[Project Code : NANB]

Nursery School Application Selection using Naive Bayes Algorithm

Project Duration : 21st Jan 2024 to 10th Feb 2024
Submission Information : (via) CSE-Moodle

Objective:
In the 1980s, there was a period when an excessive enrollment drive in Nursery schools took place in Ljubljana, Slovenia, and hence there are thousands of applications which are to be shortlisted. In the process, the rejected applications are frequently asked for objective explanations/reasons. Since the selection was of nursery school students, a deterministic performance criteria was not sufficient to fix a cutoff for the same.
In this task, you have to train a Naive Bayes-based model to evaluate a nursery application based on various factors of the candidate.

In particular, you shall be doing the following tasks:

1. Based on the dataset (described later), you will write a program to learn a **Naive Bayes Classifier.**
2. Compare the results with the results generated by the Naive Bayes classifier learning algorithm from a pre-created package such as scikit-learn.

Note: *The program can be written in C / C++ / Java / Python programming language from scratch. No machine learning /data science /statistics package / library should be used.*

DataSets: `nursery.csv`

Data Description: The **attribute** Information is given as follows.

- *parents*: usual, pretentious, great_pret
- *has_nurs*: proper, less_proper, improper, critical, very_crit
- *form*: complete, completed, incomplete, foster
- *children*: 1, 2, 3, more
- *housing*: convenient, less_conv, critical
- *finance*: convenient, inconv
- *social*: non-prob, slightly_prob, problematic
- *health*: recommended, priority, not_recom

Output Classes: recommend, very_recom, spec_prior, priority, not_recom

Your Tasks:

1. The train dataset is not divided into train and validation sets. The first task is to randomly partition the train dataset into train and test sets using 80-20 split. Use the train split for training the tree and test split for testing.
2. Naive Bayes Classifier Model:
    a. Implement naive bayes algorithm in your code and mention the same in the report. ***DO NOT use scikit-learn for this part.***
    b. Compare the results of your implemented model with the Naive Bayes Classifier from scikit-learn package.

3. Classification Report:
    a. Create a classification report in tabular form.
    b. You need to calculate precision, recall, f1-score and accuracy of the model.

Submission Details: (to be submitted under the specified entry in CSE-Moodle)

1. ZIPPED Code Distribution in CSE-Moodle
2. A brief (2-3 page) report/manual of your work
   (with your hyperparameter tuning results also presented in that report)

Submission Guidelines:

1. You may use one of the following languages: C/C++/Java/Python.
2. Your Programs should run on a Linux Environment.
3. You are **not** allowed to use any library apart from these (Also explore all these libraries if doing in Python, or equivalent of these):

```
import numpy # linear algebra
import csv # data processing, CSV file I/O
import pandas # data processing, CSV file I/O
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
from sklearn.model_selection import KFold
import operator
from math import log
from collections import Counter
```

   Your program should be standalone and should **not** use any *special purpose* library for Machine Learning for the decision tree creation algorithm. Numpy and Pandas may be used. And, you can use libraries for other purposes, such as generation and formatting of data.
4. You should submit the program file and README file and **not** the output/input file.
5. You should name your file as <GroupNo_ProjectCode.extension>.
6. The submitted program file *should* have the following header comments:
   # Group Number
   # Roll Numbers : Names of members (listed line wise)
   # Project Number
   # Project Title
7. Submit through CSE-MOODLE only.
   Link to our Course page: https://moodlecse.iitkgp.ac.in/moodle/course/view.php?id=561

You should not use any code available on the Web. Submissions found to be plagiarized or having used ML libraries (except for parts where specifically allowed) will be awarded zero marks.

For any questions about the assignment, contact the following TA:

Ayan Maity (Email: ayanmaity201@gmail.com)