DSC402 ASSIGNMENT 3

HACKATHON – PHASE I

PROBLEM: DEVELOP A CLASSIFICATION MODEL THAT WILL IDENTIFY THE COMPENSATION LEVEL OF AN INDIVIDUAL.

STEPS

Develop the classification models Decision Tree / kNN / Naïve Bayes

1. Attached is a 'sample data' and below are the features of the data. Review the sample data (Model_Data_sample.csv) and the description below. Build an understanding of the data and plan your approach to build a model

age: continuous.

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

fnlwgt: continuous.

education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, ….

education-num: continuous.

marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, ….

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

sex: Female, Male.

capital-gain: continuous.

capital-loss: continuous.

hours-per-week: continuous.

native-country: United-States, Cambodia, England, Puerto-Rico, ….Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, …..

2. Define your data exploration, imputation and visualization approach. (1 Mark)

3. You will be sent the complete data file for modeling. (Model_Data.csv)

4. Set seed for sampling (your roll number) Eg. Set.seed(17125760345)

5. Split model data into train (80%) and test (20%)

6. Build 3 Models, each using one of different type of algorithm. Send me the model building command. (1 mark each + 1 for creative “DS” think = total 4 marks)

   model1 =

   model2 =

   model3 =

7. Predict your model performance on each of the 3 models and submit (1 mark each = total 3 marks)

model1_accuracy=

model2_accuracy =

modell3_accuracy=

8. Generalization (1 mark)

9. Upload details into your account on GIT Hub (1 mark)