National University of Computer and Emerging Sciences, Lahore Campus

Course:	**Applied Machine Learning**	**Course Code:**	**EE437**
Program:	**BS (Electrical Engineering)**	**Semester:**	**Fall 2020**
Assessment Tool:	**Assignment # 1**		
Total Marks:	**50**		

Submission Deadline: Wednesday, 07 October, 2020.
Assignment type: Individual
Submission : Submit the solution codes in a .zip file on SLATE

Python Introduction

Problems:

Strings

1. Write a Python program to count and display the vowels of a given text.

Lists

2. Write a Python program to remove duplicates from a list of lists.

Sample list : [[33, 120], [40], [30, 56, 25], [33, 120], [33], [40]]

New List : [[33, 120], [30, 56, 25], [33], [40]]

Tuples

3. Write a Python program to remove an empty tuple(s) from a list of tuples.

Sample data: [(), (), ('',), ('a', 'b'), ('a', 'b', 'c'), ('d')]

Expected output: [('',), ('a', 'b'), ('a', 'b', 'c'), 'd']

Dictionary

4. Write a Python program to combine two dictionary adding values for common keys.

   dictonary1 = {'a': 100, 'b': 200, 'c':300}

   dictionary2 = {'a': 300, 'b': 200, 'd':400}

   Sample output: Counter({'a': 400, 'b': 400, 'd': 400, 'c': 300})

5. In section 4 of Python tutorial, data munging process is demonstrated. Run and analyze all the python commands on same train.csv (provided with assignment) and match histogram and distribution graphs with your output. You have seen that the missing values in loan amount column are replaced by mean values. Your task is to explore three other ways of your choice of replacing the missing values. All values should be replaced by single value (i.e you can take standard deviation and replace missing values by it). For all three methods, fill self_employed column value with YES and NO. So there will be total 6 models. All other values in missing columns will be replaced by mean value or most occurrences (in case of Boolean).

   Analyze and comment on each output. What do you think which model is more practical in the above scenario and why? Briefly explain.

6. Extreme values of loan amount and applicant income are normalized in section 4 of python tutorial by taking logs of original value. Think about some other way to normalize both these columns (applicant income and loan amount)? Compare your histogram with the provided example. How your method is more practical?

   Briefly explain how will you impute the missing values for Gender, Married, Dependents, Loan_Amount_Term, Credit_History keeping in mind the practical scenario.