SVKM'S NMIMS

MUKESH PATEL SCHOOL OF TECHNOLOGY MANAGEMENT& ENGINEERING

Academic Year: 2023-2024

Program: B.Tech	Stream : Data Science	Year: IV	Semester: VII
Subject: SNP		Time: 1 hr (8AM to 9AM)	
Date: 28/08/2023		No. of Pages: 02	
Marks: 20			

Test-I / M1

Instructions: Candidates should read carefully the instructions.
1) **Figures in brackets on the right hand side indicate full marks**.
2) **Assume Suitable data if necessary.**
3) **All questions are compulsory.**
4) **Submissions should be made in ".ipynb" formats.**
5) **Please use Google Colab to avoid loss of data.**
6) **NO Pre-trained Models, Transformers, Hugging Face Models, etc. will be allowed.**
7) **EDA is a must. List insights where ever feasible.**

<table>
<tr>
<td>CO- 1;
BL- 1</td>
<td><b>Q2</b></td>
<td>Perform sentiment analysis on the following Kaggle dataset

Link - https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

<b><u>Instructions</u></b> –
1. Use only 10k Observations for the exam. Drop the rest 40k. (Make sure to shuffle before dropping, you may invite class imbalance issues.)
2. Preprocess the data.
3. Use a library of your choice for Sentiment analysis.
4. Drop the existing label column.
5. Feature engineer a column for polarity scores.
6. Feature engineer a column for sentiment (negative, neutral and positive OR negative and positive where polarity = 0 is negative class)
7. Then train atleast 5 ML models on this and you may feel free to split the data as you wish for the training process.
8. Select best model at end. Evaluation should be done on basis of F1 Score.
9. In a single cell, pass a review of your own as a string to the model and test if it is able to capture the sentiment of your review.</td>
<td>[05]</td>
</tr>
<tr>
<td colspan="4"><b>OR</b></td>
</tr>
</table>

CO- 1; BL- 1	**Q2**	Perform text-based classification on the following Kaggle dataset Link – https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews **Instructions** – 1. Use only 10k observations for the exam. Drop the rest 40k. (Make sure to shuffle before dropping, you may invite class imbalance issues.) 2. Preprocess the data. 3. Show the use of RE or other preprocessing libraries. 4. Feature engineering is necessary. 5. Train atleast 5 ML Models on this and you may feel free to split the data as you wish for the training process. 6. Select top 3 ML Models on the basis of F1 Scores. 7. **Optional** – Amongst the 3 Perform GridSearchCV/RandomizedSearchCV to find best parameters for 1 model only.	[05]