Challenge 3

In this challenge, you are provided with the same dataset of 37,438 Twitter accounts that are either bot or have been created by humans. You're asked to use two different supervised learning algorithms, including **k-NN and SVM** to classify accounts into bot or human. Prepare a report and submit a PDF file by Tuesday (**10/13/2020 before 6 pm**) considering the below details:

1. Read the CSV file into a data structure that you think is the best. The format of this dataset is as follows:
    a. For each row (i.e., Twitter account), you have **13 different features** and a label. If the account is bot, the label in the last column is **"bot"**; otherwise, it is **"human"**, indicating that this account has been created by a human.
    b. Features are either numerical or categorical. You can convert categorical features to numerical ones by relying on various techniques. The most popular one is to obtain the list of unique values for each categorical feature, and then, assign an integer number to each value.
    c. Some features might have missing or "unknown" values. So, you may consider either replacing those cells with appropriate values or removing that specific feature.
2. Shuffle the rows of this dataset, and then, divide it into two subsets: 80% for fitting (or training), and 20% for testing.
3. Import and use k-NN and SVM supervised learning libraries of Scikit-learn to do the below tasks:
    a. Evaluate each of the above algorithms via 5-fold cross-validation using the training dataset (i.e., 80% of rows). Save the models obtained from cross-validation to use it in the testing phase. Also, report the average prediction accuracy after 5-fold cross-validation. You can use any parameters that lead to better results.
    b. Apply each of the saved models (i.e., k-NN and SVM) to the testing dataset.
    c. Discuss which algorithm do you think is the best option for this specific problem, and why. You can explain this by either leveraging the classification performance metrics or any other criteria.