|  |  |  |  |
| --- | --- | --- | --- |
| Program | MSC Data Science | Subject | Machine Learning I |
| Date | 19-Jun-2021 | Duration | 2 Hours |
| Sem/Test | Sem II / Test 2 | Marks | 45 |

**General Instructions:**

Candidates should read carefully the instructions printed on the question paper.

1. All questions are compulsory.
2. Maximum marks for respective questions given along with the questions.
3. Open book test; students can refer to any notes / books / article while answering the paper.
4. Students can access internet if required.
5. Students’ own computers are allowed.

***All The Best!!!***

**I Multiple Choice Questions [ 5**

Each MCQ carries 1 mark

Provide the right answer by high-lighting the text in bold

e.g.

0. The capital city of India is

* **Delhi**
* Mumbai
* Chennai
* Kolkatta

1. In Machine Learning, there are \_\_\_\_\_ classification models

* Two
* Three
* Many
* None of the above

1. In a Classification Model, acceptable value of accuracy of test data

* More than 95%
* More than 85% but less than 95%
* Less than 85%
* None of the above

1. In a Classification Model, apart from accuracy we should also consider

* Precision & Recall
* True Positive Rate & True Negative Rate
* False Positive Rate & False Negative Rate
* None of the above

1. A Classification model, is overfitting your training data

* When the model performs well on the training data but does not perform well on the test data.
* When the model performs well on the training data but does not perform well on the prediction data.
* When the model performs well on the training data but does not perform well on the test & prediction data.
* None of the above

1. How to impute Missing Prediction Data when Historical Data Is Not Missing

* Impute with same value that was used to impute historical data
* Impute with prediction mean
* Impute with historical mean
* None of the above

***Submission***

Please submit this doc file renamed as CLS-RNo-Name.xls (eg CLS-001-CyrusLentin.docx)

**II Classification Manual [ 10**

Refer to excel file “MaskedData.xlsx”, The historical data has been given to you. The columns are masked except for Outcome column which is the class variable. Use the historical data to classify the data given in the predict table. Classify using K Nearest Neighbor algorithm with K = 5.

You may use Excel for computation without using the Excel Euclidian Distance function.

Ensure that computation of both prediction data points are clearly shown.

***Submission***

Please submit the xlsx file renamed as CLS-RNo-Name.xls (eg CLS-001-CyrusLentin.xlsx)

**III Classification – Python [ 30**

A dataset “glassm-data.csv” is given to you for predictive analytics problem.

A dictionary “glassm-data.txt” gives you details of the “attributes” contained in the dataset.

Use “glassm-data-prd.csv” as new data for final prediction of “Type” column.

***Requirements***

Use python to provide the EDA, Transformations, VDA and select the best Classification model amongst the models you know based on Confusion Matrix Accuracy.

Based on the best model, use “glassm-data-prd.csv” as new data for final prediction of “Type” column.

***Transformations***

Outliers must be handled using the “cap” method.

Impute missing values for the columns as per default rules.

If any column has more than 30% null values, drop the column

Use performance improvement techniques to improve the accuracy of the model selected.

***Submission***

Please submit the py file renamed as CLS-RNo-Name.py (eg CLS-001-CyrusLentin.py)

Also submit the output csv file with the prd data and the predicted “Type” renamed as PRD-RNo-Name.csv (eg PRD-001-CyrusLentin.csv)

***Project Evaluation***

|  |  |
| --- | --- |
| **Particulars** | **Max Marks** |
| Read File / EDA / Transformations / VDA | 10 |
| Auto Select Best Model | 10 |
| Test Data Accuracy | 5 |
| Prediction Data Accuracy | 5 |
| ***Total*** | ***30*** |

|  |  |
| --- | --- |
| **Test Data Accuracy** |  |
| If Test Data Accuracy >= 80% | 2 marks |
| If Test Data Accuracy >= 85% | 3 marks |
| If Test Data Accuracy >= 90% | 4 marks |
| If Test Data Accuracy >= 94% | 5 marks |
|  |  |
| **Prediction Data Accuracy** |  |
| If Prediction Data Accuracy >= 85% | 2 marks |
| If Prediction Data Accuracy >= 90% | 3 marks |
| If Prediction Data Accuracy >= 95% | 4 marks |
| If Prediction Data Accuracy = 100% | 5 marks |

**Wishing You All The Best!!!**