Australian Student Performance Data Machine Learning Project

Team:

Kaur Kingsepp,

Sander Valge,

Rasmus Valk

Task 2. Business understanding

1. Business and goals

We are working with a Kaggle dataset called „Australian_Student_PerformanceData (ASPD24)“. This dataset represents student performance in higher education institutions across Australia. There are total of 51 features describing the academic, personal, demographic, socio-economic and other features of one instance. The data is anonymized and no source for it has been provided.

Our business goal is to predict the performance of students. We will do predictions on our own small dataset of people studying in Estonia, which most likely is beneficial to the identification of problem features for those people. In the end we will predict the performance class for each student (Excellent, Good, Satisfactory, Needs improvement, Poor). In addition, we will experiment predicting different features, like GPA.

2. Situation assessment

In order to do achieve our goals we have the dataset with 100 256 records and 51 features. Additionally, we will make predictions on our own dataset. In terms of tools, we will be using Python with standard ML libraries like pandas, scikit-learn to name a few. Then we will have a workstation capable of training the classification models.

Our requirements are to develop a model for student performance category and possibly other categories. We will evaluate the model using standard metrics. And then look at feature importance to understand factors influencing the performance. Or dataset is assumed to be cleaned with no missing values, which the original already is. The dataset is assumed to be internally consistent although we are not sure if it is based on real instances or partially or fully synthetic. This puts a constraint on us, which would limit our interpretability of results. We will handle personal information accordingly.

There is a risk with the unknown origin of the data, which might not represent the real world ideally. There are difficult features with many different nominal values, which should be handled properly in order to do ML or dropped.

Terminology and ML terminology:

Performance class – The categorical target variable indicating student's academic level (Excellent, Good, Satisfactory, Needs improvement, Poor).

Feature – A measurable attribute such as age, parental education, study time, or attendance.

Target – The variable the model aims to predict (Performance).

GPA – Grade Point Average; a numeric academic success measure.

We will use a Classification models – A model that predicts categories instead of numeric values.

In terms of costs there are no other than time investment for the necessary tasks, and computational costs. The benefits are that we will obtain a model for academic performance classification that helps us to determine which factors most strongly influence performance. And this will be extended to other features, like GPA. Our benefit is that we will learn more about machine learning.

3. Data-mining goals

We will build a machine learning model that can predict the performance label using the provided 51 features. This includes exploring which algorithms perform the best. Finding, which features contribute most to predictions and identifying patterns that correlate with higher or lower performance. We will also experiment predicting different features.

Our criteria are that the model should significantly outperform random guessing while achieving high accuracy and AUC. For this we will use different ML methods.

Task 3. Data understanding

1. Gathering Data

The dataset used for this project is the „Australian_Student_PerformanceData (ASPD24)", obtained from Kaggle. The data requirements for this project include a clearly defined target variable representing student performance with a diverse set of predictor features covering academic, demographic and socio-economic characteristics. These requirements are met by ASPD24, which provides a performance class label along with variables such as age, gender, parental education, study time, and many more. The dataset is free to use and downloadable in CSV format. It is sufficiently large to support machine learning models. No additional filtering was required, and the full dataset was used for analysis.

2. Describing Data

Upon loading the data, the dataset was reviewed to understand its structure and basic characteristics. It contains 100 256 records and 51 columns representing both numerical and

categorical variables. Numerical features include measures such as age and study hours, while categorical features cover aspects like gender, parental background, and various school-related indicators. The target variable “Performance” is a categorical class label suitable for classification. Initial descriptive statistics indicate that the dataset is well-organized and free of structural issues. Each feature contains values within expected ranges, and there are no apparent formatting inconsistencies. There are also some “nan” values, but these are also expected, as they describe the college level of students’ parents. However, early observations also revealed that several features display unusually balanced or uniform distributions, which is uncommon in real-world educational datasets and may indicate synthetic data generation.

3. Exploring Data

Exploratory Data Analysis (EDA) will be conducted to examine distributions, identify patterns, and gain insights into potential predictive relationships. Visualizations such as histograms, boxplots, and bar charts will be used to assess the shape of each variable. The default histograms in Kaggle show quite uniform distributions instead of natural skewness or clustering typically seen in real student populations.

4. Verifying Data Quality

Data quality was assessed in terms of completeness, consistency, and reasonableness. The dataset contains no meaningful missing values, and all variables are consistently formatted, which simplifies preprocessing. However, the complete absence of meaningful missing data, outliers, or irregularities is unusual for real educational datasets. Though these datapoints may as well have been previously removed. Distribution shapes and weak correlations further suggest artificial data patterns. Although the dataset is clean and fully usable for machine-learning experiments, its authenticity and accuracy cannot be guaranteed. For the purposes of this project, the dataset is considered suitable, but limitations arising from its possible synthetic nature are acknowledged.

Task 4. Planning the project.

Data analysis and visualization – 5 hours - Kaur

Data preparation, feature engineering – 6 hours - Rasmus

Modeling and hyperparameter tuning of different models – 20 hours - Kaur, Rasmus, Sander

Evaluation of models – 10 hours - Kaur, Rasmus, Sander

Documentation and poster – 5 hours - Kaur, Rasmus, Sander