Kaggle Project: Covid-19 Awareness and Covid-19 Cases in Ohio

Please read all of the guidelines carefully before submitting the lab. ☺ There are ***100 points*** in total. **You can work alone or in a group of two in this project.**

Due date: Sunday, April 9, 11:59 PM. Late submissions will be penalized by 5 points / day. No submissions will be accepted two days after the deadline.[1]

Deliverables:

1) The code of the project in **.ipynb** format (one file)
2) The lab report written with ***LaTeX*** and exported in **.pdf** format (one file)

Guidelines – Before You Start

1) **Please do not post any of your code or solutions online. This is a Kaggle competition.**
2) Three high-ranked teams will receive a bonus of 0.5% if they choose to present their methods and classification strategy in class.
3) You will be using the **`Python`** programming language for this project. You need to write your codes in an empty **.ipynb** file.
4) Make sure that you provide many comments to describe your code and the variables that you created.
5) Please use the ***IEEE*** journal template on ***overleaf.com***. Here is the link: https://www.overleaf.com/latex/templates/preparation-of-papers-for-ieee-sponsored-conferences-and-symposia/zfnqfzzzxghk
   To be able to work on ***overleaf.com***, you will need to register first (you can also compile your ***LaTeX*** file locally.)
6) For some of the code, you may need to do a little bit of "*Googling*" or review the documentation.

What is the Data?

The dataset provided with this assignment contains measurements on the level of awareness about Covid-19-related topics in all the counties of Ohio[2] during the pandemic, the number of Covid-19 cases and the number of deaths by county, along with some county-level social, economic, and demographic variables. Each data point represents the level of awareness on certain issues for a given day during the pandemic. The level of awareness has been measured in different ways, as will be more closely described in the section below. Unlike in a conventional ML challenge, the **`test`** set consists of %70 of all of the observations (the remaining 30% is the **`training`** set).[3] In total, there are 3141 observations in the **`training`** set, and 7331 observations in the **`test`** set.

[1] Part II of the analysis needs to be submitted by the deadline (no late submission will be accepted for Part II).
[2] https://en.wikipedia.org/wiki/Ohio
[3] Earliest data point is in November 2019. The observations in both the **`training`** and the **`test`** datasets are randomly ordered.

Data Collection

The awareness data has been extracted from tweets and has been collected by using hashtags posted by users in Ohio during the Covid-19 pandemic. To prepare the dataset, over 46,000,000 million tweets posted by over 91,000 have been collected. Here are the steps used to create the dataset:

1) The tweet data has been collected from users in Ohio during the first few months of the pandemic (starting from 2019). To select the tweets that are related to Covid-19, hashtags including the 'name of the pandemic / Covid' were used.
2) Co-occurring hashtags have been extracted from the tweets. These hashtags were then used to manually identify topics related to Covid-19.
3) Lastly, different similarity measures (Jaccard, Cosine, Intersection) were used to detect the intensity of discussion on topics related to Covid-19.

Kaggle Project – Data Dictionary

The data has been provided in the assignment folder online (**training_data.csv** and **test_data.csv**). Open the CSV files and take a look at them before you start working on the project. Individual features (columns) of the dataset have been described below:

Note: All data is county-level.

***index*:** Index associated with the observation (found only in the `test` set)
***county*:** Name of the Ohio county that is associated with the data point
***cases*:** Number of Covid-19 cases on a given day in a county of Ohio (found only in the `training` set)
***deaths*:** Number of Covid-19-related deaths on a given day in a county of Ohio
***date_index_converted*:** an anonymized index value for the day associated with awareness
***county_data_length*:** number of tweets posted on a given day to calculate awareness level
total_pop: population of the county (x 1000)
percent_25_34: percentage of the people who are between 25 and 34 years old
percent_highschool: percentage of the people who have at least finished high school
labor_force_rate: percentage of the adult people who are currently employed
unemployment_rate: percentage of unemployment in the county
median_housing_cost: median cost of a house in the county
median_household_earnings: median annual earnings of a household in the county
median_worker_earnings: median annual earnings of an employee in the county
percent_insured: percentage of the people who currently have health insurance
percent_married: percentage of the people who are currently married
***poverty_rate*:** percentage of the people who fall under the poverty line
***median_property_value*:** median value of a property
***percent_white*:** percentage of white people in the county

The variables listed below reflect the level of awareness on a given day in a county. Awareness has been measured for the following topics: core (general Covid-19 awareness), domestic issues/domestic politics, economy, education, entertainment, foreign issues/foreign politics, gender, health, ideology, illness (gen eral), nationalism, politics (love for Democrats), politics (hate for Democrats), politics (love for Republica ns), politics (hate for Republicans), race, religion, social issues, and sports. For the variables that are nam ed 'normalized', 0-1 normalization has been used.

core_cosine, core_cosine_normalized, core_intersection, core_intersection_nor
malized, core_jaccard, core_jaccard_normalized, domestic_cosine, domestic_cos
ine_normalized, domestic_intersection, domestic_intersection_normalized, dome
stic_jaccard, domestic_jaccard_normalized, economy_cosine, economy_cosine_nor
malized, economy_intersection, economy_intersection_normalized, economy_jacca
rd, economy_jaccard_normalized, education_cosine, education_cosine_normalized
, education_intersection, education_intersection_normalized, education_jaccar
d, education_jaccard_normalized, entertainment_cosine, entertainment_cosine_n
ormalized, entertainment_intersection, entertainment_intersection_normalized,
entertainment_jaccard, entertainment_jaccard_normalized, foreign_cosine, fore
ign_cosine_normalized, foreign_intersection, foreign_intersection_normalized,
foreign_jaccard, foreign_jaccard_normalized, gender_cosine, gender_cosine_nor
malized, gender_intersection, gender_intersection_normalized, gender_jaccard,
gender_jaccard_normalized, health_cosine, health_cosine_normalized, health_in
tersection, health_intersection_normalized, health_jaccard, health_jaccard_no
rmalized, health_technology_cosine, health_technology_cosine_normalized, heal
th_technology_intersection, health_technology_intersection_normalized, health
_technology_jaccard, health_technology_jaccard_normalized, ideology_cosine, i
deology_cosine_normalized, ideology_intersection, ideology_intersection_norma
lized, ideology_jaccard, ideology_jaccard_normalized, illness_cosine, illness
_cosine_normalized, illness_intersection, illness_intersection_normalized, il
lness_jaccard, illness_jaccard_normalized, labor_force_rate, median_household
_earnings, median_housing_cost, median_property_value, median_worker_earning,
nationalistic_cosine, nationalistic_cosine_normalized, nationalistic_intersec
tion, nationalistic_intersection_normalized, nationalistic_jaccard, nationali
stic_jaccard_normalized, percent_25_34, percent_highschool, percent_insure, p
ercent_married, percent_white, politics_cosine, politics_cosine_normalized, p
olitics_democratic_hate_cosine, politics_democratic_hate_cosine_normalized, p
olitics_democratic_hate_intersection, politics_democratic_hate_intersection_n
ormalized, politics_democratic_hate_jaccard, politics_democratic_hate_jaccard
_normalized, politics_democratic_love_cosine, politics_democratic_love_cosine
_normalized, politics_democratic_love_intersection, politics_democratic_love_
intersection_normalized, politics_democratic_love_jaccard, politics_democrati
c_love_jaccard_normalized, politics_intersection, politics_intersection_norma
lized, politics_jaccard, politics_jaccard_normalized, politics_republican_hat
e_cosine, politics_republican_hate_cosine_normalized, politics_republican_hat
e_intersection, politics_republican_hate_intersection_normalized, politics_re
publican_hate_jaccard, politics_republican_hate_jaccard_normalized, politics_
republican_love_cosine, politics_republican_love_cosine_normalized, politics_
republican_love_intersection, politics_republican_love_intersection_normalize
d, politics_republican_love_jaccard, politics_republican_love_jaccard_normali
zed, poverty_rate, race_cosine, race_cosine_normalized, race_intersection, ra
ce_intersection_normalized, race_jaccard, race_jaccard_normalized, religion_c
osine, religion_cosine_normalized, religion_intersection, religion_intersecti
on_normalized, religion_jaccard, religion_jaccard_normalized, social_cosine,
social_cosine_normalized, social_intersection, social_intersection_normalized
, social_jaccard, social_jaccard_normalized, sports_cosine, sports_cosine_nor
malized, sports_intersection, sports_intersection_normalized, sports_jaccard,
sports_jaccard_normalized

Part I: Descriptive Analysis (20 points)

In this part of the analysis, you will be exploring the dataset by providing some descriptive information and also creating a set of visuals.

[Important note: Use the **`training`** dataset for the descriptive analysis.]

a) Check out this page: https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Ohio. In around 250 words, summarize the Covid-19 experience of Ohio. Specifically, focus on how Ohio is different or similar to other US states in terms of the intensity of the pandemic (i), the time and the content of the different policies that have been implemented (ii), and if Wikipedia 'thinks' Ohio has dealt with Covid-19 successfully (or not) (iii). [**4 points**]

b) Find the *average* values for all the topic awareness variables. Create a **bar chart** that shows the average normalized Jaccard similarity-based awareness values for all different types of awareness topics listed above. Order the bars from the biggest to the smallest. Summarize your observations in around 100 words. [**4 points**]

c) Focus on the `core_jaccard_normalized` variable. Create a **bar chart** that shows the aggregated mean awareness value for each county. Order the bars from the biggest to the smallest. Which county has the highest awareness? Summarize your observations in around 100 words. [**4 points**]

d) Create **two county-level maps** of Ohio (an example is provided in the first page of the assignment). Using colors, show the ***number of average Covid-19 cases per capita*** and the ***number of average Covid-19 deaths per capita*** by county. What are the top-5 counties with high number of per capita cases and per capita number of deaths? Summarize your observations in around 100 words. [**4 points**]

e) Calculate the average normalized Jaccard awareness scores for every day (starting from Day 1). Create a line chart with overlapping lines in which each line represents the evolution of awareness levels for each topic.[4] The x-axis of the line chart should correspond to 'Days', and the y-axis of the line chart should represent the level of awareness. What are the trends in the graph? Summarize your observations in around 100 words. [**4 points**]

Part II: Model Creation and Prediction (50 points)

Please do not post any of your code or solutions online.

This part of the analysis needs to be submitted by the deadline (no late submission will be accepted).

In this part of the analysis, graduate and undergraduate students will be graded/ranked separately.

[4] Here is a complete list of topics: core (general Covid-19 awareness), domestic issues/domestic politics, economy, education, entertainment, foreign issues/foreign politics, gender, health, ideology, illness (general), nationalism, politics (love for Democrats), politics (hate for Democrats), politics (love for Republicans), politics (hate for Republicans), race, religion, social issues, and sports.

Please use the dataset provided to you. [We should be able to run your code with the original datasets and the additional external datasets you provide.] You cannot use any other Covid-19 data that provides number of deaths / number of cases in Ohio. You can use other datasets.

For this part of the analysis, you will need to train a model that predicts the number of cases based on all of the remaining variables in the dataset, and report the **R2-Value** of your best model. **You will mainly be graded on the R2-Value of your model** (more information provided below). Some guidelines (please also review the information shared through lectures):

- The minimum R2-value should be around ***0.5***. R2-values around ***0.5*** will be accepted as benchmark for the analysis, and therefore be considered as the minimum score you should obtain.
- You can use **any** classification/prediction model (including, but not limited to, logistic regression, decision trees, neural networks etc.)
- Your model should run on a laptop [equivalent to a modern MacBook Pro] in a reasonable amount of time (in a few hours at a maximum) for grading purposes.
- You can use **any** feature engineering method to transform your dataset, such as:
    - Dimensionality reduction methods such as PCA, t-SNE, spectral embedding
    - Logarithmic, polynomial, and other transformations
    - Different word vectorization techniques
    - Different weighting strategies
- You are free to create a new column (or a stream of data) based on the existing columns and use your new column as an independent variable.
- You are welcome to use **any** external dataset to enrich your training and test datasets.
- You are welcome to create **any** logical condition (if, else etc.) to label the target variable (if you do so, please describe why you made these choices).

Please use your real name when you sign up for the Kaggle project. To participate in the competition, please click: https://www.kaggle.com/t/e235def16ffb4591a6e295c769be024b

Model Evaluation

Your submission will be evaluated using the ***R2*** cost function. If you are unsure, please review what ***R2*** means before starting on the project. **Please also report all of your code in the .ipynb file and your confusion matrix both in the .ipynb file and in the report. Please also report *R2* in your code.**

Please use your actual name on Kaggle! [Or, please indicate your nickname in the report!]

Make sure that all of your code is running!

Save the code file you have created as "**kaggle_ohio_lab.ipynb**" in the folder you have created at the beginning.

Part III: Creating the lab report (30 points)

Write a report (minimum 2 pages) that includes you name (or your name and your group member's name), all of your findings and the visuals that you created. The report that you will write should use the *IEEE format* and include the following sections:

Abstract: A short summary of your report (This part should include a very brief summary of your methods and analysis and the answer to why you think what you have done is important)

Introduction: A summary of what you expected and did, and two-three of your most significant findings (please use some numerical results here)

Data: Introduce your descriptive findings about the dataset here

Methods: Provide a description of your strategy and the steps you took to improve your prediction model (this includes the steps you followed for data-preprocessing, setting up the model, and checking the strength of the model)

Results: A detailed discussion on the results you obtained. What is your Accuracy value? Evaluate and criticize yourself / your team.

Save the project report as **kaggle_ohio_project_report.pdf**.

Final step:

Send your code as an **.ipynb** file and the report in the **.pdf** format through *BlackBoard*.

General Rules and Grading

You will be graded based on the following criteria:

- Code: Cleanliness/understandability (i), executability (ii), format (iii) [We need to be able to run all parts of the code using the datasets provided.]
- Ranking: Ranking in the **Kaggle** competition
- Lab Report:
    - *Introduction* (i), *Data* (ii), *Methods* (iii), *Results* (iv)
    - Flow, readability, level of detail, quality of visuals/tables, adherence to the guidelines

* Three high-ranked teams will receive a bonus of 0.5% if they choose to present their methods and classification strategy in class.

And, finally, art from Ohio:

Ohio, ***Field of Corn***:
https://en.wikipedia.org/wiki/Field_of_Corn