CSE 4502 Big Data Analytics	Spring 2020
Midterm Exam	

Problem 1. (100 points) Automatic Coronavirus Test Kits

The fast automatic test kits lab is asking for help. The lab already developed a fast coronavirus test kit that can automate the testing of coronavirus. What they need help with is to improve the performance of their test kit in terms of recall and precision.

The test kit can automatically take 25 different measurements. The reading of the first measurement $x1$ is categorical. The rest of the readings $x2$ to $x25$ are numerical.

The lab provided us a training set of $30,000$ samples, and a test set of $10,000$ samples. The data sets are in the form of .csv files. The file names are train.csv and test.csv respectively. In both data sets, the last column y is the label of a sample, where value 1 means positive, and value 0 means negative. A positive label indicates the sample is from a patient infected by the coronavirus.

Our job is to use the training data to train a classifier that can achieve 100% recall and at least 70% of precision.

Once we can convinced that we have a classifier that can achieve this goal, we use the test set to test whether we indeed can achieve these requirements in the test set.

Save the code in midterm.ipynb and submit it to HuskyCT by the deadline.

The submission should include the following steps. Whenever applicable, use random_state=42.

1. Read train.csv file.
2. Obtain the high level information of the dataset.
3. Plot histograms of all the attributes.
4. Handle both categorical and numerical attributes. And use pipelines to prepare data for later steps.
5. Report the importance of each attribute by using the random forest classifier.
6. Train 3 different promising classifiers in terms of accuracy, and use these classifiers to obtain a soft-voting classifier.
    - Report the accuracy of each of these classifiers.

- Use 3-fold cross-validation for this step.

7. Explore the above 4 classifiers on the recall and precision trade-off and find a classifier that suits our purpose the most.
   - Plot the ROC curves for these classifiers.
   - Calculate the ROC AUC scores for these classifiers.
   - Plot the precision versus recall for these classifiers.
   - Find out the precision for each classifier when the recall is 1, and use these values to choose the classifier for our purpose.
8. Fine-tune the hyperparameters for the chosen classifier in terms of recall.
9. Use the best hyperparameters obtain above for the classifier to obtain the precision value when the recall is 1 using 3-fold cross-validation. If the precision is well above our target 0.70, go to the last step.
10. Use the test data from test.csv to check whether we can achieve the proposed requirement on recall and precision.