Machine Learning – Assignment 3

Due Date: 09.06.2022

Data

You have been provided with data regarding US elections, general county information, life expectancies, economic data and more. The variables description can be found in the **data-variables-description.pdf** file. In addition, the **background-information.pdf** file provides information regarding the US elections system in case you are not familiar with it.

Section A (Data Exploration and Pre-processing) *15 pts*

1. Explore the data using tables, visualizations, and other relevant methods (use at least 3 different types of graphs).
2. Apply different methods of pre-processing to the data in order to prepare it for the models you wish to apply in the next sections.
    - Provide an explanation to each method you apply. Your choice should reflect an understanding of the method and why it's needed.

Section B (Dimensionality Reduction) *20 pts*

For this section you should use only the states of California, Florida, South Dakota, and Wyoming.

At least one feature needs to be included from each of the following tables:

1. 5296US_landarea.txt
2. county_complete.csv
3. IHME_USA_LIFE_EXPECTANCY_1987_2007_Y2011M06D16.XLSX

Overall, you should include **at least 8 features**.

You may use other tables that are included while explaining the logic behind your decision. Feel free to create new features based on any of the given within the USCountyLifeExpectancyOtherDemographicData directory.

Tasks

1. Apply at least 2 dimensionality reduction algorithms.
2. Create a scatter plot for the new data and color each observation according to the state it belongs to.
   - Describe your findings (which states are most similar to one another).
   - Which features are the **most** effective to separate the four states?
     - Explain the process and findings.
     - Present the resulting clusters visually.
   - Which features are the **least** effective to separate the four states?
     - Explain the process and findings.
     - Present the resulting clusters visually.
   - **Bonus** *5 pts* – develop a method to calculate a numeric value for the goodness of separation.

<u>**Section C (Regression)**</u> *35 pts*

Voter turnout – in political science, voter turnout is the percentage of registered voters who participated in an election (often defined as those who cast a ballot).

Tasks

1. Create a measure as accurate as you can for **voter turnout percentage by state** for the years 2010 and 2012. Explain the process performed to extract a measure for voter turnout.

- This may require the use of more than one table. In the **USElectionResults19762020** folder, use the **1976-2020-house.csv** file as a basis for turnout.

2. Create **at least 6 features** to predict voter turnout, explain how you build these features in detail.
   - You may use historic data for turnout, up to 8 years back. So, to estimate turnout in 2010 you may use data up to 2002.
   - Tip – You may want to visualize the historic data by election cycle for relevant insights.
3. Using the features you built, apply at least 3 machine learning algorithms to predict **voter turnout percentage** for at least 48 states.
   - Provide feature importance for each model.
   - Visually present your predictions in 2010 and 2012 for California, Florida, South Dakota, and Wyoming.
   - The implementation should include parameter tuning.
   - Report a suitable measure to evaluate the performance (on the 48 or more states) of each model and compare the results.
4. Rank the states by how well your model predicted turnout. Specifically, write the 5 states that the turnout estimate was least successful.
   - Make modifications to your model and features so that at least 2 of the states that were in the bottom 5 will be ranked in the within the top 25 for prediction rate.

Section D (Classification) *30 pts*

Tasks

1. Use the total votes counted by state in house elections cycle to create the label. If more votes were given to the democrat candidates in total,

the state should be classified as **D**. Otherwise, if more votes were given to the republican candidates, the state should be categorized as **R**.

2. Using **at least 6 features**, apply an SVM model and at least 2 additional machine learning algorithms to predict the majority vote (D or R).
    - Provide feature importance for each model.
    - The implementation should include parameter tuning.
    - Report a suitable measure to evaluate the performance of each model and compare the results.

<u>Section E (Bonus)</u> *15 pts*

In 2022 senate elections will be held in the following states:

> Alabama, Alaska, Arizona, Arkansas, California, Connecticut, Florida, Georgia, Hawaii, Idaho, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Maryland, Missouri, Nevada, New Hampshire, New York, North Carolina, North Dakota, Ohio, Oklahoma (2 races), Oregon, Pennsylvania, South Carolina, South Dakota, Utah, Vermont, Washington, and Wisconsin.

States that are not mentioned do not have races in 2022.

<u>Note:</u> senators are in office for 6 years and each state has exactly 2 senators. The senate elections occur on even number years so typically two third of the states have a senate election on a given election year.

Task

Predict the outcome - Republican and Democrat Senator (independents count as democrats) of the 2022 midterm elections for each state.

Additional bonus points will be given to students who accurately predicted the outcome after the election results in November!

Section F (Performance - Bonus) *5 pts*

Machine learning models that outperformed other students' models for either the unit sales predictions or the rainy-day classification may get additional points as long as the non-standard methodology to obtain superior results is also explained.

- In order to get the bonus points you may want to apply multiple performance measures to ensure that we can compare your performance on an equal basis to other projects, and that you did not sacrifice performance in a specific measure to outperform in another.

Submission

- The assignment should be submitted in pairs (only one submission).
- You are required to submit two files including sections A-F. One in **.ipynb** format and one in **.html**. Both files should also include the program's outputs.
- The files' names should be of the form: **ML_HW3_#ID1_#ID2**.
- Assignments submitted late will receive a penalty of **3 points** for each day, up to one week. Later submissions will not be accepted.