DATA 201 – Project

Total marks: **20**

Due date: **11:59 p.m., Friday, October 13**.

Submit **code** and **outputs** in a **single PDF file** (name it **project.pdf**).

Background:

This project will examine vaccination, a crucial public health measure utilized to combat infectious diseases. Vaccines provide individuals with immunization, and sufficient immunization in a community can reduce the spread of diseases through *"herd immunity."*

Beginning in the spring of 2009, a pandemic caused by the H1N1 influenza virus, also known as *"swine flu,"* swept the globe. Researchers estimate that it caused between 151,000 and 575,000 deaths worldwide in its first year.

In October 2009, the H1N1 influenza vaccine became available to the public. The United States conducted the National 2009 H1N1 Flu Survey between the end of 2009 and the beginning of 2010. This telephone survey asked respondents whether they had received the H1N1 and seasonal flu vaccines, as well as personal questions. These additional questions covered the respondent's social, economic, and demographic background, as well as their perspectives on illness risks and vaccine effectiveness, as well as their actions to prevent transmission. Future public health efforts can be guided by a better understanding of the relationship between these characteristics and personal vaccination patterns.

You are required to develop a machine learning model to predict the ***probabilities*** of whether people received the H1N1 vaccine based on the background, opinions, and health behaviors they disclosed.

The model will be developed using file **train.csv** and tested on file **test.csv**. The description of the data is given below.

Data:

The target variable is **h1n1_vaccine** - whether the respondent received the H1N1 flu vaccine (`0` = No; `1` = Yes).

Other features are described below.

For all binary variables: 0 = No; 1 = Yes.

- **h1n1_concern** - Level of concern about the H1N1 flu.
    - 0 = Not at all concerned; 1 = Not very concerned; 2 = Somewhat concerned; 3 = Very concerned.
- **h1n1_knowledge** - Level of knowledge about H1N1 flu.

    - 0 = No knowledge; 1 = A little knowledge; 2 = A lot of knowledge.
- **behavioral_antiviral_meds** - Has taken antiviral medications. (binary)
- **behavioral_avoidance** - Has avoided close contact with others with flu-like symptoms. (binary)
- **behavioral_face_mask** - Has bought a face mask. (binary)
- **behavioral_wash_hands** - Has frequently washed hands or used hand sanitizer. (binary)
- **behavioral_large_gatherings** - Has reduced time at large gatherings. (binary)
- **behavioral_outside_home** - Has reduced contact with people outside of own household. (binary)
- **behavioral_touch_face** - Has avoided touching eyes, nose, or mouth. (binary)
- **doctor_recc_h1n1** - H1N1 flu vaccine was recommended by doctor. (binary)
- **doctor_recc_seasonal** - Seasonal flu vaccine was recommended by doctor. (binary)
- **chronic_med_condition** - Has any of the following chronic medical conditions: asthma or an other lung condition, diabetes, a heart condition, a kidney condition, sickle cell anemia or other anemia, a neurological or neuromuscular condition, a liver condition, or a weakened immune system caused by a chronic illness or by medicines taken for a chronic illness. (binary)
- **child_under_6_months** - Has regular close contact with a child under the age of six months. (binary)
- **health_worker** - Is a healthcare worker. (binary)
- **health_insurance** - Has health insurance. (binary)
- **opinion_h1n1_vacc_effective** - Respondent's opinion about H1N1 vaccine effectiveness.
    - 1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.
- **opinion_h1n1_risk** - Respondent's opinion about risk of getting sick with H1N1 flu without vaccine.
    - 1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high.
- **opinion_h1n1_sick_from_vacc** - Respondent's worry of getting sick from taking H1N1 vaccine.
    - 1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried.

- **opinion_seas_vacc_effective** - Respondent's opinion about seasonal flu vaccine effectiveness.
    - 1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.
- **opinion_seas_risk** - Respondent's opinion about risk of getting sick with seasonal flu without vaccine.
    - 1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high.
- **opinion_seas_sick_from_vacc** - Respondent's worry of getting sick from taking seasonal flu vaccine.
    - 1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried.
- **age_group** - Age group of respondent.
- **education** - Self-reported education level.
- **race** - Race of respondent.
- **sex** - Sex of respondent.
- **income_poverty** - Household annual income of respondent with respect to 2008 Census poverty thresholds.
- **marital_status** - Marital status of respondent.
- **rent_or_own** - Housing situation of respondent.
- **employment_status** - Employment status of respondent.
- **household_adults** - Number of *other* adults in household, top-coded to 3.
- **household_children** - Number of children in household, top-coded to 3.

The dataset was adapted from the original data below:

U.S. Department of Health and Human Services (DHHS). National Center for Health Statistics. The National 2009 H1N1 Flu Survey. Hyattsville, MD: Centers for Disease Control and Prevention, 2012.

Requirements:

- Use the area under the receiver operating characteristic curve (AUC-ROC) as the primary evaluation metric when developing your models. [**1 mark**]
- Load the dataset and explore the training set to gain insights. [**2 marks**]
- Try at least 3 different machine learning models (e.g., add pre-processing transformers, use their default hyperparameters, etc.), and estimate the performance of the models on unseen data (e.g., using cross-validation). [**3 marks**]

- Select one model, optimize it (e.g., add/remove features and/or redesign the pipeline if you wish, perform hyper-parameter tuning, etc.), and (re-)estimate the performance of the model. We call it **model A**. [**5 marks**]
- Test model A on the test set, report the AUC-ROC, and at least 4 other evaluation metrics (e.g., AUC-PR, accuracy, sensitivity, specificity, etc.). [**2 marks**]
- Select 5 or 6 features from the original set of input features to train and optimize a new model (**model B**). Report the estimated performance of model B on unseen data. Explain the way you select those features (you can create *derived features* from those features so that the performance of model B can be improved). **[3 marks]**
- Test model B on the test set, report the AUC-ROC, and at least 4 other evaluation metrics (e.g., AUC-PR, accuracy, sensitivity, specificity, etc.). [**1 mark**]
- Plot the ROC curves of models A and B when testing on the same plot. **[1 mark]**
- Plot the PR curves of models A and B when testing on the same plot. **[1 mark]**
- Include a discussion section at the end of your notebook (about what you have learned, difficulties, what has worked and not worked, future directions, etc.). [**1 mark**]

Notes:

- Write **your name and student ID** at the beginning of your notebook.
- After completing your work, use menu item ***Kernel => Restart & Run All*** in Jupyter, then print your notebook, including code and outputs, to a **PDF file** for submission. Chrome may be the best browser for this. If you are not happy with the presentation format of the PDF file, you may want to use menu *File => Download as => HTML (.html)* then print file .html to PDF. ***Make sure the code in each cell is displayed completely.***
- You can use any public Python package.
- In order to compute AUC-ROC and AUC-PR, your model needs to make predictions in probability or score (i.e., a number in [0, 1], for example, by using function `predict_proba()`). However, to compute accuracy and other metrics, you also need to predict classes (e.g., using `predict()` or thresholding the predicted probabilities).
- Use your own assumptions and judgment if you are unsure about any information in the dataset. However, remember to mention it in the discussion.
- Try to write functions for all data transformations you apply, try feature engineering (e.g., creating new features), and try to automate all the steps as much as possible (e.g., using custom transformers, etc.). You may have **bonus mark**s for this.
- **Be creative**. You may get bonus marks for your creativity as well.
- Note that your total mark, including bonus (if any), will not exceed **20**.