

Politecnico di Torino

Data Science and Engineering

Mathematics in Machine Learning

Ali Ghasemi (S289223)

Dataset: Higher Education Students Performance Evaluation Dataset Data Set

Professor: Francesco Vaccarino, Mauro Gasparini

**TABLE OF CONTENTS**

**1-Introduction**

[1]Education has vital and increasing importance almost for all countries to accelerate their development. Well-educated persons provide more benefits to their countries and for that reason, classification of students’ performance before they enter exams or taking courses is also gained an importance. Improvement of education quality must be performed during the active semester to improve students’ personal performance to response this expectation. To provide this, some of the main indicators are students’ personal information, educational preferences, and family properties. The dataset is gathered using questionnaire results that consists of these main indicators, of three different courses of two faculties to classify students’ final grade performances and to determine the most efficient machine learning algorithm for this task.

**Dataset Description**

Dataset is consisted of 33 attributes and 145 instances. Each row of the dataset represents a student and his/her/their situation.

Student ID

1- Student Age (1: 18-21, 2: 22-25, 3: above 26)

2- Sex (1: female, 2: male)

3- Graduated high-school type: (1: private, 2: state, 3: other)

4- Scholarship type: (1: None, 2: 25%, 3: 50%, 4: 75%, 5: Full)

5- Additional work: (1: Yes, 2: No)

6- Regular artistic or sports activity: (1: Yes, 2: No)

7- Do you have a partner: (1: Yes, 2: No)

8- Total salary if available (1: USD 135-200, 2: USD 201-270, 3: USD 271-340, 4: USD 341-410, 5: above 410)

9- Transportation to the university: (1: Bus, 2: Private car/taxi, 3: bicycle, 4: Other)

10- Accommodation type in Cyprus: (1: rental, 2: dormitory, 3: with family, 4: Other)

11- Mother’s education: (1: primary school, 2: secondary school, 3: high school, 4: university, 5: MSc., 6: Ph.D.)

12- Father’s education: (1: primary school, 2: secondary school, 3: high school, 4: university, 5: MSc., 6: Ph.D.)

13- Number of sisters/brothers (if available): (1: 1, 2: 2, 3: 3, 4: 4, 5: 5 or above)

14- Parental status: (1: married, 2: divorced, 3: died - one of them or both)

15- Mother’s occupation: (1: retired, 2: housewife, 3: government officer, 4: private sector employee, 5: self-employment, 6: other)

16- Father’s occupation: (1: retired, 2: government officer, 3: private sector employee, 4: self-employment, 5: other)

17- Weekly study hours: (1: None, 2: <5 hours, 3: 6-10 hours, 4: 11-20 hours, 5: more than 20 hours)

18- Reading frequency (non-scientific books/journals): (1: None, 2: Sometimes, 3: Often)

19- Reading frequency (scientific books/journals): (1: None, 2: Sometimes, 3: Often)

20- Attendance to the seminars/conferences related to the department: (1: Yes, 2: No)

21- Impact of your projects/activities on your success: (1: positive, 2: negative, 3: neutral)

22- Attendance to classes (1: always, 2: sometimes, 3: never)

23- Preparation to midterm exams 1: (1: alone, 2: with friends, 3: not applicable)

24- Preparation to midterm exams 2: (1: closest date to the exam, 2: regularly during the semester, 3: never)

25- Taking notes in classes: (1: never, 2: sometimes, 3: always)

26- Listening in classes: (1: never, 2: sometimes, 3: always)

27- Discussion improves my interest and success in the course: (1: never, 2: sometimes, 3: always)

28- Flip-classroom: (1: not useful, 2: useful, 3: not applicable)

29- Cumulative grade point average in the last semester (/4.00): (1: <2.00, 2: 2.00-2.49, 3: 2.50-2.99, 4: 3.00-3.49, 5: above 3.49)

30- Expected Cumulative grade point average in the graduation (/4.00): (1: <2.00, 2: 2.00-2.49, 3: 2.50-2.99, 4: 3.00-3.49, 5: above 3.49)

31- Course ID

32- OUTPUT Grade (0: Fail, 1: DD, 2: DC, 3: CC, 4: CB, 5: BB, 6: BA, 7: AA)

32nd attribute is the target. It represents the grade of the student.

There are classed that students must be classified into.

The dataset does not contain any NA (not available) or missing data.

A picture containing text, keyboard, electronics

Description automatically generated

Figure 1: a small representation of the dataset

**2-Data Exploration**

As the first step, data must be checked for being imbalanced or not. One of the best ways to come to a better understanding of the data, is to use visualization.

Imbalanced data refers to those types of datasets where the target class has an uneven distribution of observations. An imbalanced dataset leads to a low performance of the model

Chart, histogram

Description automatically generatedFigure 2: Number of instances of different labels across the dataset

As it can be seen in figure 2, it’s obvious that this dataset is imbalanced, and this issue must be handled.

There are various ways of overcoming the issue of an imbalanced dataset such as

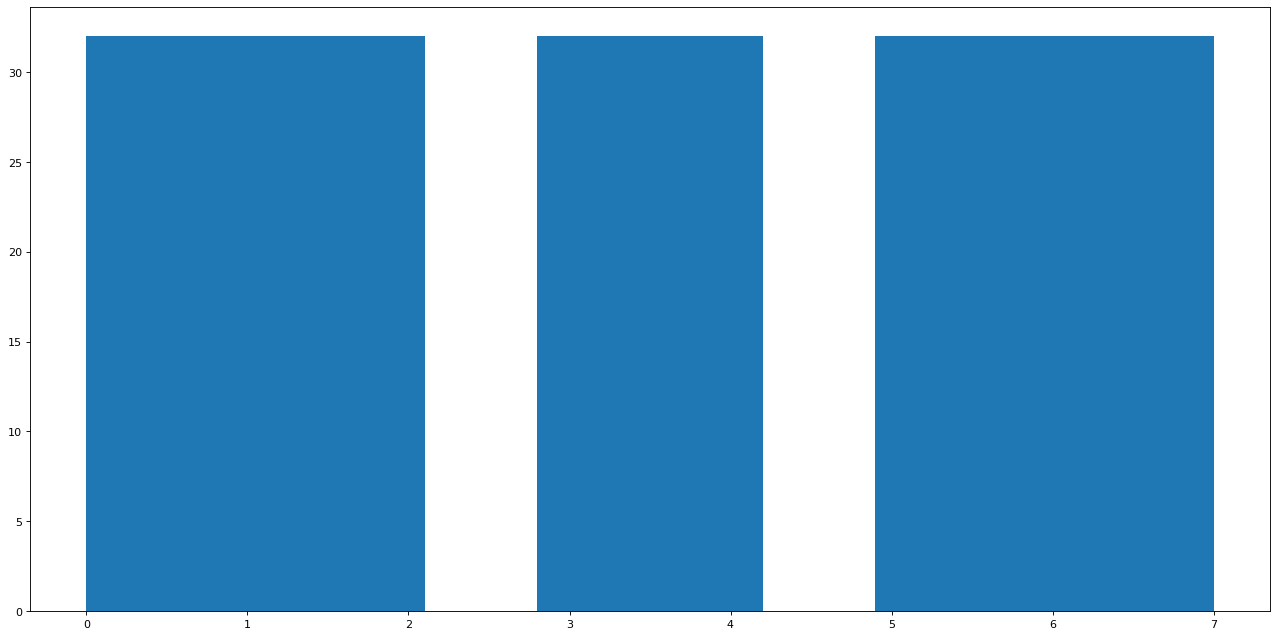
Undersampling: balances the dataset by reducing the size of the abundant class.

Oversampling: used to balance the dataset when the quantity of data is insufficient.

Etc.

The technique used in this thesina is random oversampling. Random Oversampling is a resampling method. Resampling involves creating a new transformed version of the training dataset in which the selected examples have a different class distribution. Random oversampling Randomly duplicate examples in the minority class.

Figure 3 demonstrates the labels after performing random oversampling on dataset

Figure 3: dataset labels after random oversampling

As it can be seen in figure 3, it’s obvious that dataset labels are balanced after performing random oversampling.

Also, the number of instances from each label are mentioned in figure 4:



Figure 4: instances of each label in the dataset

**3-Standardization**

Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g. Gaussian with 0 mean and unit variance).

For instance, many elements used in the objective function of a learning algorithm (such as the RBF kernel of Support Vector Machines or the L1 and L2 regularizers of linear models) assume that all features are centered around 0 and have variance in the same order. If a feature has a variance that is orders of magnitude larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected.

The result of standardization (or Z-score normalization) is that the features will be rescaled to ensure the mean and the standard deviation to be 0 and 1, respectively.

The below formula describes this method.

In which “” represent the mean of the distribution {have to check this in the slides} and “” represent the standard deviation of the distribution.

**4-Correlation**

Correlation between two variables demonstrates the relation between those two variables. One way to quantify this relationship is to use the Pearson correlation coefficient, which is a measure of the linear association between two variables. It has a value between -1 and 1 where:

-1 indicates a perfectly negative linear correlation between two variables

0 indicates no linear correlation between two variables

1 indicates a perfectly positive linear correlation between two variables-1 indicates a perfectly negative linear correlation between two variables

0 indicates no linear correlation between two variables

1 indicates a perfectly positive linear correlation between two variables

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by r. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

The formula below represents the Pearson’s Correlation Coefficient:

The numerator is the empirical estimate of the covariance of the two variables taken into exam, while the denominator is the standard deviation of the first variable times the standard deviation of the other variable.

The correlation between different variables of the dataset must be found.

Figure 4 demonstrates the correlation between different attributes of the dataset.

Highly correlated variables lead to redundant classification, and they affect each model in a different way. For linear models (e.g., linear regression or logistic regression), Multicollinearity can yield solutions that are wildly varying and possibly numerically unstable. Random forests can be good at detecting interactions between different features, but highly correlated features can mask these interactions. For sake of intuition, we can see that there is a high correlation between 30 (Expected Cumulative grade point average in the graduation) and 31 (Course ID). and for example, there's no correlation between 14 (Parental status) and 4(Scholarship type)

Based on Figure 4, there aren’t much highly correlated variables in this dataset. But still PCA will be performed to get a better result from models compared to the output of the results without applying PCA on the dataset.

**Figure 4: correlation map**Shape

Description automatically generated

**5-Outliers**

[3] Outliers once upon a time regarded as noisy data in statistics, has turned out to be an important problem which is being researched in diverse fields of research and application domains. Many outlier detection techniques have been developed specific to certain application domains, while some techniques are more generic. Some application domains are being researched in strict confidentiality such as research on crime and terrorist activities. The techniques and results of such techniques are not readily forthcoming. Several surveys, research and review articles and books cover outlier detection techniques in machine learning and statistical domains individually in great details. Outlier detection aims to find patterns in data that do not conform to expected behavior.

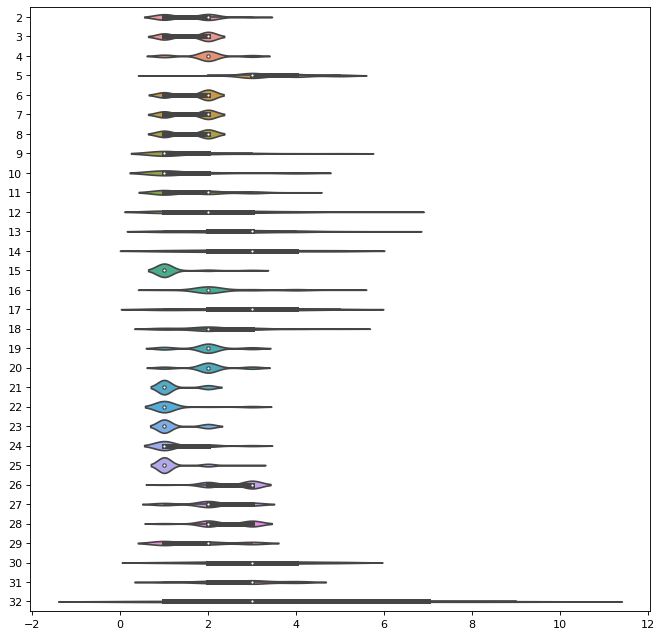
In this thesina, violin plots have been used to analyze the dataset regarding the existence of outliers. Figure 5 demonstrates the violin plots that are used to detect outliers.

One the methods that can be used to deal with the outliers’ problem is the Z-score method.

z-score (also called a standard score) gives an idea of how far from the mean a data point is. But more technically it’s a measure of how many standard deviations below or above the population mean a raw score is.

The formula for Z-score method is available below:

In which represent “standard score”, represents “observed value”, represents “mean of the sample” and, represents “standard deviation of the sample”

Figure 5

Z-score outlier removal technique removes all points located outside of normal

distribution which is higher than 3 standard deviation area. This method will eliminate around 30 rows, but no difficulties will be faced due to low amount of training data since random oversampling is used to both deal with data being imbalanced and data being of low amount.

**6-PCA**

Principal component analysis (PCA) is a technique that transforms high-dimensions data into lower-dimensions while retaining as much information as possible. PCA is extremely useful when working with data sets that have a lot of features. Common applications such as image processing, genome research always must deal with thousands-, if not tens of thousands of columns.

While having more data is always great, sometimes they have so much information in them, we would have impossibly long model training time and the curse of dimensionality starts to become a problem. Sometimes, less is more.

One of the main steps for implementing PCA on the data is to find the right number of principal components which is number of features that should be

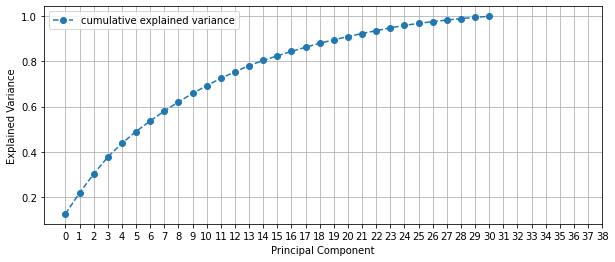
****used in the mode. One the best ways to find that is to visualize dependency of ratio from number of components which is shown in figure 6. After that, we can use the “elbow method” to understand the right number of components to use.

Figure 6: dependency of ratio from number of components

As it can be seen in figure 6, in you increase the number of components more than 14, the slope flattens, and the amount of captured variance doesn’t increase that much.

**7-Model selection and Train/Test procedure**

**7.1-Metrics and Cross Validation**

the metric used in this thesina is F1 score. The data we have in the dataset does not completely have the characteristics of the dataset which precision or recall is used on, so F1-score is the best metric for this case since it creates a balance between precision and recall.

F1-score is usually used on imbalanced datasets but the issue data of being imbalanced is solved using random oversampling.

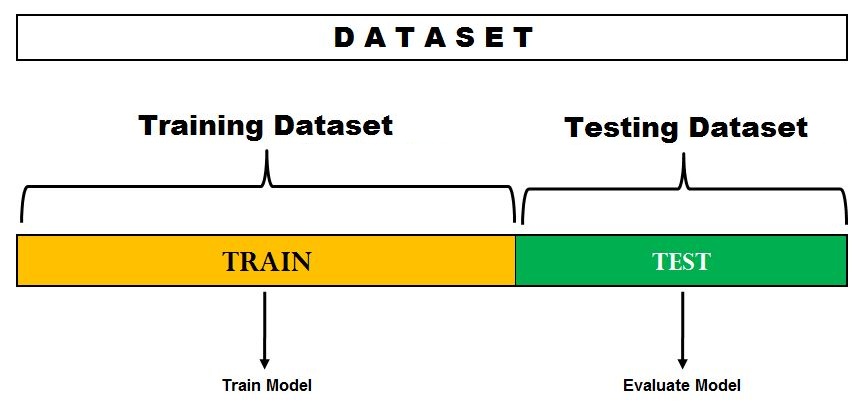
One of the other techniques used in this thesina is cross validation.

This process of deciding whether the numerical results quantifying hypothesized relationships between variables, are acceptable as descriptions of the data, is known as validation. Generally, an error estimation for the model is made after training, better known as evaluation of residuals. In this process, a numerical estimate of the difference in predicted and original responses is done, also called the training error. We need to have to data that is not seen by the model in order to completely evaluate the model, in order to do that, a technique called Cross Validation is used.

There are different ways of using Cross Validation such as the Holdout Method or K-Fold Cross Validation.

The method used in this thesina is the Holdout Method which is removing a part of the training data and using it to get predictions from the model trained on rest of the data. The error estimation then tells how our model is doing on unseen data or the validation set.

This method is demonstrated in figure 7.



**Figure 7: Holdout method**

**7.2-Models**

In this part of the thesina we’re going through different models that has been trained on the dataset.

**7.2.1-SVM**

Citations

1. Yılmaz, N., Sekeroglu, B. (2020). Student Performance Classification Using Artificial Intelligence Techniques. In: Aliev, R., Kacprzyk, J., Pedrycz, W., Jamshidi, M., Babanli, M., Sadikoglu, F. (eds) 10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions - ICSCCW-2019. ICSCCW 2019. Advances in Intelligent Systems and Computing, vol 1095. Springer, Cham. <https://doi.org/10.1007/978-3-030-35249-3_76>
2. Goldberger, Arthur S. (1991). "Multicollinearity". A Course in Econometrics. Cambridge: Harvard University Press. pp. 245–53. ISBN 9780674175440.
3. Belhadi, Asma, et al. "Machine learning for identifying group trajectory outliers." ACM Transactions on Management Information Systems (TMIS) 12.2 (2021): 1-25.