UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

Diploma Thesis

Predicting Secondary School Student Performance Using Machine Learning

()

Ime in priimek: Lucas Lorenzo Jakin
Študijski program: Računalništvo in informatika EN
Mentor: izr. prof. dr. Branko Kavšek

Koper, 2024

Ključna dokumentacijska informacija

Ime in PRIIMEK: Lucas Lorenzo JAKIN

Naslov zaključne naloge: Predicting Student Performance Using Machine Learning
Kraj: Koper

Leto: 2024

Število listov: Število slik: Število tabel:

Število referenc:

Mentor: izr. prof. dr. Branko Kavšek

Ključne besede:

Izvleček:

Key words documentation

Name and SURNAME: Lucas Lorenzo JAKIN

Title of final project paper: **Predicting Student Performance Using Machine Learning**

Place: Koper

Year: 2024

Number of pages: Number of figures: Number of tables:

Number of references:

Mentor: Assoc. Prof. Branko Kavšek, PhD

Keywords:

Abstract: The application of machine learning in educational settings has gathered significant attention, particularly in model predictions allowing businesses to make accurate guesses to the likely outcomes of a question based on historical data. My diploma thesis aims to analyze a dataset from Kaggle related to secondary school students in Portugal. By following the CRISP-DM methodology (Cross Industry Standard Process for Data Mining), I will systematically explore each phase, from data understanding and preparation to modeling and evaluation. In the modeling phase I will implement various ML algorithms, using different parameters every time, to predict students' academic success. The performance of these models will be then compared and tested to determine the most effective approach.

Zahvala

Kazalo vsebine

Kazalo tabel

Kazalo slik

Seznam kratic

...

1 INTRODUCTION

Machine Learning (ML) has become a revolutionary technology in recent years, transforming industries and changing the way data-based decisions are conducted. Machine Learning is primarly about creating algorithms that help computers learn from data and make predictions. This ability has enabled companies to discover insights, improve processes, and enhance decision-making in different areas, such as finance, healthcare, marketing, and manufacturing.

The CRISPM-DM (Cross-Industry Standard Process for Data Mining) methodology is crucial for the successful implementation of machine learning models. The structured framework provided by CRISP-DM helps guide data scientists and analysts in the complex process of developing data mining and machine learning models. This approach guarantees that projects are carried out in a structured manner, resulting in dependable and understandable results. Published in 1999 to standardize data mining processes across industries, it has since become the **most common methodology** for data mining, analytics, and data science projects.

The CRISP-DM methodology outlines six critical phases in a data mining project:

- **Business Understanding:** This initial phase focuses on understanding the objectives and requirements of the project. Establishing a strong business understanding is absolutely essential. Most of the tasks performed in this phase are foundational project management activities that are universal to most projects:
    1. **Determine business objectives:** It is crucial to first understand, from a business perspective, what the customer really wants to accomplish.
    2. **Asses situation:** Determine resources availability, project requirements, assess risks and contigencies, and conduct a cost-benefit analysis
    3. **Determine data mining goals:** Moreover, in addition to defining bussiness objectives, it is also important to define what success looks like from a technical data mining perspective.
    4. **Produce project plan:** Select technologies and tools and define detailed plans for each project phase.

- **Data Understanding:** In this phase, data is collected and explored to gain insights into its properties. It drives the focus to identify, collect, and analyze the data sets that can help you accomplish the project goals. This phase has four tasks:

  1. **Collect initial data:** Acquire the necessary data and load it into the analysis tool.
  2. **Describe data:** Examine the data and document its surface properties like data format, number of fields or field identities.
  3. **Explore data:** Query the data, visualize it, and identify relationships among the data.
  4. **Verify data quality:** Document any quality issues. How clean is the data?

- **Data Preparation:** Almost 80% of the project is data preparation.
  This phase, which is often referred to as "data munging", prepares the final data set for modeling. It has five tasks:

  1. **Select data:** Determine which data sets will be used and document reasons for inclusion/exlusion.
  2. **Clean data:** A common practice during this task is to correct, impute, or remove erreneous values. Without it, the quality of the data be innacurate.
  3. **Construct data:** Derive new attributes that will be helpful.
  4. **Integrate data:** Create new data sets by combining data from multiple sources.
  5. **Format data:** Re-format data as necessary, for example converting string values that store numbers to numeric for better mathematical operations.

- **Modeling** Even if this might be the most exciting work is also often the shortest phase of the project. Here various models will likely be built and assessed based on several different modeling techniques. This phase has four tasks:

  1. **Select modeling techniques:** Determine which algoritms to try (e.g. regression, neural network, KNN).
  2. **Generate test design:** Depending on the approach you took, there might be the need to split the data into training set, test, and validation sets.
  3. **Build model:** The selected modeling techniques are put into action to create predictive models. Beside running lines of code, this process requires careful consideration of various factors to ensure the models are robust and effective.

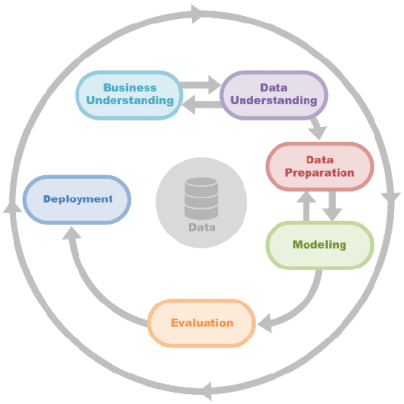
4. **Assess model:** This part is critical for determining the effectiveness and reliability of predictive models built in the previous step. It involves a detailed evaluation of the model's performance using various metrics and techniques to ensure it meets the predefined success criteria ans performs well on unseen data.

- **Evaluation** Whereas the Assess Model task of the modeling phase focuses on technical model assessment, the Evaluation phase looks more broadly at which model best meets the business and what to do next. This phase has three tasks:

  1. **Evaluate results:** Checking if the models meet the business criteria.
  2. **Review process:** Review the work accomplished. Summarize findings and correct anything if needed.
  3. **Determine next steps:** Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects.

- **Deployment** The Deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

  1. **Plan deployment:** Develop and document a plan for deploying the model.
  2. **Plan monitoring and maintenance:** Develop a thorough monitoring and maintenance plan to avoid issues during the operational phase of a model.
  3. **Produce final report:** Document a summary of the project which might include a final presentation of data mining results.
  4. **Review project:** Conduct a project retrospective about what went well, what could have been better, and how to improve in the future.



Slika 1: CRISPM-DM methodology

For this diploma thesis, a dataset will be selected to demonstrate the application of the CRISPM-DM methodology. The focus of the thesis will be to systematically apply each phase of CRISPM-DM to the chosen dataset. This will involve selecting an appropriate dataset, understanding its structure and contents, preparing it for analysis, and implementing various machine learning models to predict target values. The models will be then evaluated to determine their accuracy and effectiveness. This final project aims to provide a thorough and structured approach to data mining and predictive modeling, showcasing the practical application of these techniques on real-world data.

The dataset that I have chosen is about student performance and it is available on both Kaggle and UCI Machine Learning Repository. The dataset approaches student achievement in secondary education in two Portuguese schools.

2 Problem description

Education is a key factor for achieving long-term economic progress. During the last two decades, the portuguese educational level has improved. However, the statistics keep Portugal at Europe's tail end due to its high student failure and dropping rates. In particular, failure in the core classes of Mathematics and Portuguese (native language) is extremely serious, since they provide fundamental knowledge for the success in other remaining subjects.
The chosen data set approach student achievements in secondary education of two Portuguese schools. In Portugal, the secondary education consists of 3 years of schooling, preceding 9 years of basic education and followed by higher education. There are several courses that share core subjects as the Portuguese language and Mathematics. It includes a wide range of attributes that capture various aspects of student life and academic achievement, consisting of student grades, demographic information, social factors, and school-related features. The data was collected through school reports and questionnaires, ensuring a detailed and multifaceted view of the factors influencing student performance.

Objectives: The main objective of this thesis is to build and evaluate predictive models that can accurately forecast the final grades of students in Mathematics. The problem will be approached from two angles:

- **Classification:** Categorizing students' final grades into predefined levels or classes. In this case including multi-class classification, for example grading on a scale from A to F.

- **Regression:** Predicting the exact numerical value of a student's final grade based on the input features.

Classification and regression are two important DM goals. Both require a supervised learning, where a model is adjusted to a dataset made up of K examples, each mapping an input vector $(h_1^k, ..., x_I^k)$ to a given target y_k. The main difference is set in terms of the output representation, discrete for classification and continuos for regression.
The evaluation phase has the importance and the objective to assess the performance and reliability of the predictive models. The goal is to determine the models's accuracy

precision and robustness by employing a range of performance metrics such accuracy, Mean Squared Error (MSE), F1 Score, etc. Additionally, the evaluation will employ validation techniques such as 10-fold Cross-Validation and random TRAIN-TEST splits to ensure that the models generalize well to unseen data and are not overfitting.

Predicting student performance is of great importance to educators and students themselves. Accurate predictions can help in identifying students at risk of underperforming, allowing for timely interventions and support. By sistematically performing a thorough analysis of student performance data and develop reliable models we can achieve the above mentioned objectives.

The dataset is complemented by an article written by Paolo Cortez and Alice Silva [2], which provides an in-depth explanation of the problem and describes how data scientists approached the analysis and modeling of the data. Their work serves as a benchmark for my final project.

3 Data understanding

As mentioned before the data can be found in both Kaggle and UCI ML repositories. The UCI ML Repository provides two separate datasets: one focusing on student performance in Portuguese language and another on Mathematics. In contrast the Kaggle version of the dataset contains only the data for Mathematics. Despite the difference, the datasets are fundamentally the same in terms of structure and information. For the purpose of the thesis, the Mathematics dataset has been selected. This choice allows for a focused analysis on a single subject.

3.1 Description of the dataset

The initial version of the dataframe contained scarce information about the students, containing only the grades and the number of absences, it was then complemented with several other information such as demographic (e.g. mother's education, family income), social (e.g. alcohol consumption) and school related attributes (e.g. number of past class failures). The final dataframe consists of 33 columns (attributes) and has 395 rows (instances). Some features were discarded due to the lack of discriminative value, for instance few people answered about their family income probably due to privacy. The dataset contains three very similar attributes, which are G1, G2 and G3 that represent the grade in the first period, second period and final grade, respectively. The target attribute is G3 and is the one that I am trying to predict. G3 (also G1 and G2) contains values that go from 0 to 20, since Portugal has a 20-point grading scale, where 0 is the lowest grade and 20 is the perfect score. The target attribute G3 has a string correlation with attributes G2 and G1. This occurs because G3 is the final grade and it's value depend on the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful.

3.2 Attributes

The dataset looks like the following:

```
head(dataset)
```

```
  school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason
1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course
2     GP   F  17       U     GT3       T    1    1  at_home    other     course
3     GP   F  15       U     LE3       T    1    1  at_home    other      other
4     GP   F  15       U     GT3       T    4    2   health services       home
5     GP   F  16       U     GT3       T    3    3    other    other       home
6     GP   M  16       U     LE3       T    4    3 services    other reputation
  guardian traveltime studytime failures schoolsup famsup paid activities
1   mother          2         2        0       yes     no   no         no
2   father          1         2        0        no    yes   no         no
3   mother          1         2        3       yes     no  yes         no
4   mother          1         3        0        no    yes  yes        yes
5   father          1         2        0        no    yes  yes         no
6   mother          1         2        0        no    yes  yes        yes
  nursery higher internet romantic famrel freetime goout Dalc Walc health
1     yes    yes       no       no      4        3     4    1    1      3
2      no    yes      yes       no      5        3     3    1    1      3
3     yes    yes      yes       no      4        3     2    2    3      3
4     yes    yes      yes      yes      3        2     2    1    1      5
5     yes    yes       no       no      4        3     2    1    2      5
6     yes    yes      yes       no      5        4     2    1    2      5
  absences G1 G2 G3
1        6  5  6  6
2        4  5  5  6
3       10  7  8 10
4        2 15 14 15
5        4  6 10 10
6       10 15 15 15
```

Slika 2: Head of the dataset

There are 33 different attributes, which is a lot so let's get to know them in the following table:

Attribute	Description
school	Student's school - BINARY(GP or MS)
sex	Student's sex - BINARY(F - M)
age	Student's age - NUMERIC(15-22)
address	Home address - BINARY(Urban or Rural)
famsize	Family size - BINARY (LE3 or GT3)
Pstatus	Parent's cohabitation status - BINARY(Together or Apart)
Medu	Mother's education - NUMERIC(0-4)
Fedu	Father's education - NUMERIC(0-4)
Mjob	Mother's job - NOMINAL
Fjob	Father's job - NOMINAL
reason	Reason to choose this school - NOMINAL
guardian	Student's guardian - NOMINAL(Mother or Father or Other)
traveltime	Home to school travel time - NUMERIC(1-4)
studytime	Weekly study time - NUMERIC(1-4)
failures	Number of past class failures - NUMERIC(n if 1 GE n LT 3, else 4)
schoolsup	Extra educational support - BINARY(y-n)
famsup	Family educational support - BINARY(y-n)
paid	Extra paid classes within the course subject - BINARY(y-n)
activities	Extra-curricular activities - BINARY(y-n)
nursery	Attended nursery school - BINARY(y-n)
higher	Wants to take higher education - BINARY(y-n)
internet	Internet access at home - BINARY(y-n)
romantic	With a romantic relationship - BINARY(y-n)
famrel	Quality of family relationships - NUMERIC(1-5)
freetime	Free time after school - NUMERIC(1-5)
goout	Going out with friends - NUMERIC(1-5)
Dalc	Workday alcohol consumption - NUMERIC(1-5)
Walc	Weekend alcohol consumption - NUMERIC(1-5)
health	Current health status - NUMERIC(1-5)
absences	number of school absences - NUMERIC(0-93)
G1	First period grade - NUMERIC(0-20)
G2	Second period grade - NUMERIC(0-20)
G3	Final grade - NUMERIC(0-20)

The attributes are divided into two distinct data types, which are **Numerical** and **Nominal**. This can be seen with some further analysis by using the skim() function, a method in the R programming language, that provides a neat summary of the dataset:

```
skim(dataset)
```

Data summary

Name	dataset
Number of rows	395
Number of columns	33
_____	
Column type frequency:	
character	17
numeric	16
_____	
Group variables	None

Slika 3: Data summary

3.2.1 Categorical Att.

3 [1-10]
The dataset contains 17 categorical attributes. Most of them are binary, meaning they consist of only two possible values. The remaining categorical attributes have several values, indicating a broader range of possible categories for those features. These multi-valued attributes might include variables such as mother's job, father's job, or guardian that can occur in different forms:

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
school	0	1	2	2	0	2	0
sex	0	1	1	1	0	2	0
address	0	1	1	1	0	2	0
famsize	0	1	3	3	0	2	0
Pstatus	0	1	1	1	0	2	0
Mjob	0	1	5	8	0	5	0
Fjob	0	1	5	8	0	5	0
reason	0	1	4	10	0	4	0
guardian	0	1	5	6	0	3	0
schoolsup	0	1	2	3	0	2	0
famsup	0	1	2	3	0	2	0
paid	0	1	2	3	0	2	0
activities	0	1	2	3	0	2	0
nursery	0	1	2	3	0	2	0
higher	0	1	2	3	0	2	0
internet	0	1	2	3	0	2	0
romantic	0	1	2	3	0	2	0

Slika 4: Categorical attributes

3.2.2 Numerical Att.

numericalas ssdfv

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
school	0	1	2	2	0	2	0
sex	0	1	1	1	0	2	0
address	0	1	1	1	0	2	0
famsize	0	1	3	3	0	2	0
Pstatus	0	1	1	1	0	2	0
Mjob	0	1	5	8	0	5	0
Fjob	0	1	5	8	0	5	0
reason	0	1	4	10	0	4	0
guardian	0	1	5	6	0	3	0
schoolsup	0	1	2	3	0	2	0
famsup	0	1	2	3	0	2	0
paid	0	1	2	3	0	2	0
activities	0	1	2	3	0	2	0
nursery	0	1	2	3	0	2	0
higher	0	1	2	3	0	2	0
internet	0	1	2	3	0	2	0
romantic	0	1	2	3	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	16.70	1.28	15	16	17	18	22	▆▇▂▁▁
Medu	0	1	2.75	1.09	0	2	3	4	4	▂▃▂▇▇
Fedu	0	1	2.52	1.09	0	2	2	3	4	▂▃▇▆▇
traveltime	0	1	1.45	0.70	1	1	1	2	4	▇▂▁▁▁
studytime	0	1	2.04	0.84	1	1	2	2	4	▂▇▃▁▁
failures	0	1	0.33	0.74	0	0	0	0	3	▇▁▁▁▁
famrel	0	1	3.94	0.90	1	4	4	5	5	▁▁▃▇▇
freetime	0	1	3.24	1.00	1	3	3	4	5	▁▃▇▅▂
goout	0	1	3.11	1.11	1	2	3	4	5	▂▅▇▅▅
Dalc	0	1	1.48	0.89	1	1	1	2	5	▇▂▁▁▁
Walc	0	1	2.29	1.29	1	1	2	3	5	▇▅▃▂▁
health	0	1	3.55	1.39	1	3	4	5	5	▂▃▂▅▇
absences	0	1	5.71	8.00	0	0	4	8	75	▇▁▁▁▁
G1	0	1	10.91	3.32	3	8	11	13	19	▁▆▇▃▁
G2	0	1	10.71	3.76	0	9	11	13	19	▁▃▇▃▁
G3	0	1	10.42	4.58	0	8	11	14	20	▂▅▇▅▂

Slika 5: Image A.

```
skim(dataset)
```

Data summary

Name	dataset
Number of rows	395
Number of columns	33
Column type frequency:	
character	17
numeric	16
Group variables	None

Slika 6: Image B.

Slika 7: Image C.

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	16.70	1.28	15	16	17	18	22	
Medu	0	1	2.75	1.09	0	2	3	4	4	
Fedu	0	1	2.52	1.09	0	2	2	3	4	
traveltime	0	1	1.45	0.70	1	1	1	2	4	
studytime	0	1	2.04	0.84	1	1	2	2	4	
failures	0	1	0.33	0.74	0	0	0	0	3	
famrel	0	1	3.94	0.90	1	4	4	5	5	
freetime	0	1	3.24	1.00	1	3	3	4	5	
goout	0	1	3.11	1.11	1	2	3	4	5	
Dalc	0	1	1.48	0.89	1	1	1	2	5	
Walc	0	1	2.29	1.29	1	1	2	3	5	
health	0	1	3.55	1.39	1	3	4	5	5	
absences	0	1	5.71	8.00	0	0	4	8	75	
G1	0	1	10.91	3.32	3	8	11	13	19	
G2	0	1	10.71	3.76	0	9	11	13	19	
G3	0	1	10.42	4.58	0	8	11	14	20	

Slika 8: Numerical attributes

4 Data preparation

NO MISSING VALUES –¿ WRITE SOMETHING ABOUT ITTT

MNIST (Modified NIST) je nabor podatkov, ki vsebuje 70000 slik ročno napisanih števk. Podatke je od srednješolcev in zaposlenih na Ameriškem uradu za statisko sprva zbral NIST (National institute of standards and technology). Podatki so ločeni na učno množico in testno množico. Učna množica vsebuje 60000 slik, testna pa 10000. Pisci števk, katerih podatki so v učni množici, nimajo slik v testni množici. Vsaka slika je označena s primernim številom med 0 in 9.

Slike so velikosti 28x28, vsak piksel zavzame vrednost med 0 in 255. Vse slike so bile normalizirane glede na velikost in centrirane glede na težišče. [1]

5 Methodology

5.1 K najbližjih sosedov

kNN je primer neparametričnega, nadzorovanega strojnega učenja. Dano podatkovno točko x neznanega razreda klasificiramo tako, da najdemo k točk v naboru podatkov za učenje, ki so tej točki najbližje, nato dodelimo x tisti razred, ki je večinski v množici sosednjih točk.

(Kot neparametričen algoritem) kNN ne zgradi modela iz nabora podatkov za učenje (training data, učilni podatki? nevem), vendar shrani celoten učilni nabor, ki ga uporabi za klasifikacijo. Posledično je učenje hitrejše, sklepanje pa počasnejše v primerjavi z ostalimi algoritmi. [citat?]

Učinkovitost algoritma je močno odvisna od izbire k. Optimalna izbira je drugačna za vsak nabor podatkov. V splošnem velja, da vodijo premajhne vrednosti do prekomernega prileganja (overfit) in prevelike vrednosti do nezadostnega (?) prileganja (underfit).[citat]

Za uspešnost algoritma kNN je pomembna tudi izbira razdalje/metrike (), še posebej za podatke z visokim številom atributov. [citat] (je/ni pomembno za naš problem? curse of dimensionality?)

KNN se lahko izboljša z tako, da točke, ki so bližje x bolj upoštevamo pri klasifikaciji, torej utežimo bližnje točke glede na razdaljo. (weighted knn)

Konkretna implementacija(?)... [slika]

5.2 Metoda podpornih vektorjev

(M.Bishop, sci-kit.org) - konkretna implementacija (scikit)? - kako se izogne prekomernem prileganju, regularizacija

Metoda podpornih vektorjev (support vector machine - SVM) je ena od metod strojnega učenja. Osnovni algoritem lahko loči dva razreda, če so podatki linearno ločljivi. V fazi učenja algoritem najde hiperravnino, ki maksimizira razdaljo med razredoma. Pri določanju ravnine so pomembni le določeni vektorji (podporni vektorji), ki so najbližje ravnini (?).

Metoda se lahko prilagodi tudi za večrazredno klasifikacijo. Za to obstajata dva

načina, *ena na ena* in *eden proti vsem.*

Pri *ena na ena ...*

Pri *eden proti vsem ...*

Osnovni algoritem se odreže dobro na linearno ločljivih podatkih, vendar ne zmore ločiti nelinearnih podatkov, tudi ko obstaja hiperploskev, ki jih loči. Ta problem rešujemo tako, da podatke najprej preslikamo v višjedimenzijski prostor, kjer postanejo linearno ločljivi. Preslikava v višjo dimenzijo poveča računsko zahtevnost problema. kernel trick...

Iskanje hiperravnine je konveksen optimizacijski problem, najdena rešitev pa je globalni minimum. V tem pogledu se metoda podpornih vektorjev razlikuje od umetnih nevronskih mrež.

[slika]

5.3 Umetna Nevronska mreža

5.4 Konvolucijska nevronska mreža

6 Modeling / Classification

7 Results

8 ...

8.1 ...

9 CONCLUSION / FINAL THOUGHTS

10 LITERATURA IN VIRI

[1] MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges. *(Citirano na strani 14.)*

[2] Cortez, P., and Silva, A. Using data mining to predict secondary school student performance. *Dep. Information Systems/Algoritmi RD Centre 1* (2008), 1–8. *(Citirano na strani 6.)*