School of Computer Science Engineering and Technology

Course- BTech
Course Code- 301
Year- 2022
Date- 25-03-2022

Type- Core
Course Name-AIML
Semester- Even
Batch- IV Sem Spl

9 - Lab Assignment # No. (9.1)

Objective: Use Naïve bayes Classifier to predict whether income exceeds $50K/yr based on census data in Adult dataset. Also known as "Census Income" dataset. This dataset consists of 15 attributes and 48,842 records. The list of attributes with description is given below:

1. age: continuous.
2. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. fnlwgt: continuous.
4. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
5. education-num: continuous.
6. marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
7. occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
8. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
9. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
10. sex: Female, Male.
11. capital-gain: continuous.
12. capital-loss: continuous.
13. hours-per-week: continuous.
14. native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

Target:

15. income : >50K, <=50K

1. Load the dataset from UCI repository: https://archive.ics.uci.edu/ml/datasets/Adult (5)
2. Do the exploratory analysis of the dataset to determine the importance of each feature: (15)
   - Perform univariate analysis by plotting various charts like bar charts, distribution plots, boxplots.
   - Perform multivariate analysis
3. Impute the missing values and remove any undesirable feature from the dataset. (10)
4. Check for the outliers in the columns and treat the outliers if present. (10)
5. Handle the categorical columns. Also for target column map the income categories to numeric form such as: " >50K " to 1 and " <=50K " to 0. (10)
6. Split the dataset into train and test. (5)

7. Construct Naïve Bayes model to predict the income of a person and compare the results for train and test subsets using accuracy, precision, recall, f1 score. Also check the values in confusion matrix. (10)

8. Look for real world applications where you can apply Naïve Bayes classification model.