School of Computer Science Engineering and Technology

Course- BTech
Course Code- CSET301
Year- 2022
Date- 12-09-2022

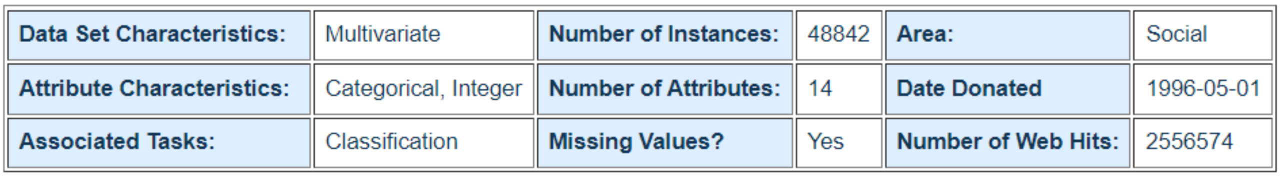
Type- Core
Course Name-AIML
Semester- Odd
Batch- V Sem

Lab Assignment 4.2_1

Exp. No.	Name	CO-1	CO-2	CO-3
4.2_1	Naïve bayes Classifier	✓	✓	--

Objective: Implement Naïve bayes Classifier model on "Census Income" dataset.

This dataset consists of 15 attributesand 48,842 records.



Data Set Characteristics:	Multivariate	**Number of Instances:**	48842	**Area:**	Social
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	14	**Date Donated**	1996-05-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	2556574

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 Columns:

income : >50K, <=50K

1. Load the dataset from UCI repository: https://archive.ics.uci.edu/ml/datasets/Adult (5)
2. Check the shape of the dataset (5)
3. Print the first 10 rows of the dataset (5)
4. Display the list of columns of the dataset (5)
5. Impute the missing values and remove any undesirable feature from the dataset. (10)
6. Check for the outliers in the columns and treat the outliers if present. (5) (Optional Part)
7. 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)
8. Split the dataset into train and test. (Ratio: 70:30, 80:20) (10)
9. Construct Naïve Bayes model (Hint: use GaussianNB model) (10)
10. Perform the prediction of test dataset (5)
11. Evaluate the performance of model on train and test subsets using accuracy, and precision. Also check the values in confusion matrix. (10)
12. Explore the different parameters while creating naïve bayes classifier model (10).

Suggested Platform: Python: Jupyter Notebook/Azure Notebook/Google Colab.