Project E8: KAGGLE - TWO PATHS TO PREDICTING SUCCESS

Team members: Marek Kalda and Nele-Liis Võhma.

Task 2: Business understanding

Background:

Academic performance is influenced by various factors, such as study hours, parental involvement, attendance, access to resources, sleep hours and more. Our Kaggle dataset includes information about 6607 students and how their attributes as social, behavioral and environmental factors affect students performance. Schools often struggle with identifying at-risk students early enough to help them get back on track before it is too late. This project explores two paths to analyse factors influencing student grades and compare which method offers the most practical insights.

Business goals:

This project does not benefit any business, but rather educational stakeholders such as teachers, students and schoolworkers overall. The goals are to better understand the drivers of academic success among students, build predictive tools that could identify students at risk and who need more support to assess which prediction model provides more practical and actionable insights.

Business success criteria:

Demonstrate that predictive models can estimate final grades or classify high-achievers and low-achievers with acceptable accuracy, more measurable metrics can not be set because synthetic data is unpredictable. Find out which predictive model provides more actionable insights for identifying students at risk of low performance.

Inventory of resources:

Kaggle dataset StudentPerformanceFactors.csv, which contains data on 6607 students and 20 features including final exam score. Data is created synthetically. Two team members, who put 30 hours of work each to finish the project by deadline. Jupyter Notebook, Python and different packages used for data science.

Requirements, assumptions and constraints:

Clean the dataset, handle missing values, encode categorical variables, build regression and classification models, compare them and finally provide visualizations, all before the deadline. Assuming the data is accurate and representative of student behaviour even when synthetic. Assuming the data does not contain missing values since it is created synthetically.

Risks and contingencies:

Since the data is synthetic there could be abnormalities or unrealistic values, only reliable features should be considered. Imbalances in the dataset and classification model is biased, apply undersampling techniques for majority class or adjust class weights. Poorly chosen performance threshold does not reflect meaningful academic cutoffs, multiple thresholds should be evaluated to find optimal threshold that provides meaningful outcome.

Terminology:

Regression model - A learning model used to predict a continuous numerical value, in this case the final exam grade.

Classification model - A model that predicts category labels, in this case whether a student is a high-achiever or low-achiever.

Threshold - A numerical cutoff that transforms continuous grades into binary classes.

Undersampling - A resampling technique used to address class imbalance by reducing the size of the majority class via selecting a subset of the class to make dataset more balanced.

Costs and benefits:

Costs include time for data exploration, cleaning, modelling, analysis and computational resources for training multiple models. Benefits include skills and knowledge for future projects, insights into academic performance factors and comparative understanding of modelling approaches.

Data-mining goals:

Process the given dataset, clean and understand the data. Build a regression model to predict the student's final grade using the factors provided in the dataset. Build a classification model by selecting a performance threshold and group low-achievers and high-achievers. Compare the approaches to identify the model that give more practigal insight. Identify most influencial factors driving student success.

Data-mining success criteria:

Regression achieves sufficiently high performance. Classification achieves acceptable classes. Clear comparison of the two paths to predict student success and identify at-risk students.

Task 3: Data understanding

Outline data requirements:

Our project aims to identify key factors influencing students academic performance and to model outcomes using both regression and classification models. Therefore, the data must have:

1. A target variable representing the academic performance, in this case a final grade which will be suitable for regression.

2. Attribute variables describing students characteristics, behaviors and environments, such as study habits, sleep hours, parental involvement, motivation, lifestyle factors, previous education, learning disabilities and attendance.
3. Sufficient sample size to support reliable modeling and cross-validation

Verify data availability:

Our dataset comes from the Kaggle project Student Performance Factors Dataset and includes 6607 student records with 19 attributes and the final exam score. The dataset is publicly accessible and downloadable as a CSV file. It already contains everything we need to reach our project's goal, so no external data sources are required.

Define selection criteria:

The dataset includes many variables, but not all may be relevant for the project. We will choose variables based on relevance to the academic performance, completeness (columns with a lot of missing values will be discarded) and interpretability (attributes must be meaningful for educational performance). Currently all variables are kept for EDA.

Describing data:

Our main target variable is named "Exam_Score" which is measured numerically. Input attributes include gender, distance from home, parental involvement, parental education level, peer influence, a lot of academic related behaviours (hours studied, attendance, school type, extracurricular activities, previous scores), psychological factors (sleep hours, motivational level, physical activity) and some environmental variable like access to internet and resources. The attributes split evenly into numeric and categorical.

Exploring data:

For initial exploration we would like to uncover patterns, distributions and relationships between some of the attributes. For numeric variables we use histograms to assess skewness. With boxplots we can detect any unusually high or low values within the attributes. For relationships group comparisons are used, such as average grades by motivation level, parental involvement, or study time. The initial findings already show that higher study time correlates positively with the final grade, whereas motivation and parental support tend to cluster with high academic outcomes and some features seem to have minimal impact on performance like gender and family income.

Verifiying data Quality:

Our checks to verify the data quality include identifying the number of missing values within the columns, checking for out of range values (for example check that exam grade falls within expected boundaries and categorical attributes only have valid labels), looking for inconsistent spelling in categorical attributes like "Yes/No" and "yes/no". At this stage, the dataset appears generally clean, no missing values were found, numeric variables didn't have any visible outliers and categorical variables seem to have no inconsistencies in spelling.

Task 4: Planning the project

To complete the project efficiently, we divided the work into five main tasks, aligning with the CRISP-DM framework. Each team member contributes 30 hours of work, resulting in a combined total of 60 working hours.

First task is to explore the dataset to understand the data better, examine attributes, identifying potential issues, understand relationships between features and the target and documenting initial findings. This is the key to meaningful outcome of the project so each member should contribute 7 hours to understand the data.

Next prepare the data for modelling (regression and classification) via cleaning the data, encoding categorical variables, handling imbalances within the classes and selecting relevant features to analyse. This task is also crutial for realistic outcome so each member should put 7 hours of work to this task.

The main task is building regression and classification models, testing various algorithms, tuning basic hyperparameters and selecting the strongest models for comparison. As building models does not take as much time as preparation, each member should contribute 6 hours.

The meaningfulness of our project comes from comparing the performance of the models based on accuracy metrics and interpretability. Identifying key factors influencing academic performance and interpreting feature importance results. Each member should contribute 5 hours to this task.

Finally, to give the project its full value, we dedicate time to produce all final deliverables. This includes writing the final report, creating visualizations, preparing a PDF poster and presentation slides if needed. These outputs must communicate our findings effectively and to ensure this, each member should put approximately 5 hours of work to make the project presentable.

Using Python, Jupyter Notebook, pandas, scikit-learn, matplotlib/seaborn, train-test split, regression models (e.g. Linear Regression, Random Forest), classification models (e.g. Logistic Regression, Decision Tree, Random Forest) and class balancing techniques (e.g. undersampling).

Link to the repository: https://github.com/liisnele/IDS2025-performace-factors