only 1 file: **evaluation.xlsx**
- *readme*
  This subfolder should contain all the necessary explanations that support the submitted data science solutions. You may place as many documents as necessary into this subfolder, but only in PDF format.
- *Assignment Cover Sheet* (signed)

Finally, compress the folder as a .zip file (e.g. **group1.zip**) and submit it via Canvas.

Each Excel file above may contain as many worksheets as necessary. **Important!** The content of the files should clearly show all the working and the Excel functions that have been used. Messy file and data organisation may incur penalty up to 3%.

Originality: The University takes plagiarism issues seriously. The Unit Outline explains what constitutes plagiarism. Students may lose the entire mark for this assignment for submitting plagiarised work in the face of sufficient evidence that suggests so.

I. Data Preparation (3%)

Datasets

Three .csv files are given:

- *train.csv* (891 instances; training set)
- *test.csv* (418 instances; test set)
- *titanic_full_survivor_list.xls* (1309 instances)

Note that these datasets are the same data that we have previously experimented with in **Practice 4-1**.

Tasks

1. Import *train.csv* and *test.csv* into new files named **train.xlsx** and **test.xlsx**, respectively.
2. The file *titanic_full_survivor_list.xls* contains the actual survival value of every passenger listed in *test.xlsx*. Use the MATCH and INDEX/OFFSET functions to locate and pull the actual survival values from the *titanic_full_survivor_list.xls* into the *test.xlsx* file. Where applicable, employ TRIM and CLEAN functions.

II. Model Building (15%)

Tasks

3. Once the data preparation above is completed, build a **nominal Naïve Bayes model** based on the data in the training set using Microsoft Excel. The model shall predict two classes: 0 (not survived) and 1 (survived).
4. At the minimum, the built model must include the following components:
   a. One or more frequency count tables. Each table lists the frequency count of every nominal value of a single feature against each class. These tables must be constructed using PivotTable.
   b. Two (2) probability tables; one for each value of the class. The probability values in these tables must be calculated from the values in the frequency count tables by employing the MATCH and INDEX/OFFSET functions.
   c. The predicted survival value must be determined using an IF function.

Requirements for higher grades than PASS: In addition to the requirements above, your group must pay careful considerations to the following aspects when building the Naïve Bayes models:

- Discretization of numerical values into nominal values (if applicable)
- Laplace Smoothing
- Dealing with missing values (if applicable)
- Normalisation of the class probability scores into percentages
- Use log probabilities to deal with potentially very small probability values

III. Model Deployment (4%)

Task

5. Deploy the model built in Part II to predict the survival of all passengers in the test set (*test.xlsx*).

IV. Model Evaluation (8%)

Tasks

6. Evaluate the performance of the Naïve Bayes model of the training set and the test set. For each dataset, you must:
   a. Build a confusion matrix. Employ the COUNTIFs function.
   b. Calculate the following evaluation metrics: TP, TN, FP, FN, Recall, Precision, Type I Error Rate, Type II Error Rate, and Accuracy. Use Excel formulas to perform the calculation instead of manual calculations.
7. Compare the performance of the Naïve Bayes model against the model previously built for **Practice 4-1** on both the training and test sets. Similarly, you should construct confusion matrices and compute all evaluation metrics mentioned above.

Requirements for higher grades than PASS: In addition to the requirements above, your group must pay careful considerations to the following In addition to the requirements above, you should also:

- Discuss why the models' performance on the training set tend to be better than their performance on the test set.
- Describe the overfitting and underfitting problems and discuss several ways to alleviate them.
- Perform one additional cross validation method for evaluating the performance of the Naïve Bayes model. Explain your finding.

Marking Rubric

Grades	1. Data Preparation (3%)	2. Model Building (15%)	3. Model Deployment (4%)	4. Model Evaluation (8%)
High Distinction 80 - 100%	Include elements of data cleaning using TRIM and/or CLEAN.	Built at least 3 different versions of the Naïve Bayes model, either by: ▪ trying different discretization techniques, OR ▪ performing feature/variable selection	All 3 versions of the built Naïve Bayes model are deployed on the test set.	Select one version of the built NB models and perform one (1) cross validation method for evaluating the performance of the selected model. Explain the finding.
Distinction 70 – 79%		The model satisfies the independence assumption of Naïve Bayes. The model implements standard Laplace Smoothing for handling nominal values with zero frequencies. The class probability scores are normalised.	Nominal values with zero frequencies are handled correctly.	Describe the overfitting and underfitting problems and discuss several ways to alleviate them. The answers are supported by at least 3 reliable references.
Credit 60 – 69%	All survival values are correctly extracted.	The rationale behind the discretization technique is explained. All calculations are correct. The model correctly handles missing values.	Effective use of the appropriate Excel functions.	Discuss why the models' performance on the training set tends to be better than their performance on the test set.
Pass 50 – 59 %	Use of MATCH and INDEX/OFFSET functions Some mistakes in the extracted survival values.	Meet the minimum components of model building as stated in the assignment description. All features are transformed into nominal features. Minor miscalculations are present.	Correct application of the model on the test set.	Meet the minimum components of model building as stated in the assignment description.
Fail <50%	Does not meet the basic requirements of each task, OR Excel functions not visible/did not employ any of the prescribed Excel functions, OR The required files are missing from the submission/submitted files are corrupted			