CSCE822 Homework1

Attached melb_data.csv file is the Snapshot of Tony Pino's Melbourne Housing Dataset. Do the following data preprocessing and apply KNN and RandomForest algorithms to classify the property prices.

1. Fill the missing values in the dataset using imputation approaches as we talked in class. You can use the scikit-learn's module

```
from sklearn.impute import SimpleImputer

my_imputer = SimpleImputer()

data_with_imputed_values = my_imputer.fit_transform(original_data)
```

The default imputer use mean values to fill the missing values. You can try other imputation method as well.

2. Replace the categorical/nominal attributes with one-hot-encoding

You can use Category Encoders package for use with scikit-learn in Python

Read this blog for more approaches for data encoding

https://towardsdatascience.com/smarter-ways-to-encode-categorical-data-for-machine-learning-part-1-of-3-6dca2f71b159

3. Install Weka system on your computer

Sort all the property samples by the property prices and divide the samples equally into 5 categories/classes: Top value, High value, medium value, low value, bottom value. Change the labels to the classes so it becomes a 5-class classification problem.

Install Weka https://www.cs.waikato.ac.nz/ml/weka/ software

Step1: You need to split the whole dataset into training (75% samples), 10% for validation, and 15% for testing datasets

Step2: Apply the KNN algorithm of Weka with K=5 to 10 to classify the property instances into 5 classes. Calculate the accuracy for each K values on the validation set, pick the best K value, $K_{best}$and then evaluate its performance on the test set.

Step3: repeat step1 and step2 10 times to get 10 test set performance and calculate mean and standard deviation.

4. Now do this same experiments using the RandomForest algorithm, but this time, we use scikit-learn package and report the performance.

Test set performance.

	K=5	K=6	K=7	K=8	K=9	K=10
KNN	Average accuracy.	…				
RandomForest	Average accuracy					

Write report to discuss the performances of KNN and randomforest on test sets. You are encouraged to compare the performance of different missing value imputation methods or the categorical encoding methods.

Hints:

a) random split function from scikit-learn

from sklearn.model_selection import train_test_split
xTrain, xTest, yTrain, yTest = train_test_split(x, y, test_size = 0.2, random_state = 0)

b) random forest example of scikit-learn can be easily found online.
c) For problem3, you can also use sci-kit learn if you prefer.

Submission:

Zip your code and the report and upload to https://dropbox.cse.sc.edu