LUND UNIVERSITY

SCHOOL OF ECONOMICS AND MANAGEMENT

Master's Programme in Data Analytics and Business Economics

Analyzing Predictors of Schools' Performance using Machine Learning Methods

Case of Primary and Secondary Schools in Slovakia

by

Miriama Sokoláková

&

Gabrielė Lelkaitė

Master's Essay - DABN01
May 2024
Supervisor: Jonas Wallin
Word Count: 13 828

Acknowledgements

Gabrielė Lelkaitė

I would like to express sincere gratitude to my parents and my brother for supporting me throughout this challenging yet rewarding academic journey. Their encouragement and belief in me have been a constant source of motivation. I will remain forever grateful for their unwavering support in pursuing my dreams.
I would also like to thank my co-author, Miriama, for her hard work and valuable feedback, which have been instrumental in the completion of this thesis. I am grateful to have had the opportunity to conquer this challenge together.

Miriama Sokoláková

I am deeply grateful to my classmates and friends. Without their support, I couldn't have imagined graduating. My heartfelt thanks also go to my parents and family, whose constant support has been invaluable throughout my academic journey. Lastly, I'm thankful to Gabrielė. I greatly appreciate her unwavering support, patience, punctuality, and willingness to embark on this journey with me.

Typeset in LaTeX. Text is fine-tuned and grammar-checked in Notion, Grammarly, and ChatGPT.

Abstract

This thesis explores the determinants of educational performance in Slovak schools using advanced machine-learning (ML) techniques. It identifies key factors influencing academic outcomes and evaluates the effectiveness of various ML models, including Random Forest, Gradient Boosting, and Neural Networks, among others. This study compiled a complex dataset of 1409 primary and 656 secondary schools and matched it to a variety of demographic and economic characteristics. Results indicate that ensemble tree methods, particularly XGBoost, outperform other models in terms of predictive accuracy. These models consistently identify the higher-educated population in the region, the ratio of teachers to students, and the number of pupils in a school as the most significant predictors of academic performance.

Contents

List of Figures

List of Tables

1

Introduction

Education, and consequently educational performance, are key predictors of societal quality. Understanding the factors that influence academic performance has long been a priority for educational researchers and policymakers. With the arrival of machine learning (ML) techniques, there is an opportunity to study these factors more deeply and predict educational outcomes with greater accuracy. Inspired by similar papers by Chen and Ding (2023) and Masci et al. (2018), this thesis explores the determinants of educational performance at the school level using advanced ML techniques, aiming to enhance our understanding and provide actionable insights for improving educational policies and practices. The central research question of this thesis is: *What are the key factors influencing academic performance in Slovak schools, and how effectively can machine learning techniques predict these outcomes?*

The motivation behind this research stems from the growing recognition that traditional statistical approaches, while valuable, may not fully capture the complex and non-linear relationships inherent in educational data. Machine learning offers a robust alternative that is capable of handling large, complex datasets and uncovering patterns that might be overlooked by conventional methods (Nafea, 2018). This study leverages a dataset comprising school-level data from Slovakia, including variables such as school characteristics, socio-economic status (SES) of pupils, ICT usage, demographic factors, and many others. By applying a range of ML techniques, including Random Forest, Gradient Boosting, Light GBM, XGBoost, Support Vector Machines (SVM), Neural Networks, K-Nearest Neighbors (kNN), and Kernel Ridge Regression (KRR), the research aims to identify the most effective models and the key predictors of academic performance. Results indicate that ensemble tree methods, particularly XGBoost, outperform other models in terms of predictive accuracy. These models consistently identify the higher-educated population in the region, the ratio of teachers to students, and the number of pupils in a school as the most significant predictors of academic performance.

The data and methodology involve a very complex and manual process of data collection, preprocessing, model selection, and evaluation. Our dataset comprises 1409 primary schools and 656 secondary schools. We used national tests by National Institute of Education and Youth as a reference for educational achievement. Besides that, the dataset includes various factors that potentially influence educational outcomes sourced from various Slovak public databases. Preprocessing techniques were employed to handle missing data, standardize variables, and transform skewed

distributions. The selected ML models were then trained and validated using cross-validation techniques to ensure robust and reliable predictions. The study acknowledges several limitations. The dataset does not include some schools in Slovakia that do not report the dependent variable - the standardized test or some important independent variables. This may affect the generalizability of the findings to other contexts. Additionally, while ML models offer high predictive accuracy, their interpretability can be challenging, complicating the translation of results into actionable policy recommendations.

Following this introduction, Chapter 2 - Literature Review examines existing research on the determinants of educational performance and the application of ML techniques in educational research. This review contextualizes the current study and highlights gaps that this thesis aims to address. Chapter 3 - Data and Chapter 4 - Methodology details the data sources, preprocessing steps, and ML models used in the analysis, along with the criteria for evaluating their performance. Chapter 5 - Results presents the findings, including the performance of different models and the importance of various predictors. Comparative analyses identify the most effective models and the key factors influencing academic performance. Section 5.3 - Discussion interprets the results in light of existing literature and explores their implications for educational policy and practice. The strengths and limitations of the study are critically examined, and recommendations for future research are provided. The conclusion summarizes the key findings, reiterates the contribution of our study to the field of educational research, and highlights the practical implications of the results. It concludes with a discussion of potential future directions for research in this area.

In summary, this thesis aims to enhance our understanding of the determinants of educational performance using advanced ML techniques. By identifying key predictors and evaluating the effectiveness of various models, this study seeks to provide actionable insights for policymakers and educators. The findings highlight the potential of ML to improve educational outcomes. Through a comprehensive analysis and a clear articulation of its implications, this thesis aspires to contribute meaningfully to the field of educational research and policy.

2

Literature Review

This chapter aims to evaluate the latest research on forecasting academic performance and analyzing educational attainment determinants at the school level using machine learning techniques. Our approach for this literature review involved investigating relevant literature and papers to compare these studies' methodologies and corresponding findings. We filtered the studies based on the specifics of our data and then also examined the limitations of these studies and discussed how their results can be applied in the context of this Master's Essay. Lastly, we highlight how this Master's Essay builds upon and contributes to the existing research.

2.1 School Performance Factors

The determinants of students' and schools' educational performance have received significant interest in recent years. Researchers agree that personal circumstances and external or school-related factors influence student performance. According to the literature, there is a variety of factors that can influence educational performance. The effects might vary based on the regional context or socioeconomic setting, yet we decided to list the most important ones below.

Consistently shown as a decisive factor, socioeconomic status has been found to have a negative effect on educational performance (Carlisle and Murray, 2015). A student's family's financial and social capital can also heavily influence the educational resources, from study materials to the opportunity for private tutoring (Amini et al., 2015). The nature of the school itself, whether public or private, has been documented to affect educational efficiency, though this effect can be either negative or positive (Cherchye et al., 2010). The geographical setting of a school also plays a significant role. Urban versus rural locations can negatively or positively impact efficiency, as each comes with unique challenges and advantages regarding resources, student population, and teacher availability (Bouck, 2018).

Surprisingly, the presence of competition, measured by the number of competing schools in the vicinity, positively impacts a school's efficiency. This phenomenon is attributed to the market-like dynamics where schools strive to improve to attract and retain students (Agasisti, 2009). The experience and salaries of teachers are directly associated with higher school efficiency. Experienced teachers tend to have more refined teaching methodologies and classroom management skills, contributing to better student outcomes. Likewise, competitive salaries can attract higher-quality teaching staff and indicate a school's investment in human capital (Britton and

Propper (2016); Hanushek and Rivkin (2007)).

Smaller class sizes have been associated with better educational outcomes as they allow for more individualized attention (Fredriksson et al., 2013). Similarly, the size of the school itself can have a positive impact, with smaller schools often fostering a more focused educational environment (Werblow and Duesbery, 2009). The academic background of parents has a positive effect on school performance. Parents with higher levels of education are often more equipped to support their children's educational journey (Afonso and Aubyn, 2016). On the other hand, parental pressure has a negative relationship with efficiency, possibly due to the additional stress it places on students (Toraman et al., 2022).

Based on these findings, where possible, we try to add all these features in some comparable form to our dataset.

2.2 Application of Machine Learning in Education

The rapidly evolving intersection of machine learning and educational research also gained serious popularity and provides a fresh perspective into the various determinants affecting academic performance across schools (Korkmaz and Correia, 2019). Educational data is known to exhibit often complex and, therefore, non-linear relationships, something that traditional linear models cannot adequately capture or interpret (Nafea, 2018). This complexity in the educational data leads to switching to using more advanced analytical techniques such as machine learning (Wu, 2020). The development of ML in educational research is mainly due to the increasing availability of large-scale datasets. These datasets allow for the comparative analysis of academic performance across various demographics and geographies. The opportunity to leverage these large datasets creates a new playground for transformative research in education, potentially leading to more nuanced algorithms for understanding school performance factors (Jordan and Mitchell, 2015; Hilbert et al., 2021).

Classical statistical models and machine learning models differ significantly in their approaches to analyzing educational data. Classical statistics, such as linear regression, ANOVA, and logistic regression, rely on assumptions about data distribution and are simpler and easier to interpret, requiring fewer computational resources. These methods are well-documented for their effective handling of structured data within specific parametric frameworks (Korkmaz and Correia, 2019; Wu, 2020).

In contrast, machine learning models like Decision Trees, Neural Networks, and Support Vector Machines are favored for their ability to manage large, complex datasets and model non-linear relationships, which often results in higher accuracy. These models excel in environments with ample data and diverse variable interactions, enhancing predictive performance through advanced learning algorithms (Nafea, 2018; Jordan and Mitchell, 2015).

However, the benefits come with challenges. Classical models may struggle with complex patterns and require strict data conformity to assumptions. Machine learning models, although powerful, can be opaque ("black boxes"), making them difficult to interpret and requiring significant computational power, which can be a barrier

in resource-constrained settings (Hilbert et al., 2021; Jordan and Mitchell, 2015).

2.3 Predictive Models and Educational Outcomes

As shown in Table 2.1, we conducted a thorough literature review for papers carrying out similar research on educational outcomes using predictive models. We excluded papers with small sample sizes or those not utilizing machine-learning methods. Our review resulted in a table of 16 research papers or articles focusing on this specific research area.

It is noteworthy that all studies are relatively recent, and most use individual students rather than entire schools as the unit of interest. The research geography is diverse, with papers from around the world. Additionally, some studies are naturally classification problems or translated into classification problems.

Table 2.1 illustrates the evolution of predictive modeling in education, reflecting advancements in data availability and machine learning techniques. Early studies primarily used simpler models like Decision Trees and Logistic Regression. However, recent research increasingly incorporates more sophisticated algorithms such as Random Forest, Neural Networks, and Support Vector Machines. These models provide more robust predictions and can handle larger, more complex datasets.

This subset of research using machine learning to determine educational outcomes typically follows two main approaches. One approach focuses on the individual student, while the other compares entire schools based on other collected features (Khan and Ghosh, 2021). The latter approach is less common, with researchers tending to use students as the base unit. Studies using students as a unit base tend to have slightly higher accuracy measurements.

In terms of data size, the samples ranged from smaller (e.g., 105 schools in Tunisia by Rebai et al. (2020)) to very large (e.g., 1.2 million students in Australia by Cornell-F. and Garrard (2020)). While the studies also used a wide range of features, some of the most common predictors across the research included student demographics (Mousa and Maghari, 2017; Masci et al., 2018; Carlos et al., 2021; Chen and Ding, 2023; Naicker et al., 2020), prior academic achievement (Chung and Lee, 2019; Zafari et al., 2021; Yağcı, 2022), and school characteristics (Rebai et al., 2020; Cruz-Jesus et al., 2020). However, the results varied significantly across studies, emphasizing the complex and multi-dimensional nature of school performance.

One of the more comparable pieces of literature from Table 2.1 is the paper by Chen and Ding (2023), titled "A Machine Learning Approach to Predicting Academic Performance in Pennsylvania's Schools." This paper uses ML across educational datasets in Pennsylvania with various other variables to predict student performance and identify at-risk schools. The relevance of this work to the current thesis lies in its comprehensive analysis of ML algorithms, such as Random Forests, Support Vector Machines, Decision Trees, and Neural Networks, which this thesis aims to use as its methodology. Secondly, this paper uses variables that are very similar to ours, making the studies very comparable.

Table 2.1: Overview of Relevant Literature for Prediction of Educational Performance

	Characteristics				Methods & Accuracy [%]						
Author (year)	Country	Sample	Unit	Feature imp.	DT[1]	RF[2]	SVM[3]	kNN[4]	LR[5]	NN[6]	NB[7]
Nghe et al. (2007)⋆	Thailand	21 534	students	not reported	72	-	-	-	-	-	-
Anuradha et al. (2015)⋆	India	≈500	students	not reported	-	-	-	63	-	-	70
Mousa and Maghari (2017)⋆	Palestine	1 036	students	SES[8]	93	-	-	89	-	-	92
Masci et al. (2018)	OECD (9 c.)	3 600	schools	SES[8,9]	14-60[9]	-	-	-	-	-	-
Harvey and Kumar (2019)	USA (MA)	403	schools	not reported	60	-	-	-	-	-	71
Chung and Lee (2019)	South Korea	165 715	students	absenteeism	-	93	-	-	-	-	-
Cruz-Jesus et al. (2020)	Portugal	110 627	students	school size	79	79	51	79	81	77	-
Cornell-F. and Garrard (2020)	Australia	1.2 mil	students	not reported	77	83	-	-	84	83	-
Naicker et al. (2020)⋆	USA	1 000	students	race, gender, lunch	88	-	97	-	97	-	74
Rebai et al. (2020)	Tunisia	105	schools	gender, school size	-	n/a	-	-	-	-	-
Yildiz and Börekci (2020)⋆	Turkey	421	students	not reported	92	90	90	86	78	-	82
Zeineddine et al. (2021)⋆	UAE	1 491	students	not reported	69	76	74	69	71	71	72
Carlos et al. (2021)⋆	Colombia	163 030	students	SES[8]	-	-	-	-	-	82	-
Zafari et al. (2021)	Iran	459	students	absence	-	72	73	-	76	83	-
Yağcı (2022)⋆	Turkey	1 854	students	midterm	-	75	70	74	72	75	71
Chen and Ding (2023)⋆	USA (PA)	8 129	schools	SES[8]	48	54	51	-	50	60	-

Notes: 1 - Decision trees, 2 - Random Forest, 3 - Support Vector Machines (Regression), 4 - k-Nearest Neighbours, 5 - Logistic Regression, 6 - Neural Nets, 7 - Naive Bayes, 8 - Socially and economically disadvantaged students, 9 - Depending on the country, ⋆ - Classification study.

Another more complex and complicated approach is presented by Masci et al. (2018) in a paper titled "Student and school performance across countries: A machine learning approach," where a two-step ML tree-based method for data is utilized. The PISA 2015 test scores dataset is from nine countries: Australia, Canada, France, Germany, Italy, Japan, Spain, the UK, and the USA. This research aims to identify which student and school characteristics significantly correlate with test scores and how school value-added (measured at the school level) is associated with school-level variables. Key findings from this study reveal that several student and school-level characteristics are significantly related to students' achievements. However, considerable differences in predicted variability (the measure used by the study) across different countries were observed.

Moreover, the study by Khan et al. (2022), "Student Performance Prediction in Secondary School Education Using Machine Learning," offers a comparative perspective that is highly relevant to this thesis. By contrasting traditional statistical methods with deep learning models, this study illustrates the superior predictive capabilities of ML in educational research with accuracy reaching 94%. This comparison underscores the potential of employing advanced ML algorithms to gain deeper insights into school performance metrics beyond what traditional econometric methods can reveal.

2.4 Contribution

The studies shown in Table 2.1 clearly illustrate the powerful applications of ML methodologies in educational research. This increasing interest from various international contexts underlines a growing agreement about the importance of advanced analytics in this area. For Slovakia, specifically its primary and secondary education sectors, using ML techniques offers the potential for new insights. Our research aims to combine the precision of ML algorithms with traditional econometric methods, enhancing the analytical capabilities to understand better what drives educational success.

A major contribution of our study is using a public dataset that includes all primary and secondary schools in Slovakia with good reporting capabilities (most of them). Unlike many studies that rely on samples or subsets, our approach allows for a more detailed and complete analysis of educational dynamics nationwide. This enhances the generalizability of our findings and provides a robust framework for evaluating the entire educational system.

Moreover, this study investigates a geographical area that has not been previously examined at all. By focusing on Slovakia, a region often overlooked in global educational research, we contribute a unique perspective to the literature. This is supplemented by our data source - standardized national tests - which provide a consistent, reliable measure of academic achievement within the country.

Our approach also differs from most of the other papers by using schools as the primary unit of analysis instead of individual students. This shift allows us to include geographic and socioeconomic characteristics at the school level, creating the opportunity to study the importance of these features. This perspective is essential for understanding systemic factors affecting school performance, including regional disparities and resource allocation, often overlooked in student-centric analyses.

3

Data

In this chapter, we introduce the dataset and its features. The dataset includes results from 7 years (2014-2022; COVID years excluded due to cancellation of the test) of 1409 primary schools and 656 secondary schools in Slovakia. To avoid bias, we only included regular schools, excluding those for pupils with special needs. Only primary schools participating in the NIVAM (National Institute of Education and Youth) tests are included. Schools with only four grades do not participate in these tests and, therefore, are not part of our dataset. We also ensured the schools included have good reporting capabilities for variables such as the number of pupils and teachers. As for secondary schools, we included only those with no or a minimal number of missing values in the panel data. This means we excluded schools established later or ones that had closed.

We start by introducing the dependent variable - the results and percentiles from standardized NIVAM tests (Testing 9, Maturita) that every pupil in each school must take unless excused for serious reasons such as long-term inability to attend school or a condition that prevents attendance. These results were obtained from the National Institute of Education and Youth.

Next, we introduce the independent variables. Specific details for each school (type, founder, number of students, number of teachers, number of students from socially disadvantaged backgrounds, etc.) were obtained from Institute for Economic and Social Reforms. For town-specific data, we used population percentages and higher education percentages from Statistical Office of the Slovak Republic. District-specific data, such as unemployment rates, were also sourced from Statistical Office of the Slovak Republic, while crime rates were obtained from Ministry of Interior of the Slovak Republic through Open Data request.

The shapefiles of Slovakia used in the map figures were acquired from Geodetic and Cartographic Office of Bratislava.

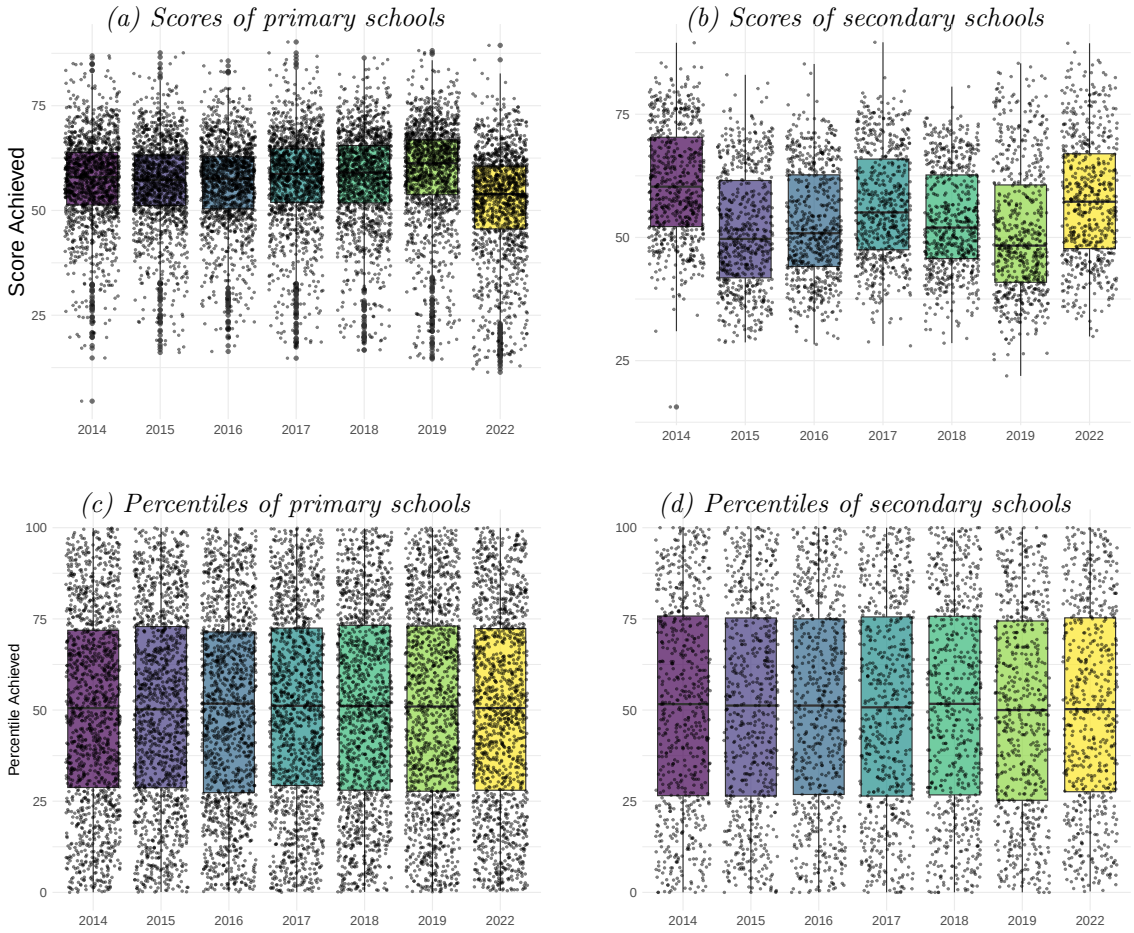
3.1 Standardized Tests as Dependant Variable

We use the percentiles of NIVAM tests as our dependent variable. These are norm-referenced national tests in Slovakia, similar to PISA tests. Their main objective is to assess the strategic competencies of students. Hence, these tests offer a good representation of educational achievement among individual schools in Slovakia.

There are three nationwide tests that NIVAM administers, but we are utilizing only two for our research:

- Testing 9 (nationwide from the school year 2004/2005) - mandatory tests in Mathematics and the Slovak language (or alternatively Hungarian) for primary school students in 9th grade.
- Maturita (nationwide from the school year 2006/2007) - mandatory tests in the Slovak language (or alternatively Hungarian) and optional (occasionally obligatory) tests in additional languages (English, German, Spanish, or French) and Mathematics for secondary school students in the school-leaving grade.

Figure 3.1: NIVAM test results between years 2014-2022



Source: National Institute of Education and Youth

For this Master's Essay, we created a special variable to measure the educational achievement of schools. We calculated the average percentiles of Testing 9 in Mathematics and Slovak (or Hungarian) language for primary schools. For secondary schools, we calculated the average percentiles of Maturita in Slovak (or Hungarian) language and English language.

In Table 3.1 and Figure 3.1, we provide descriptive statistics and graphical representations for both primary and secondary schools. The graphs in Figure 3.1 explain our decision to use percentiles instead of standard test scores. As shown in the Figure 3.1a and the Figure 3.1b, the tests are not consistently reliable due to significant fluctuations in the mean. These fluctuations could be due to varying test difficulty over the years, deletion of certain questions, or granting all students points for poorly formulated questions that were deemed misleading post-testing.

However, these factors do not affect percentiles, which continue to rank schools relative to each other. Therefore, using percentiles as the dependent variable proved to be a better option, as it offers a more reliable variable for machine learning models.

Table 3.1: NIVAM Test Score - Percentiles: Descriptive Statistics

	Primary Schools						Secondary Schools					
Year	**N**	**Min**	**Med.**	**Mean**	**Max**	σ	**N**	**Min**	**Med.**	**Mean**	**Max**	σ
2014	1400	0.00	50.53	50.22	99.89	26.75	656	0.00	51.62	51.06	100.00	28.61
2015	1402	0.08	50.18	50.30	99.92	26.85	656	0.00	51.62	51.06	100.00	28.61
2016	1400	0.18	51.75	50.55	99.81	26.81	656	0.00	51.23	50.96	100.00	28.48
2017	1400	0.04	51.17	50.49	99.86	26.86	636	0.00	50.74	50.65	100.00	28.61
2018	1409	0.11	51.11	50.48	99.82	27.01	654	0.00	51.71	50.99	100.00	28.57
2019	1409	0.15	50.97	50.51	99.93	27.30	655	0.00	50.00	49.86	100.00	28.15
2022	1396	0.41	50.55	50.24	100.00	27.15	610	0.26	50.24	50.66	100.00	27.95

Source: National Institute of Education and Youth (2023).

3.2 Independent Variables

This Master's Essay stands out due to the numerous specific independent variables that the authors have sourced and matched at the most precise level possible. To illustrate the specificity, data from six organizations and two censuses were utilized. In this section, we describe all the variables included in specific categories. First, we discuss numerical variables specific to each school. Next, we report on dummy variables that, while not numerical, are still school-specific. Finally, we describe variables reported at the town and district level, focusing on those specific to each area.

School level variables

Table 3.2: Descriptive Statistics for School Variables by School Type

	Primary Schools					Secondary Schools				
Variable	**Min**	**Med.**	**Mean**	**Max**	σ	**Min**	**Med.**	**Mean**	**Max**	σ
Ratio of Teachers	2.63	8.17	8.90	100.00	2.76	0.95	10.39	12.39	333.33	9.26
ICT (in %)	4.00	85.00	76.55	100.00	25.68	0.00	73.45	68.78	100.00	28.57
Number of Pupils	15.00	244.00	300.59	1410.00	189.79	7.00	304.00	328.03	1164.00	188.51
Number of SES Pupils	0.00	0.00	16.51	1024.00	53.55	0.00	0.00	1.20	312.00	11.93

Note: Ratio of Teachers - number of teachers per pupil, ICT - Usage of interactive and communications technologies.
Source: Institute for Economic and Social Reforms.

School statistics often form the basis of the research in cases like this one; our main reference papers (Chen and Ding (2023); Masci et al. (2018)) also include these; hence, we include four relevant statistics for this case. These are the ratio of teachers to students, ICT (usage of interactive and communication technologies), the total number of pupils, and the number of socially and economically disadvantaged (SES) students. By incorporating these variables, we aim to provide insight into how these characteristics influence the results, such as whether the ratio of teachers or school size is more significant.

For instance, from the Table 3.2, it is evident that SES students are likely concentrated in certain schools, as indicated by the median and mean. This concentration

may be due to the significant Roma minority in some regions of Slovakia, which could be a crucial factor in determining the educational outcomes of specific schools.

Dummy variables

Another interesting statistic to consider includes dummy variables indicating whether the school is public or private, church-established, and the language of instruction. This is particularly relevant in a regional context, as many schools in Slovakia use Hungarian as the language of instruction due to a large Hungarian minority, primarily in the south. These students then also take Hungarian tests in conjunction with tests from other subjects. This may introduce bias, making it beneficial to include this information. You can find these statistics in the Table 3.3.

Table 3.3: Dummy variables: descriptive characteristics

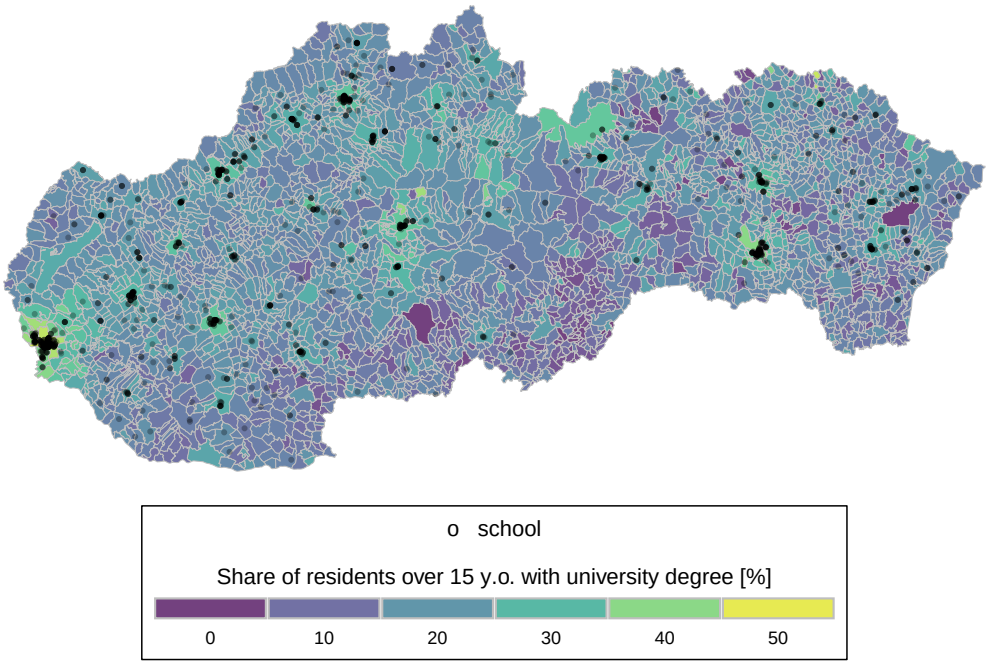
	Primary schools	**Secondary schools**
Type of establisher	**N**	**N**
Public	1305	499
Private	31	119
Church	98	74
Language of instruction	**N**	**N**
Slovak	1293	540
Hungarian	124	30
Bilingual	17	122

Source: Institute for Economic and Social Reforms

Town and district level variables

In this section, we include the most interesting variables at the town level that could potentially influence students' and pupils' perceptions of their area. For instance, we include variables such as unemployment and crime rates, including youth crime rates, which could affect the overall environment of the school's location. We also consider factors like the ratio of highly educated individuals, the number of divorces per new marriage, and out-of-wedlock births, which may provide insight into the population structure. The importance of these factors is evident in Figure 3.2, where schools that outperform are mainly located in areas with a high proportion of university-educated individuals. We also consider the number of school canteens, libraries, and extracurricular institutions per school-aged person in the region, as these factors can be significant.

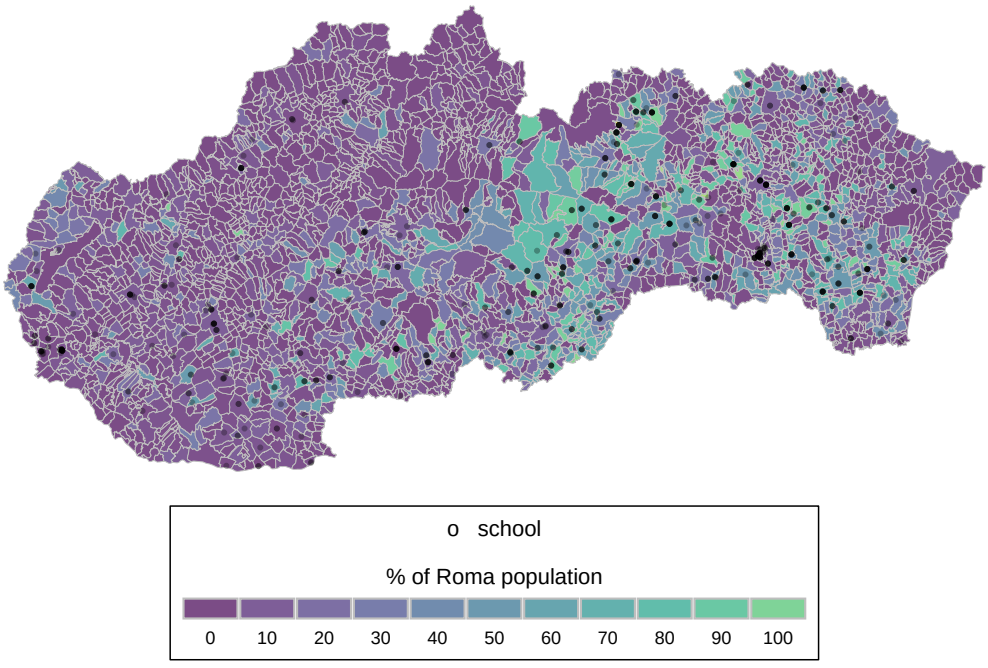
Figure 3.2: Overperforming schools above 90th percentile in NIVAM tests with respect to university-educated population



Source: Authors' own visualization based on public datasets cited in Chapter 3.

In the context of Slovakia, it makes sense to include the ratio of ethnic populations. This is apparent in Figure 3.3, where many underperforming schools over the years have been in areas with a high concentration of the Roma population.

Figure 3.3: Underperforming schools below 5th percentile in NIVAM tests with respect to Roma population



Source: Authors' own visualization based on public datasets cited in Chapter 3.

A summary of all the variables included and their descriptive statistics can be found in Table 3.4.

Table 3.4: Descriptive Statistics for Town and District Variables by School Type

	Primary Schools					Secondary Schools				
Variable (level)	**Min**	**Med.**	**Mean**	**Max**	σ	**Min**	**Med.**	**Mean**	**Max**	σ
Population (T)	163	3526	15480	113215	22881	570	25492	35888	113215	27586
Roma Population (T)	0.00	0.36	2.48	78.28	6.74	0.00	0.49	1.40	28.12	2.79
Hungarian Population (T)	0.00	0.26	9.71	92.36	22.36	0.00	0.41	6.89	81.59	16.87
Slovak Population (T)	0.00	90.77	81.81	100.00	22.93	0.00	88.24	82.63	98.31	16.53
Religious Population (T)	15.33	77.04	74.63	99.28	13.75	28.07	63.49	64.61	94.91	10.89
Higher Educated Population (T)	7.13	15.71	17.36	57.43	6.83	7.18	12.44	12.93	42.90	3.63
Divorces per New Marriage (T)	0.00	33.33	36.33	700.00	31.70	0.00	34.57	35.65	200.00	12.03
Out-of-Wedlock Births (T)	0.00	38.46	40.35	100.00	18.14	0.00	37.61	40.17	100.00	11.99
School Canteens (T)	0.00	0.00	0.02	0.28	0.04	0.00	0.03	0.05	0.28	0.06
School Libraries (T)	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.02	0.00
Extracurricular Institutions (T)	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00
Unemployment (T)	0.00	0.035	0.046	0.34	0.03	0.00	0.03	0.04	0.20	0.02
Crime Rate per Thousand (D)	0.43	1.05	1.20	5.01	0.60	0.43	1.07	1.23	5.01	0.62
Youth Crime Rate per Thousand (D)	0.00	0.06	0.08	0.50	0.07	0.00	0.06	0.08	0.50	0.06

Source: Statistical Office of the Slovak Republic; Ministry of Interior of the Slovak Republic. Note: T - town level, D - district level.

Missing Data Handling

As shown in Figure 3.4, our dataset has missing values due to some schools not reporting certain variables. The total number of pupils and socioeconomically disadvantaged (SES) pupils data is missing at random (MAR) for the year 2014 - unreported by about half of the schools. This is because these statistics only started being collected that year, so the assumption is that only some institutions reported it on a voluntary basis that year.

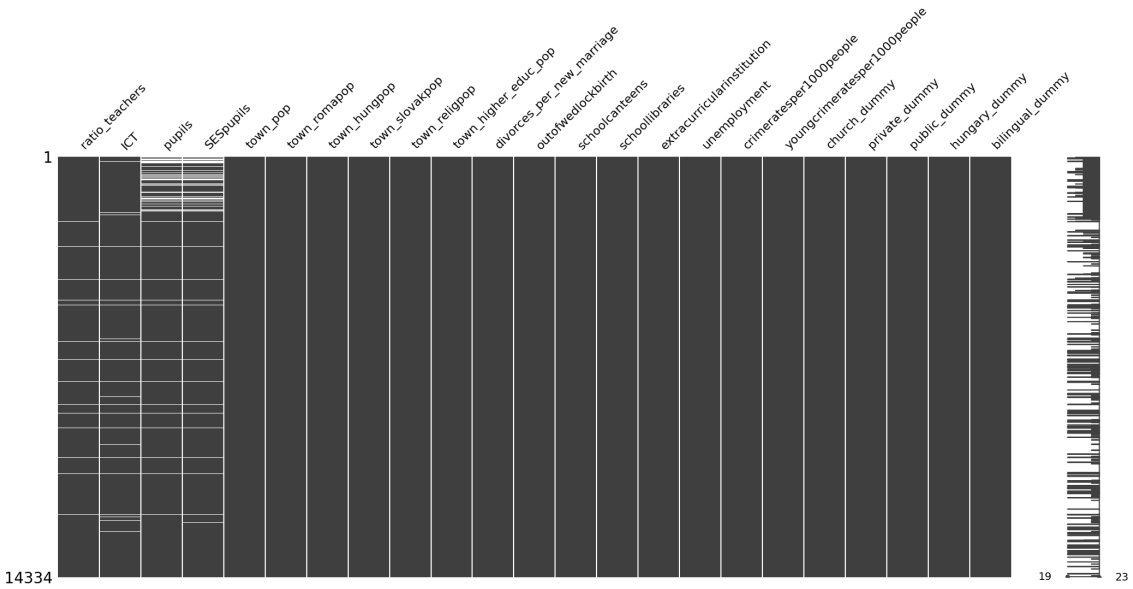


Figure 3.4: Missing Data

To address this, we imputed the MAR 2014 data for the number of pupils and SES pupils using linear imputation, taking into account the individual schools' trends over the following years. The rationale behind choosing linear imputation is based on the assumption that the number of pupils in a school tends to be relatively stable year-to-year, with only minor fluctuations. By leveraging the enrollment data from

subsequent years, we can estimate the missing values more accurately. This method assumes that the changes in the number of pupils follow a linear trend, which is reasonable given the typically stable nature of school enrollments. However, this approach has limitations. It assumes that there are no significant events (e.g., new school openings, major demographic shifts) that would cause abrupt changes in student numbers, which should not be the case for the general amount of schools. Despite these limitations, linear imputation is a practical method for handling missing data in this context.

On the other hand, a few schools systematically fail to report the teacher-student ratio and ICT usage data. These schools consistently failed to report this data over the years, so we removed them from the dataset due to inconsistent reporting.

Finally, in our preprocessing pipeline, we removed the remaining 345 NA values. **This leaves us with 13,989 observations for analysis with our models.**

Preproccessing of Census Data

In addition to the imputation for missing data in 2014, we also performed a more complex linear imputation for several variables derived from the census data of 2011 and 2021. These variables include the Roma population, Hungarian population, Slovak population, religious population, and higher-educated population at the town level. The rationale behind this imputation was the recognition of the significant value these variables add to our analysis, as they provide critical demographic and socio-cultural context that can influence educational outcomes.

We observed the trends at the district level for these variables and found them to be linear over the census periods. This linearity suggests that the demographic changes in these populations are relatively stable and predictable over time. Consequently, we applied a linear imputation method to estimate the values for the intervening years at the town level. By doing so, we assumed that the linear trends observed at the district level are representative of those at the town level, allowing us to fill in the gaps in the dataset.

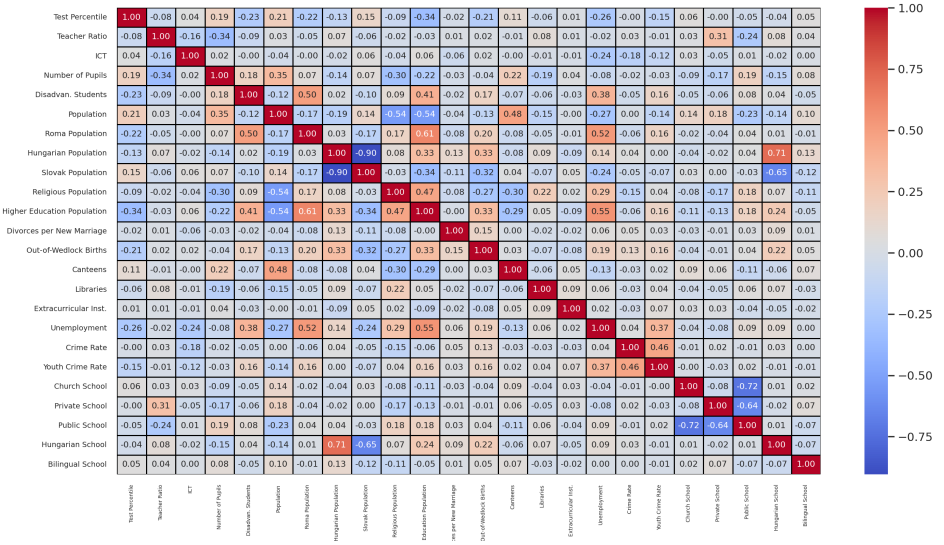
This approach, while more drastic, is justified due to the critical nature of these variables. The presence and proportions of different ethnic groups, religious adherence, and the level of higher education within a town can have profound impacts on the social and educational environment. For instance, areas with a higher proportion of the Roma population may face unique educational challenges, while towns with a higher percentage of higher-educated residents might benefit from a more supportive learning environment. Similarly, the linguistic composition (Hungarian and Slovak populations) and religious demographics can influence school culture and community engagement with education.

By retaining these variables through linear imputation, we ensure that our analysis incorporates these vital contextual factors. Despite the inherent assumptions and potential limitations of this method, such as the possibility of non-linear changes in smaller communities, the importance of these variables to our overall research objective justifies their inclusion. Thus, this imputation enables us to maintain a richer and more informative dataset, hopefully supporting a more interesting analysis of educational outcomes.

3.3 Correlation Analysis

In our analysis, we found some strong correlations with the dependent variable, as can be seen in Figure 3.5. There is a strong positive correlation between the dependent variable and the number of pupils, the population of the town, the Slovak population of the town, and the number of school canteens. This implies that larger schools in more populous towns with a higher number of school canteens tend to have better educational outcomes. On the other hand, there is a strong negative correlation between the dependent variable and the number of socioeconomically disadvantaged pupils (SES), the Roma and Hungarian population, the higher education population in the town, out-of-wedlock births, unemployment rates, and youth crime rates. This could suggest that schools with a higher percentage of disadvantaged pupils or those located in areas with high unemployment, crime rates, or out-of-wedlock births tend to have lower educational outcomes.

Figure 3.5: Correlation Matrix



In addition, we observed other strong positive correlations not directly tied to the dependent variable. For example, there is a strong positive correlation between the Roma population in the town and the SES pupils, suggesting that schools in towns with higher Roma populations tend to have more socioeconomically disadvantaged students. There's also a strong correlation between the Hungarian dummy variable and the Hungarian population in the town, which is logical as the Hungarian dummy variable is likely to be activated in towns with a higher Hungarian population. Lastly, we noted a strong positive correlation between unemployment and the higher education population in the town, which could be due to the higher competition for jobs in areas with a highly educated population.

We also found strong negative correlations between certain variables. For instance, there is a strong negative correlation between the Slovak population in the town and the Hungarian population, indicating that towns with a higher Slovak population tend to have fewer Hungarian inhabitants. Lastly, the Hungarian dummy variable has a strong negative correlation with the Slovak population in the town,

implying that towns with a higher Slovak population tend to have fewer Hungarian schools. These correlations provide insightful context for the analysis, and the occurrence is very understandable.

While it is true that a high correlation between predictor variables can sometimes pose problems in statistical analyses, it is important to note that this is not always a cause for concern. In many cases, a high correlation between predictor variables is expected and logical based on the nature of the variables themselves. For example, an increase in the population of a town might naturally lead to an increase in the number of libraries in that town. In such instances, retaining both variables in the analysis can be justified because they each provide unique and valuable information. Therefore, we decided to keep all of these variables despite some high correlation.

4

Methodology

After reviewing the relevant literature and describing the data, we move on to discuss the methodology employed in our analysis. This section begins by introducing our approach and briefly revisiting the advantages of machine learning over traditional statistical models, particularly in handling our complex, nonlinear dataset. After that, we provide a detailed description of each stage in our machine learning pipeline, from data preprocessing and model selection to evaluation and interpretation techniques, as well as the tools utilized for analysis.

4.1 Approach

To begin, this study aims to identify the key factors influencing educational outcomes in Slovak schools and determine the suitability of machine learning models for such analysis. Our study will concentrate on evaluating model performance as well as obtaining feature importance scores from advanced machine learning models and analyzing their implications. By doing so, we intend to identify at-risk groups and provide decision-makers with data-driven insights that facilitate early intervention strategies to prevent sub-optimal educational outcomes.

As previously discussed, machine learning methods offer numerous advantages over traditional statistical approaches. What makes them particularly suited for our study, however, is their unmatched ability to handle complex, high-dimensional data with skewed and imbalanced predictors. Moreover, features like automated hyperparameter tuning and the capability to quickly adapt to new data make machine learning methods highly advantageous in educational research. This strategic choice not only accommodates the complex nature of our data but also broadens the applicability of our findings for real-world scenarios. Consequently, our choice to employ machine learning algorithms leads to a different approach to the methodology section - instead of focusing on a specific estimable equation as in traditional analyses, our methodology revolves around a description of an extensive machine learning pipeline, which will be the topic of the following section.

4.2 Machine Learning Pipeline

Preprocessing

Data preprocessing is an important first step in any machine learning analysis because it ensures compatibility with the models and aligns data better with their underlying assumptions. While not all models require preprocessing to achieve good predictive accuracy - for example, ensemble trees can usually handle relatively raw data - it also offers faster convergence and more consistent feature importance scores. Many different data preprocessing techniques exist, each optimal for various models; therefore, choosing an appropriate transformation is important to achieve the best results.

The preprocessing that we chose for our data was twofold. First, we applied the Yeo-Johnson transformation (detailed in Equation 4.1), the purpose of which is to make skewed numerical variables more normally distributed and reduce the effect of outliers. This is especially beneficial for algorithms that rely on distance calculations, such as K-Nearest Neighbors, as it ensures that no single variable affects the outcome disproportionally due to its scale. Furthermore, the Yeo-Johnson transformation aligns data better with the key statistical assumptions underlying many predictive models, thus increasing robustness and improving the interpretability of feature importance by stabilizing the variance. Second, we standardized the transformed variables, as defined in Equation 4.2. This process adjusts the variables to have a mean of zero and a variance of one. This way, the predictors contribute equally to model training and features with larger variances do not dominate during the modeling process, which is crucial for scale-sensitive algorithms like Support Vector Machines and Neural Networks. Additionally, standardization is one of the tools that is known to improve convergence speed. It is also important to note, that these preprocessing techniques were applied only to the numerical predictors.

$$X'(\lambda) = \begin{cases} \frac{(X+1)^{\lambda}-1}{\lambda} & \text{for } X \geq 0 \text{ and } \lambda \neq 0, \\ \log(X+1) & \text{for } X \geq 0 \text{ and } \lambda = 0, \\ -\frac{(-X+1)^{2-\lambda}-1}{2-\lambda} & \text{for } X < 0 \text{ and } \lambda \neq 2, \\ -\log(-X+1) & \text{for } X < 0 \text{ and } \lambda = 2. \end{cases} \tag{4.1}$$

$$Z = \frac{X' - \mu}{\sigma} \tag{4.2}$$

Where:

- X' is the Yeo-Johnson transformed variable.
- λ is the transformation parameter determined based on maximizing the log-likelihood function.
- μ is the mean of the Yeo-Johnson transformed variable.
- σ is the standard deviation of the Yeo-Johnson transformed variable.
- Z is the standardized variable.

By combining these techniques, we ensure that our models benefit from both normalized feature distributions and fair feature scaling. This, in turn, helps to minimize the effects of skewness and adjusts the scale to prevent features with large variance from dominating the training process. Moreover, this combination of preprocessing techniques not only improves convergence but also aids in obtaining comparable feature importance scores, which is a key goal of our analysis. In contrast, in the literature we reviewed, including Chen and Ding (2023) and Masci et al. (2018), standardization alone was applied; however, we chose to combine both the Yeo-Johnson transformation and standardization due to their complementary nature, this way enhancing the quality of data and potentially improving the algorithms discussed in other studies.

Model Selection

After discussing data preprocessing, we now proceed to describe the model selection process. This step is critical in our data analysis framework, as each model has its strengths and limitations, which are particularly important to evaluate given our context of skewed features and imbalanced dummy variables. Although careful preprocessing can somewhat mitigate these issues, selecting the right models to handle these characteristics effectively remains essential. In light of this, we explored a wide variety of models to analyze our data from different angles. The main strengths and weaknesses of our key models are summarized in Table 4.1 and will be discussed in detail below.

Ensemble tree methods are central to our analysis. By combining multiple trees, they offer higher robustness and an improved predictive performance compared to individual decision trees. Random Forest, for example, excels in its ability to handle skewness and imbalances by constructing multiple trees simultaneously and aggregating their outputs. However, its speed can deteriorate as the number of trees grows. We also implemented numerous boosting tree methods, such as AdaBoost, Gradient Boosting, Light GBM, and XGBoost, which excel in their ability to handle complex, non-linear relationships and skewed variables by building trees sequentially to correct previous errors. While these algorithms are similar, they also have their distinctive strengths: Light GBM, for example, is renowned for its efficiency and training speed, whereas XGBoost includes regularization features to help prevent overfitting. AdaBoost can handle skewed variables and imbalanced dummies well due to its focus on difficult cases and adaptive weighting, though it may face difficulty in handling outliers. Notably, while tree methods offer significant advantages, they also share a vulnerability of overfitting, which is important to address by careful hyperparameter tuning and specific countermeasures like cross-validation (CV).

To make our approach more diverse, we also explored other models, like SVM, which relies on support vectors to define a separating hyperplane in a higher-dimensional space, making it well-suited for our high-dimensional data. However, its performance may degrade with a large, skewed dataset unless the kernel is chosen carefully. Neural Networks (NN) are flexible and capable of capturing complex patterns but require careful architecture design to avoid overfitting. kNN is a model appealing due to its simplicity - it offers an assumption-free method by using the nearest neighbors for prediction, though it might struggle with high-dimensional data containing irrelevant features. Lastly, Kernel Ridge Regression extends ridge

regression with the kernel trick to model non-linearities but may face challenges with large datasets.

Table 4.1: Strengths and weaknesses of various models in the context of skewed numerical and imbalanced dummy predictors.

Model	Strengths	Weaknesses
RF	Handles skewness and predictor imbalance	Slower with many trees, complex model
AdaBoost	Robust to skewness, adaptive weighting	Sensitive to outliers
GB	Robust to skewness	Overfitting without careful tuning
Light GBM	Robust to skewness, fast, scalable	Overfitting without careful tuning
XGBoost	Robust to skewness, offers regularization	Computationally demanding
SVM	Effective in high-dimensional space	May struggle with size and skewness
NN	Flexible, can capture complex relationships	Requires large data and tuning
kNN	Simple, no distribution assumption	Sensitive to irrelevant features
KRR	Non-linear modelling	Not ideal for very large datasets

To conclude, each model was carefully selected for its potential to offer unique insights into our data. By employing diverse models, we can utilize the distinct strengths and mitigate the inherent weaknesses of individual models, resulting in a comprehensive analysis of educational outcomes. Our approach expands upon methodologies discussed in the literature review. While our foundational paper by Chen and Ding (2023) employs decision tree, Random Forest, logistic regression, SVM, and Neural Network models, we explore a wider variety of algorithms by incorporating additional boosting models, along with kNN and KRR. Similarly, while Masci et al. (2018) combines tree-based methods with traditional econometric approaches, our study extends this by incorporating a wider variety of machine learning models to ensure a diverse evaluation of the data.

Model Training, Validation and Hyperparameter Tuning

Following the selection of our models, we move on to describe our approach to model training, validation, and hyperparameter tuning - essential stages for optimizing model performance and ensuring the accuracy of our predictions on school performance.

We began by partitioning our data into training and test subsets with a 75/25 split ratio, which allows for comprehensive learning while still putting aside a substantial subset for unbiased evaluation of our models on the test set. In addition to the initial split, we further segmented the training data to apply 5-fold CV, aiming to combat overfitting - a condition where a model fits the training data too well

but performs poorly on unseen data. CV is a method that systematically cycles the dataset through multiple training and validation phases, ensuring that each data point is used for both training and validation. This approach not only improves the model's ability to generalize to new data but also maximizes the use of available data for training, providing a robust estimate of model performance.

One efficient method for applying CV is through GridSearchCV, which also performs automated hyperparameter optimization. This technique exhaustively searches the hyperparameter space to find the optimal model configurations by evaluating various combinations on the training set and assessing their performance on the validation set. Given the wide variety of models we utilized, ranging from tree-based algorithms to more complex architectures, each model required specific hyperparameters that we carefully tuned to achieve optimal performance. Careful selection of hyperparameters also helps to control overfitting, as certain models incorporate built-in regularization parameters, while others allow for limiting tree depth or selecting node purity criteria.

In the Random Forest model, we adjusted the number of trees, maximum tree depth, and the criteria for node purity to enhance prediction robustness and manage model complexity. For AdaBoost, we fine-tuned the number of consecutive trees to build and the learning rate while also selecting the most effective loss function that determines how the algorithm weights misclassified data points, optimizing iterative adjustments. The Gradient Boosting model underwent a similar tuning process with particular attention to the loss function, choosing between squared error and Huber loss to find the ideal error correction approach. Light GBM adjustments involved tuning the maximum number of leaves per tree to directly control model complexity, and the L1 regularization term on weights to promote sparser models and improve generalization. For XGBoost, we tuned parameters like gamma, min_child_weight, and colsample_bytree, all of which are aimed at combating overfitting. For SVM, we optimized the penalty parameter C to control the trade-off between achieving lower errors on the training data and minimizing the complexity of the model, and epsilon, which defines a margin of tolerance where no penalty is given to errors, crucial for effectively handling noise in the data. In the kNN model, we tuned the number of neighbors, and the distance metric, which influences how distances are calculated between data points. Lastly, in KRR we focused on the parameter alpha for regularization, the kernel type to enable non-linear modeling, and the gamma parameter to adjust model flexibility. For a full list of tuned hyperparameters, please refer to the Table 5.1.

Another approach was taken towards Neural Networks. While we employed 5-fold CV as with other models, we did not utilize GridSearchCV for hyperparameter optimization due to the different nature of NN. Our architecture comprised a sequence of densely connected layers. The first hidden layer had 256 neurons with ReLu activation, followed by batch normalization to stabilize the learning process and a dropout rate of 10% to reduce overfitting. This configuration was repeated with subsequent hidden layers of 128 and 64 neurons, each followed by batch normalization and dropout. The final hidden layer had 32 neurons followed by a dropout of 10% before leading into a single-neuron output layer for predictions. The model was compiled using the Adam optimizer with mean squared error as the loss function. To avoid overfitting, we implemented early stopping by monitoring the validation loss and terminating the training process if no improvement was noted after 10

epochs. Training was conducted over 100 epochs with a batch size of 32, incorporating a 20% validation split to continuously monitor model performance against unseen data during the training phase.

It is important to note, however, that model complexity comes at a cost of computational demands. Therefore, balancing high model performance with computational efficiency was a key consideration in our process, ensuring the training is effective while still managing the available computational resources and time constraints.

Performance Metrics and Model Evaluation

After discussing model training, validation, and hyperparameter tuning, we now focus on evaluating model performance. This section details the three different types of performance metrics that we employed to assess the effectiveness of our models, discusses the actual vs predicted plots used for visual validation, and describes the model evaluation process.

1. **Training Metric - Mean Squared Error (MSE)**: We selected MSE as the primary metric to guide model optimization by minimizing prediction errors. MSE, detailed in Equation 4.3, measures the average squared difference between predicted and actual values, this way providing a clear measure of model accuracy during the training phase.

2. **Validation Metric - Negative Mean Squared Error (Neg MSE)**: For model validation and hyperparameter tuning with GridSearchCV, we used Neg MSE, an adaptation of MSE that aligns with optimization algorithms designed to maximize outcomes. For neural networks, MSE also supported early stopping mechanisms by monitoring validation loss, thus preventing overfitting during model training.

3. **Evaluation Metrics - Root Mean Squared Error (RMSE) and R-squared ($\mathbf{R^2}$)**: For the final evaluation of all models, we chose RMSE and R^2. RMSE, defined as the square root of MSE, offers direct interpretability by measuring model error in the same units as the target variable. R-squared, defined in equation 4.4, measures the proportion of variance in the dependent variable that is explained by the independent variables, indicating how well our models explain the variability in educational outcomes.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{4.3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{4.4}$$

The variables in the formulas are defined as follows:

- Y_i are the actual values of the dependent variable
- $\hat{Y}_i$ are the predicted values estimated by the model
- $\overline{Y}$ is the mean of all actual values of the dependent variable.

To visually supplement these metrics, we examined actual vs. predicted value plots for each model. These plots provide a straightforward method to assess the accuracy of the models visually, showing how predicted values compare directly with actual outcomes. By aligning the predicted values with actual values on a plot, we can observe the degree of variance from the line of perfect prediction (the diagonal), which further helps in understanding the effectiveness of each model at different data ranges.

The selection of our performance metrics - MSE for training, Neg MSE for validation, and RMSE together with R^2 for final evaluation - was influenced by the skewness and imbalances in our data. Furthermore, implementing Yeo-Johnson transformation and standardization aided in making these metrics more reliable by normalizing data features and mitigating the impact of extreme values. Since we consider the outliers to be important data points rather than typing errors or anomalies, this combination of performance metrics and preprocessing techniques is especially effective for our skewed data because MSE emphasizes large errors, ensuring that the outliers adequately influence the model training process.

Our evaluation process varied slightly across model types. For all models except Neural Networks, GridSearchCV played an important role in tuning and validating the models. After identifying the optimal hyperparameters, these models were trained on the full training dataset and then evaluated on the test set to confirm accuracy and explanatory power. For Neural Networks, our evaluation approach included a validation split to continuously monitor performance and early stopping to prevent overfitting, followed by performance assessment on the test set using RMSE and R-squared. All in all, our evaluation process confirms the effectiveness of the models and highlights their reliability in handling the specific challenges of our dataset.

Interpretation and Explanation

Given the complex, often black-box nature of machine learning models, integrating effective interpretation tools is crucial. These tools help make model predictions understandable and actionable, which is particularly important in decision-making contexts like education. Since our primary goal is to identify key factors affecting exam scores, this section discusses the interpretation tools we utilized, adapted to the specific model types we employed.

Our interpretation framework varied based on the model architecture. For our tree-based models like Random Forest, Gradient Boosting, and others, we utilized direct feature importance measures. These measures are calculated based on the average reduction in the model's prediction error when a feature is used in the trees. Higher importance values indicate that modifying the feature's values significantly alters the model's accuracy, thus highlighting its critical role in affecting predictions. This method provides a first look at which features are driving the model's decisions. For models where direct feature importance is not inherently available, such as SVM, Neural Networks, and others, we used permutation importance. This method involves randomly shuffling individual features and observing the effect on model accuracy. A significant change in model performance upon shuffling a feature indicates its importance in the predictive process. Additionally, permutation importance is model-independent, allowing for the comparison of feature significance

across various model types.

To visualize these importance metrics, we utilized horizontal bar plots, ranking the features by their importance. This visual representation helps in quickly identifying the features that most significantly impact model predictions, facilitating a straightforward comparison across models.

While these interpretation methods provide valuable insights, they also come with limitations. Direct feature importance may misrepresent the influence of highly correlated features, potentially overstating the importance of one feature over another. Permutation importance, while useful for capturing the overall influence of features across models, does not account for interactions between features and can be computationally demanding, especially with large datasets or complex models. Furthermore, both methods generally assume that features influence model performance independently, which might not capture the complex interdependencies in real-world data. However, despite the limitations, we opted for these interpretation methods because they provide clear, actionable insights and are relatively straightforward to implement. As a further improvement to our methodology, advanced techniques like LIME (Local Interpretable Model-agnostic Explanations) or SHAP (SHapley Additive exPlanations) could be used to improve model transparency and offer deeper insights into the nuanced relationships and interactions among features.

All in all, effective interpretation techniques are essential for stakeholders who need to trust and understand machine learning-driven decisions, especially in education where such insights can inform targeted interventions that directly impact student outcomes. By applying these methods, we not only make our models more interpretable and transparent but also enhance the credibility and utility of our machine learning solutions.

Tools

To conclude the methodology, in this section we briefly discuss the key tools employed for data analysis, model construction, and performance evaluation.

Our research primarily utilized Python and R, which offer extensive libraries that are particularly well-suited for machine learning. We used Python as our primary platform, executing our code in Google Colab's cloud-based environment. We relied on several Python libraries: NumPy and Pandas for data manipulation; Matplotlib and Seaborn for visualization; Scikit-learn for modeling and evaluation; and TensorFlow along with Keras for advanced machine learning models. While Python was the key tool that we used, R played a crucial role in creating data visualizations and maps, primarily using the ggplot2 and rgdal packages. This combination of tools and platforms enabled us to effectively handle our complex dataset and implement advanced machine learning algorithms, this way supporting our research objectives.

4.3 Conclusion

In conclusion, our methodology used advanced machine learning techniques to explore the complex factors influencing exam results in Slovak schools. We began our analysis with careful data preprocessing, using the Yeo-Johnson transformation and standardization to align our dataset with the assumptions of advanced modeling techniques. We then selected a diverse array of models, from ensemble tree methods

to Neural Networks, each chosen for its ability to reveal different aspects of the data. Furthermore, we employed hyperparameter tuning and CV to optimize each model's performance, thus enhancing their predictive accuracy and reliability. By integrating advanced modeling techniques with thorough data management, we aimed to obtain results that are statistically robust and practically applicable, contributing meaningful insights to the field of educational data analysis. Looking ahead, our methodology could be further improved by incorporating advanced outlier detection methods to compare model performance with and without extreme values, conducting geospatial analysis to explore regional educational trends, and integrating interpretative techniques such as LIME or SHAP to deepen understanding of feature importance.

5

Results

In the previous section, we provided an extensive overview of the methodology for our thesis, detailing the strategic choices made in model training, validation, and hyperparameter tuning. Building on this foundation, this section evaluates model performance from multiple perspectives to determine their effectiveness and applicability in educational settings.

First, we discuss the results of hyperparameter tuning, focusing on the optimization strategies used for each model. Next, we evaluate model performance by comparing quantitative metrics - RMSE and R^2 - and examining actual vs. predicted value plots to visually assess each model's accuracy and fit. After this quantitative assessment, we evaluate feature importance scores to identify key factors influencing exam results. This analysis not only highlights the most important predictors of educational outcomes but also provides practical insights for improving educational strategies in real-world settings.

Lastly, we discuss how our findings fit into the existing body of research, pointing out where our results agree with or differ from previous studies. We also outline potential improvements and directions for future research that could enhance our evaluation framework and model performance.

5.1 Model performance

Hyperparameter Selection

This section discusses the hyperparameters selected for each model using GridSearchCV, emphasizing our strategic approach to optimizing model performance by carefully balancing accuracy, computational efficiency, and the ability to generalize. The specific parameters chosen are detailed in Table 5.1 and will be discussed in detail below.

Random Forest (RF)

For the Random Forest model, we selected 400 trees (n_estimators: 400). While a higher number of trees generally reduces variance by averaging multiple decision paths, it also increases computational demands and may lead to overfitting. The parameters for minimum samples per split and minimum samples per leaf were set at 2 and 1, respectively, enabling the model to identify subtle differences between data

points. To prevent overfitting while allowing the model to learn detailed patterns, the maximum tree depth was capped at 20.

AdaBoost

The AdaBoost model utilized 50 trees (n_estimators: 50) to achieve a good balance between computational efficiency and model accuracy. A learning rate of 0.1 was chosen to strike a balance between preventing overfitting and allowing the model to capture complex data patterns effectively. The linear loss function was selected for its effectiveness in simplifying the process of weight adjustment and focusing on balanced error reduction.

Gradient Boosting (GB)

Similarly to AdaBoost, Gradient Boosting was configured with 100 trees (n_estimators: 100), and the learning rate of 0.1. Minimum samples per split and per leaf were set at 2 and 5, respectively. Due to overfitting concerns, tree depth was capped at 10. Moreover, the squared error loss function was chosen for its effectiveness in handling outliers.

Light GBM

Similar to the other boosting models, Light GBM configuration included 100 trees (n_estimators: 100) and a learning rate of 0.1. The number of leaves was set to 90, allowing for detailed yet computationally efficient tree growth. The model's depth was unrestricted (max_depth: -1) allowing it to capture intricate details within the data but increasing the risk of overfitting. To counteract this, L1 regularization was applied at a moderate level (reg_alpha: 0.5) to penalize large coefficients. Moreover, the minimum number of samples per child was set to 10 (min_child_samples: 10), ensuring reliable decisions while further mitigating overfitting. Additionally, the column-wise building algorithm (force_col_wise: True) was used to optimize the processing of our high-dimensional dataset.

XGBoost

XGBoost was configured with 300 trees (n_estimators: 300) and a low learning rate (0.01), combined with a maximum tree depth of 15. This setup supported robust learning by enabling detailed tree structures and enhancing the model's ability to capture complex non-linear relationships while mitigating overfitting by limiting tree depth. The gamma value was set to 0, reducing the regularization of leaf nodes and allowing for greater flexibility in constructing tree structures. A minimum child weight of 5 ensured each leaf node represented several observations, reducing sensitivity to individual data points. Additionally, a colsample_bytree value of 0.8 limited the model's complexity by using 80% of the features for building each tree, further protecting against overfitting.

Support Vector Machines (SVM)

For SVM, a penalty parameter C of 70 was selected to control the trade-off between maximizing the margin and minimizing the training error. A higher value of C

leads to a model that prioritizes a close fit to the training data thus increasing the risk of overfitting. A larger epsilon, set to 15, increases the model's tolerance for deviations from the predicted regression line, which can help to achieve a more robust and generalized solution by allowing for certain errors without penalty. The combination of a high C and a large epsilon was chosen to balance achieving a close fit to the data while managing the risks of overfitting. Additionally, the RBF kernel was chosen to handle non-linear data patterns effectively, offering flexibility to capture complex relationships.

K-Nearest Neighbors (kNN)

The kNN model used 20 neighbors to ensure robust averaging of results, with distance weighting to give closer points more influence. A low leaf size of 1 was chosen to optimize the efficiency of the algorithm in searching for nearest neighbors. Additionally, the Manhattan metric was employed, which is particularly effective in high-dimensional data settings by emphasizing differences across individual dimensions and improving the relevance of distance calculations.

Kernel Ridge Regression (KRR)

KRR was configured with an alpha of 0.1, applying mild L2 regularization to balance model complexity with a good fit to the training data. An RBF kernel was used to handle non-linear relationships in the data effectively. As the gamma parameter was not adjusted, it defaulted to a preset value. This default setting helps control the model's sensitivity to data variations, ensuring stability and preventing overfitting without requiring manual tuning.

Neural Networks

For Neural Networks, hyperparameters were not optimized using GridSearchCV. For details on the parameters and architecture chosen for Neural Networks, refer to Model Selection in Section 4.2.

In this section, we explored the strategic selection of hyperparameters for various predictive models using GridSearchCV. This systematic approach enabled us to find an optimal balance between accuracy, computational efficiency, and generalizability to suit our data and objectives. Hyperparameters for each model were adjusted to align with its unique characteristics, enhancing predictive accuracy and minimizing the risk of overfitting. These carefully chosen hyperparameters contributed to the robustness and effectiveness of our predictive models in achieving the research objectives of our thesis.

Model performance evaluation

Following the optimization of hyperparameters, this section evaluates and compares the performance of various machine learning models, linking back to the discussions on model selection in Section 4.2. We assess each model's accuracy and fit to the data using both quantitative performance metrics and visual comparisons.

Table 5.1: Hyperparameter ranges and the best parameters selected through GridSearchCV for various predictive models.

Model	**Hyperparameter Range**	**Parameters Chosen by GridSearch**
RF	n_estimators: [200, 300, 400] max_depth: [10, 20, None] min_samples_split: [2, 10] min_samples_leaf: [1, 2]	n_estimators: 400 max_depth: 20 min_samples_split: 2 min_samples_leaf: 1
AdaBoost	n_estimators: [30, 50, 70] learning_rate: [0.01, 0.1, 1.0] loss: linear, square, exponential	n_estimators: 50 learning_rate: 0.1 loss: linear
GB	n_estimators: [100, 200] learning_rate: [0.1, 1.0] max_depth: [5, 10, None] min_samples_split: [2, 5] min_samples_leaf: [1, 5] loss: squared error, huber	n_estimators: 100 learning_rate: 0.1 max_depth: 10 min_samples_split: 2 min_samples_leaf: 5 loss: squared error
Light GBM	n_estimators: [50, 100, 200] learning_rate: [0.01, 0.1, 0.2] num_leaves: [70, 90, 110] reg_alpha: [0.0, 0.5, 0.7] max_depth: [-1, 5] min_child_samples: [5, 10, 15] force_col_wise: True	n_estimators: 100 learning_rate: 0.1 num_leaves: 90 reg_alpha: 0.5 max_depth: -1 min_child_samples: 10 force_col_wise: True
XGBoost	n_estimators: [100, 200, 300] learning_rate: [0.01, 0.1, 1.0] max_depth: [5, 10, 15] gamma: [0, 0.5] min_child_weight: [1, 5] colsample_bytree: [0.5, 0.8, 1.0]	n_estimators: 300 learning_rate: 0.01 max_depth: 15 gamma: 0 min_child_weight: 5 colsample_bytree: 0.8
SVM	C: [50, 70, 90] epsilon: [5, 10, 15] kernel: linear, poly, rbf	C: 70 epsilon: 15 kernel: rbf
kNN	n_neighbors: [15, 20, 25] weights: uniform, distance leaf_size: [1, 3, 5] metric: euclidean, manhattan, chebyshev, minkowski	n_neighbors: 20 weights: distance leaf_size: 1 metric: manhattan
KRR	alpha: [0.01, 0.1, 1] kernel: linear, poly, rbf gamma: [None, 1, 10]	alpha: 0.1 kernel: rbf gamma: None

For quantitative analysis, we utilized RMSE and R^2 to evaluate the accuracy and explanatory power of our models. The results, presented in Table 5.2, rank the models based on these metrics.

XGBoost demonstrated the best overall performance with the lowest RMSE (21.717) and the highest R^2 (0.360), suggesting it was the most effective at predicting outcomes and explaining variance in the dataset, likely due to its regularization features that prevent overfitting. Light GBM followed closely with an RMSE of 21.950 and an R^2 of 0.346, demonstrating its strong predictive accuracy and efficient data handling capabilities. Random Forest and Gradient Boosting also showed strong performances, reflecting their ability to manage skewness and imbalances by aggregating outputs from multiple decision trees. In contrast, KRR, SVM, Neural Networks, and kNN, despite careful preprocessing and model-specific optimizations, demonstrated moderate performance, consistent with the challenges of our high-dimensional and skewed dataset. AdaBoost struggled the most, with the highest

RMSE (24.870) and the lowest R^2 (0.161), likely due to its sensitivity to outliers.

Table 5.2: Comparison of Machine Learning Models

Model	**RMSE**	**R^2**
XGBoost	21.717	0.360
Light GBM	21.950	0.346
Random Forest	22.016	0.342
Gradient Boosting	22.036	0.341
KRR	23.132	0.274
SVM	23.165	0.272
Neural Networks	23.326	0.262
kNN	23.336	0.261
AdaBoost	24.870	0.161

Note: Models are trained with 5-fold CV and ranked by performance based on RMSE and R^2. Lower RMSE and higher R^2 indicate better performance.

Additionally, to complement our quantitative metrics, we conducted visual comparisons of actual versus predicted values for each model, providing an intuitive representation of their accuracy, as shown in Figure 5.1.

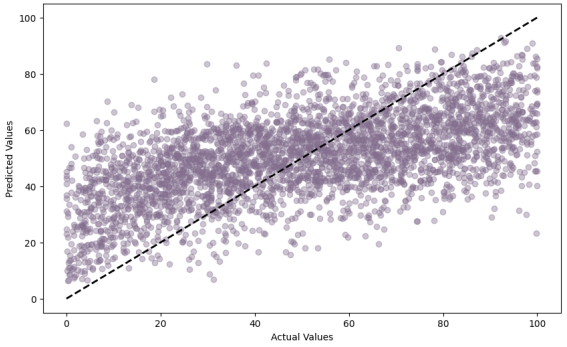
XGBoost, Light GBM, and Random Forest demonstrated tight clustering around the diagonal, exhibiting high accuracy and consistent performance across different value ranges. Gradient Boosting also showed good alignment along the diagonal, though with slightly more scatter. Nevertheless, the plots emphasize the robust predictive accuracy of our ensemble tree methods. In contrast, SVM and KRR displayed reasonable scatter but generally maintained alignment with the diagonal, suggesting solid but less precise predictive accuracy. Similarly, kNN showed comparable performance with good clustering around the diagonal. Neural Networks presented more variability in predictions across all value ranges, suggesting potential challenges in model calibration or issues related to overfitting. AdaBoost consistently underestimated or overestimated values, evidenced by horizontal patterns deviating from the diagonal. This suggests difficulties in modeling complex patterns, which might be caused by the model's sensitivity to outliers. These visual evaluations support our quantitative results and provide valuable insights into each model's strengths and limitations, emphasizing the necessity for careful model selection and hyperparameter optimization, especially in handling high-dimensional and skewed datasets.

In conclusion, our findings confirm expectations about the effectiveness of ensemble tree methods in handling non-linear characteristics, skewness, and imbalances in the dataset. This effectiveness was supported by meticulous preprocessing, strategic hyperparameter tuning, and 5-fold cross-validation - all key elements emphasized in our methodology. Among the models evaluated, our analysis highlighted XGBoost as the most effective model, showcasing good predictive capabilities and adaptability across various dataset characteristics.

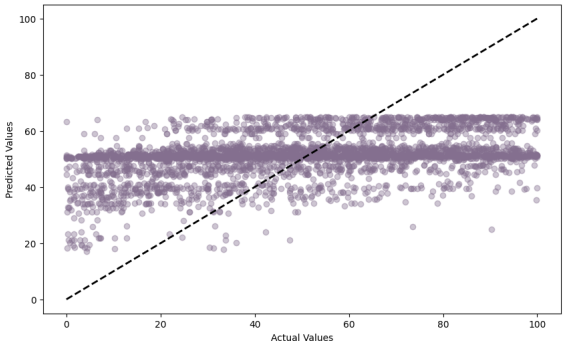
This analysis not only reflects the outcomes anticipated from our methodological setup but also highlights areas for future research improvements. Firstly, the potential for overfitting within some models emphasizes the need for deeper investigation.

Figure 5.1: Actual vs Predicted Values Graphs

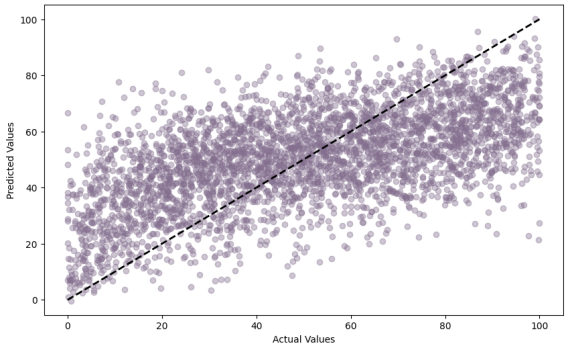
(a) Random Forest - Actual vs Predicted Graph



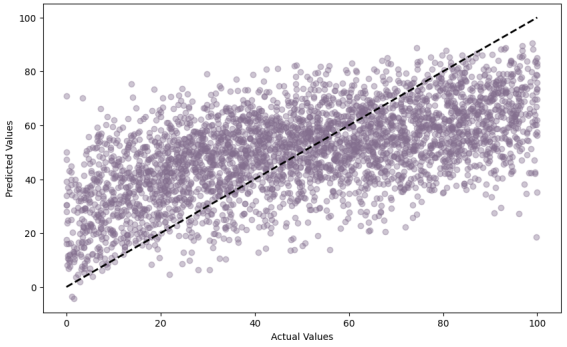
(b) AdaBoost - Actual vs Predicted Graph



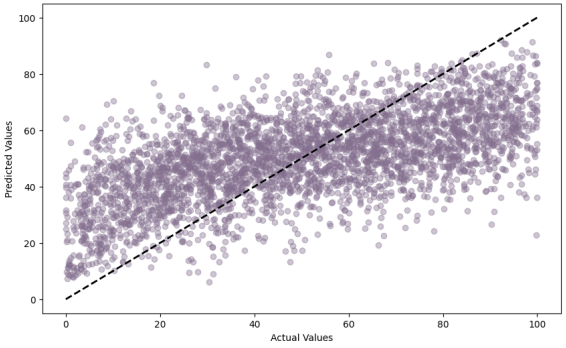
(c) Gradient B. - Actual vs Predicted Graph



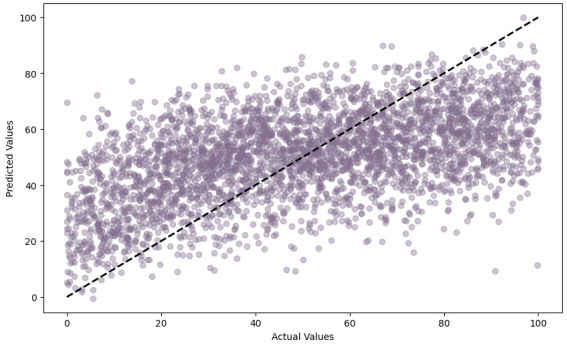
(d) Light GBM - Actual vs Predicted Graph



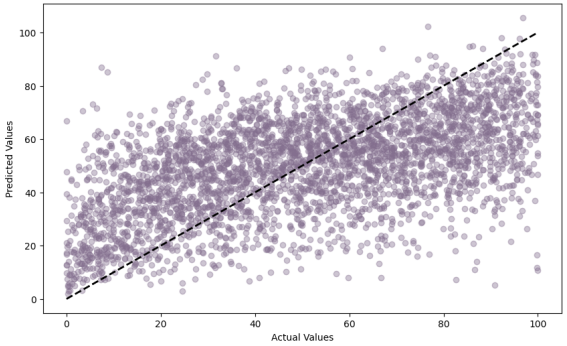
(e) XGBoost - Actual vs Predicted Graph



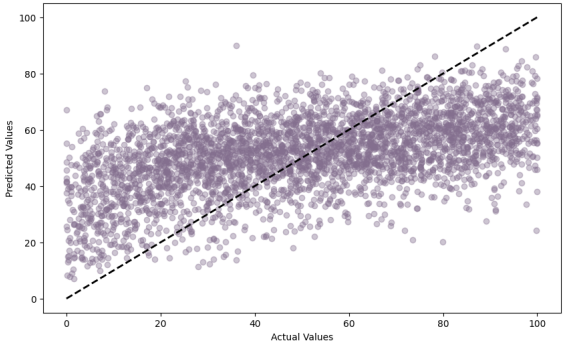
(f) SVM - Actual vs Predicted Graph



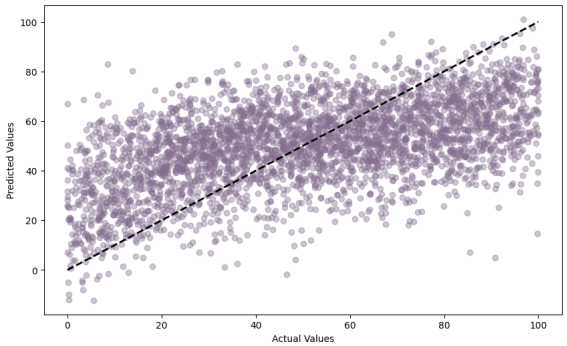
(g) Neural Nets - Actual vs Predicted Graph



(h) kNN - Actual vs Predicted Graph



(i) KRR - Actual vs Predicted Graph



Second, it is important to note that our comparisons rely on observed metrics without statistical validation due to time constraints. Integrating statistical tests like ANOVA can be used to compare the performance metrics of different models to assess if the differences in performance are statistically significant or simply due to random variations in data. Lastly, future research could focus on advancing ensemble techniques by integrating different algorithms, exploring additional hyperparameters, and further investigating deep learning models. All of these suggestions could lead to improvements in predictive accuracy and model interpretability, this way enhancing our methodology.

5.2 Feature importance

After discussing model performance and comparing their effectiveness, this section explores the importance of various features derived from predictive models, aiming to highlight the factors that are most impactful in influencing educational outcomes. Our analysis employs two main methods, as outlined in the Methodology section (refer to Model Interpretation and Explanation in Section 4.2): direct feature importance for tree-based models and permutation importance for non-tree models.

This section is structured into three distinct parts: an initial analysis of top-performing models, followed by an examination of less optimal models, and concluding with a comprehensive analysis across all models. The results are presented in Table 5.3, which provides a comparative ranking of feature importance across the models used in our study. The table is organized with features listed vertically and models horizontally and includes R-squared scores to facilitate easier comparison of model performance. Each cell contains a numerical ranking of the feature's relative importance within that model; a ranking of '1' indicates the highest influence on prediction outcomes for that model, with higher numbers indicating decreasing importance. Features with negligible impact are not ranked and are marked with a dash.

Additionally, we have conducted a visual analysis, providing feature importance plots for each model. These plots visually represent the relative importance of each feature, further illustrating how different attributes impact model predictions. For a detailed view of these graphical representations, please refer to Figure A.1 in the Appendix.

Feature Importance in Top-performing Models

We begin by discussing the feature importance for the top-performing models: XGBoost, Light GBM, Random Forest, and Gradient Boosting. Each model highlighted different factors significantly influencing educational outcomes.

XGBoost

Our top-performing model identified *Higher Educated Population* as the most significant predictor. It uniquely prioritized *Private Dummy* as the second most crucial factor, and ranked *SES Pupils* as the third most important feature. This highlights the model's sensitivity to educational attainment, the type of schooling, and socioeconomic background. Interestingly, XGBoost assigned relatively high significance

to the dummy variables, while other models either ranked these variables lower or found them completely insignificant.

Light GBM

Contrary to the other models, Light GBM assigned the highest importance to *Pupils*. Furthermore, it ranked *Ratio of Teachers* second and *School Canteens* third. The features picked by Light GBM as the most significant indicate a focus on school size, educational resources, and infrastructure.

Random Forest and Gradient Boosting

These models provided very similar feature importance scores, with the top five features being identical, and others in slightly varying order. Like XGBoost, these models rated *Higher Educated Population* as the most important feature. Both also placed *Ratio of Teachers* and *Pupils* as the second and third most important features, respectively, highlighting the critical roles of teacher availability and school size in predicting educational outcomes.

Across all top-performing models, a consistent emphasis on higher education within the town suggests a strong link between educational attainment in the region and exam scores. The importance of *Pupils* and *Ratio of Teachers* across various models (except for XGBoost, which ranks these features lower) indicates a shared valuation of school and class sizes. Variations in the importance assigned to features such as *ICT* and *SES Pupils* highlight differing model sensitivities, which may guide their application in specific educational contexts.

Feature Importance in Less Optimal Models

While the top-performing models provided valuable insights into the predictive factors influencing educational outcomes, exploring the feature importance of models that performed less optimally can provide insights into alternative predictors and offer a broader perspective on the factors influencing educational outcomes that may not be as important in the leading algorithms.

AdaBoost

Consistent with the top-performing models, AdaBoost identified *Higher Educated Population* as the most important feature. It placed substantial emphasis on *SES Pupils* and *Out-of-Wedlock Births*, suggesting that family dynamics and socio-economic factors impact education.

SVM

Contrary to top-performers, this model positioned *Town Population* as its most important feature. It also emphasized *Pupils* and the ethnic composition of student populations, indicating its sensitivity to the demographic and social contexts in predicting exam scores.

Neural Networks

Neural Networks ranked *Pupils* and *Town Population* as top features, reflecting a strong alignment with SVM. Additionally, it shared SVM's emphasis on ethnic demographics, such as *Town Roma Population* and others

kNN

kNN stood out by ranking *Ratio of Teachers* as its top feature, suggesting it values teacher availability most out of all the models. It also ranked *Pupils* and *SES Pupils* highly, suggesting a focus on school size and socioeconomic status.

KRR

This model shared similarities with SVM and NN in valuing the same features for the top 4. Particularly, it agreed with SVM on *Town Population* being the most important feature. Interestingly, it placed the highest significance on *Town Religious Population* out of all the models, indicating consideration of religious contexts in its predictions.

These models illustrate that town population, along with broader societal influences such as ethnic composition and cultural contexts, are significant predictors, providing a complementary perspective to the top-performing models. However, despite highlighting these different features, most models consistently identify higher educated population, ratio of teachers, and number of pupils as the most influential features. This indicates a strong agreement between all algorithms on the importance of these key features.

Comprehensive Analysis of Feature Importance

Having discussed the most important features for each model separately, we now present a comprehensive analysis of feature importance across various machine learning models. This analysis reveals which attributes consistently play crucial roles, which are less influential, and which display varying degrees of influence in predicting educational outcomes, helping to inform strategic model application and educational policy development.

Top Influential Features Across Models

Higher Educated Population

This feature consistently ranked highly across almost all models, particularly XGBoost, Random Forest, Gradient Boosting, and AdaBoost, indicating its strong predictive power. This consistency suggests that the level of educational attainment in a region is crucial for predicting educational outcomes. Higher educational attainment can positively influence school exam scores by creating an environment that supports academic achievement, provides role models, and encourages a culture that values education. Additionally, regions with higher educational attainment would likely have more highly skilled teachers, leading to better educational outcomes. Conversely, less educated regions may experience a shortage of teachers, especially in STEM fields, which would adversely affect education.

Pupils

The size of the student body is a critical feature in most of the models, but especially emphasized by Light GBM, SVM, and Neural Networks, where it ranked as the most significant or near the top. This suggests that the scale of educational institutions impacts their effectiveness, potentially due to resource allocation challenges and the ability to provide individual attention.

Ratio of Teachers

Frequently appearing as a top feature in most models, particularly Random Forest, Gradient Boosting, Light GBM, and kNN, this indicates the importance of teacher availability relative to student numbers. It aligns with educational theories advocating for smaller class sizes, which enable more personalized instruction and, consequently, lead to better educational outcomes.

Features with Varying Influence

SES Pupils

The socioeconomic status of pupils showed significant variability in its impact, ranking highly in models like XGBoost, kNN, and AdaBoost but less so in others.

ICT

The importance of ICT varied across models as well. For instance, it ranked higher in Light GMB, kNN, Gradient Boosting, and Random Forest, while being insignificant or very close to the bottom of the list for others.

Ethnic demographics

Variables such as *Town Roma Population*, *Town Slovak Population*, and *Town Hungarian Population* also showed variability in importance across different models. *Town Roma Population* ranked in the top 5 for some models but was low in others like Light GBM and XGBoost. *Town Slovak Population* was notably important in models like SVM and Neural Networks but less so in others. Similarly, *Town Hungarian Population* ranked highly in Light GBM and kNN but was less significant in other models.

Least Influential Features

School libraries and Extracurricular Institutions

These features generally ranked low in influence across the models. This lower ranking could indicate that, while these factors contribute to a rich educational environment, their direct impact on measurable educational outcomes may not be as significant as other more directly linked educational factors such as teacher-to-student ratios or school size. This suggests that models may not fully capture the nuanced benefits these resources provide, such as fostering lifelong learning habits or improving student engagement, which may not immediately result in higher academic performance.

Unemployment, Divorces per New Marriage, Crime Rates

These community factors generally appeared less influential across the models. This might suggest that while they affect the broader socio-economic environment, their direct impact on educational outcomes is limited compared to more immediate educational factors.

Table 5.3: Feature Importance Ranking Across Models

Model (R^2)	**RF** (0.342)	**AdaB.** (0.161)	**GB** (0.341)	**L.GBM** (0.346)	**XGB** (0.360)	**SVM** (0.272)	**NN** (0.262)	**kNN** (0.261)	**KRR** (0.274)
Higher Educated Population	1	1	1	9	1	3	3	6	3
Ratio of Teachers	2	4	2	2	12	4	4	1	4
Pupils	3	5	3	1	11	2	1	2	2
Town Population	4	6	4	8	8	1	2	12	1
SES Pupils	5	2	5	16	3	8	9	3	12
Out-of-Wedlock Births	6	3	7	11	9	13	12	14	10
ICT	7	-	9	4	20	16	18	5	16
Town Roma Population	8	7	6	13	10	5	5	4	5
School Canteens	9	-	8	3	13	15	11	-	9
Town Religious Population	10	-	10	10	18	7	7	-	6
Town Slovak Population	11	-	11	7	17	6	6	8	7
Town Hungarian Population	12	-	12	6	15	10	8	7	8
Unemployment	13	-	14	5	19	11	14	-	15
School Libraries	14	-	13	12	16	14	13	9	13
Divorces per New Marriage	15	-	17	15	23	19	-	-	-
Youth Crime Rates per Thousand	16	-	15	14	21	-	17	-	18
Crime Rates per Thousand	17	-	18	17	22	18	-	-	-
Extracurricular Institution	18	-	16	18	14	12	15	-	14
Bilingual Dummy	-	-	-	-	4	17	16	13	17
Private Dummy	-	-	-	-	2	-	-	-	-
Public Dummy	-	-	-	-	7	-	-	10	-
Hungary Dummy	-	-	-	-	6	9	10	-	11
Church Dummy	-	-	-	-	5	-	-	11	-

Note: Each cell represents the rank of importance for the feature in each model. We do not report ranking if the variable is too insignificant - significance less than 10% compared to the value of the most important variable.

In conclusion, this analysis has revealed the complexity of factors influencing educational outcomes. The consistent significance of *Higher Educated Population*, *Pupils*, and *Ratio of Teachers* across various models highlights their critical impact on academic performance. This indicates that the immediate educational environment and resources within schools play a more significant role in shaping educational outcomes than broader community socio-economic conditions, with the notable exception of higher education within the community.

However, the variability in the importance assigned to socio-economic, infrastructural, and ethnic demographic features across different models points to the need for strategic focuses that accommodate specific regional and demographic contexts. While our models consistently highlight the significance of school-specific factors and educational attainment, further research is needed to better understand the impact of socioeconomic factors and demographic compositions. This research could help develop targeted policies aimed at improving educational outcomes.

5.3 Discussion

This section connects our analysis to its broader implications for educational policy and model selection. We begin by aligning our findings with the existing literature, highlighting both consistencies and deviations to validate or challenge prevailing

views on educational outcomes. This is followed by a discussion on the implications for educational policy and model selection in educational research. Lastly, we address the limitations of our study and propose directions for future research to further explore the factors influencing educational outcomes.

Synthesis of Findings

Integrating our findings with existing literature reveals several interesting patterns and deviations that contribute to a broader understanding of educational outcomes. Our study confirms the impact of socioeconomic factors, specifically higher educated populations and the socioeconomic status of pupils, aligning with the conclusions of Britton and Propper (2016), Hanushek and Rivkin (2007), Carlisle and Murray (2015) and Amini et al. (2015). These factors consistently emerged as top predictors across multiple models, with higher educated populations being universally significant and socioeconomic status being important for top-performing models, underscoring their critical role in influencing educational performance. This consistency with prior research reinforces the importance of socioeconomic context in educational outcomes and highlights the robustness of these predictors across different methodological approaches.

However, our results challenge some of the established views in the literature. For instance, while Bouck (2018) emphasized the significant impact of geographical settings (urban vs. rural), our models showed that town population and specific demographic variables, such as town Roma population, were not uniformly significant across all models. This divergence suggests that while geographical factors do play a role, their influence might be context-dependent and less universally applicable. Additionally, the varying importance of features like ICT and extracurricular institutions suggests that these factors, although beneficial, may not directly translate into measurable academic performance improvements as consistently as other factors such as teacher-student ratios or school size.

Comparing our RMSE and R^2 results can be challenging. This is mainly because many of the reference studies in Table 2.1 are classification studies that report accuracy, making them not directly comparable. Another factor is the unit of prediction; studies that use an individual as a unit typically have better results than those using schools. However, our results can be compared to some studies, such as Masci et al. (2018), which reported a proportion of variability explained ranging from 14 to 60, depending on the country of measurement. Although the measure used in this study is not directly comparable to ours, the results are still relevant to our best models. Our top-performing model, XGBoost, achieved an RMSE of 21.717 and an R^2 of 0.360. While these figures may be slightly worse, they still demonstrate substantial predictive power. These discrepancies may be due to differences in data characteristics or model configurations. For instance, studies utilizing large-scale individual-level data might capture more nuanced patterns, leading to higher accuracy. Furthermore, the chosen best models and their most important features varied significantly across studies, with some emphasizing demographic factors while others focused on school infrastructure and teacher quality. This highlights the complex and multifaceted nature of educational performance determinants, suggesting that model selection and feature importance can vary greatly depending on the specific context and dataset.

Implications for Educational Policy and Model Selection

The insights gained from analyzing feature importance across models are invaluable for informing educational policies and interventions. As discussed in Section 5.2, key predictors such as *Higher Educated Population*, *Pupils*, and *Ratio of Teachers* significantly impact academic performance. Higher education consistently emerges as a strong predictor because educated communities create an environment that supports academic success by providing role models, highly skilled teachers, and a strong emphasis on the value of education. The number of students in a school is an important predictor because it affects resource allocation and individual attention, with smaller student bodies likely enabling better interactions and more personalized support. Lastly, a higher teacher-to-student ratio allows for more individualized instruction, leading to improved educational outcomes. Consequently, the findings of our study lead to several key recommendations for educational policy:

1. **Focus on Educational Attainment:** Policies should aim to raise educational levels within communities. This could involve adult education programs, parental engagement initiatives, and community support for education, creating an environment that enhances overall educational outcomes for children. Additionally, prioritizing higher education can lead to more qualified teachers, further improving student outcomes.

2. **Optimize School Size and Teacher Ratios:** It is important for schools to maintain manageable sizes and optimal teacher-student ratios. Therefore, policies could include funding for hiring more teachers, reducing class sizes, and improving teacher training. Additionally, offering ongoing professional development courses for teachers to stay updated with the latest teaching methodologies can further enhance the quality of education.

3. **Improve technological access:** Although *ICT* was not important in all models, it was relatively significant in our top-performing algorithms. Therefore, policies should aim to integrate technology in classrooms to enhance learning and support personalized education. This could include investing in digital tools, training teachers to use technology effectively, and making sure all students have access to necessary technological resources.

4. **Customized Strategies for Socio-Economic and Demographic Contexts:** The varying importance of socio-economic and ethnic demographic features across models suggests the need for further research. However, our evidence still supports the need for region- and demographic-specific strategies. For instance, policies in socio-economically disadvantaged areas could focus on meeting basic needs and re-evaluating support systems.

While these suggestions can help policymakers and educational leaders develop strategies that target the most important factors, it is also crucial to consider potential challenges such as budget constraints, resistance to change, and difficulties in hiring and training more teachers. These issues can be addressed through strategic planning and effective communication of the benefits of these policies. Additionally, it is important to regularly monitor and evaluate the implemented policies to assess

their effectiveness and, if necessary, make adjustments, since keeping policies flexible to adapt to changing educational needs and environments helps maintain their relevance and effectiveness over time.

In addition to these policy recommendations, our findings have significant implications for model selection in educational research. Since the importance of different features varies across algorithms, using multiple models can provide a more comprehensive understanding of the factors influencing exam scores. For example, ensemble modeling techniques can offer a wider range of insights by combining the strengths of individual models. This approach helps address both direct and indirect factors affecting student success, ensuring more robust and reliable predictions and leading to more effective educational strategies.

The implications for educational policy and model selection derived from this analysis highlight the importance of targeted interventions and strategic resource allocation to improve educational outcomes. By utilizing insights gained from multiple models and focusing on key predictive factors, policymakers and educational leaders can develop more effective strategies to address educational inequalities and enhance student performance across diverse contexts.

Limitations and Future Research

To provide a clear understanding of our findings and offer directions for future research, this section acknowledges the limitations of our analysis and aims to guide further exploration and validation of the factors influencing educational outcomes. Throughout our thesis, we have identified various limitations and future research directions, discussing them in detail within each section. This summary consolidates these points to emphasize their importance.

1. **Sample Size and Data Quality**

   The dataset used in this study, while comprehensive, is limited to schools in Slovakia and may not fully represent global educational contexts. Additionally, data quality issues such as missing values and potential reporting inaccuracies by the schools could have influenced the results. Therefore, investigating why some schools fail to report ICT usage and teacher-student ratios could help ensure the highest data quality. Furthermore, future research in different country contexts could deepen the understanding of factors influencing educational outcomes.

2. **Model Limitations**

   This study employed a wide range of machine learning models, each with its strengths and weaknesses; however, no single model can capture the full complexity of educational outcomes. Additionally, the potential for overfitting, despite careful hyperparameter tuning, remains a concern and requires deeper investigation. Future research could explore integrating additional models, such as deep learning techniques, and investigating a broader range of hyperparameters to enhance predictive accuracy and robustness. Additionally, exploring ensemble model techniques could offer a more holistic view of the factors influencing educational outcomes.

3. **Feature Selection and Importance**

   Our study highlighted the significance of certain features, such as a higher educated population, pupil numbers, and teacher ratios. However, the variability in the importance of socio-economic, infrastructural, and ethnic demographic features across models suggests that further research is needed to better understand these dynamics. Future studies should investigate these less consistently important features to uncover hidden impacts or interactions not immediately apparent through single-model analyses. Additionally, longitudinal studies that capture changes over time could offer more targeted insights for policy interventions by revealing how educational outcomes and influencing factors evolve.

4. **Methodological Improvements**

   While our study utilized advanced machine learning techniques, there is always room for methodological improvement. Future research could incorporate more sophisticated outlier detection methods and geospatial analysis to explore regional educational trends. Advanced interpretative techniques like LIME and SHAP could offer deeper insights into feature importance and interactions. Additionally, statistical tests such as ANOVA could be used to compare the performance metrics of different models to determine the statistical significance of differences in model performance.

   Addressing these limitations and exploring the suggested areas for future research could lead to improving the understanding of educational outcomes and creating more effective educational policies and interventions. By enhancing the quality of the data, integrating insights from multiple predictive models, exploring the complexities of socio-economic and demographic factors, and refining methodology, future studies can provide a more holistic view of the factors that influence educational success.

All in all, this discussion has compared our study's findings against existing literature, revealing both consistencies and deviations that contribute to a deeper understanding of educational outcomes. We have highlighted significant implications for educational policy and model selection, emphasizing the need for targeted interventions and methodological advancements. Additionally, we addressed the limitations and future research directions that have the potential to enhance the robustness of educational analyses and support the development of more effective policies. Our study contributes valuable insights to the field, highlighting the importance of comprehensive and context-specific approaches in educational research.

5.4 Conclusion

In this comprehensive discussion of the results, we have carefully analyzed model performance, examining hyperparameter selection, model performance metrics, and feature importance to uncover the key factors influencing educational outcomes. Our study highlighted the strong predictive power of higher educated population, number of pupils, and the student-teacher ratio, emphasizing the critical role of these features in shaping academic performance. The variability in the importance

of socio-economic, infrastructural, and ethnic demographic features across different models emphasized the complexity of educational outcomes and the need for further research to fully understand these dynamics.

Our findings align with existing literature on the significance of educational attainment within the population, while challenging some established views regarding geographical settings and the direct impact of certain infrastructural elements. These insights have significant implications for educational policy, suggesting targeted interventions to raise educational attainment, optimize school sizes and student-teacher ratios, and effectively integrate technology. Furthermore, the variability observed in feature importance across models indicates the necessity for customized strategies tailored to specific regional and demographic contexts.

We have also identified several limitations in our study, including the representativeness of the dataset and the potential for overfitting in our models. Addressing these limitations through future research in different country contexts, integrating additional models such as deep learning techniques, and employing more sophisticated methodologies could further enhance the understanding of educational outcomes.

In summary, our study highlights the key factors that influence educational success, emphasizing the need for targeted, evidence-based educational policies. By combining findings from multiple models and focusing on key predictive factors, policymakers and educators can create more effective strategies to improve student performance. Our research highlights the importance of comprehensive and context-specific approaches in education, paving the way for future studies to build on these findings and help develop better educational policies and interventions.

6

Conclusion

This study aimed to identify the key factors influencing academic performance in Slovak schools and evaluate the effectiveness of machine learning techniques in predicting educational outcomes. We explored how machine learning techniques can capture the complex, non-linear relationships inherent in educational data, where traditional statistical approaches might fall short. Utilizing an extensive dataset from 1,409 primary schools and 656 secondary schools, which included information on school characteristics, socio-economic status, ICT usage, demographic factors, and more, we applied various machine learning models. These models included Random Forest, Gradient Boosting, Light GBM, XGBoost, Support Vector Machines, Neural Networks, K-Nearest Neighbors, and Kernel Ridge Regression.

Our methodology involved several complex processes, including data collection and careful preprocessing to handle missing values, standardize variables, and transform skewed distributions to make them more normal. We meticulously selected a variety of models, each chosen for its ability to reveal different aspects of the data. These models were trained and validated using cross-validation techniques to ensure robust and reliable predictions, with hyperparameters tuned to find optimal configurations and mitigate overfitting. To evaluate model performance, we chose a range of metrics, including quantitative measures and actual vs. predicted value plots. Given that machine learning models are often considered black boxes, we integrated effective interpretation tools, such as feature importance scores from tree models and permutation importance for other algorithms, to provide insights into their decision-making processes.

Our findings demonstrate that ensemble tree methods, particularly XGBoost, consistently outperform other models in terms of predictive accuracy. The models we utilized identified the higher educated population in the region, the ratio of teachers to students, and the number of pupils in a school as the most significant predictors of academic performance. This confirms the importance of educational attainment within the community, teacher availability, and school size in shaping educational outcomes.

In light of these results, we have successfully answered our research question, demonstrating that advanced machine learning techniques can effectively predict academic performance and highlight critical factors that influence learning outcomes. Our findings have significant implications for educational policy, suggesting that efforts should focus on raising educational attainment within communities, optimizing school sizes and teacher-student ratios, and integrating technology effec-

tively to enhance learning. Moreover, the variability in feature importance across different models indicates the need for context-specific strategies that take regional and demographic characteristics into account. This variability highlights the complex and multifaceted nature of educational performance determinants, indicating that a one-size-fits-all approach may not be effective.

The findings of our thesis align with existing literature in several key areas. Consistent with some studies, we found that socio-economic factors, particularly the level of education within the community, play a crucial role in academic performance. However, our study challenges some established views that emphasize the significant impact of urban versus rural settings. The results suggest that while geographical factors are important, their influence might be context-dependent and less universally significant.

In addition, several limitations must be acknowledged to provide a clear understanding of our findings and offer directions for future research. The dataset, while comprehensive, is limited to Slovak schools and may not fully represent global educational contexts. Data quality issues, such as missing values and potential reporting inaccuracies, could also influence the results. Additionally, despite careful hyperparameter tuning, the potential for overfitting remains a concern. Future research should explore different country contexts, integrate additional models such as deep learning techniques or ensemble methods, and employ more sophisticated methodologies to enhance predictive accuracy and robustness.

In conclusion, this research contributes valuable insights into the factors that influence educational outcomes, emphasizing the need for targeted, evidence-based policies to address educational inequalities and improve student performance. By combining insights from multiple models and focusing on critical predictive factors, policymakers and educators can develop more effective strategies to enhance educational success. Future studies should continue to explore these dynamics and build on our findings to support the creation of more effective educational policies and interventions. Through ongoing research and methodological advancements, a better understanding of the diverse factors impacting education can be achieved, enabling the development of comprehensive solutions to improve educational outcomes worldwide.

Bibliography

Afonso, M. G. and Aubyn, M. S. (2016). Early, Late or Never? When Does Parental Education Impact Child Outcomes? *Econometric Modeling: Microeconometric Studies of Health*. http://dx.doi.org/10.2139/ssrn.2203273.

Agasisti, T. (2009). The Efficiency of Italian Secondary Schools and the Potential Role of Competition: a Data Envelopment Analysis using OECD-PISA 2006 Data. *Education Economics*, 21:520–544. https://doi.org/10.1080/09645292.2010.511840.

Amini, M., Nasrabadi, H. B., and Heydari, M. (2015). Education and Social Capital. *Research on humanities and social sciences*, 5:98–104. https://www.semanticscholar.org/paper/Education-and-Social-Capital-Amini-Nasrabadi/c935ff03952497a04049f66d4304f8cd75275c1e.

Anuradha, C., Velmurugan, T., and Velmurugan, T. (2015). A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance. *Indian Journal of Science and Technology*. https://doi.org/10.17485/ijst/2015/v8i15/74555.

Bouck, E. C. (2018). How Size and Setting Impact Education in Rural Schools. *The Rural Educator*. https://doi.org/10.35608/RURALED.V25I3.528.

Britton, J. and Propper, C. (2016). Teacher Pay and School Productivity: Exploiting Wage Regulation. *Journal of Public Economics*, 133:75–89. https://doi.org/10.1016/j.jpubeco.2015.12.004.

Carlisle, B. L. and Murray, C. (2015). Effects of Socio-Economic Status on Academic Performance. *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, pages 43–48. https://doi.org/10.1016/B978-0-08-097086-8.23054-7.

Carlos, F., Rodríguez-Hernández, Musso, M., Kyndt, E., and Cascallar, E. (2021). Artificial Neural Metworks in Academic Performance Prediction: Systematic Implementation and Predictor Evaluation. *Computers and Education: Artificial Intelligence*, 2:100018. https://doi.org/10.1016/j.caeai.2021.100018.

Chen, S. and Ding, Y. (2023). A Machine Learning Approach to Predicting Academic Performance in Pennsylvania's Schools. *Social Sciences*, 12(3):118. https://doi.org/10.3390/socsci12030118.

Cherchye, L., Witte, K., Ooghe, E., and Nicaise, I. (2010). Efficiency and Equity in Private and Public Education: A Nonparametric Comparison. *Eur. J. Oper. Res.*, 202:563–573. https://doi.org/10.1016/j.ejor.2009.06.015.

Chung, J. Y. and Lee, S. (2019). Dropout Early Warning Systems for High School Students using Machine Learning. *Children and Youth Services Review*, 96:346–353. https://doi.org/10.1016/j.childyouth.2018.11.030.

Cornell-F., S. and Garrard, R. (2020). Machine Learning Classifiers Do Not Improve the Prediction of Academic Risk: Evidence from Australia. *Communications in Statistics: Case Studies, Data Analysis and Applications*, 6(2):228–246. https://doi.org/10.1080/23737484.2020.1752849.

Cruz-Jesus, F., Castelli, M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M., and Rosa-Louro, A. (2020). Using Artificial Intelligence Methods to Assess Academic Achievement in Public High Schools of a European Union Country. *Heliyon*, 6(6). https://doi.org/10.1016/j.heliyon.2020.e0408110.

Fredriksson, P., Ockert, B., and Oosterbeek, H. (2013). Long-Term Effects of Class Size. *European Economics: Labor and Social Conditions eJournal*. http://dx.doi.org/10.2139/ssrn.1906182.

Geodetic and Cartographic Office of Bratislava (2022). Shapefiles of Slovakia. https://www.geoportal.sk/en/zbgis/download/.

Hanushek, E. and Rivkin, S. G. (2007). Pay, Working Conditions, and Teacher Quality. *The Future of Children*, 17:69–86. https://doi.org/10.1353/foc.2007.0002.

Harvey, J. L. and Kumar, S. A. (2019). A Practical Model for Educators to Predict Student Performance in K-12 Education using Machine Learning. In *2019 IEEE symposium series on computational intelligence (SSCI)*, pages 3004–3011. IEEE. https://doi.org/10.1109/SSCI44817.2019.9003147.

Hilbert, S., Coors, S., Eb, K., Bischl, B., Frei, M., Lindl, A., Wild, J., Krauss, S., Goretzko, D., and Stachl, C. (2021). Machine Learning for the Educational Sciences. *Review of Education*. https://doi.org/10.31234/OSF.IO/3HNR6.

Institute for Economic and Social Reforms (2023). Ranking for Primary and Secondary Schools. https://skoly.ineko.sk/metodika/.

Jordan, M. I. and Mitchell, T. (2015). Machine Learning: Trends, Perspectives, and Prospects. *Science*, 349:255–260. https://doi.org/10.1126/science.aaa8415.

Khan, A. and Ghosh, S. K. (2021). Student Performance Analysis and Prediction in Classroom Learning: A Review of Educational Data Mining Studies. *Education and Information Technologies*, 26(1):205–240. https://doi.org/10.1007/s10639-020-10230-3.

Khan, M. I., Khan, Z. A., Imran, A., Khan, A. H., and Ahmed, S. (2022). Student Performance Prediction in Secondary School Education Using Machine Learning. pages 94–101. https://doi.org/10.1109/ITT56123.2022.9863971.

Korkmaz, C. and Correia, A. (2019). A Review of Research on Machine Learning in Educational Technology. *Educational Media International*, 56:250–267. https://doi.org/10.1080/09523987.2019.1669875.

Masci, C., Johnes, G., and Agasisti, T. (2018). Student and School Performance across Countries: A Machine Learning Approach. *European Journal of Operational Research*, 269(3):1072–1085. https://doi.org/10.1016/j.ejor.2018.02.031.

Ministry of Interior of the Slovak Republic (2023). Crime Rates in Slovakia.

Mousa, H. and Maghari, A. (2017). School Student's Performance Prediction using Data Mining Classification. *International Journal of Advanced Research in Computer and Communication Engineering*, 6(8):136–141. https://www.researchgate.net/publication/321920848_School_Students'_Performance_Predication_Using_Data_Mining_Classification.

Nafea, I. (2018). Machine Learning in Educational Technology. *Machine Learning - Advanced Techniques and Emerging Applications*. https://doi.org/10.5772/intechopen.72906.

Naicker, N., Adeliyi, T., and Wing, J. (2020). Linear Support Vector Machines for Prediction of Student Performance in School-Based Education. *Mathematical Problems in Engineering*, 2020:1–7. https://doi.org/10.1155/2020/4761468.

National Institute of Education and Youth (2023). Dataportal. https://vysledky.nucem.sk/.

Nghe, N. T., Janecek, P., and Haddawy, P. (2007). A Comparative Analysis of Techniques for Predicting Academic Performance. In *2007 37th annual frontiers in education conference-global engineering: knowledge without borders, opportunities without passports*, pages T2G–7. IEEE. https://doi.org/10.1109/FIE.2007.4417993.

Rebai, S., Yahia, F. B., and Essid, H. (2020). A Graphically Based Machine Learning Approach to Predict Secondary Schools Performance in Tunisia. *Socio-Economic Planning Sciences*, 70:100724. https://doi.org/10.1016/j.seps.2019.06.009.

Statistical Office of the Slovak Republic (2023). Various Datasets. https://slovak.statistics.sk.

Toraman, C., Aktan, O., and Korkmaz, G. (2022). How Can We Make Students Happier at School? Parental Pressure or Support for Academic Success, Educational Stress and School Happiness of Secondary School Students. *Shanlax International Journal of Education*. https://doi.org/10.34293/education.v10i2.4546.

Werblow, J. and Duesbery, L. (2009). The Impact of High School Size on Math Achievement and Dropout Rate. *The High School Journal*, 92:14 – 23. https://doi.org/10.1353/hsj.0.0022.

Wu, J.-D. (2020). Machine Learning in Education. In *2020 International Conference on Modern Education and Information Management (ICMEIM)*, pages 56–63. https://doi.org/10.1109/ICMEIM51375.2020.00020.

Yağcı, M. (2022). Educational Data Mining: Prediction of Students' Academic Performance using Machine Learning Algorithms. *Smart Learning Environments*, 9(1):11. https://doi.org/10.1186/s40561-022-00192-z.

Yildiz, M. and Börekci, C. (2020). Predicting Academic Achievement with Machine Learning Algorithms. *Journal of educational technology and online learning*, 3(3):372–392. https://doi.org/10.31681/jetol.773206.

Zafari, M., Sadeghi-Niaraki, A., Choi, S.-M., and Esmaeily, A. (2021). A Practical Model for the Evaluation of High School Student Performance Based on Machine Learning. *Applied Sciences*, 11(23):11534. https://doi.org/10.3390/app112311534.

Zeineddine, H., Braendle, U., and Farah, A. (2021). Enhancing Prediction of Student Success: Automated Machine Learning Approach. *Computers & Electrical Engineering*, 89:106903. https://doi.org/10.1016/j.compeleceng.2020.106903.

Appendix A

Additional Graphs and Plots

Figure A.1: Feature Importance Graphs

(a) Random Forest - Feature Importance

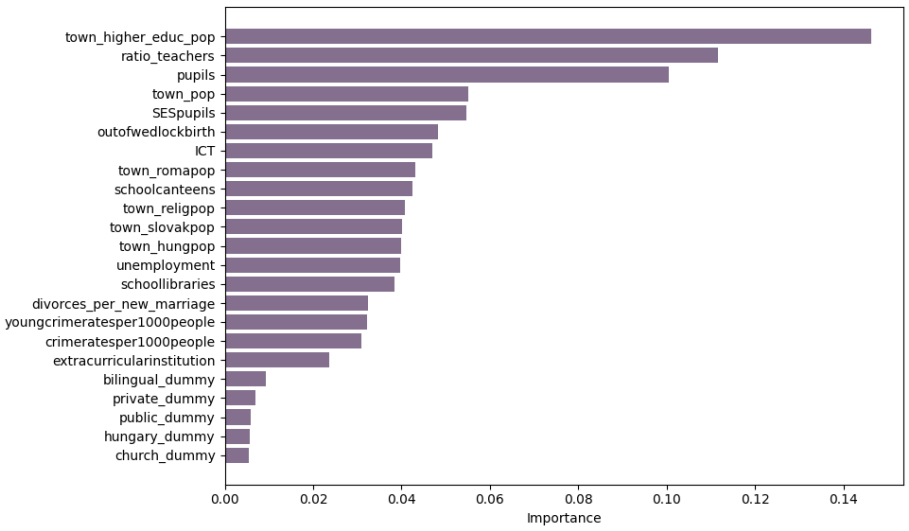
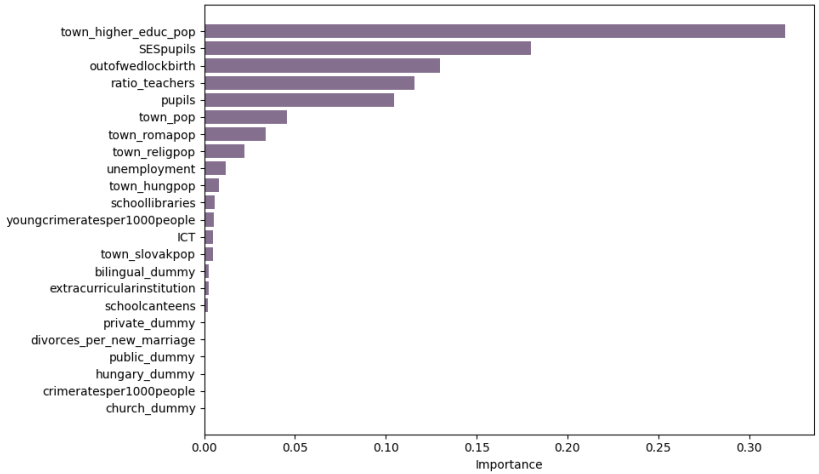
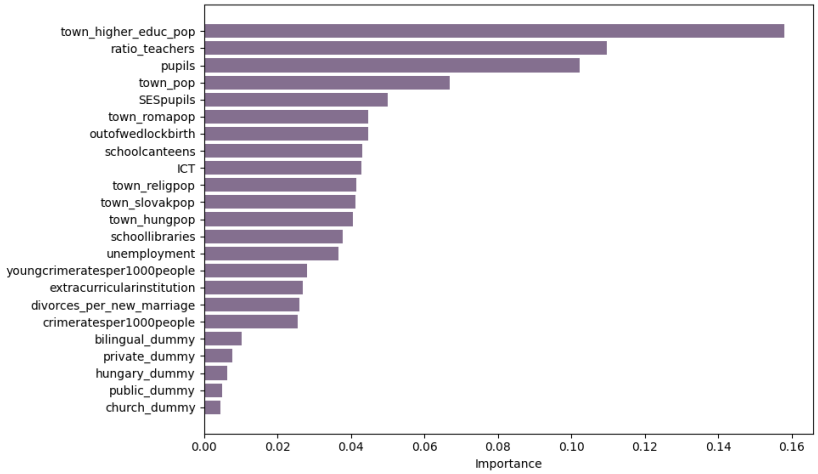


Figure A.1: Continuation of Feature Importance Graphs

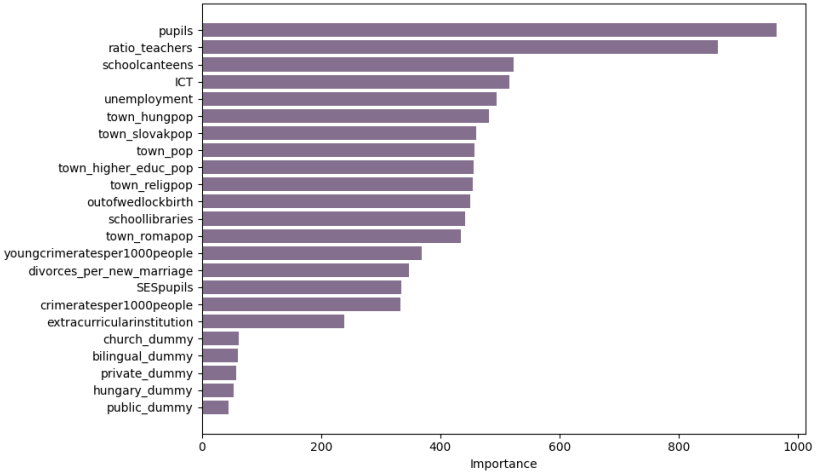
(b) AdaBoost - Feature Importance



(c) Gradient Boosting - Feature Importance



(d) LightGBM - Feature Importance



(e) XGBoost - Feature Importance

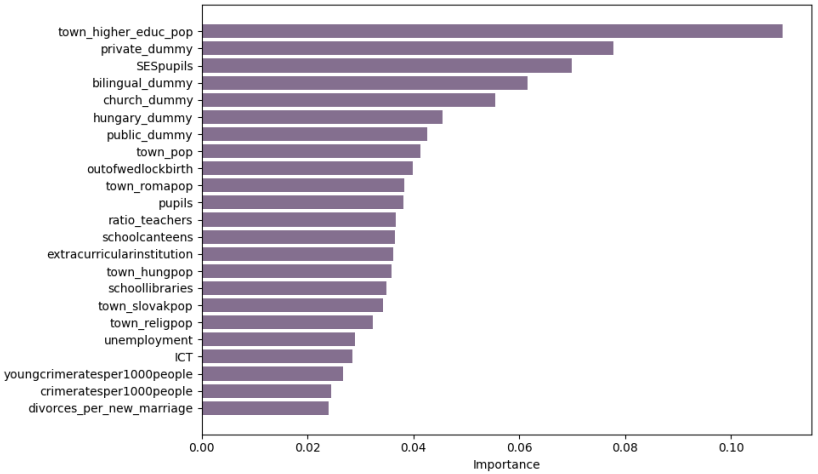
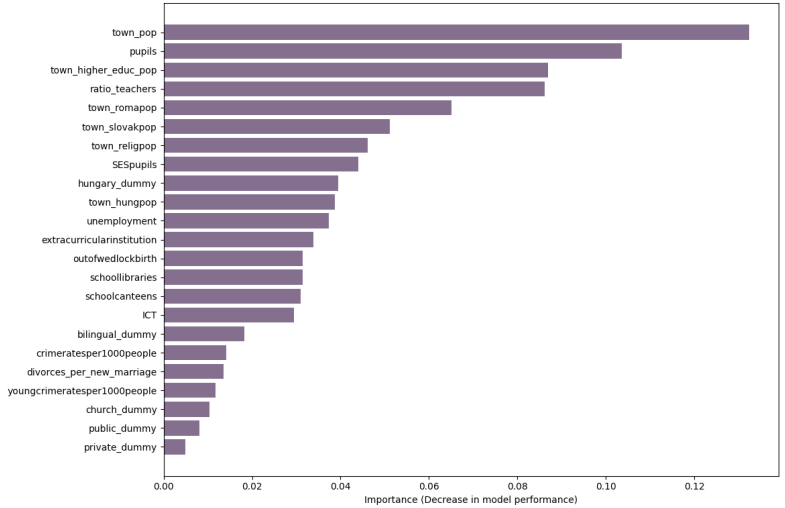
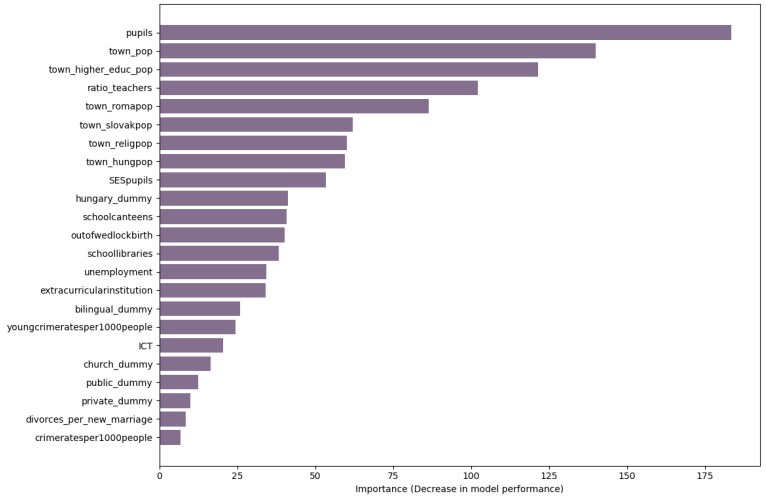


Figure A.1: Continuation of Feature Importance Graphs

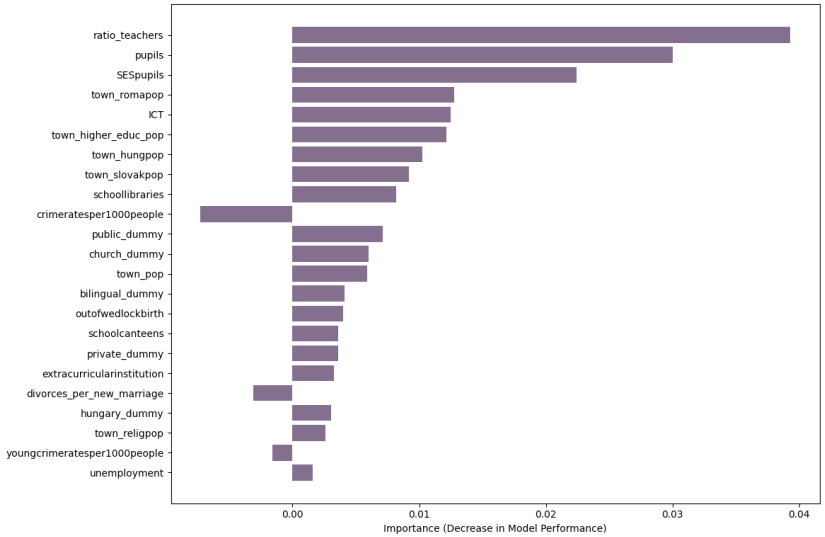
(f) SVM - Feature Importance



(g) NN - Feature Importance



(h) KNN - Feature Importance



(i) KRR - Feature Importance