201A Foundation of Data Science (Level 6, 15 Credits) Assessment 2

Weighting within the course: 40%

Assessment information

- You can always ask your online tutor if you need further explanation or if the instructions are not clear.
- The purpose of this assessment is to assess your knowledge. As part of your academic and professional integrity, you must work alone on this assessment. In the event Yoobee suspects collusion, this will be addressed. For more information on plagiarism, please refer to the Student Handbook.
- Submit your completed assessment online in the correct space provided.
- Marks and feedback will be returned within 15 days of the submission date.

Learning Outcomes (LO)

On successful completion of this assessment, students have demonstrated they are able to:

LO2: Demonstrate an understanding of the core concepts of machine learning algorithms and models.

LO3: Create and apply machine learning pipelines to solve common practical problems.

Tasks

Task 1: Implement a project in machine learning (LO2, LO3) (90 %)

1. Select a text-based dataset from https://www.kaggle.com/datasets or another appropriate source. Select a dataset that is meant for a multi-class classification problem. the dataset should contain a column indicating the label/class of the samples.
2. Use pre-processing techniques covered in the learning material to handle out-of-range values, missing data, categorical data, and data normalization.
3. Train your model using the following classifiers: KNN, logistic regression, decision tree, and SVM. Perform a suitable train-test split of your data, while ensuring that this split remains uniform across all classifiers. This means that the same training samples should be used to generate the classification models, and the same test samples should be used to evaluate the performance of these models. You have the freedom to select the parameters for some of these models.
4. Calculate the confusion matrix for each of the classification models you trained and apply them to the test dataset.
5. Choose one performance metric (e.g., accuracy, precision, F1 score, etc.) and compare the result of the four models using a bar graph.

Task 2: Report writing (LO2, LO3) (10 %)

Write a report that covers the following points:

1. Provide details on the dataset chosen for the project, including some meta-data information about the dataset, and explain why this dataset was chosen.
2. Describe the pre-processing steps that were carried out on the dataset and explain the reasons behind each step taken.
3. Explain how the models were trained and provide some insight into the parameters that were chosen.
4. Comment on the process used to evaluate the performance of the models and explain the metric that was selected for this purpose.

Deliverables

Please submit the following via Yoobeeonline,

- A .ipynb (jupyter notebook) format file containing the coding tasks.
- A .docx or .pdf file containing your report.

Performance Criteria

PROJECT IMPLEMENTATION (90 %)				
Task 1	**A-, A, A+**	**B-, B, B+**	**C-, C, C+**	**D**
1. Selection of dataset (5 %)	Selecting a dataset for multiclass classification with reasonable number of samples. Aspects to look for: 1- Multiclass dataset 2- Existence of missing data instances 3- Fields with unnormalized examples	Selection of multiclass dataset with either: 1- Small samples 2- No missing data	Selection of dataset with either: 1- Small samples 2- No missing data 3- Normalized data 4- Binary classes	Selection of inappropriate dataset e.g., specialized datasets (images, sound, etc.) or regression/clustering dataset OR minimal or no attempt to complete this task
2. Pre-processing (15 %)	Appropriate preparation of data including: 1- Proper data normalization 2- Handling out-of-range data 3- Handling missing data 4- Proper handling of categorical data	Reasonable preparation of data including: 1- Proper data normalization 2- Handling out-of-range data	Poor preparation of data including: 1- Only proper data normalization	No normalization of data whenever required
3. Train-test process and model creation (50 %)	Good process of data splitting and model training including model parameters. Good implementati	Reasonable process of data splitting and model training including model parameters. Implementati	Poor process of data splitting and model training including model parameters. Implementati	No splitting of data into train and test instances and little to no implementation of models

	on of all four algorithms (both training and testing phases)	on of only 2 or 3 models.	on of only one model.	
4. Evaluation process (10 %)	Good use and implementation of evaluation metrices. This includes comparison with other models	Reasonable use and implementation of evaluation metrices. This includes partial comparison with other models	Poor use and implementation of evaluation metrices. This includes inadequate comparison with other models	Little to no implementation of the task
5. Data visualization (10 %)	Appropriate selection and implementation of data visualization	Reasonable selection and implementation of data visualization	Poor selection and implementation of data visualization	Little to no implementation of the task
PROJECT DOCUMENTATION (10 %)				
Task 2	**A-, A, A+**	**B-, B, B+**	**C-, C, C+**	**D**
Quality of report (10 %)	Comprehensive details and structure which includes: 1. Dataset details and justification of selection 2. Providing good details on the normalization and other pre-processing steps 3. Good model learning and testing details 4. Logical analysis and performance evaluation along with appropriate comparison details	Reasonable details and structure which includes: 1. Brief dataset details and justification of selection 2. Brief details on the normalization and other pre-processing steps 3. Some model learning and testing details 4. No logical analysis or comparison	Poor details and structure which includes: 1. Poor dataset details and justification of selection 2. Little details on the normalization and other pre-processing steps 3. No model learning and testing details 4. No logical analysis or comparison	Little to no implementation of the task