Liisi Koitjärv
Mihkel Hain
Group 7

Project G5: KAGGLE-STUDENT PERFORMANCE

GitHub: LiisiKoitjarv/G5-StudentPerformance

Business understanding

Identifying your business goals

- Background

  There are many factors that could affect a student's academic performance. We aim to find and spread awareness on factors that affect it the most.
  The Kaggle dataset we're working with provides useful information about various students that potentially affect their performance on the final exam.

- Business goals

  The primary goal is to build a regression model to accurately predict students' results on the final exam. We also want to find out by analyzing our data, what factors influence students' final exam results the most.

- Business success criteria

  The project will be successful if our model can predict the students' exam results accurately, at least in the range of +- 10 points (e.g. the student scored 91 on the exam, and the model predicts anywhere from 81-100).

Assessing your situation

- Inventory of resources

  Dataset: Kaggle dataset containing an overview of various factors affecting student performance in exams and their final exam results. (link: Student Performance Factors)
  Tools: Jupyter Notebook, Python, Pandas, Seaborn, scikit-learn, NumPy, laptops
  Team: Two students with some data science experience.

- Requirements, assumptions, and constraints

  The dataset represents typical students with accurate variables.
  The project must be done by December 8th 2025, at 12pm.
  The dataset is small (641.95 kB) so that may limit the complexity of the model.
  The dataset is largely unbalanced so some sort of balancing needs to be done. Low variability of exam scores could make predictions less accurate.

Liisi Koitjärv
Mihkel Hain
Group 7

- Risks and contingencies
  Risk: Data quality, exam results are unbalanced and have low variability, some missing values present.
  Contingency: Clean the data of missing values, balance the data.
  Risk: Predicting the precise exam results (points) may prove to be impossible.
  Contingency: Don't expect an exact prediction, give the model some wiggle room to make approximate predictions (e.g. 10 points).

- Terminology
  Regression model - a model, that predicts a numerical value (exam score)
  Features - attributes / factors that affect student performance
  Key features - the most influential attributes that affect final exam results
  MAE - mean absolute error, the average size of errors between predictions and actual values.
  RMSE - root mean squared error, the square root of the average of squared errors, an evaluation metric for regression models, sensitive to big errors.
  R^2 - coefficient of determination, measures how well a model predicts the outcome.

- Costs and benefits
  The main cost is time spent by the team on the project, including understanding and cleaning the data, training the model, analyzing the results, finding correlations between certain factors and the final exam results.
  The project could benefit educators, students or their parents by potentially spreading awareness on the most impactful factors on student performance.

Defining your data-mining goals

- Data-mining goals
  Build a regression model that predicts final exam scores. Identify key features that affect student performance the most.

- Data-mining success criteria
  Achieve accurate predictions of exam results.

Liisi Koitjärv
Mihkel Hain
Group 7

Data understanding

Gathering data

- Outline data requirements
  The data requirement is to be able to identify the factors that lead to better exam scores. The dataset is to contain multiple factors about different students studying habits and their socio-economic situation.

- Verify data availability
  For this we're gonna use StudentPerformanceFactors.csv for the data itself, the data contains 20 variables such as Hours_Studied, attendance etc.. In the dataset there are over 1000 records contained.

- Define selection criteria
  Inside the selected dataset, all of the attributes will be used and considered for analysis. In the future if any of the attributes are determined to be, they will be removed. Right now the primary attribute is Exam_score, which gives us a target variable to test against.

Describing data

The dataset contains the records of 1000 students and attached to those 20 students are 20 attributes that might be related to the exam score the students achieved.
The data itself contains multiple numerical and ordinal or nominal categorical variables

The numerical variables are: Hours_Studied, Attendance, Sleep_Hours, Previous_Scores, Tutoring_Sessions,Physical_Activity and Exam_Score

While the categorical values are: Parental_involment, Access_to_Resources, Extracurricular_Activies,Moitvation_Level,Internet_Access, Family_income, Teacher_Quality,School_Type, Peer_influnce, Learning_Disabilities, Parental_Education_level, Distance_from_home,Gender

The data inside the dataset should be enough for us to see what values lead to better exam scores and be able to predict the exam scores of students.

Liisi Koitjärv
Mihkel Hain
Group 7

Exploring data

After exploring the data i will bring fourth the ranges and correlation to the exam_score of each of the attributes
Numerical Values-
Hours_Studied- Value Range: 1-44, Correlation 0.445
Attendance- Value Range: 60-100, Correlation 0.581
Sleep_Hours- Value Range: 4-10, Correlation -0.017
Previous_Scores- Value Range: 50-100, Correlation 0.175
Tutoring_Sessions- Value Range: 0-8, Correlation 0.157
Physical_Activity- Value Range: 0-6, Correlation 0.028
Exam_Score- 55-101, This is the target variable, but it is interesting the maximal value is 101.
Categorical Values-
Parental_involvement: values- low 20%, medium 51%, high 29%
Access_to_resources : values-low 20% ,medium 50%,high 30%
Extracurricular_Activies: values- no 40%, yes 60%
Motivation_Level: values- low 29%,medium 51%,high 20%
internet_access: values- No 8%, yes 92%
family_income: low 40%,medium 40%,high 19%
Teacher_quality: low 10%, medium 60%, high 30%, missingvalues 1.2%
School_type private 30%,public 70%
peer_influence: negative 21%, neutral 39%, positive 40%
Learning_disabilites: no 89%, yes 11%
Parental_education_level: high_school 49%, college 30%,
postgraduate 20%, Missing values 1.4%
Distance_from_home: Near 59%, Moderate 30%,Far 10%, missing values 1%
Gender: Female 42%, Male 58%

Verifying data quality

Overall data quality is good with only 3 attributes having any nan values.
There are concerns about outliers influencing the final result for example a Exam_score of 101 exists but overall the value ranges are within expected boundaries.
Attributes that have missing values are: Teacher_Quality 1.2% missing,

Liisi Koitjärv
Mihkel Hain
Group 7

Parental_Education_level 1.4% missing and distance_from_home 1% missing.
Due to the low number of missing values, to fix missing values using some sort of data imputation is applicable.

Planning your project

Make a detailed plan of your project with a list of tasks

Task 1: Data understanding and cleaning.
Both team members should understand the data that we're working with and we need clean data to train our model.
Estimated time: About 3 hours each.

Task 2: Find key features.
Analyze which factors affect student performance the most
Estimated time: About 3 hours each.

Task 3: Train and evaluate the model.
Train a regression model that accurately predicts the exam scores of students. Evaluate the model's accuracy (RMSE, R^2, MAE)
Estimated time: About 10 hours each

Task 4: Visualization of results.
Use plots to visualize the model's performance, show correlations between final exam results and key features.
Estimated time: About 4 hours each

Task 5: Making the poster.
Make a poster for the final presentation detailing the results of the project.
Estimated time: About 5 hours each.

List the methods and tools that you plan to use

We plan to use regression algorithms, Jupyter Notebook, Python, Pandas, Seaborn, scikit-learn, NumPy, cross-validation.