Project E2: Ann Marleen Varul, Karina Pinajeva, Jana Kotšnova

Project repository:
https://github.com/kkerychka/KAGGLE-E2-INSIGHTS-INTO-STUDENT-PERFORMANCE-

INSIGHTS INTO STUDENT PERFORMANCE: *Predicting Outcomes using Machine Learning and Identifying Risk Profiles by clustering*

1. BUSINESS UNDERSTANDING

1.1 Background

Educational institutions currently rely on reactive approaches to student support, usually identifying struggling students only after poor performance such as failed exams or dropout rates. This project develops a predictive system that helps with early identification of at-risk students and provides data-driven insights for intervention strategies.

The project is mainly aimed at people involved in education: educational institutions (like public and private schools, as well as the Ministry of Education). More precisely it benefits the educators and policymakers (teachers, academic advisors, and education policymakers). Additionally, students and their families benefit by receiving personalized suggestions on improving academic performance.

1.2. Business goals

Goal 1: Support teachers in identifying students who may need extra attention.

- The prediction model aims to highlight which students might struggle on the final exam so teachers can plan support or interventions more efficiently.

Goal 2: Give school officials an overview of student groups and potential risk patterns.

- Clustering is meant to reveal trends, such as groups of students who show similar academic or behavioural patterns. Since schools and ministries operate with limited budgets for student support services, this could help schools and decision-makers decide where additional resources or support programs might be needed. By accurately

categorizing students into risk profiles, institutions can target their limited resources more effectively.

Goal 3: Provide students with simple and personalised improvement tips

- Students benefit from specific, actionable feedback rather than generic advice. The small LLM aims to generate short feedback messages that help students reflect on their habits and see what changes might improve their performance. Each student will be able to receive tailored recommendations based on their individual patterns and circumstances, which could empower students to improve their academic performance.

Goal 4: Offer a data-driven foundation for discussions or policy considerations

- Educational institutions need to identify students at risk of failure while there's sufficient time for meaningful intervention. This requires concrete data about the relationship between student backgrounds (socioeconomic factors, living conditions, etc) and academic outcomes. The results of such analysis (overall patterns, risk profiles, key factors) can be used as an empirical foundation for developing inclusive educational policies.

1.3. Business Success Criteria

The success of this project will be measured by how well our machine learning models perform. For the regression model that predicts final exam scores, we will split our data into training and testing sets (80/20) to check how accurate the predictions are. We're aiming for an R^2 score of at least 0.75, which would mean the model can explain most of the variation in student scores. We'll also look at how far off our predictions are from actual scores on average. As for clustering, that groups students into risk categories, we'll test its accuracy by coloring the data points based on their assigned groups and checking if students with similar characteristics actually end up in the same clusters. We'll use fuzzy clustering techniques to see how clearly separated our groups are. Good clustering should show distinct groups (high-performing, moderate-risk, and at-risk) without too much overlap between them. The language model will be tested on whether it can generate helpful and relevant feedback messages that match each student's predicted risk level and study patterns. To make sure our models work well with new data and not just the data we trained them on, we'll use cross-validation: testing the models multiple times with different data splits to confirm they give consistent results.

1.4 Assessing your situation

Learning institutions can be said to commonly adopt response actions when dealing with struggling students, such as when students perform poorly in exams or when they show lack of commitment or indiscipline. It should be noted that almost every school uses computer-based records to capture students' academic and pre-academic details but rarely exploits these records to establish risks or predict problems that students can face academically in future.

There are several problems arising from this gap:

- Late identification of students who require academic or psychological support.
- Inefficient allocation of limited educational resources, such as tutoring, counseling, or targeted interventions.
- Absence of personalized feedback that could guide students toward improving their learning habits.
- Lack of analytical evidence to support strategic decisions at the school or ministry level, especially when planning educational policies or resource distribution.

It directly resolves these shortcomings because it incorporates predictive analytics into the education system. It ensures that early identification of at-risk students occurs so that necessary measures can be taken with efficiency. It not only improves educational outcomes but also ensures well-informed decisions with objective insights because it is evidence-based. It boosts efficiency at educational institutions and improves fairness and effectiveness at supporting students.

For building an accurate predictive model, it is important to assess factors such as:

- In what ways contemporary processes of education could benefit predictive analytics and if instructors and school leaders can properly utilize these models.
- Representativeness of Kaggle's student's dataset to real-life academic setting and how closely the structure of the insights satisfies real-life factors influencing students' performances.
- What kinds of Data quality concerns could affect your project, like incomplete Data, Biased Data or Data that lack detail?
- What are the roles of different participants in the learning environment like teachers, school administrators, policymakers, students, and parents, and how will these different parties' individual needs be met?

1.5 Inventory of Resources

For a predictive system to be successfully built, it would be important to evaluate the resources at your disposal and under which circumstances you would be implementing your project.

Data resources

The project requires use of a dataset with 6,590 students and incorporates multiple factors for analysis using various models and statistical techniques.

The dataset contains:

- Learning behavior variables:
  Hours_Studied, Attendance, Access_to_Resources, Motivation_Level
- Family and socio-economic background:
  Parental_Involvement, Family_Income, Parental_Education_Level
- Personal well-being indicators:
  Sleep_Hours, Physical_Activity, Learning_Disabilities
- School environment factors:
  Teacher_Quality, School_Type, Peer_Influence, Distance_from_Home
- Academic performance:
  Previous_Scores and final Exam_Score

These aspects can provide a robust basis for regression modeling, clustering techniques, or even customized recommendation systems because they encapsulate different dimensions of a student's academic life.

Technical resources

- Python environment with libraries Scikit-learn, Pandas, NumPy for modelling
- Python Packages Supporting LLM Training, Clustering Methods, and Regression Modelling
- Visualization tools Matplotlib, Seaborn for exploratory analysis
- Jupyter Notebook for running experiments, visualizing results, and documenting modelling steps.

- GitHub for version control, collaboration, and storing code, notebooks, and model outputs.
- Cloud or local storage for processing and safeguarding student data

Human resources

- Student Data Scientists - are part of the project team involved in creating regression models, clustering techniques, preprocessing of data, and analysis related to model evaluation.
- Student Data Analysts - team members involved in exploratory data analysis and creation of visualizations to explain variable relationships and outcomes.
- Instructor - serves as expert guidance for methodology and ensures project meets analytical/ethical best practices.
- Project Team Members - develop the system jointly, document processes, test models, incorporate LLM into the system, and achieve project objectives.

1.6 Requirements, Assumptions, and Constraints

Requirements

The model should be able to forecast scores on the final exam using regression techniques with input variables such as those provided for the dataset, which range from Hours_Studied to Sleep_Hours and other factors related to academic performance. It would be preferable if it could yield an R-squared of at least 0.75.

Apart from performing regression analysis, the project should contain clustering techniques that can segment students into logical groups such as high performers, moderate risk, and at-risk students based on factors such as Motivation_Level, Access_to_Resources, Family_Income, Physical_Activity and Peer_Influence.

A language model (LLM) should create concise messages with recommendations related to individual students' profiles (recommendations related to sleep for students with low Sleep_Hours).

The system should also offer visualizations of important insights like importance or characteristics of features or clusters. Lastly, the workflow should support adding new student data (new exam cycles or new cohorts) so that new predictions can be made.

From a non-functional point of view:

The system must be interpretable so that teachers can know how individual variables such as Hours_Studied or Parental_Involvement affect or lead to the final outcomes. The models should work well on several splitting techniques and ensure that basic ethics are applied, especially while dealing with sensitive data such as Gender, Family_Income, or Learning_Disabilities. It should be easy to use and scalable if deployed among a larger set of students.

Assumptions

It is assumed that the given set of data comes close to representing actual classes of students and that these measures of student engagement and activity patterns reflect their actual performances. It can also be safely assumed that entries or measures recorded under factors such as “Motivation Level”, “Peer Influence”, “Teacher Qualifications” or “Learning Disabilities” provide accurate enough data for analysis purposes.

We propose that teachers would be ready to utilize outputs of this system provided that there would be clear delivery of predictions and recommendations. Additionally, we suggest that students would find a benefit provided that recommendations would be positive and helpful and suited to improvement purposes at an educational level. Lastly, we propose that the dataset was adequate enough for machine learning models to not require too much imputation.

Constraints

There exist challenges or limitations to the project. For instance, certain features such as Peer_Influence and Parental_Involvement can be subjective or recorded in an irregular manner and therefore can impact model accuracy. There exist concerns or measures related to ethics that entail sensitive features like Family_Income or Learning_Disabilities.

There could be technical constraints if either learning institutions or schools lack enough computing power to execute complex models rather than LLMs. There could be financial constraints if scaling or sustaining the system becomes too expensive. Lastly, organizational constraints could arise if teachers show hesitation or unwillingness to adopt new technologies when making decisions.

1.7 Risks and contingencies

The following are some of the risks that we have to consider when building this predictive system: first, the model will not always predict accurately. This might happen because, in this case, there is a certain subjectivity concerning some variables like Teacher_Quality or Peer_Influence, which are not very precisely measured. In order to reduce this problem, it is important to properly encode these variables and enhance the features used in the model. The second kind of risk is bias. It may well occur that factors like Family_Income or Gender will make the model biased. To counter this, periodic checks should be made for possible bias in the model and corresponding adjustments made.

Of course, there are also risks related to the dataset. Values, like Sleep_Hours or Physical_Activity, can be wrong or not realistic, which will affect the quality of our predictions. This should be solved by data cleaning and the modification or substitution of bad values. Additionally, categories like Learning_Disabilities or Access_to_Resources may not be representative of the actual situation-a fact that the model cannot capture fully. In the future, this might require further enlargement of the dataset to solve.

Third, we need to put ourselves in the place of people who may use or be affected by the system. For example, it is unlikely that teachers will take action based on predictions like, "this student is at risk because of low motivation" unless they understand how the model arrived at such a conclusion. Explanations and simple visualizations can help engender trust. Parents may also worry if sensitive variables like Family_Income appear to influence the outcome on exams. Here, we reassure them that this is a model designed to provide support, not a judgment or label applied to students.

Finally, some of the schools may not have the technical capacity to store and process data for more than 6,590 students securely. In order to avoid this problem, the project can make use of encrypted cloud storage or lighter versions of the system that require fewer resources.

1.8 Terminology

- Academic Performance – a student's overall academic performance, measured by Previous_Scores and Exam_Score.
- Regression Model – a machine learning model that predicts a numerical metric, such as the final Exam_Score.
- R-squared – a metric showing how well the model explains grade variations.
- Clustering – a method for grouping students by similar characteristics (e.g., high-performing, moderate-risk, or at-risk).
- Fuzzy Clustering (Fuzzy C-Means) – a method in which a student can partially belong to multiple clusters.
- Risk Profile – a risk level indicating the likelihood of a student experiencing academic problems.
- LLM (Large Language Model) – a language model that creates personalized text recommendations for students.
- Feature – any metric from the dataset (e.g., Hours_Studied, Motivation_Level, Sleep_Hours).
- Feature Importance – an assessment of which features have the greatest impact on the final Exam_Score.
- EDA (Exploratory Data Analysis) – analyzing data through graphs and statistics to identify patterns.
- Categorical Variable – categorical data: School_Type, Gender, Family_Income, Access_to_Resources.

- Numerical Variable – numerical data: Hours_Studied, Attendance, Sleep_Hours, Physical_Activity.
- Bias – bias or unfairness of the model that arises when using sensitive features (Gender, Family_Income).
- Model Interpretability – how clear is it to users how the model makes its predictions.
- Data Imputation – the process of replacing missing or incorrect values.
- Predictive Analytics – using data and models to predict future performance (e.g., exam results).

1.9 Costs and benefits

In developing this predictive system, different costs and benefits need to be considered. Although the project is carried out by university students with no large financial investment, it still involves certain resource demands. The main costs include those pertaining to computing, time, and organizational effort. On the technical side, training the models – particularly the LLM – may require the use of cloud services or GPU support; for a small expense, paid platforms would be applicable. Storing the dataset and project files securely, particularly if cloud storage is chosen, might also incur minor costs. Most software libraries applied in the project are free, but managing the different tools, notebooks, and packages requires careful setup and maintenance.

Another significant cost is human effort. Team members have to invest time in data cleaning, developing regression and clustering models, their evaluation, and preparing visualizations. And, of course, studying how to fine-tune or apply an LLM requires extra study. The instructor who led this project invested time in ensuring that the methodology was correct and ethical considerations were upheld. Being a student project, a major component of the cost is just time and effort to learn and apply a new tool.

Besides, there are organizational costs such as data preparation, creating GitHub repositories for the code, organizing teamwork, and testing different modelling approaches. Additional work will be needed to prepare documentation, interpret results, and validate assumptions.

Despite these costs, the system provides substantial benefits. For students, the model can identify risks early in a student's academic career and provide personalized recommendations on how to improve study habits, sleeping routines, or motivation. Students who then make changes in advance of exams will get a better understanding of which variables are affecting their performance. Educators and advisors also receive

objective insights into the reasons why some of their students might be struggling. Instead of following their intuition, educators will see which variables – such as Hours_Studied, Motivation_Level, or Access_to_Resources – play the most important role. This helps educators target support more effectively and free up resources for where it is needed most.

Schools can also benefit more generally through a better understanding of broader patterns in student behavior and performance. The analysis underlines how certain groups of students might need extra support, showing trends related to socio-economic background, family involvement, or school environment. These can enable schools to plan their intervention more strategically and could improve academic attainment across the board.

This would be of benefit to the student team itself, as well: it gives hands-on experience with real data, machine learning, clustering, and LLMs. By collaborating and using GitHub, the team acquires practical experience that will be useful in future academic or professional work. It deepens understanding regarding how predictive analytics can be used in real educational settings; it offers both technical and social learning outcomes.

1.10 Defining your data-mining goals

CRISP-DM defines data mining as the process of using statistical and machine-learning techniques to discover patterns in data and generate models that support business decision-making. Thus, data-mining methods are necessary. Regression will be used to predict each student's final exam score based on behavioural, socio-economic, and academic features. Clustering will be used to discover meaningful risk groups that show similar learning patterns. Together, these data-mining tasks support early-risk identification, targeted interventions, and the generation of personalised feedback through the LLM. As mentioned above, the data-mining phase is successful if the regression model achieves strong predictive performance (target $R^2 \geq 0.75$), the clustering produces stable and interpretable student groups with clear separation (validated through silhouette and fuzzy-membership scores), and the extracted patterns are consistent enough to be used for generating reliable LLM-based recommendations for students and educators.

2. DATA UNDERSTANDING

After defining our goals we needed to ensure that our chosen dataset contained the necessary information and was compatible with our planned analysis tools in Python. Since the overall goal of the project is to explore the factors that influence student success and develop predictive insights, this stage focuses on verifying that the chosen dataset aligns with the project's objectives.

2.1.1. Outline data requirements

Based on the project goals, we needed data that describes different factors related to student success. The required data included demographic information (such as age and gender), academic performance (grades or exam results), and behavioral factors (study habits, attendance, motivation, etc.).

In addition, the data must cover enough instances to allow meaningful statistical and machine learning analysis. The dataset includes 6607 records which is enough for this project.

2.1.2. Verify data availability

The dataset we used for this project is a single CSV file downloaded from Kaggle. We kept all the columns and all the student records because every variable might be useful later when exploring which factors influence student success. Using the whole dataset ensures that we do not accidentally remove any information that could help with analysis or building models.

2.1.3. Define selection criteria

Since the dataset already fits the project goals, we decided to use the entire dataset exactly as provided on Kaggle. The data comes in a single CSV file, which makes selection straightforward and avoids complications with multiple tables or databases. The dataset represents a cross-sectional observational sample, meaning it captures many students at one point in time without tracking them over a period. The dataset does not include variables such as age, country, or school grade level, so the exact population source is not specified. Based on the structure, it appears to be a synthetic or anonymized educational dataset rather than data collected from a specific real-world school system.

2.2. Gathering data

The dataset contains 6 607 student records, each representing one individual. Each row represents one student, and each column represents a variable, such as demographics, study behavior, exam performance, or personal factors. It includes a mix of numerical and categorical variables. Overall, the structure is clean and easy to understand, and the dataset includes all the types of variables needed for examining student success.

2.2.1. Categorical Variables

Several variables have clear, limited categories:

- Parental Involvement: High (1908), Medium (3362), Low (1337)
- Access to Resources: High (1975), Medium (3319), Low (1313)
- Extracurricular Activities: Yes (3938), No (2669)
- Motivation Level: High (1319), Medium (3351), Low (1937)
- Internet Access: Yes (6108), No (499)
- Family Income: High (1269), Medium (2666), Low (2672)
- Teacher Quality: High (1947), Medium (3925), Low (657)
- School Type: Public (4598), Private (2009)
- Peer Influence: Positive (2638), Neutral (2592), Negative (1377)
- Learning Disabilities: Yes (695), No (5912)
- Parental Education Level: High School (3223), College (1989), Postgraduate (1305)
- Distance from Home: Near (3884), Moderate (1998), Far (658)
- Gender: Male (3814), Female (2793)

2.2.2. Numerical Variables

Key numerical features (mean ± std):

- Hours_Studied: 19.98 ± 5.99 (range 1–44)
- Attendance: 79.98 ± 11.55 (range 60–100)
- Sleep_Hours: 7.03 ± 1.47 (range 4–10)
- Previous_Scores: 75.07 ± 14.40 (range 50–100)
- Tutoring_Sessions: 1.49 ± 1.23 (range 0–8)
- Physical_Activity: 2.97 ± 1.03 (range 0–6)

- Exam_Score: 67.24 ± 3.89 (range 55–101)

The values fall within realistic ranges, except for the Exam_Score max of 101, which likely comes from extra credit or rounding but can be treated as a minor outlier. These distributions look reasonable and balanced enough for analysis.

2.3. Exploring data

During exploration, we looked at basic summaries of the variables. For numerical fields, we checked ranges and averages. For categorical fields, we checked the frequency of each category. This helped us understand the overall shape of the data, spot unusual values, and form early ideas about which variables might influence student success.

2.4. Verifying data quality

To assess data completeness, we calculated the number of missing values for each variable. Most of the dataset is fully complete, with no missing values in the majority of the fields. Only three variables contain missing entries:

- Teacher_Quality has 78 missing values
- Parental_Education_Level has 90 missing values
- Distance_from_Home has 67 missing values

These missing values are moderate in size and can be handled during data preparation, for example by using imputation (mean, median, or most frequent value) or by dropping rows if necessary. No column shows extreme or unusable levels of missing data, and there are no corrupted or unreadable values. Overall, the dataset is clean and suitable for further analysis.

3. LIST OF TASKS

Task 1: Data quality assessment

We`re going to evaluate the Kaggle dataset (6,590 students) for quality issues, missing values, and outliers. Additionally, we`ll clean the data, handle missing data, normalizing features like study hours and attendance percentage. Create appropriate train/test splits and validate data distributions. Tools: Python (pandas, numpy), Jupyter notebooks for exploratory data analysis.

Task 2: Regression Model Development (35 hours per member)
We`re going to build and test regression models to predict final exam scores using available features (study hours, attendance, parental involvement, sleep hours, etc.). Compare different algorithms (linear regression, random forest, gradient boosting) and select the best performer. Tools: Python (scikit-learn, XGBoost), cross-validation for model selection.

Task 3: Clustering Implementation (30 hours per member)
We`re going to develop a clustering model to group students into risk categories based on academic and behavioral patterns. Test k-means, hierarchical clustering, and fuzzy c-means to find optimal groupings. Validate clusters using silhouette analysis and visualization. Tools: Python (scikit-learn, fuzzy-c-means library), matplotlib/seaborn for visualization.

Task 4: Language Model Training (25 hours per member)
We`re going to fine-tune a small language model to generate personalized feedback messages based on student profiles and predicted outcomes. A training dataset will be created with example messages for different risk levels. Tools: Hugging Face Transformers, GPT-2 or similar small model.

Task 5: Model Evaluation and Integration (20 hours per member)
We`re going to test all models using cross-validation, calculate performance metrics (R^2, silhouette scores), and create visualizations of results. Tools: Streamlit for demo interface, Python for integration scripts.

Task 6: Documentation and Reporting (15 hours per member)
We`re going to write technical documentation, create a final report with findings and policy recommendations, and prepare presentation materials. Tools: LaTeX/Word for reports, PowerPoint for presentations.

Teachers/mentors who will help evaluate the usefulness of the findings:

- Lectures/Practice/Homework documentation will help achieve project goals
- CRISP-DM documentation, which forms the basis for the entire project