Project 1: Feature extraction and visualization

02450 Introduction to Machine Learning and Data Mining

Authors

Niels Torp Grønskov, **s204510**
Yiming Zhang, **s232896**
Frederik Danielsen, **s214718**

Contributions

	Intro + Section 1	Section 2	Section 3	Section 4	Exam questions
s204510	25%	20%	45%	40%	33.3%
s232896	45%	50%	10%	30%	33.3%
s214718	30%	30%	45%	30%	33.3%

October 3, 2023

Introduction

Higher education is quite vital to employment, economic growth, social justice, and many other areas. Thus, the problem of dropout has become a matter of concern for most higher education institutions. There is no generally well-accepted definition of dropout. In reference paper, Realinho et al., 2022 defines dropout from a micro perspective, where changes in fields and institutions are treated as dropout independent of when these changes occur. Compared to macro-perspective, this definition leads to much higher dropout rates. This is because the macro-perspective only takes into account students who leave the higher education institutions without a degree.

Additionally, early prediction of student dropout and completion rates has attracted more researchers in recent years. However, although the Universities has generated a great deal of research data, we still need to collect more and better administrative data, including dropout and transfer reasons.

In this project, we will use the same data set as Realinho et al., 2022, created by a higher education institution. This data set is sourced from students with undergraduate degrees in a variety of fields, such as agronomy, design, education, nursing, journalism and technology. It contains 4424 records with 36 attributes, each of which represents information about one student and can be used for data analysis as well as training in the field of machine learning. This data set also includes information about the students at the time of enrolment as well as the students' academic achievements at the end of their first and second semesters. The problem is expressed in terms of three types of categorical tasks (dropout, enrolled and graduate) at the end of the normal duration of the program. In addition, the data set

The rest of this paper is organized as follows. Section 1 provides the details of the data set. Section 2 presents the types of attributes and basic summary statistics . Section 3 shows the application of PCA and data visualization with a brief data analysis. Section 4 presents our discussion about the data.

1 Data set description

The data set contains information about 4424 students (incidents) collected up until the end of their second semester as well as a target column: whether or not the students dropped out, graduated from their educations, or remained enrolled past the normal duration of the course (Realinho et al., 2022).

As described by Realinho et al., 2022, the data was collected in an effort to reduce academic dropout, by allowing for the development of models that are able to detect a potential dropout at an early stage. This would allow schools to initiate measures, in a timely manner, that could increase the success rate of students.

The data set consists of 36 features that describe each of the 4424 students. These features include, among other things, information about a student's academic path, demographic, and social factors collected at the time of enrollment, as well as academic performance metrics collected at the end of the first and second semesters. The features are divided into six categories as seen in Table 2.

The data set was obtained from the UC Irvine Machine Learning Repository and was created by Valentim Realinho, Mónica Vieira Martins, Jorge Machado, and Luís Baptista. They describe the data in their paper "Predicting Student Dropout and Academic Success" from October 2022 (Realinho et al., 2022).

1.1 Original source paper results

In their paper, Realinho et al., 2022 find, among other things, the feature describing "the number of curricular units approved in the 2nd semester" to be the most important for predicting the target. The features describing "the number of curricular units approved in the 1st semester" and whether or not the tuition fees are up to date make up the second most important features. This was determined by considering the Permutation Feature Importance of the features using four different types of machine learning models (Random forest, XGBoost, LightGBM, and CATBoost), hence the shared *second place.* Realinho et al., 2022 draw attention to the fact, that the target column is not evenly distributed (see Table 1), which might result in models having a higher accuracy when predicting the majority target class and a poor accuracy when predicting the minority class - simply as this would be the case with a random model. If one were to pick a student from the data set at random, picking a student who has graduated is more than twice as probable as choosing a student who is still enrolled. Furthermore, Realinho et al., 2022 find collinearity between some features in the data set. This is mostly seen between features from the same category.

Graduate	**Dropout**	**Enrolled**
50%	32%	18%

Table 1: Frequency table of target categories (Realinho et al., 2022).

1.2 Goals of the project

The goals of this project are twofold. Firstly we seek to train a classification model to predict whether a student will graduate, drop out, or stay enrolled past the normal course duration (the three categories of the target column). Secondly, it could be interesting to train a regression model to predict a feature regarding the student's grade e.g. the GPA. When doing this, however, we should consider which features to exclude from the model. For example, it would defeat the purpose of the model to include the target feature in a model that predicts the GPA since this feature solely is available after a GPA is available. I.e. it does not make sense to predict a GPA when a student already has a GPA, so the chronological aspect of the features should be considered carefully.

The model will be trained on the presented data set, however, at the moment, the target column contains the three different categories as strings. To overcome this, the column will be transformed using the one-out-of-K fold method, generating the three numeric/binary columns:

Graduate	**Dropout**	**Enrolled**

Furthermore, the columns will be standardized to ensure the equal importance of each feature and outlying values will be examined. Fortunately, the data set has already been preprocessed in the sense that there are no missing values. Lastly, we will explore the performance of the models when only including the most important features (as found be Realinho et al., 2022).

2 Attribute description

2.1 Attribute types

The types of attributes are as shown in Table 2 by class.

Category	**Attribute**	**Type**
Demo-graphic data	Marital status	nominal/discrete
	Nationality	nominal/discrete
	Displaced	nominal/binary
	Gender	nominal/binary
	Age at enrollment	ordinal/discrete
	International	nominal/binary
Socio-economic data	Mother's qualification	nominal/discrete
	Father's qualification	nominal/discrete
	Mother's occupation	nominal/discrete
	Father's occupation	nominal/discrete
	Educational special needs	nominal/binary
	Debtor	nominal/binary
	Tuition fees up to date	nominal/binary
	Scholarship holder	nominal/binary
Macro-economic data	Unemployment rate	ratio/continuous
	Inflation rate	ratio/continuous
	GDP	interval/continuous
Academic data at enrollment	Application mode	nominal/discrete
	Application order	ordinal/ordinal
	Course	numeric/discrete
	Daytime/evening attendance	nominal/binary
	Previous qualification	nominal/discrete
	Previous qualification (grade)	ordinal/continuous
	Admission grade	ordinal/continuous

Academic data at the end of 1st semester	Curricular units 1st sem (credited)	ratio/discrete
	Curricular units 1st sem (enrolled)	ratio/discrete
	Curricular units 1st sem (evaluations)	ratio/discrete
	Curricular units 1st sem (approved)	ratio/discrete
	Curricular units 1st sem (grade)	ordinal/continuous
	Curricular units 1st sem (without evaluations)	ratio/discrete
Academic data at the end of 2nd semester	Curricular units 2nd sem (credited)	ratio/discrete
	Curricular units 2nd sem (enrolled)	ratio/discrete
	Curricular units 2nd sem (evaluations)	ratio/discrete
	Curricular units 2nd sem (approved units 2nd sem)	ratio/discrete
	Curricular units 2nd sem (grade)	ordinal/continuous
	Curricular units 2nd sem (without evaluations)	ratio/discrete

Table 2: Overview of features and their types.

2.2 Basic summary statistics

Since the data set has been pre-processed, there are no missing values or corrupted data. However, from our observations, the data set for some of the features still need to be processed (the processing method is based on the appendix of Realinho et al., 2022) The basic summary statistics for all attributes are shown in the tables below. The tables include the mean, median, dispersion, and minimum and maximum values for each attribute value.

Attribute	**Mean**	**Median**	**Dispersion**	**Min.**	**Max.**
Marital status	1.179	1	0.606	1	6
Nationality	1.255	1	1.748	1	21
Displaced	0.548	1	0.498	0	1
Gender	0.352	0	0.478	0	1
Age at enrollment	23.265	20	7.587	17	70
International	0.025	0	0.156	0	1
Father's qualification	16.455	14	11.044	1	34
Mother's qualification	12.322	13	9.025	1	29
Father's occupation	7.819	8	4.856	1	46
Mother's occupation	7.318	6	3.997	1	32
Educational special needs	0.012	0	0.107	0	1
Debtor	0.114	0	0.317	0	1
Educational special needs	0.012	0	0.107	0	1
Tuition fees up to date	0.881	1	0.324	0	1
Scholarship holder	0.248	0	0.432	0	1
Unemployment rate	11.566	11.1	2.664	7.6	16.2
Inflation rate	1.228	1.4	1.383	-0.8	3.7
GDP	0.002	0.32	2.27	-4.06	3.51
Application mode	6.887	8	5.298	1	18
Application order	1.728	1	1.314	0	9
Course	9.899	10	4.331	1	17
Daytime/evening attendance	0.891	1	0.312	0	1
Previous qualification	2.531	1	3.963	1	17
Previous qualification(grade)	43.814	44	17.279	1	101
Admission grade	266.33	259	133.717	1	620

Curricular units 1st sem (credited)	0.71	0	2.36	0	20
Curricular units 1st sem (enrolled)	6.271	6	2.48	0	26
Curricular units 1st sem (evaluations)	8.299	8	4.179	0	45
Curricular units 1st sem (approved)	4.707	5	3.094	0	26
Curricular units 1st sem (grade)	10.641	12.286	4.843	0	18.875
Curricular units 1st sem (without evaluations)	0.138	0	0.691	0	12
Curricular units 2nd sem (credited)	0.542	0	1.918	0	19
Curricular units 2nd sem (enrolled)	6.232	6	2.196	0	23
Curricular units 2nd sem (evaluations)	8.063	8	3.948	0	33
Curricular units 2nd sem (approved)	4.436	5	3.014	0	20
Curricular units 2nd sem (grade)	10.23	12.2	5.21	0	18.571
Curricular units 2nd sem (without evaluations)	0.15	0	0.754	0	12
Target		Graduate	1.02		

Table 3: Basic statistics information about the target attributes

3 Data visualization and PCA

Understanding the structures and relationships within a dataset is essential to building good predictive models. In this section, various plots representing the transformation of our data through PCA will be showcased, illuminating the inherent relationships and structures within the dataset.

3.1 Normal distributed attributes

As can be seen in Figure 1, the distribution of some of the attributes seems to be bell-shaped. This might indicate that those attributes are normal distributed. This includes the attributes in Table 4. The attributes in bold are the most clearly bell-shaped, and all have to do with grades, which is certainly something that is likely to be normal distributed. The ”age at enrollment” attribute seems to follow a skewed distribution like a gamma distribution or possibly an inverse Gaussian distribution.

- Previous qualification (grade)
- Admission grade
- Curricular units 1st sem (enrolled)
- Curricular units 1st sem (evaluations)
- Curricular units 1st sem (approved)
- Curricular units 1st sem (grade)
- Curricular units 2nd sem (enrolled)
- Curricular units 2nd sem (evaluations)
- Curricular units 2nd sem (approved)
- Curricular units 2nd sem (grade)

Table 4: List of bell-shaped attributes.

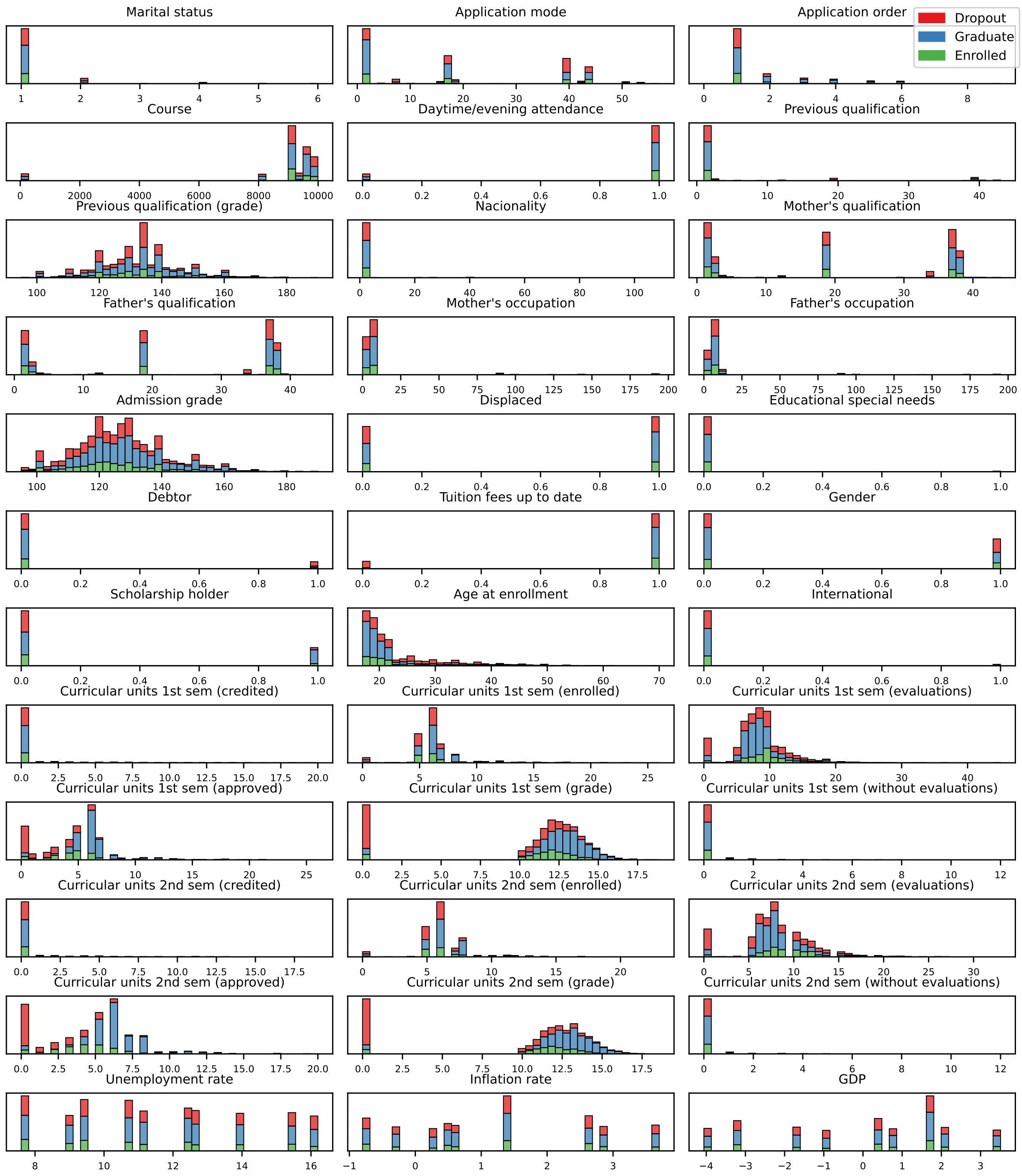


Figure 1: Distribution visualization (histogram) of each feature in the data set. The colors indicate the fraction of each bin taken up by each target category; red: dropout, blue: graduate, green: enrolled.

3.2 Outliers

To detect outliers we consider the *ratio* and *ordinal* attributes (excluding the attribute 'application order' since the distribution of this attribute does not look bell-shaped) and assume these to follow a normal distribution $\mathcal{N}(\mu, \sigma^2)$. This includes the attributes in Table 4. In this regard, we seek to determine the probability that an observation $Z \sim \mathcal{N}(\mu, \sigma^2)$ falls outside an interval $I = [\mu - d, \mu + d]$ where $d \in \mathbb{R}$:

$$P(Z < \mu - d) + P(Z > \mu + d) = 1 - P(\mu - d \leq Z \leq \mu + d).$$

Due to the symmetry of the normal distribution around the mean μ we know that $P(Z \leq \mu + d) = P(Z \geq \mu - d)$. Thus it is the case that

$$\begin{aligned} P(\mu - d \leq Z \leq \mu + d) &= 1 - 2(1 - P(Z \leq \mu + d)) \\ &= 2P(Z \leq \mu + d) - 1 \\ &= 2\mathrm{cdf}_{\mathcal{N}}(\mu + d) - 1. \end{aligned}$$

This means that the probability that $N \in \mathbb{N}$ observations fall within $I = [\mu - d, \mu + d]$ is

$$P(\mu - d \leq Z \leq \mu + d)^N$$

and so the probability that at least one observation falls outside I is

$$1 - P(\mu - d \leq Z \leq \mu + d)^N = 1 - (2\mathrm{cdf}_{\mathcal{N}}(\mu + d) - 1)^N.$$

Now, to determine which values are real outliers (errors) we want the probability of such a value appearing in our data naturally to be low; for example 1%. This would mean that

$$1 - (2\mathrm{cdf}_{\mathcal{N}}(\mu + d) - 1)^N = 0.01.$$

Solving this equation for $d = d^*$ we get

$$d^*(N) = \mathrm{erf}^{-1}\left(\sqrt[N]{1 - 0.01}\right)\sigma\sqrt{2},$$

and the acceptance region

$$I^* = [\mu - d^*, \mu + d^*].$$

Thereby we know that the probability that all naturally occurring (not due to error) values fall within I^* is $1 - 0.01 = 99\%$. Thus we now know that the probability that one or more values that are natural occurrences fall outside this range is less than or equal to 1%. Therefore we decide that values outside this region are outliers. Applying this to the attributes in Table 4 we get the result in Table 5. Figure 2 visualizes the respective acceptance intervals for each of the attributes. Evaluating whether or not these outliers are indeed errors is difficult since they are perfectly possible. However, in light of the low probability (1%) that any value outside the acceptance interval is a natural occurrence, we choose to remove these values from the data set. Note that this is under the assumption that the data of these attributes is normal distributed. Furthermore, note that in some calculations the zeroes of the attribute have not been included since these do not seem to be part of the distributions. An example is the curricular units in 1st and 2nd semester (grade). In other words, these zeroes seem to be an absence of values rather than real observations.

The remaining non-nominal attributes that do not seem to have a bell-shaped distribution are:

- Age at enrollment
- Unemployment rate
- Inflation rate
- GDP

In regard to these, we have chosen to define outliers as points that are more than four standard deviations from the mean (see the results of this in appendix Table 7). The nominal attributes are categorical and therefore bounded in a sense that does not allow for outliers in terms of extreme magnitudes.

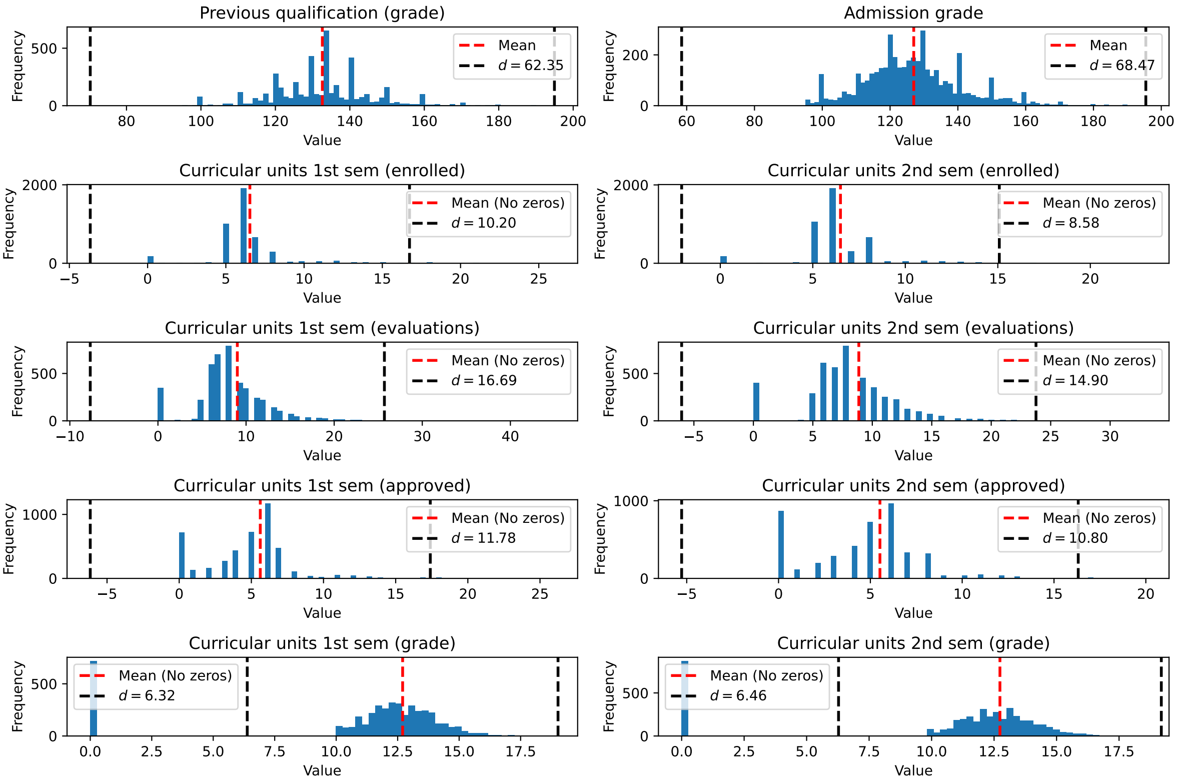


Figure 2: Histograms of attributes assumed to be normally distributed. The red vertical lines indicate the means of the respective attributes. The black vertical lines indicate the acceptance intervals ($[\mu - d, \mu + d]$). The legend indicates whether zero values were included in the calculation of the mean and standard deviation of the data.

Attribute	**Outliers**	**No. outliers**	**No. observations**
Previous qualification (grade)	None	0	4424
Admission grade	None	0	4424
Curricular units 1st sem (enrolled)	18, 21, 17, 18, 17, 19, 18, 18, 18, 18, 18, 18, 17, 17, 17, 17, 17, 17, 21, 18, 18, 18, 17, 18, 17, 23, 21, 17, 17, 19, 21, 17, 18, 23, 18, 18, 17, 18, 21, 18, 18, 17, 26, 17, 18, 21	**46**	4244
Curricular units 2nd sem (enrolled)	19, 17, 17, 17, 17, 17, 16, 17, 23, 17, 19, 17, 17, 23, 17, 18, 21, 18, 17, 17, 19	21	4244
Curricular units 1st sem (evaluations)	45, 45, 26, 29, 29, 26, 36, 32, 27, 31, 26, 28, 33, 27, 26	15	4075
Curricular units 2nd sem (evaluations)	26, 27, 24, 28, 24, 26, 25, 33, 27, 26, 24	11	4023
Curricular units 1st sem (approved)	18, 21, 18, 19, 18, 18, 18, 18, 18, 18, 21, 18, 20, 20, 19, 21, 20, 18, 18, 18, 18, 18, 26, 18, 21	25	3706
Curricular units 2nd sem (approved)	19, 17, 17, 17, 17, 17, 20, 17, 19, 17, 20, 18, 18, 17, 19	15	3554
Curricular units 1st sem (grade)	None	0	3706
Curricular units 2nd sem (grade)	None	0	3554

Table 5: Outliers in various columns. Note that the number of observations is less than 4424 in some columns because zeros were not included in the calculation of these.

3.3 Correlation

To find the correlation, we typically use correlation matrix and its visualization. Figure 3 shows a correlation heatmap of all the attributes. It is clear that the attributes having to do with curricular units comprise a *cluster* of correlation. Furthermore, we see that the features 'International' and 'Nationality' are highly correlated and that the features 'Daytime/evening attendance' and 'Age at enrollment', interestingly, are somewhat negatively correlated. The most correlated attributes are listed in Table 6.

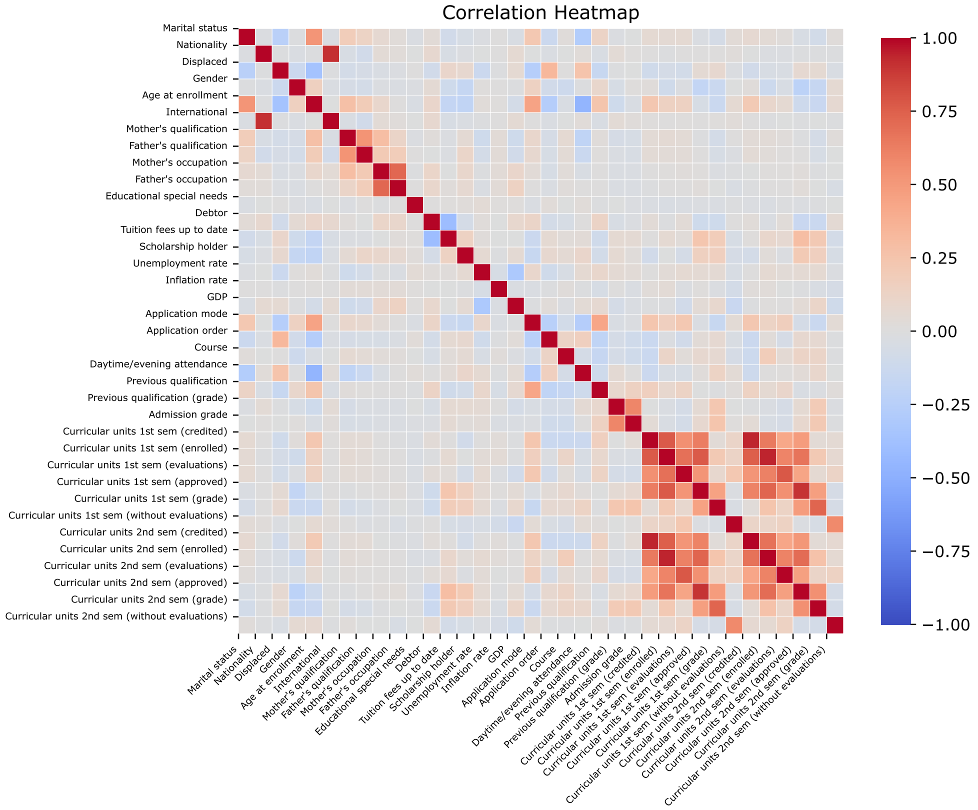


Figure 3: Data correlation heatmap (Pearson correlation). Red corresponds to positive correlation and blue corresponds to negative correlation.

Attribute	**Collinearity with**	**Correlation**
Curricular units 1st sem (credited)	Curricular units 2nd sem (credited)	**0.9448**
	Curricular units 1st sem (enrolled)	0.7743
Curricular units 1st sem (enrolled)	Curricular units 2nd sem (enrolled)	**0.9426**
	Curricular units 1st sem (approved)	0.7691
	Curricular units 2nd sem (credited)	0.7537
Nationality	International	**0.9117**
Curricular units 1st sem (approved)	Curricular units 2nd sem (approved)	**0.9040**
	Curricular units 1st sem (enrolled)	0.7338
Curricular units 1st sem (grade)	Curricular units 2nd sem (grade)	0.8372
Curricular units 1st sem (evaluations)	Curricular units 2nd sem (evaluations)	0.7789
Curricular units 2nd sem (approved)	Curricular units 2nd sem (grade)	0.7608
Mother's occupation	Father's occupation	0.7240
Curricular units 2nd sem (enrolled)	Curricular units 2nd sem (approved)	0.7033

Table 6: Collinearity between features with Pearson correlation coefficient greater than 0.7

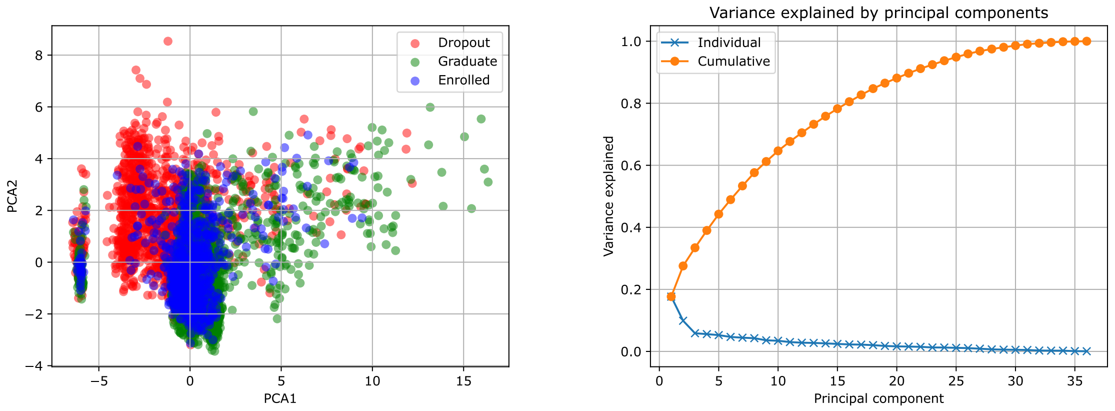
3.4 Modelling feasibility

In Figure 1 we see that the representation of each target category is not equal everywhere. For example, looking at subplot 'Curricular units 1st sem (grade)' in Figure 1, the mean with respect to the category 'enrolled' is clearly different from the mean with respect to 'graduate'. This means that it is somewhat possible to predict the target category given the attribute 'curricular units 1st semester'. The same can be said for other features such as 'Age at enrollment' where dropouts seem to be relatively prominent with older ages and most graduates or enrolled students are around the age of 20. This indicates that it should at least be somewhat feasible to create a working classification model. Similarly, as seen in Figure 3, some features are clearly correlated, which indicates that some features could be employed to predict others with a regression model. Thus accomplishing the goal of creating such models seems possible.

3.5 Principal component analysis

A plot of the first and second principal components is presented in Figure 4a. Each point in the plot corresponds to an observation from the dataset after it has been transformed onto PC1 and PC2. To provide a sense of the data's classification, the target value is plotted in separate colors: Dropout, Graduate, Enrolled. This plot gives the opportunity to better understand the structure and relationship within the dataset when reduced to two dimensions. Observations that are close together in this space are similar in terms of the features that have the highest weights in these two principal components.

When conducting PCA, one can consider the variance explained by the individual principal components or the cumulative amount of variance explained as a function of the number of principal components included. Figure 4b displays these two aspects of the PCA in blue and orange respectively. From the plot, we can see that the first 2 components explain approx 28% of the variance. PC1 explains about 18% and PC2 explains about 10%. As we include more components, the additional variance explained decreases as expected.



(a) Data transformed onto PC1 and PC2

(b) Variance explained by principal components

Figure 5 represents the principal directions of the first two principal components in regards to each feature. The principal directions are shown simply by plotting their coefficients. The distance of each point from the origin reflects its significance in the given principal component.

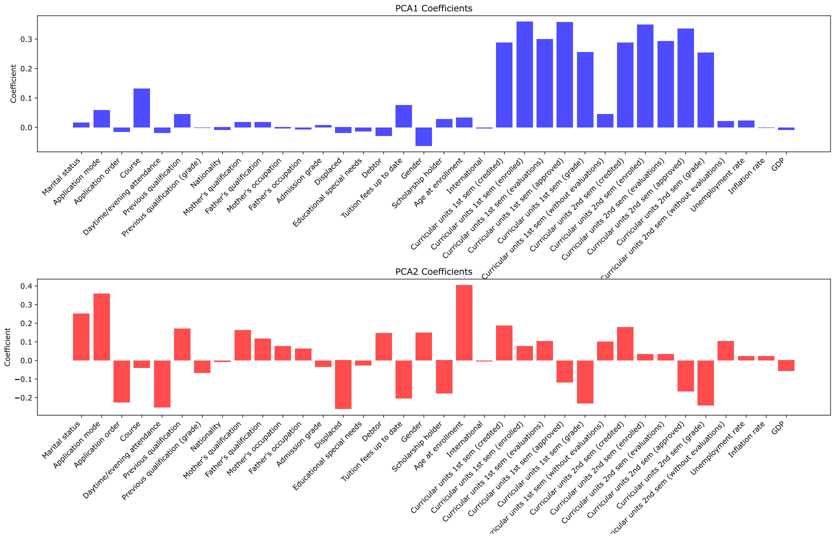


Figure 5: Coefficients of PC1 and PC2

4 Discussion

As mentioned, some of the attributes' distributions seem to be bell-shaped, which we utilize to derive a meaningful acceptance interval. This is, however, only possible under the assumption that these attributes are truly normal distributed. If this is not the case, we cannot trust the probabilities. Having said that, the visualizations of the acceptance intervals are reassuring as they seem not to deviate much from what one could expect them to be.

For the remaining non-nominal attributes it is not as straightforward to determine which values are outliers. In these cases, one could attempt to derive the value d with respect to the distribution at hand (as done for the bell-shaped distributions). This is, however, very elaborate as one might imagine, which is why we have decided, for said attributes, to define outliers as values further than four standard deviations from the mean of the data. It is more difficult to determine whether these outliers are in fact errors, as this kind of *filter* is not as directly linked to the probability of an outlier being an error.

As mentioned earlier, we see some promise in the way the data is distributed in regard to the target categories and the correlation between some features. Thus we deem it worthwhile to train a classification and regression model.

Using a correlation matrix and heatmap, we find many co-linear attributes. It's crucial to address the collinearity as it can adversely affect the analysis and interpretation of a data set. In this project, we applied PCA to convert the original features into a set of uncorrelated features. This will work well to solve the colinearity and prepare for future project.

While being great for future models, the PCA also reveals some underlying structuring of the data. Especially, when projected onto the first to principal components, slight clustering becomes apparent. The explained variance in the data was interesting as the first two principal components explained 28% of the total variance and about 15 principal components were necessary to explain 80% of the variance. Interesting patterns in the coefficients of the principal components also emerge. Here PC1 for instance shows a great relation in the features regarding "1st semester" and "2nd semester" where all of the coefficients are relatively high. The second principal component is shown to have more varied coefficient accross the features.

References

Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success [Number: 11 Publisher: Multidisciplinary Digital Publishing Institute]. *Data*, *7*(11), 146. https://doi.org/10.3390/data7110146

Appendix

Attribute	**Outliers**	**No. Outliers**	**No. Observations**
Age at enrollment	55, 70, 60, 54, 61, 58, 58, 59, 55, 54, 54, 59, 55, 57, 54, 60, 54, 54, 62, 57, 59, 58, 55, 54, 55	**25**	4424
Unemployment rate	None	0	4424
Inflation rate	None	0	4424
GDP	None	0	4424

Table 7: Outliers of non-nominal attributes that do not seem to have a bell-shaped distribution. These are values that are more than four standard deviations from the mean of the data.

5 Exam questions

Question 1. Spring 2019 question 1

We see that all attributes from x_2 to x_7 are ratios. This is because they are all counts of something. These attributes being zero means the absence of what is being measured: in this case, the number of traffic lights (x_6), run over accidents (x_7), etc. Furthermore, we see that x_1, which is the time of day measured in 30-minute intervals, must be of the type *interval*. This is because the attribute is ordered and there is a fixed distance between consecutive data points. Subtraction and addition mean the same for all observations. Lastly, we see that the target y is ordinal because the level of congestion is a kind of ranking from low to high. Thus **the correct answer is D**: x_1 (*Time of day*) : interval, x_6 (*Traffic lights*) : ratio, x_7 (*Running over*) : ratio, y (*Congestion level*) : ordinal.

Question 2. Spring 2019 question 2

First of all, we know that the limit of the p-norm distance between two vectors $x, y \in \mathbb{R}^n$ as $p \to \infty$ is the Chebyshev distance: $\lim_{p\to\infty} \|\mathbf{x} - \mathbf{y}\|_p = \max_{i=1}^n |x_i - y_i|$. This means that given $\mathbf{x}_{14} = [26\ 0\ 2\ 0\ 0\ 0\ 0]^\intercal$ and $\mathbf{x}_{18} = [19\ 0\ 0\ 0\ 0\ 0\ 0]^\intercal$ we have that $d_\infty(\mathbf{x}_{14}, \mathbf{x}_{18}) = |26 - 19| = 7$, which means that the **answer option A is correct**. Furthermore, we see that the remaining answer options are incorrect: $d_3(\mathbf{x}_{14}, \mathbf{x}_{18}) = \sqrt[3]{351} = 7.054 \neq 3.688$, $d_1(\mathbf{x}_{14}, \mathbf{x}_{18}) = 9 \neq 1.286$, $d_4(\mathbf{x}_{14}, \mathbf{x}_{18}) = \sqrt[4]{2417} = 7.012 \neq 4.311$.

Question 4. Spring 2019 question 4

To test the hypothesis of each statement, we project an observation vector onto the PC in question. This is done by calculating the dot product of the observation onto the PC. Here a "high value" in the observation vector is set to 1, a "low value" is set to -1 and the rest is set to 0. The 5 principal directions are provided in each column of the matrix V.

A)
$-2.14 = \text{DotProduct}(\begin{bmatrix}-1 & -1 & 1 & 1 & -1\end{bmatrix}^\intercal, \begin{bmatrix}0.52 & 0.33 & -0.62 & -0.24 & 0.43\end{bmatrix}^\intercal)$
Since the value is -2.14 the projection will not typically have a positive value and this statement is **false**.

B)
$-1.33 = \text{DotProduct}(\begin{bmatrix}0 & 0 & -1 & 1 & 0\end{bmatrix}^\intercal, \begin{bmatrix}0.08 & -0.01 & 0.43 & -0.9 & 0.03\end{bmatrix}^\intercal)$
Since the value is -1.33 the projection will not typically have a positive value and this statement is **false**.

C)
$1.86 = \text{DotProduct}(\begin{bmatrix}-1 & 1 & -1 & 0 & -1\end{bmatrix}^\intercal, \begin{bmatrix}-0.49 & 0.71 & -0.25 & -0.19 & -0.41\end{bmatrix}^\intercal)$
Since the value is 1.86 the projection will not typically have a negative value and this statement is **false**.

D)
$1.76 = \text{DotProduct}(\begin{bmatrix}-1 & 1 & 1 & 0 & 1\end{bmatrix}^\intercal, \begin{bmatrix}-0.5 & 0.23 & 0.23 & 0.09 & 0.8\end{bmatrix}^\intercal)$
Since the value is 1.76 the projection will typically have a positive value and this statement is **true**.

Question 6. Spring 2019 question 27

According to Table 2: Probability of observing particular values of $\hat{x}_2$ and $\hat{x}_7$ conditional on y.
We can know:
$p(\hat{x}_2 = 0, \hat{x}_7 = 0|\hat{y} = 2) = 0.81$ and $p(\hat{x}_2 = 0, \hat{x}_7 = 1|\hat{y} = 2) = 0.03$
easily,
$p(\hat{x}_2 = 0|\hat{y} = 2) = p(\hat{x}_2 = 0, \hat{x}_7 = 0|\hat{y} = 2) + p(\hat{x}_2 = 0, \hat{x}_7 = 1|\hat{y} = 2) = 0.84$
Thus, the correct answer is **Option B**