PA02: Support Vector Machine, Probabilistic Models, Trees

Machine Learning

Problem

부산대학교 컴퓨터공학과 학생인 산지니는 학과 사무실에서 근로 장학생을 하고 있다. 산지니는 학과 학생들의 정보를 정리하던 중에 이러한 데이터가 머신러닝에 사용될 수 있을지 궁금해졌다. 하지만 산지니는 머리가 영 좋지 않아서 친구인 당신에게 도움을 요청했다. 불쌍한 산지니를 위해 저번 수업에서 배운 모델을 활용하여 도움을 주자.

- 입력
  - 컴퓨터공학과 각 전공의 학생들의 정보가 담긴 csv 파일 “student-AI.csv”와 “student-Computer.csv”가 주어진다.

- 출력
  - Csv 파일 안에 있는 학생들의 개인 정보들을 적절히 활용하여 해당 학생의 전공을 예측하자.

- 제한
  - 이번 과제의 대상인 SVM, Naïve bayes, Decision Tree 모델만을 사용해야 한다.

Dataset

- “student-AI.csv”
  - 컴퓨터공학과 AI 전공 학생들에 대한 정보가 담겨 있습니다.
  - 395 rows x 23 columns

- “student-Computer.csv”
  - 컴퓨터공학과 컴퓨터 전공 학생들에 대한 정보가 담겨 있습니다.
  - 649 rows x 23 columns

Dataset

- Dataset preview.

school	sex	age	address	famsize	Mjob	Fjob	reason	traveltime	studytime	failures	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
SS	F	22	U	GT3	at_home	teacher	course	2	2	0	yes	no	4	3	4	1	1	3	6	C0	C0	C0
SS	F	27	U	GT3	at_home	other	course	1	2	0	yes	no	5	3	3	1	1	3	4	C0	C0	C0
SS	F	25	U	LE3	at_home	other	other	1	2	3	yes	no	4	3	2	2	3	3	10	C+	C+	B0
SS	F	25	U	GT3	health	services	home	1	3	0	yes	yes	3	2	2	1	1	5	2	B+	B+	B+
SS	F	26	U	GT3	other	other	home	1	2	0	yes	no	4	3	2	1	2	5	4	C0	B0	B0
SS	M	26	U	LE3	services	other	reputation	1	2	0	yes	no	5	4	2	1	2	5	10	B+	B+	B+
SS	M	26	U	LE3	other	other	home	1	2	0	yes	no	4	4	4	1	1	3	0	B0	B0	B0
SS	F	27	U	GT3	other	teacher	home	2	2	0	yes	no	4	1	4	1	1	1	6	C0	C0	C0
SS	M	25	U	LE3	services	other	home	1	2	0	yes	no	4	2	2	1	1	1	0	A0	A+	A+
SS	M	25	U	GT3	other	other	home	1	2	0	yes	no	5	5	1	1	1	5	0	B+	B+	B+
SS	F	25	U	GT3	teacher	health	reputation	1	2	0	yes	no	3	3	3	1	2	2	0	B0	C+	C+
SS	F	25	U	GT3	services	other	reputation	3	3	0	yes	no	5	2	2	1	1	4	4	B0	B0	B0

Dataset

- Dataset info.
  - **school** - student's school (binary: 'SS' - 수시 or 'JS' - 정시)
  - **sex** - student's sex (binary: 'F' - female or 'M' - male)
  - **age** - student's age (numeric: from 22 to 27)
  - **address** - student's home address type (binary: 'U' - urban or 'R' - rural)
  - **famsize** - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
  - **Mjob** - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
  - **Fjob** - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
  - **reason** - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
  - **traveltime** - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
  - **studytime** - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

Dataset

- Dataset info.
  - **failures** - number of past class failures(F) (numeric: n if 1<=n<3, else 4)
  - **internet** - Internet access at home (binary: yes or no)
  - **romantic** - with a romantic relationship (binary: yes or no)
  - **famrel** - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
  - **freetime** - free time after school (numeric: from 1 - very low to 5 - very high)
  - **goout** - going out with friends (numeric: from 1 - very low to 5 - very high)
  - **Dalc** - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
  - **Walc** - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
  - **health** - current health status (numeric: from 1 - very bad to 5 - very good)
  - **absences** - number of school absences (numeric: from 0 to 93)
  - These grades are related with the course subject, Math or Portuguese:
  - **G1** - first period grade (numeric: from D0 to A+)
  - **G2** - second period grade (numeric: from D0 to A+)
  - **G3** - final grade (numeric: from D0 to A+)

Todo

- 주어진 2개의 데이터셋을 모두 활용하여 데이터 EDA를 진행하고, 모델 학습을 위한 적절한 형태로 가공해주세요.
  - 불필요한 행 또는 열을 삭제해도 좋고, 필요한 열을 추가하여도 좋습니다.

- 가공한 데이터셋에 SVM, Naïve bayes, Decision Tree 모델을 각각 적용해주세요.
  - 이때 데이터를 재가공을 하거나 모델의 하이퍼파라미터를 바꾸는 등 다양한 방법을 사용해 모델의 성능을 향상시키는 경험을 해보시면 좋을 것 같습니다.

- 3가지 모델을 적용하여 실험한 결과를 잘 정리하여 보고서를 작성해주세요.

- 작성한 최종 보고서 및 코드를 PLATO에 제출해주세요.

Submission

- Submission form
  - 코드 구현 및 실행 후 실험에 대한 실행 결과가 저장된 형태인 ipynb 파일.
  - 그림, 표 등을 포함하는 보고서를 작성하여 제출하나, 분량은 그림, 표 등을 제외한 텍스트 기준 최소 워드 2장 이상의 분량으로 제출.
  - 각각의 알고리즘에 대한 성능 평가 및 결과 분석 내용 포함.
  - (Optional) 현재 머신러닝 수업을 수강중인 동료에게 도움을 받았을 경우, 동료의 학번/이름 작성.
    - 추후 표절 시비를 방지하기 위함.
  - 과제 제출 기한은 2024.04.28(일) 23:59 까지 입니다.
    - 추가 제출은 데드라인 이후 1일 이내이며, 점수의 50%가 감점됩니다.

Thank you