CH5019 – Mathematical Foundations of Data Science

Term Project

January-May 2020 Semester assigned on **14th April 2020**

(Total Marks 100)

Instructions:

1. *Make group of 5 members.*
2. *Use either MATLAB or Python for coding.*
3. *Report should be submitted in pdf format.*
4. *Report should include figures, brief explanation, references (wherever required) and should not exceed 15 pages. Use font size of 11 pts.*
5. *Submit all your codes along with the report separately in a zipped folder with your Names, Roll numbers as the file name.*
6. *Upload the zipped file only in Moodle (as **Roll_NO.zip** extensions).*
7. *You can make visualizations wherever required. Extra marks for creativity and further insights. For visualizations, if you use Tableau or any toll, zip those files too.*

Question 1: [30 marks]

Grayscale images of 15 subjects under 10 different conditions were obtained and are given in the file Dataset_Question1. Due to storage limitations, only one representative image can be stored for each subject in the database for future automated facial recognition purposes. Perform SVD on all the images for a subject to identify characteristic features that will be stored for each subject. Only this information should be used in the future automated facial recognition tasks. In the report, please show the representative images of all the subjects. Given an image, the facial recognition method is based on the smallest norm between the image and the representative images in the database. Determine the number of images out of 150 that you are able to correctly identify based on this approach in terms of accuracy. MATLAB image processing functions or python image processing libraries are not to be used for this assignment.

Question 2: [30 marks]

This data set describes operating conditions of a reactor and contains class labels about whether the reactor will operate or fail under those operating conditions. Your job is to construct a logistic regression model to predict the same.

Dataset_Question2.xlsx: The data contains a 1000 X 6 data matrix. The first five columns are the operating conditions of the reactor. The sixth column provides necessary annotation:

- **Temperature:** 400-700 K
- **Pressure:** 1-50 bar
- **Feed Flow Rate:** 50-200 kmol/hr

- **Coolant Flow Rate:** 1000-3600 L/hr
- **Inlet Reactant Concentration:** 0.1-0.5 mole fraction
- **Test:** fail/pass. Whether the reactor will operate or fail under the corresponding operating conditions.

Using the above datasets, make a report on the following: **Note:** *Any assumptions made should be properly mentioned in the report.*

1. Describe the statistics of the data.
2. Partition your data into a training set and a test set. Keep **70%** of your data for **training** and set aside the remaining **30%** for **testing**.
3. Fit a logistic regression model on the training set. Choose an appropriate objective function to quantify classification error. **Manually code for the gradient descent procedure** used to find optimum model parameters. (**Note:** You may need to perform multiple initializations to avoid local minima)
4. Evaluate the performance of above model on your test data. Report the **confusion matrix** and the **F1 Score**.

Question 3: [40 marks]

Coronaviruses are a large family of viruses which may cause illness in animals or humans. In humans, several coronaviruses are known to cause respiratory infections ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS).

The number of new cases are increasing day by day around the world. This dataset has information from the states and union territories of India at daily level.

Dataset:	**AgeGroupDetails.csv**	**Covid_19_india.csv**	**HospitalBedsIndia.csv**
About the file:	Age group details of affected cases.	Number of covid-19 cases in India at daily level	Number of hospital beds in each state in India
	Sno: Serial number	**Sno:** Serial number	**Sno:** Serial number
	AgeGroup: Age group	**Date:** Date of observation	**State/UT:** State or Union Territory
	TotalCases: Number of cases in the age group	**Time:** Time of observation	**NumPrimaryHealthCenters_HMIS:** Number of primary health care centers
	Percentage: Percentage of cases in the age group	**State/UnionTerritory:** Name of the State / Union territory	**NumCommunityHealthCenters_HMIS:** Number of community health care centers

		ConfirmedIndianNational: Cumulative number of confirmed Indian nationals	**NumSubDistrictHospitals_HMIS:** Number of sub district health care centers
		ConfirmedForeignNational: Cumulative number of confirmed foreign nationals	**NumDistrictHospitals_HMIS:** Number of district hospitals
		Cured: Cumulative number of cured people	**TotalPublicHealthFacilities_HMIS:** Total number of public health centers
		Deaths: Cumulative number of death cases	**NumPublicBeds_HMIS:** Total number of beds in public health care centers
		Confirmed: Cumulative number of confirmed cases	**NumRuralHospitals_NHP18:** Number of Rural hospitals
			NumRuralBeds_NHP18: Number of beds in rural hospitals
			NumUrbanHospitals_NHP18: Number of urban hospitals
			NumUrbanBeds_NHP18: Number of beds in urban hospitals

ICMRTestingDetails.csv	**IndividualDetails.csv**	**Population_india_census2011.csv**
Number of COVID testings at daily level from ICMR	Individual case level details	Population of different states in India
SNo: Serial number	**Id:** ID column	**Sno:** Serial number
DateTime: Date and time in IST	**government_id:** Government ID of the case	**State / Union Territory:** Name of the State / Union territory
TotalSamplesTested: Total number of samples tested	**diagnosed_date:** Date of diagnosis	**Population:** Total population

TotalIndividualsTested: Total number of individuals tested	**age:** Age	**Rural population:** Population in rural areas
TotalPositiveCases: Total number of positive cases	**gender:** Gender	**Urban population:** Population in urban areas
Source: Source of data	**detected_city:** City in which the case is detected	**Area:** Total area
	detected_district: District in which the case is detected	**Density:** Population density
	detected_state: State in which the case is detected	**Gender Ratio:** Gender ratio
	nationality: Nationality of the case	
	current_status: Current status of the case	
	status_change_date: Status change date	
	notes: Notes	

The dataset can also be taken from https://www.kaggle.com/sudalairajkumar/covid19-in-india.

Using the above datasets, make a report on the following points: **Note:** *Any assumptions made should be properly mentioned in the report.*

1. Which age group is the most infected?
2. Plot graphs of the cases observed, recovered, deaths per day country-wise and state-wise.
3. Identify the positive cases on a state level. Quantify the intensity of virus spread for each state.*
   Intensity here means number of positive cases/population density.
4. List places in the country which are active hotspots/clusters as on 10.04.2020.* Hotspot is defined as an area in a city where 10 or more people have been tested positive.
5. Which states have the maximum change (consider increase and decrease separately) in number of hotspots on weekly basis from 20.03.2020 to 10.04.2020 (3 weeks).
6. For the given data, identify cases with international travel history (primary case), personal contact with primary case (secondary case). Cases which do not fall in the primary and secondary fall into tertiary case. Quantify them based on the percentage for the top 5 states with maximum cases till 10.04.2020.*

7. Find out the number of additional labs needed from the current existing labs (assume 100 tests per day per lab) with an increase rate of 10% cases per day from 11.04.2020 - 20.04.2020. List out any further assumptions considered.
8. Plot the number of cases starting from 1st March - 10th April. Based on this plot can you comment on the popular notion of 'flattening the curve'.*
9. As we know, social distancing is the best option to avoid the spread. Based on the time series data (**covid_19_india.csv**), can you suggest how successful the 21 days lockdown has been?

* - Visualization required