Ministry of Science and Higher Education of the Republic of Kazakhstan

Suleyman Demirel University

Diana Bairamova

Identification of Students at Risk of Not Completing the Course Using Machine Learning

THESIS

Presented in Partial Fulfilment for the

Master of Technical Sciences Degree in Computer Science

(degree code: 7M06102)

Department of Computer Science

Faculty of Engineering and Natural Sciences

Supervisor: **PhD Moldagaliyev Birzhan**

Kaskelen 2023

Suleyman Demirel University
Faculty of Engineering and Natural Sciences
Department of Computer Science

Dean of Faculty

Associate Professor

PhD Zhamanov A.

06 06 2023

Topic of the thesis:

Identification of Students at Risk of Not Completing the Course Using Machine Learning

Thesis submitted as part of the requirements for the award of the MSc in "7M06102 - Computer Science" SDU, 2021-2023

Head of Department _____ Assistant Professor, PhD Mukash Zh.

Academic Supervisor _____ PhD Moldagaliyev B.

Master student _____ Bairamova D.

Kaskelen 2023

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Diana Bairamova

2023

Acknowledgements

I would like to express my gratitude to my supervisor **Birzhan Moldagaliev** for all the guidance and valuable advice that made it much easier for me to conduct the research, find the necessary information for my work, and improve my results every time. Throughout the entire research process, my supervisor provided me with valuable advice on how to improve my work and assisted me in selecting the appropriate methods to achieve the best research outcomes.

I also express my gratitude to my family and relatives who have been supporting me in my academic and professional development to this day. In particular, I am grateful to my little sister **Amina Bairamova** for her great moral and physical support during the entire academic path of the master's degree.

In addition, I am very grateful to the teachers and administration of our Suleiman Demirel University during my studies, who help in all matters and also contribute to my academic and professional development. I express my gratitude to all my friends and fellow students who helped to increase my mood and desire to study and develop further, without stopping there.

Dedication

This thesis is dedicated to:

My parents, Birzhan Moldagaliyev, Kamila Orynbekova,Inkar Shoganova, Gulnar Baktybayeva, Ualikhan Sadyk and many other for their support, help, sense of humour and useful comments for improving this project.'

Abstract

This dissertation is dedicated to the topic of identifying students at risk of not successfully completing a course at an early stage of their education using machine learning algorithms. In this study, the final exam score, academic performance category, and the risk group of students failing to complete the course are determined for accurate and detailed identification of students at risk. Research shows that machine learning algorithms such as LightGBM Regressor (for the final exam score prediction), Logistic Regression (for identifying two groups of students - those who will complete and those who will not complete the course), and K-Means (for identifying the academic category) can help identify students who need assistance from teachers with high accuracy of prediction. Detecting this group of students at an early stage of their education can enhance students' motivation for further learning and assist teachers in individually and timely identifying which student requires help.

Аңдатпа

Дипломдық жұмыс машиналық оқыту алгоритмдерін қолдана отырып, курсты сәтсіз аяқтау қаупі бар студенттерді бастапқы сатысында анықтау тақырыбына арналған. Бұл зерттеу курсты сәтсіз аяқтау ықтималдылығында тұрған студенттерді дәл және егжейтегжейлі анықтау үшін емтиханның қорытынды балын, оқу үлгерімінің санатын және курстан өтпеген студенттердің тәуекел тобын анықтайды. Зерттеулер көрсеткендей, LGBM регрессоры (емтиханның қорытынды балын болжау үшін), логистикалық регрессия (студенттердің екі тобын анықтау үшін - курсты аяқтайтындар және оны аяқтамайтындар) және K-орташа мәндер (академиялық санатты анықтау үшін) сияқты Машиналық оқыту алгоритмдері мұғалімдердің көмегіне мұқтаж студенттерді жоғары болжау дәлдігімен анықтауға көмектеседі. Оқушылардың осы тобын олардың оқуының басында анықтау оқушылардың одан әрі оқуға деген ынтасын арттырып, мұғалімдерге қай оқушының көмекке мұқтаж екенін жеке және уақтылы анықтауға көмектеседі.

Аннотация

Дипломная работа посвящена теме выявления студентов, подверженных риску неуспешного прохождения курса, на ранней стадии их обучения с использованием алгоритмов машинного обучения. В этом исследовании для точной и детальной идентификации студентов, подверженных риску, определяются итоговый балл экзамена, категория академической успеваемости и группа риска студентов, не прошедших курс. Исследования показывают, что алгоритмы машинного обучения, такие как регрессор LGBM (для прогнозирования итогового балла экзамена), логистическая регрессия (для определения двух групп студентов: тех, кто завершит курс, и тех, кто не завершит его) и K-средние (для определения академической категории), могут помочь выявить студентов, нуждающихся в помощи учителей с высокой точностью прогнозирования. Выявление этой группы учащихся на ранней стадии их обучения может повысить мотивацию учащихся к дальнейшему обучению и помочь учителям индивидуально и своевременно определить, какой ученик нуждается в помощи.

Abbreviations

AI - Artificial Intelligence

ML - Machine Learning

LGBM- Light Gradient Boosting Machine

IQR - Interquartile range

R2 - Coefficient of Determination

GOSS - Gradient-Based One Side Sampling

GBDT - Gradient-boosted decision trees

KNN - K-Nearest Neighbour

SMOTE - Synthetic Minority Oversampling Technique

DART - Data Addition and Rcmoval Trees

SVM - Support Vector Machine

MLP - Multilayer Perceptron

BIRCH - Balanced Iterative Reducing and Clustering using Hierarchies

GridSearchCV - GridSearch cross-validation

MSE - Mean Squared Error

MAE - Mean Absolute Error

MAPE - Mean Absolute Percentage Error

Table of Contents

Chapter 1

Background and motivations

1.1 Introduction

1.1.1 Motivation

Special attention at the moment should be paid to the development of the learning system through the use of Artificial Intelligence technologies, because education is the foundation for the development of the future generation.

Modern technologies through the use of Artificial Intelligence help us in all aspects of life [1], including transport and logistics, healthcare, security, in various studies and sciences, and also help to improve the education system. Thanks to modern methods of Artificial Intelligence, there is an opportunity to improve education, namely, to create various adaptive programs for each student, which can help to give timely feedback on students' understanding of various disciplines.

Since the recent times after the pandemic, many educational systems have transitioned to an online format [2], whereas previously it was not quite acceptable for many modern educational institutions. After the pandemic, many universities and colleges support a blended learning format, and now it is perfectly normal if education takes place outside the walls of the university or college. The shift in the learning format has impacted students' education, specifically their motivation in learning, where some students responded positively to the change in the learning

format, while others struggled to embrace online learning due to its drawbacks.

It is no secret that at all times the support of students' motivation in their studies is an important part in the educational process. There are several important reasons why motivation is considered important [3] for teaching students:

1) **Effectiveness in education**. Motivation is the key to achieving success in education. A motivated student will make increasing efforts to enhance their knowledge in all subjects studied and strive for the best outcomes during their educational journey.

2) **Enthusiasm from acquired knowledge.** Education is a process in which only motivated students derive pleasure from the learning process itself. They ask questions, show interest, actively participate in lessons, and seek new ways to acquire knowledge for their personal development.

3) **Creating a conducive learning environment.** When students are motivated in their education, they create favorable conditions for successful and effective learning, benefiting both themselves and the teacher.

4) **The desire to achieve a goal.** In most cases, motivated students are motivated not only in their studies but also in other aspects of life. They understand the importance of education and can relate successful learning to their future career growth, which is why they show interest not only in learning but also in other values in life.

5) **Sustainability in learning.** In the process of learning, each of the students faces difficulties in mastering any skills or knowledge. More motivated students have to overcome such difficulties much easier, because they realize the value of the knowledge they have received, so they are ready to overcome difficulties in learning something and show resilience in learning.

The above points give an idea of the importance of motivating students in the educational process. The motivation of even one student in a lesson can provide an incentive for other students.

With the help of machine learning algorithms that are part of Artificial Intelligence, it is possible to continuously improve the quality of the education system

and enhance student motivation [4]. By utilizing the capabilities of machine learning, early in the learning process, it is possible to find students who are at risk of not successfully pass the course and require timely support from the teacher.

Knowing which students are at risk of not finish the course, the teacher can use adaptive teaching methods for each student in the form of additional educational programs for such students. Also, students themselves, knowing in a timely manner which group they belong to, those who complete the course successfully or to the one where students fail the course, can start improving their academic performance on the course in a timely manner. In most cases nowadays, students do not pay due attention to learning from the very beginning of the course and can only start learning by the end of the subject when it is already too late to correct their situation.

1.1.2 Aims and Objectives

This study is dedicated to identifying students who risk not completing the course successfully before the course ends, in order to increase the level of motivation in teaching students in a timely manner before the course ends. For a more successful and accurate identification of students at risk of not completing the selected course well and those who complete the course successfully, this study examines several issues such as:

- **Final Exam Score** - prediction of the final exam score, ranging from 0 to 100, that a student will achieve on the final examination.
- **"Pass/Not Pass"** - Prediction of successful course completion or non-completion.
- **Academic Performance Category** - prediction of the category of academic performance of students (A, B, C or D):

  1) A - the given group consists of students who actively participate in classes and consistently perform above-average on course assignments, earning high grades (does not need help from the teacher).

  2) B - this group includes those students who regularly take part in classes

and complete course assignments below average, receiving not the highest grades (some assistance may be required).

3) C - this group includes those students who take part in classes irregularly, and also receive grades below average (needs help).

4) D - this group includes students who rarely attend classes and also receive low grades. (needs help).

As shown in the figure 1.1, each of the four groups of students is determined using two parameters, which in turn consist of small components (predictors), namely from the student's performance, which in turn reflects the student's grades on various tasks and midterm exams, as well as using the participation indicator, which consists of studying materials, reading various notifications from the teacher on the learning platform.

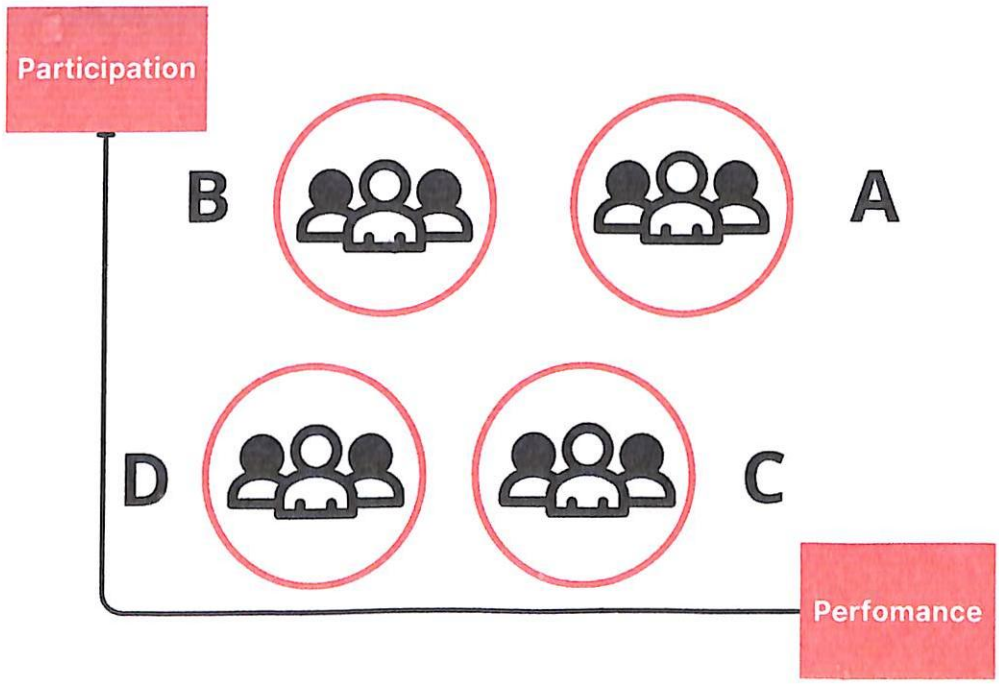

Figure 1.1: Definition of four categories of academic performance

This study shows which machine learning algorithms help to determine with high accuracy such academic indicators at an early stage of training as: final score for the exam, risk groups for successful and unsuccessful completion of the course and categories of academic performance (A,B,C,D). Each of these indicators has

its own value, however, the application of all three indicators of students together is able to give a detailed description and the necessary information as shown in Table 1.1 that the teacher and the student can use to improve their academic performance.

The first indicator which shown in Table 1.1 helps to guess what grade the student will receive on the final exam, but the forecast of the exam score cannot give us an answer whether the student will pass the course or not, for this reason, this study predicts the classification task, namely failure or successful completion of the course. At the same time, the assessment received for the exam and also which group the student belongs to does not give a complete picture of what exactly students have problems in studying the subject, for this reason, one of the 4 categories is predicted in this study, which helps to determine for what reason the student may not take the course, and more precisely what exactly students have gaps that negatively affect their successful completion of the subject. According to the above reasons, **only the use of each of the indicators together** gives a complete picture individually for each student.

Table 1.1: Category Detection

№	Category	Determination	Benefit	Assistance
1	A	Successful both on the "Perfomance" axis as well as on the "Participation" axis.	1. Increased motivation 2. The ability to help fellow students. 3. The possibility of professional and career growth associated with the course.	The help of the teacher and other students is not needed. The student is independent.
2	B	Student is successful as on the "Participation" axis, but not successful on the "Perfomance" axis.	1. The possibility of a concept where there are gaps. 2. Elimination of gaps in undeveloped topics. 3. Striving to get high marks	The student can improve his performance by himself, but they may need some help from a teacher or other students.
3	C	The student is half successful on both indicators of academic performance. Can get sometimes high marks, sometimes low.	1. Awareness of what may not pass the course. 2. Addressing knowledge gaps in a timely manner. 3. Search for additional materials for easy assimilation.	A student needs the help of a teacher and can also ask for help from other students.
4	D	Student is not successful both on the "Participation" axis and on the "Perfomance" axis.	1. The opportunity to complete the course in case of timely intervention. 2. Organization of individual teaching methods for better assimilation. 3. Awareness of the moment when the gaps in knowledge began.	A student really needs the help of a teacher and can also ask for help from other students.

Identifying students who risk not completing the course successfully before

the end of the course will help in the following points:

1) **Avoiding student dropout.** Identifying course students who may not pass the course well at an early stage avoids the loss of motivation of students, which contributes to the dropout of the student as a whole from the educational process. Students who have lost motivation in learning are able to reduce their academic performance in various subjects studied and may quit their studies altogether.

2) **Getting support from a teacher.** Understanding exactly which students are experiencing learning difficulties, teachers can help such students by including, for example, any additional classes, mentoring, or providing students with additional resources to help them learn the subject.

3) **Effective time saving.** Early identification of students at risk and timely support to students will help save time for which a student can spend repeatedly for training if he fails the course. Support at an early stage of learning is much more effective and expedient than later intervention in the student's education.

4) **Improving academic performance.** If a student is identified at an early stage of training and is supported by a teacher or his classmates, he will be able to change the situation and finish the course successfully. When students' problems in learning any subjects are solved at an early stage of training, they have a better chance of completing the course successfully and coping with learning difficulties.

Early identification of students who may not finish the course is **important topic** for research for the following reasons:

- **COVID-19 and it's impact.**

  One of the first and most important reasons is the period of the **COVID-19 pandemic**, the consequences of which led to the fact that many educational systems around the world switched not to the traditional format of education [2], namely to remote online learning and continue to study online. This provoked various learning difficulties such as various technical problems,

misunderstanding of the subject and, most importantly, a decrease in the level of motivation of students.

- **Individual approach for students.**

  An equally important and relevant topic is the individual approach in teaching each student. Identifying students at an early stage who are at risk of not completing the course allows us to determine the characteristics of each student and helps to understand exactly what kind of support each student needs. Teachers can create their own methods or techniques of an individual approach in understanding students of certain topics or other not a few important things in increasing the motivation [3] of learning a subject .

- **Efficiency in the use of resources.**

  The relevance is also reflected in the understanding that it is very necessary to effectively use the resources of educational institutions, as well as the time devoted by students and teachers themselves to studying any courses. Identifying students at risk will help avoid the problem of student dropout and motivate students to study during, before the course ends.

- **Regular improvement of the quality of education.**

  Continuous improvement of the quality of education. If institutions are able to identify students at an early stage who are at risk of not completing certain courses, then the quality of education will constantly improve for the whole institution and for the teacher. Providing support to students in a timely manner can improve students' understanding of the subject, determine which points are not clear to students and help provide additional resources and support that will help not only the student but also teachers to understand what it is worth focusing on for many students in the further education of students.

Summarizing the above, we can give the urgency of the problem in order to effectively manage the resources of both institutions and students, constantly improve the quality of education by providing timely assistance to students, an individual approach to teaching for each student, effectively respond to providing

support to students in teaching various topics.

1.1.3 Thesis Outline

This thesis follows a specific order:

1) **Chapter 1.**

   The first chapter presents an introduction to the research problem, specifically discussing the goals and importance, relevance, and novelty of the study.

2) **Chapter 2.**

   The next Chapter 2, provides an introduction to the theoretical framework of the research, mentioning various terms used in the study, as well as an introductory description of the machine learning algorithms used to successfully identify students who may not successfully finish the subject.

3) **Chapter 3.**

   Chapter 3 describes previous research on the topic, including key findings from previous works and recommendations put forth by the authors of those studies.

4) **Chapter 4.**

   In the subsequent Chapter 4, the data preparation process, research methods, and results are described in detail.

5) **Chapter 5.**

   The final Chapter 5 concludes the research, summarizes the conducted work and the obtained results, and provides recommendations for future research.

1.2 Background

1.2.1 Dropout of students at the initial stages of the educational process.

In our time, the issue of early student dropout from courses affects any educational institution. Student attrition at an early stage occurs due to various reasons, providing an opportunity to make changes before the student's early departure from the course happens. The problem of early student dropout from courses lies in the fact that students may discontinue their studies in certain subjects or, in general, terminate their educational journey before completing a particular course or even before finishing their education at the educational institution.

This type of student attrition is significant because it occurs at an early stage of student education. It can happen within a month or even six months after the start of the course, but it occurs before the completion of any specific course. It can have a considerable impact on students' overall academic success as well as the overall effectiveness of the educational system.

The reasons why students may not successfully complete a course and drop out at an early stage of their education can be diverse [5]. Some of these reasons include:

1) **Insufficient academic preparedness.**

   Some students may be unprepared for specific courses due to their inadequate study skills. Not every student, before entering higher education, understands where exactly they are headed and what they need to know in order to successfully complete academic courses and become a sought-after professional in their chosen career.

2) **Inappropriate educational environment.**

   It is important to understand that not all students may fit certain requirements in the educational environment of academic institutions, and some students may not meet specific criteria [5] or align with a particular type of

education in educational institutions.

3) **Disappointment with acquired knowledge.**

One of the reasons for early student dropout can be the disappointment students experience with the knowledge they acquire during their education compared to their expectations. If students do not receive the desired knowledge during their studies, it can lead to a loss of motivation for further learning, causing them to abandon the course at an early stage.

4) **Insufficient engagement and participation in class.**

Many students mistakenly assume that involvement in class and active participation do not play any role in successfully mastering a subject and cannot influence early student dropout from a course. It is essential for every student to feel like an active participant engaged in the learning process of the course [5]. If a student does not recognize the importance of being active, they may lose motivation to learn the subject. Consequently, the lack of interaction with teachers or fellow classmates in the course contributes to this.

5) **Financial difficulties.**

Financial challenges can arise for students, preventing them from accessing the essential resources and educational opportunities [6] they need. The steep expenses associated with tuition fees, textbooks, accommodation, and transportation can act as obstacles, particularly for students who have limited financial means [6].

6) **Inadequate course selection.**

Students might not fully comprehend the level of difficulty or fail to align their interests properly when making course choices, resulting in a decline in motivation and a higher likelihood of dropping out.

7) **Insufficient support.**

Students, especially those who encounter difficulties adjusting to a new environment or have personal issues, may feel a lack of support from the

educational institution. The absence of guidance, mentorship, or access to resources can diminish their motivation and hinder their ability to overcome challenges.

8) **Personal circumstances.**

Students may encounter personal problems, such as family obligations, financial hardships, or health issues, that can result in the interruption of their studies.

It is crucial that the students initially enrolled in any course are actively engaged in the lessons and understand their attitude towards the subject. It is important to identify if they are facing any difficulties and which specific topics they are struggling to comprehend. In addition to the students themselves, the teacher, with the help of Artificial Intelligence methods, can understand these issues at an earlier stage and provide timely support to the students before their education concludes.

The possibilities of using Artificial Intelligence can help to fix the problem of early dropout of students at the early stages of education [4]. Starting with data analysis and collection, Artificial Intelligence methods allow us effective work with very large amounts of data that a person himself will not be able to master. Namely, thanks to machine learning algorithms, it is possible to process data with a large size and different types [7]. In particular, for the problem of early dropout of students, various academic indicators of students are processed and analyzed with the help of machine learning, such as engagement in the lesson, attendance of classes by students, various points received or grades on quizzes, completed projects, various tests , as well as grades on previous exams. The use of these factors allows Artificial Intelligence methods to determine at the early stages of training the risk that a student will not complete the course successfully and may lose motivation for further academic development.

In addition to processing and analyzing large amounts of data with various types of data [4], Artificial Intelligence methods can identify students at an early stage can not finish the course. That is, with the help of machine learning techniques, Artificial intelligence can identify students who have a very high risk of

not successfully completing the course. Identifying such students at an early stage of learning will allow teachers to understand which students need their support and can take the necessary measures at the early stages of teaching the subject. After identifying students at risk of not completing courses, Artificial Intelligence can also suggest adaptive learning for each student. Machine learning algorithms allow you to determine which type of training is suitable for each student, that is, adaptive machine learning methods built for each student allows the teacher to provide various educational programs or additional resources that are necessary for a particular student in mastering the learning material and overcoming various difficulties for successful completion of the course.

The main task of applying the adapted teaching method for various students is to optimize the educational process and increase academic performance in the classroom among students. Earlier, the authors of [1] in their study describe how the Moodle system can be used to apply adapted teaching methods among students. Namely, the authors give examples based on all the functions of the system when the system analyzes all data about students using Artificial Intelligence and can automatically offer certain individual tasks for each student as well as additional sources or resources for teaching a particular topic. This feature also allows you to determine at what point of learning students have learning difficulties. It is also a great opportunity for students to focus on certain topics and get more support from the teacher, because thanks to this opportunity, the teacher will see at what stages the student finds it difficult to master the material successfully. In addition, the authors describe in detail exactly what situations a user may encounter when using a management system for studying various subjects to determine adapted teaching methods suitable for each student.

The use of these Artificial Intelligence capabilities can significantly increase the level of motivation of students [4]. Artificial intelligence methods can provide both for the student and for the teacher detailed personalized information about the student's learning ability of various courses as well as various motivational materials as well as his predictive indicators for the course that will help students work harder and gain knowledge for the successful completion of the subject.

All of the above features of Artificial Intelligence on the problem of early

dropout of students from the course (analysis of various factors of students with large volumes and different types of data, identification of students who can not finish the course successfully, as well as the application of adaptive teaching methods to various students) they to promote the involvement of students in the lesson and also increase the chances of successful completion of any course in the selected disciplines.

The system for identifying students with a risk group is presented by scientists in their study [1], where they help determine which students at an early stage of study will not successfully graduate from a higher educational institution. The system developed by scientists uses complex Artificial Intelligence algorithms based on artificial neural networks [1] that accurately identify students at risk of not successfully graduating from higher education.

Author describe in detail how artificial neural networks were used, which are machine learning methods for processing a large amount of data about students and subsequent analysis of various academic indicators obtained during the period of study by students, their activities and other various parameters to successfully identify students who can not finish the course successfully [1]. The author note that the neural networks used are able to identify various relationships and complex patterns, which is the most effective way to eliminate student failures in the successful teaching of subjects and successful ways to predict early dropouts of students from educational institutions.

One of the important objective of the study [1] is to develop a system for classifying students into risk groups for not successfully graduating from an educational institution. With the help of this system, the managers of the education system in the educational institution itself can identify such students and provide timely support to students in the form of additional classes, providing additional resources so that students can improve their situation for the better and successfully graduate from higher education.

Author focus on how the model is trained based on machine learning algorithms using neural networks [1]. The machine learning with neural network is trained on the basis of historical data about each student, including students' academic

performance, and builds models for predicting the probability of a student dropping out of training at a given educational institution. In this way, the machine learning algorithm helps to identify such students from the risk group and helps educational institutions to take various measures to improve the academic performance of students at the time. The author emphasize what opportunities can be achieved by using machine learning methods to eliminate the problem of early dropout of students from training. It is noted that educational institutions using the developed system can see personalized information for each student and improve the effectiveness of the academic performance [1].

1.2.2 Machine Learning with Python

Machine learning is currently used in many everyday tasks [8] and is able to simplify the routine work of mankind many times. Learning from the provided historical data, machine learning algorithms identify various dependencies between data and patterns, and ultimately solve the problem much faster and more efficiently than the person himself. In addition, machine learning methods are able to cluster themselves into groups or identify patterns in a large amount of data without providing historical data. Big data processing and analysis is an important link in any field of activity [7], when introducing any business and without using machine learning methods, it is very difficult for people to identify patterns in the available data on their own and find solutions to their questions.

Machine learning is only one of the subsections of Artificial Intelligence, which is an integral part of all the available capabilities of Artificial Intelligence. This subsection relates to the creation of various forecasting models and algorithms that are able to extract the necessary information for computers [8], learn from this knowledge provided and produce the desired result without any difficulties for humans.

Nowadays, this area is of particular importance in all areas for the following reasons:

1) **Processing and subsequent analysis of available data.**

With the appearance of the Internet and its development, a large amount of

information is generated every second, which has to be processed and analyzed to conduct any business or for various other purposes in any existing activity at the moment. Machine learning can help analyze data [7], deduce patterns, and help us draw conclusions. The model created on the basis of trained data can be constantly improved and used if necessary.

2) **Optimization and automation of routine processes.**

Machine learning algorithms can automate any processes or tasks, namely what a person previously needed to do, for example, recognize an image to access something, or an example when it is necessary to classify data. If a lot of data is generated and this process is dynamic, managing such a volume can take a long time for a person, for this reason, the use of machine learning methods is necessary in such cases [7]. Examples of using machine learning to automate and optimize complex or everyday routine tasks may be enough for various purposes.

3) **Prediction of various things.**

With the use of ML, it has become possible to create models that will be able to determine something in advance for the successful implementation of any business. One example is our daily weather forecast, on the basis of which people can rely on air temperature and weather, planning their day. Another very good example is the use of ML-based predictions in the field of medical diagnostics, a diverse number of diagnoses are predicted based on machine learning models [8]. In addition, it is used for our daily recommendations in various aspects of life.

4) **Adaptability to different users.**

Machine learning models can adapt to each user and create, based on the available data, their recommendations for examples on choosing clothes to buy or watching movies. The use of various personal data for model training allows models to expand their capabilities and adjust to the taste and choice of each user/client, offering a certain category of what may be liked or may be suitable for a person [7]. This flexibility allows us to apply and train

artificial intelligence more and more nowadays.

In the application of machine learning, in addition to the fact that it is necessary to know where and in what cases models trained by artificial intelligence can be used, it is important to learn to master the language in which such models can be created. One of these languages is Python, where it is very popular in the field of machine learning for a number of reasons [9]. In general, this language is very easy to learn, which is its main advantage over other existing languages.The syntax of the language is very simple compared to other languages.

The code itself written in Python can be easily read and legible not only for one person who is the author of the code, but also for all other users [9]. This greatly simplifies the ways of communication between different members of projects created using machine learning. In addition to its simplicity, it is very rich in various ready-made libraries with the help of which it becomes very convenient to solve any tasks of varying complexity. It enjoys high power and an extensive ecosystem of various built-in tools. All these features help us choose this particular language to solve machine learning problems, which is common for many developers facing machine learning.

In addition to this, the language is so popular that it has been able to contribute to the creation of a large community and to the creation of large ecosystems of libraries [10]. Due to the great demand between users of the language in solving their tasks, at the current time, the necessary tools and libraries are being created more and more, which are used in the future by other users using the possibility of artificial intelligence in tasks.

The creation of an entire ecosystem of certain tools and libraries makes it possible to simplify the process of solving problems based on the use of artificial intelligence, which in general entails facilitating the development of machine learning all over the world.

An additional feature of Python is that it can easily integrate with other programming languages and tools. This feature allows to use this language in various existing systems to solve problems based on machine learning [9], as well as apply it in various technology stacks.

In general, the high performance and efficiency of using the language is manifested in the fact that it has a large number of built-in optimized libraries that are constantly expanding and thereby provide high performance compared to other languages.

The language has become widespread not only because it is possible to solve problems based on artificial intelligence, but also because it solves problems of varying complexity and volume, so it is very popular not only among machine learning tasks [10], but also for all other cases. This feature allows users of the language to apply the acquired skills and knowledge in solving various other tasks in the same language. For this reason, many language users consider choosing this particular language compared to other languages.

Chapter 2

Literature Review

2.1 Student dropout prediction.

Earlier in his research, scientist Cameron Cooper, using machine learning algorithms, namely using neural networks, identified students at risk of not completing one of the prerequisites - "Introduction to computer programming" successfully. Earlier, before conducting this study, scientists [1] studied the problem of calculating the exact percentage of underachieving students in the introduction to programming courses, as well as identifying factors that could affect the successful completion of the course and vice versa on the failure of the course. The researchers [1] put forward several different hypotheses in their study according to which it was possible to reveal the main reasons for the failure of students in the course of introduction to programming, namely, they identify such predictors as the complexity of the material being studied in the course, that students might not have been prepared for this course and did not have any previous experience in this subject, as well as not the best and necessary support from teachers in this subject.The results of this study [11], show that only 67% of students successfully complete their studies in the subject - "Introduction to Computer programming" in the period of 7 years in education centers. On the other hand, the study has limitations because the results are based on data obtained from various universities and this may not reflect the full picture of the failure of students of the course on programming for a variety of educational institutions and in all contexts. The

authors also note [11] how much % of students fail in the course introduction to programming, however, the authors do not offer ways or techniques that could be applied and changed to improve the situation according to the students' rights.

The main purpose of this research [1] conducted by the scientist Cameron Cooper was to find students who may be at great risk of not completing their studies or experiencing difficulties in the programming course, which is an introductory course for all novice programmers.

To conduct the study, the scientist collected data on students' academic performance, their demographic indicators, educational history before they started studying this course for 7 years of studying this subject Introduction to Computer Programming [1], namely, he collected 592 data about each student to teach his model. The scientist uses machine learning algorithms and analyzes all the collected data about students. After teaching the model to the students' past historical data, the author creates a predictive model that allows to successfully find students who can not finish this course successfully.

The results of this study [1] showed how much machine learning capabilities allow you to correctly identify students at risk with a forecast accuracy of 99%. The author uses 25 different neural networks. The author shows that of all the neural networks used, the best result according to the forecast turned out to be a Probabilistic neural network. This model has achieved high forecast accuracy, where the author encourages teachers to use the forecast model for timely intervention if a student is at risk of not completing the course successfully. Author in the research [1] emphasizes the great potential for using this student prediction model, noting that there is an opportunity to actively use machine learning models in order to support students to complete a programming course very well, while noting that further research can contribute to improving the efficiency of higher education students.The author shows how sensitivity analysis is used to identify the most important periods in students' learning [1], which help teachers correct the situation in time and help improve students' academic results. The results of the study [1] showed how with the using ML algorithms it is possible to increase the level of completion of a programming course, namely, this study contributed to an increase in academic performance by 23% through the use of

an alert system.

On the other hand, the authors [1] note that the conducted research and its model can be used only for students for a two-year community college, namely for the course "Introduction to Computer programming", which students take when entering the computer science specialty. This is a limitation of this study, since it is suitable only for students with two-year studies and the specifics of the study are tied to only one subject, which may negatively affect the application of the model in other disciplines or use in other educational institutions.

In another study [12], the author studied this problem of identifying students can not finish the course, but with a different approach. Author divided the period of study of students into 3 different stages of learning a subject. To solve this problem, the author used only such predictors as grades for various tasks, as well as the number of visits to the classes themselves by students [12]. The three different stages into which the method of identifying such students is divided are divided in the following way:

- **Stage 1.**

  The first stage includes all visits and evaluation of the very first task for the first 4 classes of the course.

- **Stage 2.**

  The second stage includes the analysis and revision of grades for the first and second tasks in aggregate, as well as the number of visits over the last 8 weeks of training.

- **Stage 3.**

  The last third stage is similar to the previous two stages, but it contains only the grades of three tasks and the number of visits by students for 10 completed weeks of study, and it also contains an assessment for the intermediate exam before the end of the course.

Author in his research [12] says that he does not use all the assessments and visits by students in a row for all 10 weeks of training on the course, since the

teacher cannot always conduct all the lessons, that is, in some weeks lessons may not be conducted due to various circumstances. Further, after being divided into three different stages, the scientist uses different types of machine learning algorithms and teaches at all three stages separately. In addition to using algorithms at all three stages of training, the author uses a methodology previously derived by scientists in their study [13] in which researchers note that using a combined result from various algorithms gives a better result than just using one algorithm to solve a problem.

For this reason, the authors uses a combination of several algorithms [13] at the same time. Namely, the author explains his solution to the problem using three different schemes, namely in such a way that the student belongs to the risk group if at least 1 of the 3 algorithms used classifies the student to the risk group, this is included in the first scheme proposed by the author of the study.The following scheme refers to the result obtained by two algorithms, which means that of all the algorithms used, the student will be considered at risk only if two algorithms determine the student as at risk of not finishing. The most recent third scheme for detecting students at risk is determined in such a way that a student will one of the students at risk only if all three algorithms [13] classify students as at risk, this will mean that students really may not graduate successfully.

After carrying out all the methods, the authors shows the result that the use of the second method when the result is used when at least two of all the algorithms used show that the student is in a group who can not finish the course with the most higher accuracy rate - 85.33% [13], only in this case the student will be determined to the group who will not successfully complete the course on the subject.

In his experiments [12], the author used three different algorithms to identify students at risk, namely, such as the Decision tree, Instance-based learning classifier and Naive Bayesian algorithm. The results of the experiments have shown that of all 3 machine learning algorithms such as the Instance-based learning Classifier, Decision Tree and Naïve Bayes, the K-star algorithm showed the best result according to the forecast [12]. In addition, the author does not use predictors such as age or gender, as he points out that in such studies it is necessary to use only

time-varying predictors that have the ability to change and not be statistical [12]. For example, the predictors used as grades and attendance are time-varying data about students.At the same time, the author compares his obtained results on the use of only time-varying data with the results of previously obtained studies, where he notes that his results are better than in the study where the author [14] also use only time-varying predictors and lower than in the study of other scientists [13] according to the same logic of using time-varying predictors.

Comparing with the studies of other authors [13] on this topic, the difference is that the author used exactly time-varying data on students and shows us that for such studies, only time-varying data can be taken and this will be sufficient for an accurate forecast. However, on the opposite side, the limitation in this study is that the author uses data collected from online courses and also the author does not conduct dependence and analysis on the influence of the collected predictors for this study on the model obtained by experimenting with various machine learning algorithms. Author [12] also calls in his work for the further development of the experiments done by the author, namely, he considers it necessary to carry out similar work only using learning algorithms based on instances [12], such as KNN, to obtain accurate results on the prediction that the student is in poor condition in the subject and may not finish the course successfully.

In their study [15], the scientists devoted time to studying the topic of the problem of potential dropout of students in the field of STEM in colleges. A distinctive feature from other researchers is that the authors [15] tried to find the most stable factors (predictors) of data on college students that can accurately lead students to a loss of student interest in learning as well as to early dropouts from education in general. The authors note that it is at the beginning of training that students are at risk of early dropout from students, the author notes that more than 60% of students are able to drop out at early stages of training. For this reason, the authors [15] note that it is extremely important to detect students at risk of not completing their studies at the early stages of their studies, in order to maintain motivation for students to study by college administrators.Authors [15] compare various factors to determine the most significant for identifying such students as:

- **GPA level taken from high school.**
- **GPA level during college**
- **Age**
- **Gender**
- **Race**
- **Date of admission to college**
- **Different SAT exam scores (math, verb, etc.)**
- **Duration of training**
- **Specialty of study**
- **Final semester of study**

Authors consider 9 different specialties with different academic requirements, as well as various difficulties faced by students or teachers themselves during the study of the subject. Authors consider 9 different specialties [15] (Mathematics, Physics, Chemistry, Biology, Computer Science, Information Technology, Civil and Infrastructure Engineering, AIT(Applied Information Technology) and Psychology) with different academic requirements, as well as various difficulties faced by students or teachers themselves during the study of the subject. The results of a comparative analysis between 9 specialties in the students' cohabitation in these specialties show that the process of teaching students in different 9 specialties takes place in students in different ways [15], and scientists also say how much the courses and curriculum may differ from each other, as well as differences in student performance in these specialties.

Of all the 9 listed specialties for which a comparative analysis was conducted, most of all students are delayed and graduate successfully in the Information Systems specialties, then these are Computer Science specialties and other engineering specialties. The analysis showed [15] that in the specialty of Physics, students are most at risk of not completing their studies. If a student studies successfully for the first two years of study, then he will have a better chance

of completing his studies. Author noted [15] this pattern in his analysis by the fact that physics is one of the most difficult specialties and those who enter this specialty are much older than those who enter other specialties mainly due to the fact that they approach the choice of their specialty very deliberately and without nonsense that young people can allow by choosing their specialty thoughtlessly. Students who enrolled in the specialty of Physics who do not make efforts to improve their academic performance in most cases give up and drop out at the early stages of their studies.

In addition, authors [15] conducted experiments on teaching machine learning algorithms separately for different 8 semesters of training in any specialty. In addition to machine learning algorithms, Survival Analysis was also present in the study, the author took algorithms such as **Logistic Regression (LR), Naïve Bayes (NB),Decision Tree (DT), Random Forest (RF), and Adaboost (AB)** methods. The results of the study [15] showed that the duration of study at a higher educational institution as well as the level of academic performance by semester (GPA) are important predictors for calculating the risk among students of not completing their studies.

Author notes that machine learning methods [15] did the best job of predicting student dropout compared to methods of survival analysis. Of all the algorithms and analysis methods used, machine learning algorithms such as Logistic Regression and AdaBoost methods have shown the best result. They are the most accurate methods for finding students can not finish their studies. The logistic model showed the best predictions for the 4th semester of training with the accuracy of the F1 score = 95% model, also the AdaBoost method showed a high result of the accuracy of the model for the 4th semester of training with the accuracy of the F1 score = 95 model%[15] The limitation of the work carried out is that the data on students enrolled in college used in the work such as the GPA level in higher education [15], as well as various grades on SAT exams may not be available to all educational institutions around the world, so this is a limitation on the study. In addition, the author made an experiment only in certain disciplines, while earlier deductions, although in a small amount, may also be in other specialties and is also an important problem.

2.2 Prediction of students' final grades.

Linear Regression

Scientist Mahesh Gadhavi [16] conducted experiments to predict the final grade of students on the course, which can be used in the future in educational fields. Linear regression was used to analyze the relationship between independent variables and the final final course score. In this study [16], linear regression is used to predict a student's final grade based on various factors such as student attendance at courses, grades for previous exams, the number of hours passed in class, which can affect academic success. Using data on attendance, grades, as well as the number of study hours spent, a linear regression model was built that takes into account the relationship between predictors and the final grade. To verify the accuracy of the obtained model, the researcher made an analysis comparing the estimates obtained from the forecast with the average values of the real final grades of students. The scientist [16] notes that in order to obtain more accurate results, the predictors used in the study were normalized to 100% to obtain a more correct assessment of the course.

The results of the work [16] done have shown that the model based on the usual linear regression is very effective for building a model based on the forecast of the course assessment.

Author notes [16] that the study may be suitable for various educational institutions, according to how teachers and managers of universities or colleges can help students make decisions on student support, as well as identify students at risk of not graduating from an educational institution, and can also help teachers develop a system for teaching students individually for each student for more successful completion of the course.

However, it should be noted that the results of the study [16]are limited by the data used and the only one linear regression model used in the research. For more accurate forecasting and analysis, it may be necessary to use other algorithms, not only Linear Regression and a wider data set.

Generalized Linear Model

After the introduction of a mixed format of education in connection with COVID-19, the problem with a decrease in motivation in learning among students began to develop even more among students than it was before COVID-19 appeared. In his research [17], the scientists analyzed the collected data on students during the introduction of online learning, where he tried to reveal the topic of predicting the final grade for the course, which the student receives at the end of the whole course during the first semester.

In certain experiments, the authors notes that academic performance data were collected by monitoring academic performance in the form of various grades received on the subject for the period from 2020-2021 during the first semester. Authors uses 35 predictors to analyze [17] the forecast of estimates, namely:

1) **Knows about MS Teams**
2) **Participation in other laboratories**
3) **Grades for other courses**
4) **Uncertainty in understanding of the subject**
5) **Readiness for the final on the topic of the subject**
6) **How many times have there been technical difficulties**
7) **The tasks were evaluated and re-returned for revision**
8) **Time spent studying the subject**
9) **Hours spent on training on weekdays**
10) **How well the various completed tasks are evaluated**
11) **Assessment based on notes written in the classroom**
12) **The impact of the tests on the knowledge of the theory**
13) **The need to apply visualizations for a good understanding**
14) **Reach in the study of complex topics**
15) **Usefulness in the use of graphic images**

16) **The student likes this subject more compared to others**

17) **Eliminating gaps in the topics covered**

18) **The maximum limit of exhaustion from the subject**

19) **Hours spent on all types of work**

20) **Number of hours of attendance on the learning platform**

21) **Complaints about late grades for the lesson**

22) **Marking gaps in conducting online lectures**

23) **Basic knowledge of office programs**

24) **The presence of completed tasks related to future work**

25) **The probability of good performance of tasks related to future work**

26) **Sustainability in learning**

27) **Academic performance level by topic**

28) **Physics grades on the initial exam**

29) **The number of activities on the part of the student**

30) **The number of students who dropped out**

31) **Number of lectures attended**

32) **Activity in online classes**

33) **The class group to which the student belongs**

34) **Type of completed high school student**

35) **Final assessment of the course**

After conducting a statistical analysis of all 35 predictors, the authors[17] identifies and subsequently utilizes only the factors that influence the academic performance of students in the online course during the first semester of a specific

subject within the Natural Sciences of the Engineering Faculty for the purpose of grade prediction.

Using the influential factors for the course grade, the authors [17] creates a hybrid model using machine learning. The hybrid model consists of a generalized linear model, where the errors are used as input for a neural network. In other words, initially, the course grade passes through a model using a regular linear regression, which is typical for predicting values. Then, the obtained error results enter the neural network model, which is composed of 70% for train data, 15% for testing data, and the remaining 15% as the validation. The use of a hybrid model [17] that can continuously learn from its errors has allowed the authors to achieve excellent results in predicting the grade a student can obtain upon completing the course. In their findings, the authors note that the utilization of the hybrid model (linear regression and neural network) specifically led to automatic grade improvement with a high coefficient of determination, R1 = 1[17]. On the flip side, the study has several limitations. Firstly, the predictors used were only based on online student sessions and not in a blended learning format. Additionally, the authors only utilized data from a single semester, specific to one subject. This means that the findings of this study [17] can only be applied in a limited context, exclusively for the first semester of online learning. For future research, it is advisable to include data not only from online learning but also from different semesters. The authors [17]recommend exploring the selection of the 35 predictors for subsequent courses in various subjects and in a blended learning format for future studies.

The tables below 2.1, 2.2 show a comparison of current work as well as limitations on previous studies aimed at predicting various kinds of academic performance, such as grades, dropout of students from the course.

Table 2.1: Limitations and comparison of other studies with this study.

№	Research work	Methods/Research	Limitations	Comparison with this research
1	Using Machine Learning to Identify At-risk Students in an Introductory Programming Course at a Two-year Public College [1]	25 different Neural Networks (Probabilistic Neural Network, Regression MLP, Classification MLP, Reg Gen Feedforward) using Backward Elimination for input data	Limited sampling of data for only one subject	Sampling in various subjects, predictors are universal for different courses.
2	Identifying At-Risk Students Using Machine Learning Techniques: A Case Study with IS 100 [12]	Comparison of 3 different algorithms: Instance-based learning classifier, decision tree and naive Bayesian algorithm for predicting students at risk, comparison of time-varying attributes and time-invariant attributes such as hunter or age	The data collected only from online courses does not carry out dependence and analysis of the influence of the collected predictors on the model	The analysis of the influence of predictors on the forecast, data collected from both online and offline formats, has been carried out.
3	Dropout prediction in e-learning courses through the combination of machine learning techniques [13]	3 different schemes: 1 scheme - student belongs to a risk group even if only 1 algorithm assigns him to a risk group, 2 scheme - if at least 2 algorithms show that this student belongs to risk group, scheme 3 - if all 3 algorithms determine a student as from a risk group	Applicable only for courses with online training, there is no analysis of the influence of factors on the forecast of the model.	Applicable for both training formats, a comparison of the obtained factors for the forecast is given.

Table 2.2: Limitations and comparison of other studies with this study.

№	Research work	Methods/Research	Limitations	Comparison with this research
4	Student Engagement Level in an e-Learning Environment: Clustering Using K-means [18]	K-means was used as an unsupervised learning clustering algorithm to determine the level of student engagement in the lesson. Three clustering models are considered: 2-level, 3-level and 5-level clustering (Very Low, Low, Medium, High, Very High).	The possibility of using it for online courses as the attendance on the course platform is applied, the time spent on the platform. Does not give a complete picture of the student's progress, only involvement in the lesson	The opportunity to see complete and detailed information about the student's academic performance, helping to improve his academic performance.
5	Applications of machine learning to student grade prediction in quantitative business courses [19]	Three different algorithms are used, such as: naive Bayes method, KNN, SVM to determine the final grade of the course.	The data were not normalized to improve the forecast indicators, and the impact of forecast factors was not analyzed.	Data preprocessing was carried out to improve the results of forecasting, as well as the impact of data on the forecast.

In the conclusion of previous works [1], [12], [13], [14], [15], [11], [16],[17],[18] related to predicting various academic indicators and student attrition, the main value and distinction of this study is that it categorizes students into four groups, making it easier for teachers to identify the most appropriate support for each student. Additionally, this study examines two cases of incomplete learning: student dropout and student receiving an F grade in the course.

Chapter 3

Methodology

3.1 Data description and preprocessing.

3.1.1 Feature Selection

Returning to the main purpose of the research, which is to provide a comprehensive picture of academic performance (final exam scores, dropout prediction, and one of the four categories of academic achievement in a subject [20]) for each student taking a particular course, the researchers [18] obtained data from the North American University [18] as predictors for conducting experiments. The data used in this study consists of the academic performance of students in the natural sciences course who were enrolled in the second year of their undergraduate program at the University of Western Ontario in Canada.The full description of each of the predictors used in this article is described below:

1. **Visits**: How many times does a student attend classes.
2. **Study of materials**: How many times has the student read the course material in the learning platform.
3. **Reading notifications**: The number of readings of notifications from the teacher on the learning platform.
4. **Quiz**: The score obtained from the first quiz

5. **Assignment 1**: Evaluation of the very first completed task.
6. **Midterm**: The score obtained from the first midterm.
7. **Assignment 2**: Evaluation of the second completed task.
8. **Assignment 3**: Evaluation of the third completed task.
9. **Assignments submission delay**: Missing a deadline.
10. **Final Exam**: The score obtained from the final exam.
11. **Participation in the discussion**: Number of participants in the lesson.

The F1 score classification quality score, which measures the accuracy and completeness of data, is calculated as follows in Equation(3.1.1) :

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.1.1}$$

In addition to the F1 score evaluation for assessing the accuracy of positive predictions, the "Precision" evaluation in Equation(3.1.2) and "Recall" which measures how well the model detects all positive examples in Equation(3.1.3) is used based on the following calculation.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{3.1.2}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{3.1.3}$$

The accuracy evaluation of the model's predictions was also calculated using the following Equation(3.1.4)

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ number\ of\ Predictions} \tag{3.1.4}$$

The collected data contains 486 rows about each student with 11 predictors, where the distribution of data is as follows:

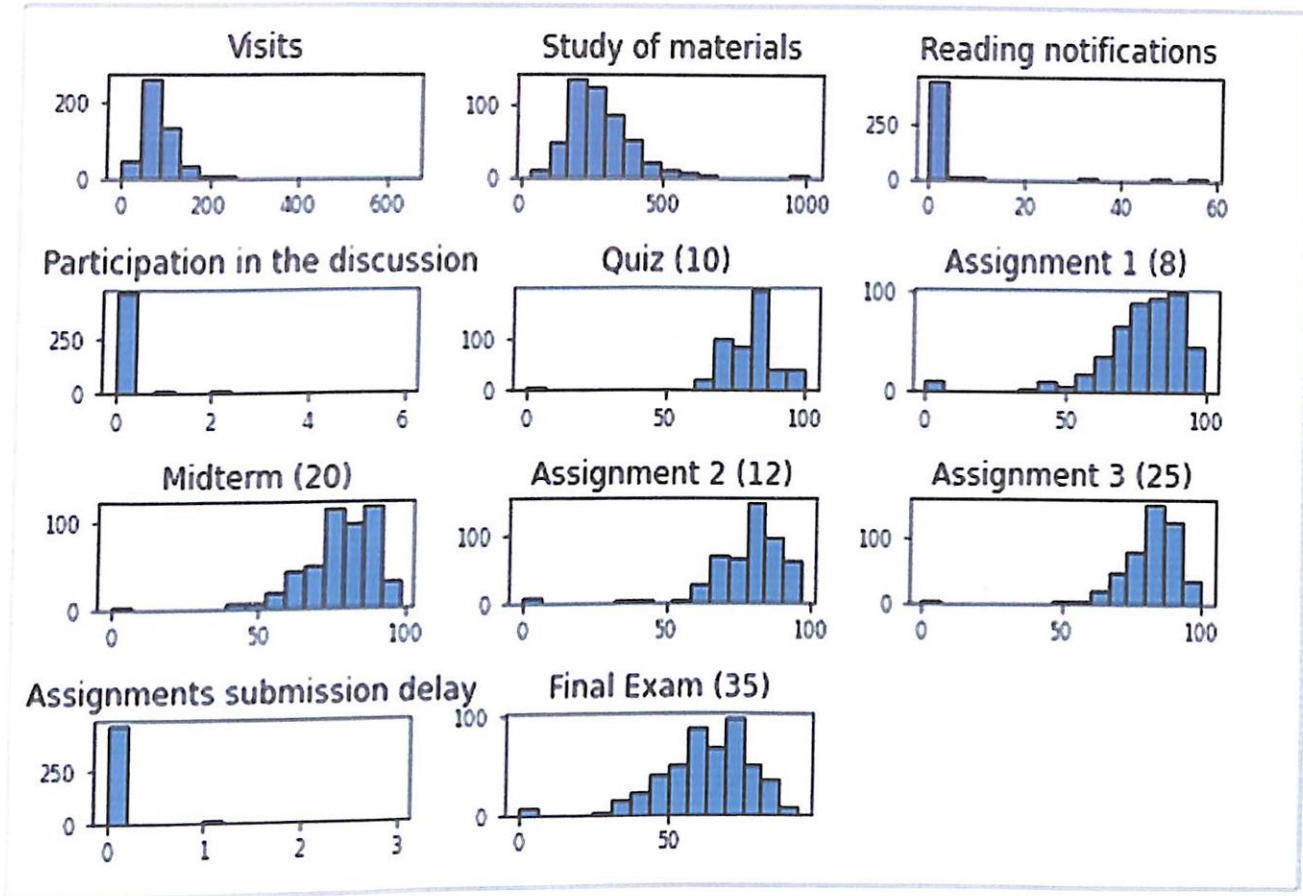

Figure 3.1: Data distribution

Before training the models, an analysis of the collected data served as a mandatory part of this study in order to understand whether there are any dependencies in the data or perhaps some factors do not affect the forecast of academic performance and can be removed for further work. To identify correlations between predictors in this study, a correlation heatmap and a statistically significant test are used.

The figure below 3.2 shows a correlation heatmap of all predictors in this study, where a warmer shade (red color) indicates a dependency between variables and represents the degree of correlation between them.

After applying the correlation table, we can conclude that we have dependent data among themselves, namely, these are three different indicators of students' grades with a maximum correlation coefficient equal to 0.56 (Table 3.2).

For a more visual representation of the relationship between correlating data in

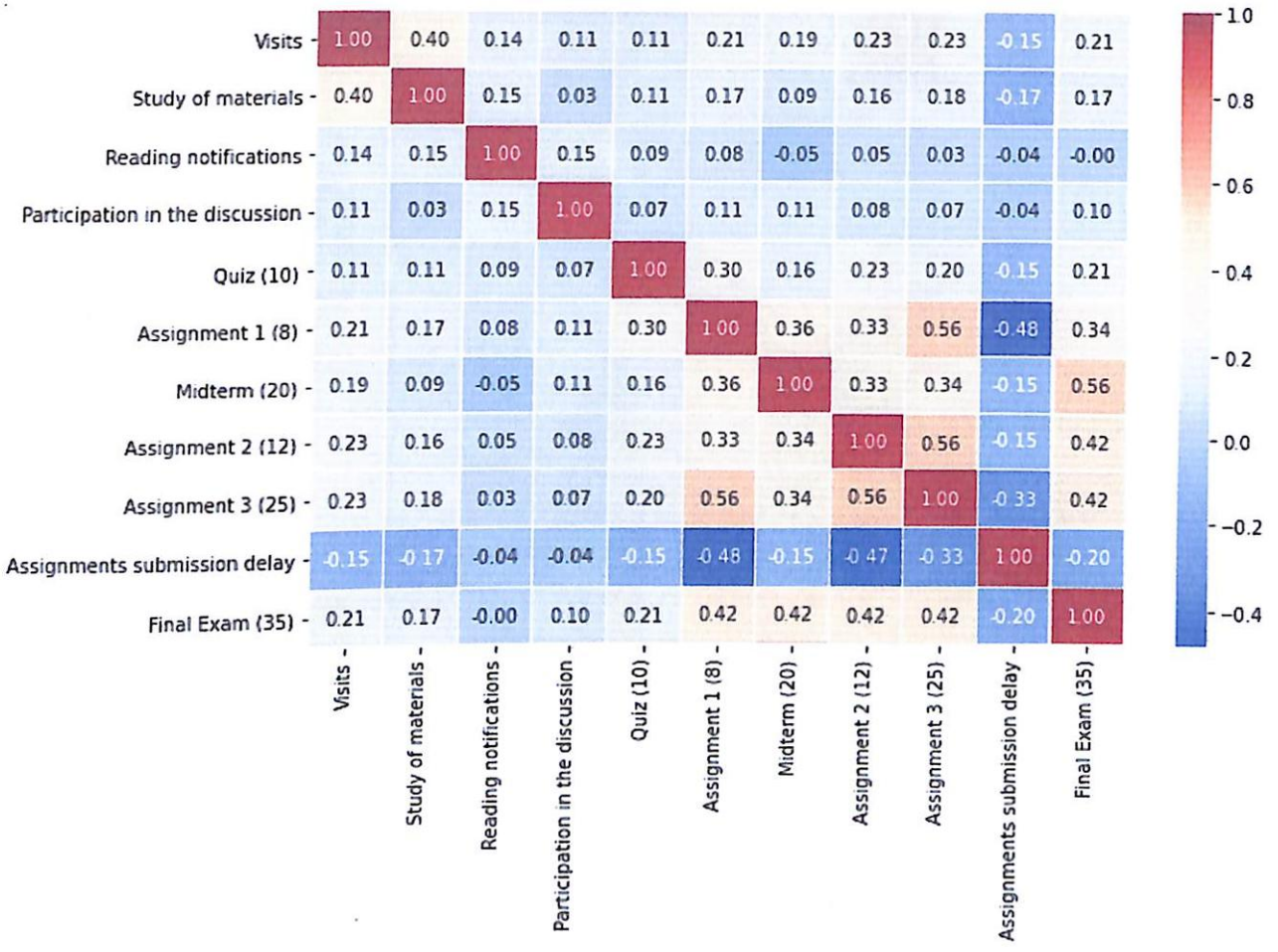

Figure 3.2: Correlation heatmap between predictors

order to understand why there is a dependence in these data, as well as the overall distribution of data on these correlating predictors, a paired data distribution graph was constructed 3.3, where you can see exactly how the data is distributed across the three predictors.

As shown in the graph below 3.3, in fact, there is a dependence in our data and we can also conclude that if a student receives a good grade on the first task, then he is more likely to receive a good grade for the next two tasks and vice versa, those who received bad grades for the first task also receive low grades for subsequent tasks. This may be one of the reasons that students are already demotivated after receiving a bad grade on the first task and do not want to study the subject properly further.

In addition to the correlation between predictors, another factor was added

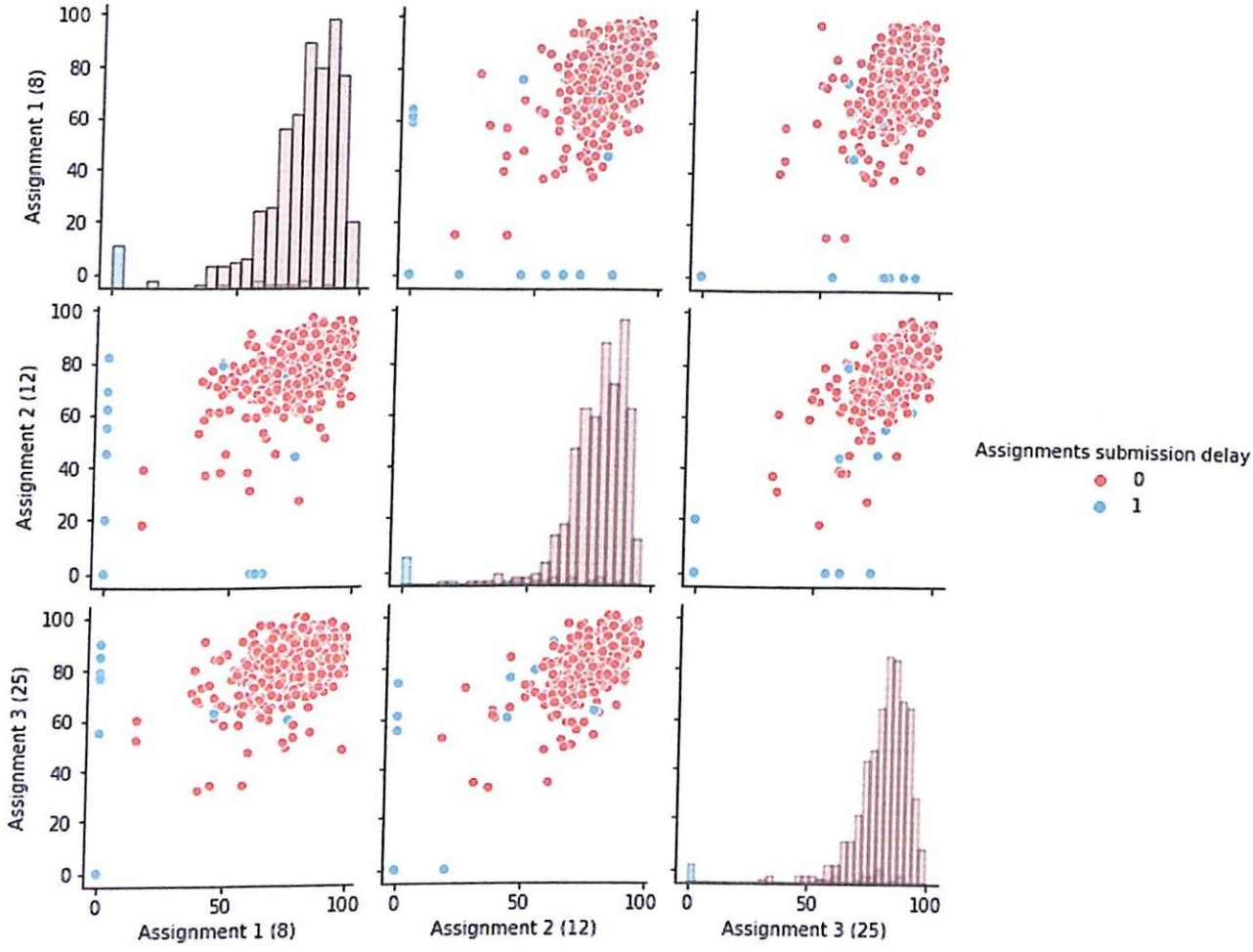

Figure 3.3: Assignments predictors Pairwise Plots

regarding the timely submission of assignments. Students who have missed at least one assignment submission are represented by the blue color, while the pink color 3.3 indicates that the student did not miss any assignment submission. By observing the above graph, we can conclude that those students who missed submissions generally received very low scores on the assignments.

Based on the visual representation of the relationship between predictor variables, it was crucial to select only significant factors that influence students' academic performance for further investigation. To achieve this, a statistically significant test was additionally employed in this study, where the p-value represents the probability of obtaining the observed data if the null hypothesis is true. The null hypothesis, in turn, indicates that there is no statistically significant relationship between the variables. If the p-value is less than 0.05, we can reject the null hypothesis and conclude that there is a difference between the vari-

ables, indicating that these predictors should be considered for further analysis.As shown in the figure below, the statistical method Ordinary Least Squares shows detailed information for checking the relationship between the dependent variable (in this study, this is the final exam score) and independent multiple variables. The result of OLS regression 3.4 shows various indicators about standard errors of coefficients, t-values, p-values and other statistical indicators.In the figure below, the p-values associated with the t-values indicate the probability of observing coefficient estimates if there is no relationship between the independent variables and the dependent variable. The value of predictors with p less than 0.05 indicates statistical significance that the independent variable has a significant effect on the dependent variable, for this reason the variable should be used for further research.

OLS Regression Results

Dep. Variable:	Final Exam (35)	R-squared:	0.575
Model:	OLS	Adj. R-squared:	0.571
Method:	Least Squares	F-statistic:	129.9
Date:	Wed, 26 Apr 2023	Prob (F-statistic):	7.58e-87
Time:	10:19:31	Log-Likelihood:	-1658.2
No. Observations:	5486	AIC:	3328.
Df Residuals:	5475	BIC:	3353.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.1637	0.107	29.592	0.000	2.954	3.373
Visits	0.0621	0.010	6.513	0.000	0.043	0.081
Study of materials	0.0340	0.005	6.995	0.000	0.024	0.043
Reading notifications	0.0062	0.018	0.344	0.000	-0.029	0.041
Participation in the discussion	0.0043	0.021	0.207	0.000	-0.036	0.045
Quiz (10)	0.0005	0.008	0.062	0.000	-0.016	0.017
Assignment 1 (8)	0.0069	0.009	0.759	0.000	-0.011	0.025
Midterm (20)	0.3037	0.009	34.606	0.000	0.286	0.321
Assignment 2 (12)	0.1122	0.010	11.256	0.000	0.093	0.132
Assignment 3 (25)	0.0907	0.010	9.358	0.000	0.072	0.110
Assignments submission delay	-0.2826	0.045	-6.317	0.000	-0.370	-0.195
Study_of_materials	0.0340	0.005	6.995	0.000	0.024	0.043

Figure 3.4: OLS Regression Results

The results obtained using OLS regression 3.4 indicated that all independent variables have an effect on the dependent variable, since the p value for all independent variables was less than 0.05, for this reason all factors were used to predict academic performance in the future.

3.1.2 Outliers identification

In addition to identifying influencing factors for predicting students' academic performance, this study analyzes the detection of outliers in the data [21]. Detecting outliers in data is an important process before starting model training, this is due to the following reasons:

1) **Distortion of statistical forecasting models.** If there are outliers in the collected data and they are not processed properly, this may lead to incorrect estimates of the model parameters.

2) **Inaccuracy of data.** The appearance of outliers in the collected data may occur due to inaccuracy in filling in the data, if the data were collected manually, there may often be errors in the data. The accuracy in the data will help to build a more accurate model that we expect to get.

3) **Influence on the result obtained.** Outliers can significantly distort the predicted results and only replacing or deleting outliers in the data will help to avoid this situation.

4) **Normal distribution in the data.** For some forecasting models, it is assumed that the data has a normal distribution. Outlier detection helps to observe the normal distribution in the data, for accurate data predictions.

5) **Incorrect conclusions from forecasts.** Distortion of emissions leads to the fact that general final conclusions can lead to incorrect conclusions that should contribute to solving any problems.

The points listed above give an understanding of exactly how emissions affect the prediction of models, as well as how much emissions can distort the results obtained and lead to erroneous conclusions. Before starting to remove or replace emissions, the task was to determine whether there are emissions in the data. To do this, graphical visualizations were initially used, which help to visually see if there are outliers in the data, and a function was also used with which it was possible to find outliers in all data columns at the same time. Graphical visualizations were built using the available graphical ways to visualize the distribution of data:

- **Box plot.**

  This is one of the types of graphs that displays static data of indicators. The graph has a rectangular box shape where its boundaries are the first quartile (Q1) and the third quartile (Q3). In addition to the rectangular shape of the box, vertical lines are visible that indicate the range of data, and what is outside the range is considered outliers and is displayed as dots.

- **Histogram.**

  One of the types of graphical representation, where data is distributed columnally and the X-axis indicates the distribution of data, and the Y-axis indicates the frequency of such distribution. With histogram graph, you can see how many times a value occurs in the data, and in addition to the main distribution in the data, you can see outliers or other anomalies that require analysis.

Visual graph with distribution of data help to understand exactly where the outliers are, however, for further rapid analysis of outlier detection in this research, a function based on a statistical indicator - IQR (Interquartile Range) [21], which is based on the distribution of quartiles in the data set, was used. The IQR method sorts the data in a certain order and then finds three quartiles in the data. The first quartile Q1 denotes the value below which 25% of the data is located, the second quartile Q2 denotes the accurate value below which 50% of the data is located, and the last third quartile Q3 is the value below which 75% of the data is located. Further, to calculate this indicator, the upper limit is subtracted from the lower one, and then the upper and lower emission limits are determined, that is, everything that will be outside these limits is emissions. This IQR was used to determine the presence of outliers in the data , which determines in which column how many outliers there are.

3.1.3 Data Normalization

Data normalization is extremely important for model training, since it is the normalized data that helps to achieve high results of model prediction accuracy.

Its function is to compare the data in such a way that the data is not very scattered, that is, so that for all the available signs (predictors), the data are compared. This is of particular importance because if the features in the data with large values can significantly affect the model, and those predictors where possible are data with smaller values, they will not affect the model at all. Data normalization helps to bring each of the predictors into one range, that is, it scales the data in such a way that there will not be one where one predictor has large values, another factor in the data will have very small values that will not have a significant impact on the model. Currently, there are various ways to normalize data, where the main ones are:

1) **Min-Max Scaling.**

   Min-Max Scaling (this is a way of scaling data by minimum value and maximum), that is, bringing all the values of each predictor to a certain interval, usually from 0 to 1.

2) **Z-Score Standardization.**

   Z-Score Standardization (standardization of available Z-score data), in this method, the values are converted in such a way that they have an average value of 0 and a standard deviation of 1.

3) **Logarithmic transformation.**

   In this case, to reduce various distortions and biases in the distribution of values, the application of a logarithmic function to the data is used.

In this study, the Min-Max method was used [21], which is one of the fastest and most effective ways of normalization, the task of which is to first determine the minimum and maximum values in the predictors and then convert each value from 0 to 1 so that all distributions in the data remain the same, but the minimum and maximum for each predictor there were the same ones.

3.1.4 Imbalance of the available data

An imbalance in the available data occurs if one class or one category in the data significantly exceeds its share over another category or forecast class. In this study, the class of students with a "PASS" mark [20], indicating that a student has successfully completed a course in a subject, exceeds its share over the class of a minority of "NOT PASS" students who have not successfully completed the course. The class of students who passed made up about 79%, while the class of students who did not pass made up 21% of all students.

Such an unbalanced balance between classes or other categories leads to the fact that the model that will be trained on the provided data will only be able to give correct results for further forecasting of the majority class. At this time, when the model is well trained in the majority class, the class of students who have not completed the course in further forecasting will be determined in most cases incorrectly and with low prediction accuracy, that is, this suggests that in a larger case the model will be able to correctly identify only those students who complete the course successfully.To eliminate this situation, 2 methods were used in this study [20] to regulate the disequilibrium in the data: Near Miss, SMOTE.

The SMOTE method is used to increase the minority class, that is, based on the data on the minority class, the method tries to increase the set based on synthetic data to the point that the class with more data equates to the class with less data. SMOTE automatically selects its nearest neighbors for each of the minor class values, and then new synthetic data is created between the selected value and its one of the closest neighbors. This process is repeated for each of the data values and stops until a balance is reached between the data of the class of students who have completed the course successfully [20] and those who have not completed the course successfully.

Near Miss is the opposite of the SMOTE method, however, both methods have one main function to balance data between prediction classes. It is the exact opposite because it tries to reduce the data set from the majority class until the data of one class exceeds its number of the other class. For each of the data set of the minority class [20], the nearest neighbors are selected, and then the data

from the majority class that are close to the neighbors are selected and then they are removed from the data set. The process is dynamic and stops until it equates the two classes.

3.2 Proposed Method

After working with data preprocessing, the next necessary task for the study was to train models for multidisciplinary prediction of students' academic performance, namely, to determine:

1. **Final Exam Score** - determination of the final grade on the final exam of the course.
2. **"Pass/Not Pass"** - determining which of the two groups "Pass", "Not Pass" refers to each student of the course.
3. **Academic Performance Category** - determining which performance group a student belongs to in the course [20]:

    1) A

    2) B

    3) C

    4) D

Final Exam Score

Determining the final grade for the course is a regression task, where any value is predicted by a number of influencing predictors on the model. In this study, to predict the final grade for the course, models of conventional **Linear regression** and **LGBM regressor** with an increase in gradient were experimented. The gradient Increasing algorithm (LGBM Regressor) [21] was used with the three most well-known parameters as: GOSS, GBDT, DART. To determine the best boosting parameter, the GridSearchCV method of automatic search for optimal parameters was used, which selects the best parameter of all of them set for a more accurate and better prediction.

"Pass/Not Pass"

Identification of students at risk is one of the main tasks of this study [20], where 7 most well-known algorithms for solving classification problems were used to identify students at risk of not completing the course successfully at an early stage of training : **DecisionTreeClassifier, Naïve Bayes, Logistic Regression, SVM Classifier, MLPClassifier, BaggingClassifier, KNN**. As hyperparameters for more successful prediction, cross-validation or in other words cross validation with 30-fold repetition was used, as well as the GridSearchCV method for optimal search of the best hyperparameters of the model.

Academic Performance Category

To solve the clustering task for predicting one of the 4 groups of academic performance in a course that a student belongs to, the following algorithms were used: **K-Means, BIRCH and MiniBatch K-Means**. At the same time, the analysis of the main components or PCA was used for each of the algorithms, which was necessary because the data contains more than two predictors. The main task for the analysis of the main components is to reduce the dimensionality of the data so that the loss of important information for forecasting is sufficiently minimal. Before applying various clustering algorithms, the elbow method (Elbow Rule) was used in this study, which automatically selects the required number of clusters in the data. In addition, to check how well the observations were clustered, the silhouette coefficient was used, in cases where the silhouette coefficient is the highest compared to others, this means that choosing the number of clusters with this number will cluster the data better.

Chapter 4

Result

As a result of data preprocessing, totally 34 outliers [20]were identified in the all data and all outliers replaced by the average values of each predictor. The graphs below show outliers as dots of the "Quiz" predictor.

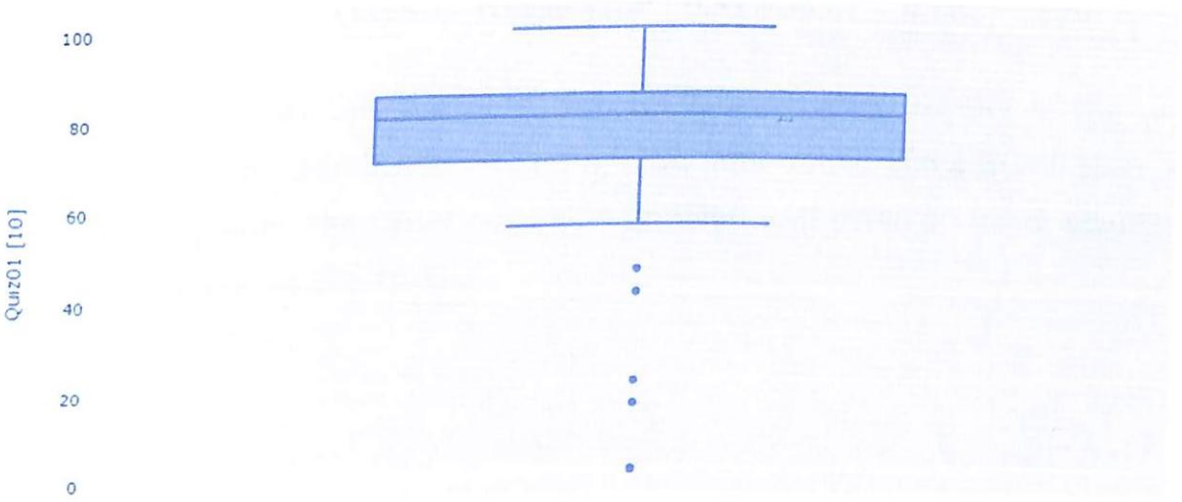

Figure 4.1: Outliers for 'Quiz' predictor.

4.1 Final Exam Score

The results of identifying the final grade on the final exam showed that the ensemble method of increasing the gradient - LGBM Regressor [21], copes much better with the prediction of the final exam grade than using conventional Linear regression. In addition, the automatic selection of the best hyperparameters (GridSearchCV)[21] determined that the type of the GOSS boosting model surpasses the results in predicting the estimate.

As shown in the table below, the results of the boosting model method surpassed Linear Regression with an estimate of the coefficient of determination R2 = 0.81, which is 0.24 more than the estimate of the determination of Linear regression (more details in the table 4.1)

Table 4.1: Final Exam Score Prediction Results.

Linear Regression vs LGBM(GOSS) Regressor				
Algorithm Name	R2	MSE	MAE	MAPE
Linear regression	0.52	78.47	4.93	3.69
LGBM(GOSS) Regressor	0.81	33.57	4.06	3.07

The visual representation (Fig.4.2) of the prediction results of the two algorithms in the figures below makes it clear how much the LBGM Regressor algorithm surpasses the usual Linear Regression and gives accurate results on the final assessment of the course.

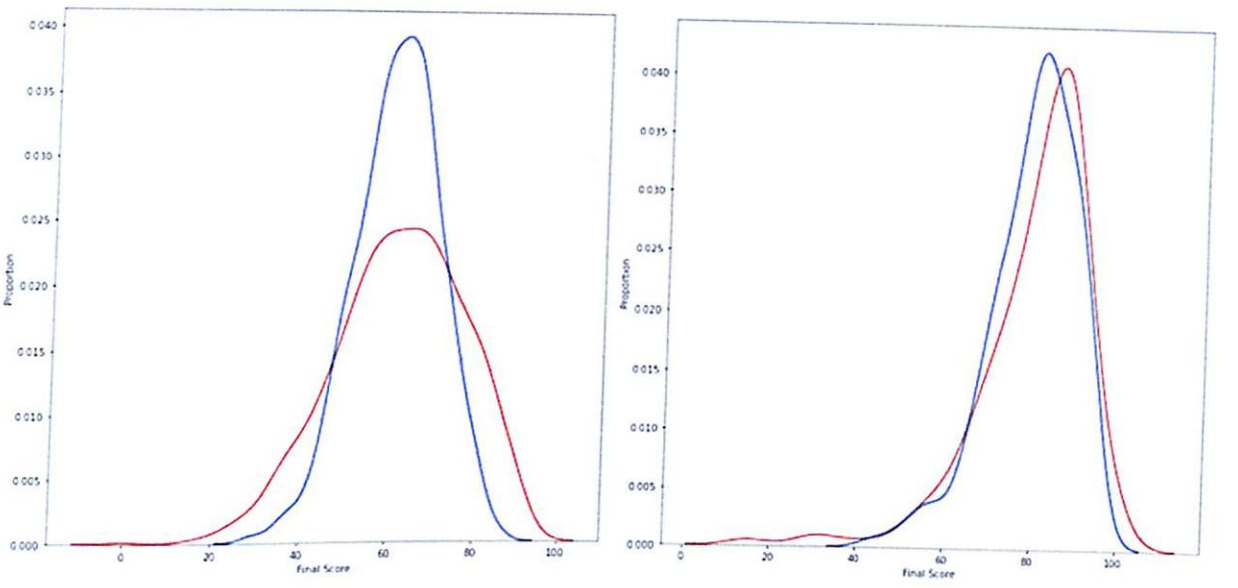

Figure 4.2: Actual vs Forecast data of Linear Regression (left) and LBGM Regressor(right).

4.2 "Pass/Not Pass"

The application of 7 different classification algorithms [20] for predicting students who can not finish the course successfully and those who complete the course successfully showed that using Logistic regression with a cross 30-fold check determines 81% successfully those students who complete the course successfully and 53% successfully those who do not complete the course. Compared to the other 6 algorithms, Logistic Regression showed better results (Table 4.2) in predicting two classes of students, those who successfully complete the course and those who do not well complete the course.

In addition, this study used two methods of balancing data between the majority class (those students who will complete the course) and the minority class (students who will not complete the course successfully: SMOTE and Near Miss. In comparison with the Near Miss method, the SMOTE method contributed to increase the prediction accuracy for the minority class of students who did not successfully complete the course by 11% [20], where the recall rate of the Near Miss method is 2% less compared to the SMOTE method.

Table 4.2: Classification Results.

Class	Not Pass			Pass		
Name	Precision	Recall	F1-score	Precision	Recall	F1-score
Logistic Regression	42%	72%	**53%**	91%	73%	**81%**
MLPClassifier	37%	78%	51%	92%	65%	76%
DecisionTreeClassifier	34%	52%	41%	85%	73%	78%
BaggingClassifier	43%	53%	47%	87%	81%	84%
KNN	37%	74%	50%	85%	71%	75%
SVM Classifier	43%	53%	47%	90%	71%	79%
Naïve Bayes	42%	72%	50%	86%	71%	79%

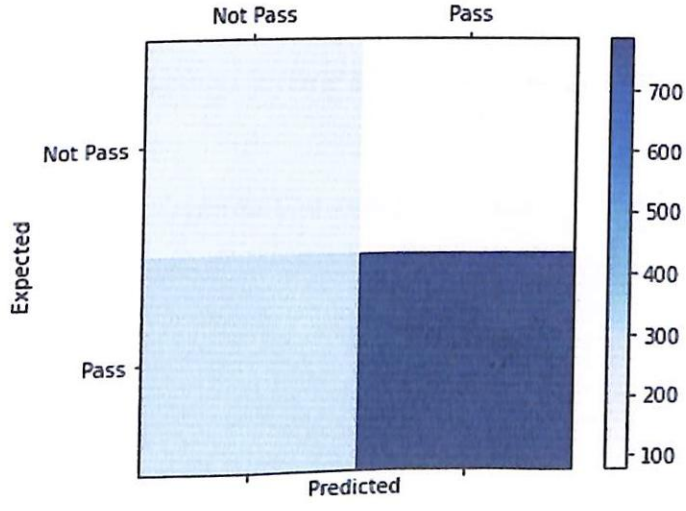

Figure 4.3: The results of Logistic Regression - Confusion Matrix.

4.3 "Academic Performance Category"

The results of using the elbow method showed that the most optimal value of K for identifying groups of students with academic performance is 4 clusters. In the figure below (Fig.4.4), a dashed dotted line indicates the K value, which is the most optimal value for splitting data into groups.

The "Elbow" graph helps to visually assess the optimal number of clusters or the value of the parameter, which, as we see in the figure below, is equal to 4. On

the graph, it is noticeable how the bend begins exactly when K = 4 and then it is reduced and becomes less noticeable.

Figure 4.4: Distortion Score Elbow for K-means Clustering

The graphs below (Fig.4.5) show the results of three different cluster algorithms, where each of the four groups of students is painted in its own specific color:

1. **Birch** (Fig.4.5)
2. **K-Means** (Fig.4.6)
3. **MiniBatch K-Means** (Fig.4.7)

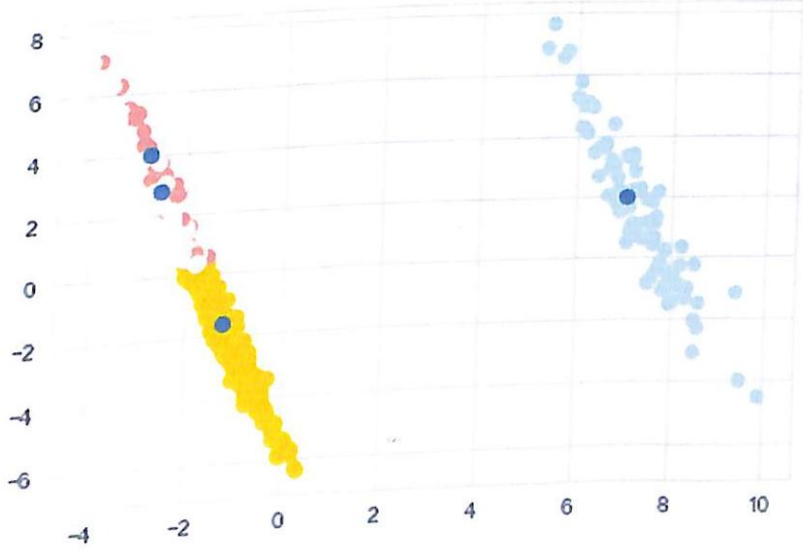

Figure 4.5: BIRCH Clustering Result.

Comparing the clustering of students into 4 groups of the K-Means algorithm and BIRCH algorithm models, we can conclude that at the moment both algorithms have approximately equally divided students into 4 groups of academic performance. This gives us an understanding that both algorithms are very similar to each other, however, to determine the accuracy in clustering, it was necessary to check the silhouette coefficient. Compared to the previous two algorithms (Birch, K-Means), MiniBatch showed a completely different result. As shown in the graph (Fig.4.7), both on the right side of the graph and on the left side of the graph, the students were divided into completely different groups, even when they are very close to each other.

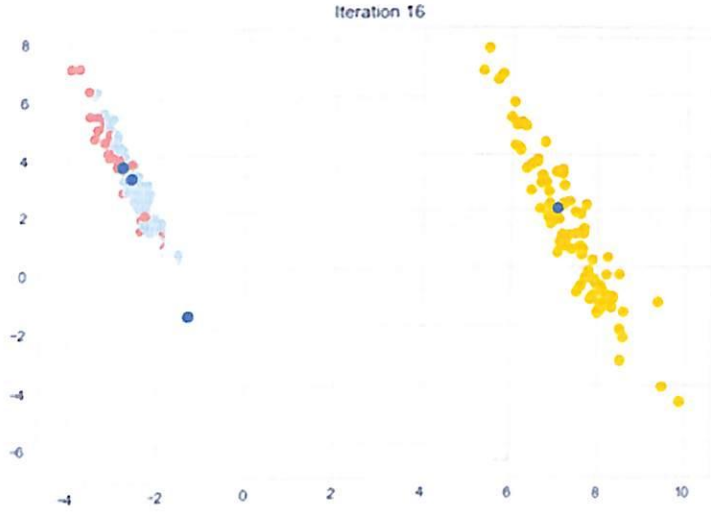

Figure 4.6: K-Means Clustering Result.

To compare the accuracy of the clusters obtained in this study, the **silhouette coefficient** was used, the task of which is to determine how well each of the algorithms coped with the division of students into groups of academic performance. In turn, the coefficient of determination is an indicator that measures how much the values within one cluster are similar to each other compared to the values of data from other clusters.

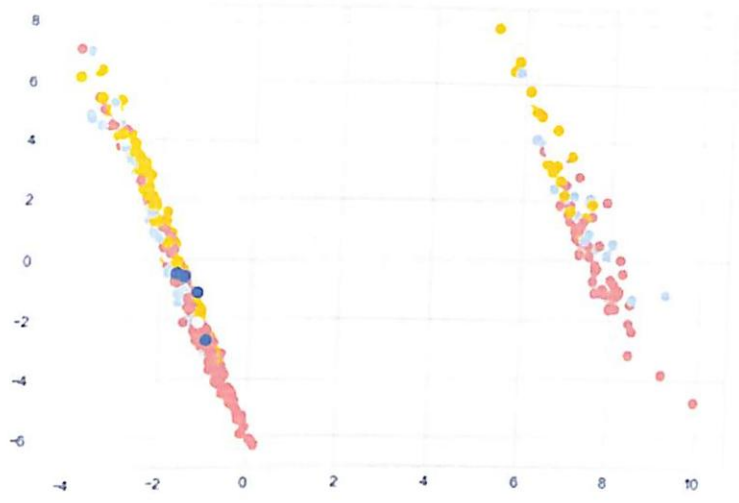

Figure 4.7: MiniBatch K-Means Clustering Result.

As a result, the higher the value of the silhouette coefficient for each of the algorithms, the better the clustering quality. The indicator helps to choose the optimal number of clusters or evaluate the effectiveness of the clustering algorithm. The graphs below (Fig.4.5) show the results of the silhouette coefficient for each of the clustering algorithms.

Birch Clustering

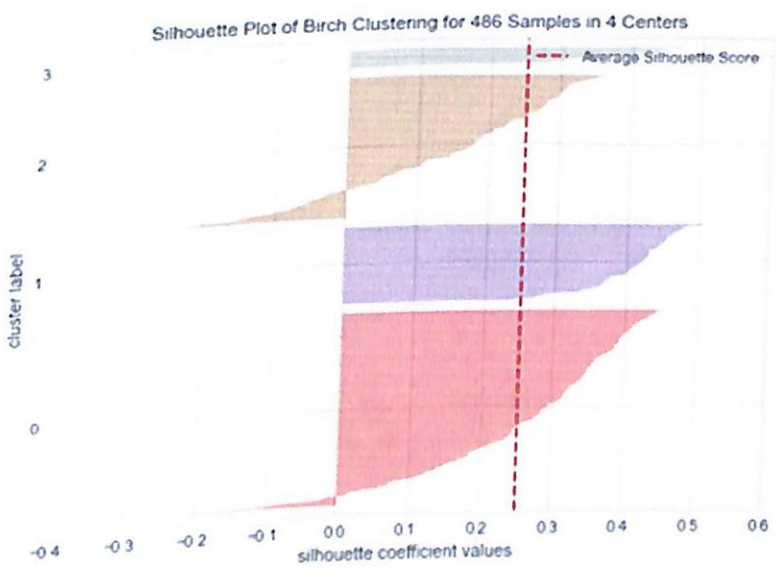

Figure 4.8: Silhouette Plot of Birch Clustering.

K-Means Clustering

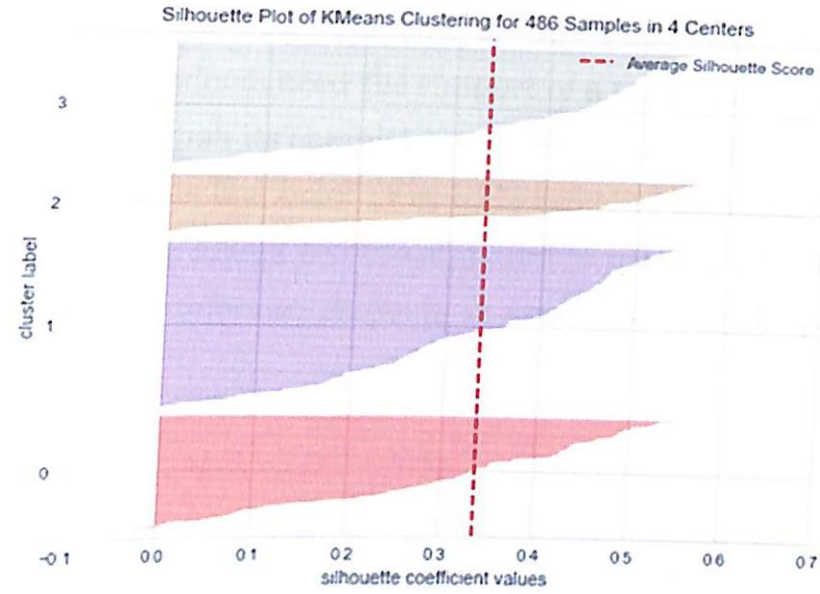

Figure 4.9: Silhouette Plot of K-Means Clustering.

Silhouette coefficients for the K-Means algorithm showed better results compared to the Birch algorithm. For the K-Means algorithm, the silhouette coefficient is 0.58 (Fig.4.9), while for the Birch algorithm 0.51 (Fig.4.8), it also has negative coefficient values, which indicates incorrect clustering of data.

MiniBatch K-Means Clustering

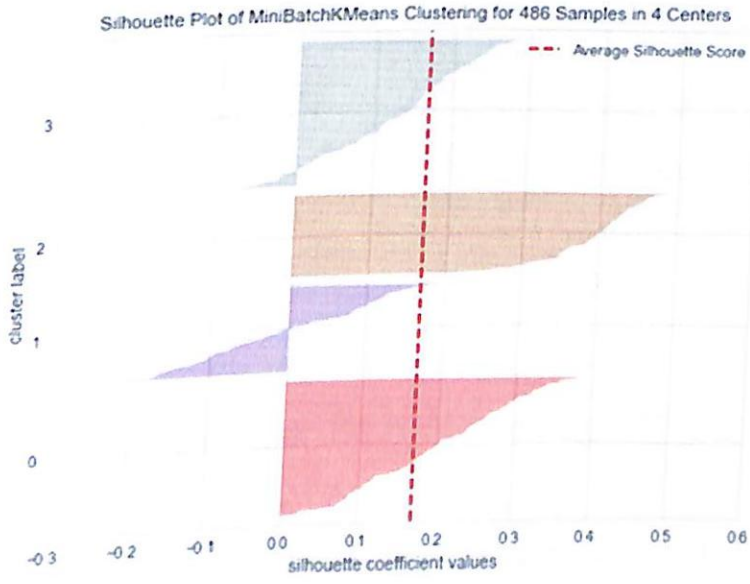

Figure 4.10: Silhouette Plot of MiniBatch K-Means Clustering.

always attend classes, but these students do any tasks while receiving not high scores. This group of students need the support of a teacher and may not complete the course successfully upon its completion. The last fourth group of students with a mark "3" with a purple line color (Fig.4.11) refers to students who are late for assignments, receives the lowest scores compared to other students on many types of assignments and also refers to students in need of teacher help.

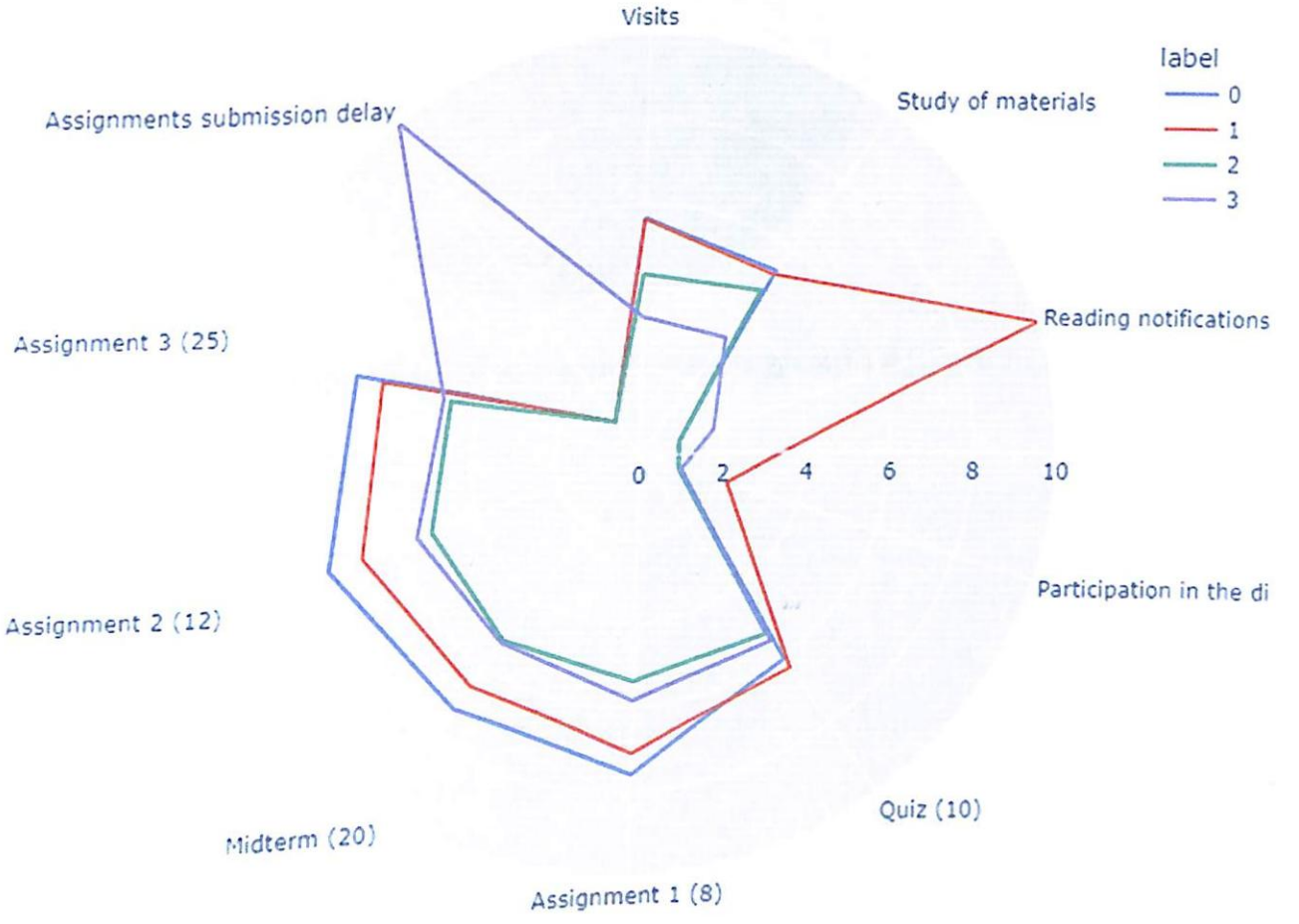

Figure 4.11: Line Polar for K-Means Algorithm Result.

In general, the K-means data clustering algorithm divided the data into 4 groups, where the group of students who successfully complete the course and have high academic performance amounted to 63.8% (group A and B), where the rest are students with group C and D (Fig.4.12).

C and D group of students make up the remaining 36.2% of students in need of teacher assistance and are students from the group who can not finish the course.

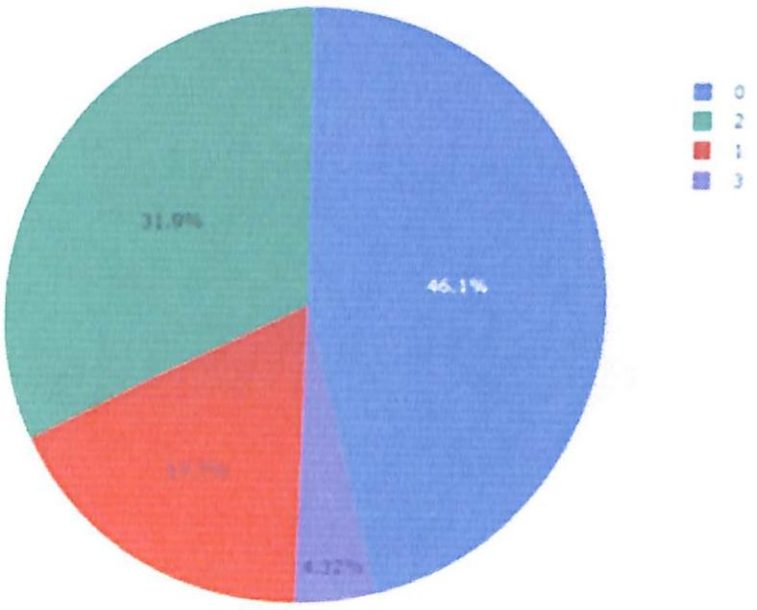

Figure 4.12: Proportion of students in each group (A(0),B(1),C(2),D(3)).

Chapter 5

Conclusions and future work

5.1 Conclusions

This dissertation is devoted to the topic of multidisciplinary prediction using various machine learning algorithms for the course being studied in order to determine the full picture of academic performance in the context of each student. The main purpose of this study is to successfully determine at an early stage of training that a student belongs to a group of students who can not finish the course at an earlier stage of training for timely intervention and support of the teacher as well as timely elimination of gaps in the study of the subject by the student himself. Identifying students can not finish the course successfully can increase the level of motivation in learning a subject at an earlier date, before the course ends.

In addition, a complete picture of a student with important academic indicators can help course teachers develop various individual teaching methods for each student or groups of students with similar academic indicators for a better understanding of the course by students and also increase motivation among students for a more successful understanding and completion of the course.

The full picture of the academic performance of each student is determined by the following indicators, which were analyzed in detail in this study: the final grade that the student will receive on the exam, the group to which the student

belongs (will pass the course successfully or will not pass successfully), as well as one of the 4 categories of students to which the student with various academic indicators for completing tasks belongs, attendance of lessons, activity in the lesson, as well as various points obtained as a result of the course.

In comparison with previously performed works, this study differs in that each student as well as the teacher himself can see the full picture of students' academic performance in the context of various academic indicators that can make it clear exactly where students have gaps in knowledge and how exactly students can be helped to complete the course successfully and increase their motivation in studying the course during, in order for the student to have time to be active and involved in the course before the end of the subject.

To display a complete picture of a student's academic performance, various machine learning algorithms were used, which helped to see complete information about each student. First of all, before training the models, various data preprocessing methods were used, which in turn helped to avoid various problems with inaccurate classification and prediction.

In addition, various methods were used in this study to determine the significance and influence on the predictive indicators of students.After preliminary data processing, 7 classification algorithms were used to determine that a student belongs to a group that completed the course unsuccessfully or unsuccessfully, where the logistic regression algorithm shows the accuracy of the forecast for a class of students who will graduate the course successfully by 81%, as well as the accuracy of the forecast by 53% for a minority class of students who will not graduate the course successfully. To predict the final grade for the final exam, two algorithms were used from the regression type as: Linear regression and LGBM Regressor with the type of boosting model "GOSS", which showed the coefficient of determination of Linear Regression by 0.24 more.

Also, one of the distinctive parts of this study is the prediction of one of the 4 groups (A, B, C, D) of the classization of students according to various academic indicators. Each of the cluster groups indicates the level of attraction, motivation of students in learning the course, as well as the need for help from the teacher

to provide timely support for the successful completion of the course.This study used three different algorithms for clustering students as K-Means, MiniBatch K-Means and Birch algorithm. The K-Means algorithm with the optimal number of clusters = 4 showed the most accurate result for clustering students. Using all three different academic indicators, students and teachers can freely and easily pay attention to gaps in knowledge, improve their performance in the early stages of learning, as well as increase efficiency in studying any course.

5.2 Future Work

In the future, using these indicators of academic performance, it would be better if we introduce this system of forecasting academic indicators into each of the studied disciplines and check the effectiveness of using the system for each of the subjects. In addition, in order to predict students from the minority group who will not pass the course successfully, it is necessary to improve the quality of forecasts using various other machine learning algorithms. Also, one of the subsequent research works could be the introduction of this prediction into online platforms for teaching students.

Bibliography

[1] Cameron I. Cooper. Using machine learning to identify at-risk students in an introductory programming course at a two-year public college. Journal of Ohio State University, 2(2):407–421, July 2022. URL https://www.oajaiml.com/uploads/archivepdf/97051127.pdf.

[2] Tenzin Doleck David John Lemay, Paul Bazelais. Transition to online learning during the covid-19 pandemic. Computers in human behavior reports, 4 (100130), August - December 2021. doi: 10.1016/j.chbr.2021.100130. URL https://pubmed.ncbi.nlm.nih.gov/34568640/.

[3] Claire Paul R. Pintrich, Donald R. Brown. Student motivation, cognition, and learning. Taylor and Francis Group, 1994. doi: https://doi.org/10.4324/9780203052754. URL http://surl.li/icmqv.

[4] Charles Fadel Wayne Holmes, Maya Bialik. Artificial intelligence in education. promise and implications for teaching and learning. The Center for Curriculum Redesign, 2019. doi: 10.1007/978-3-030-10576-1_107. URL http://udaeducation.com/wp-content/uploads/2019/05.

[5] Scott D. Willging, Pedro A.; Johnson. Factors that influence students' decision to dropout of online courses. Journal of Asynchronous Learning Networks, 13(3):115–127, Oct 2009 2022. doi: https://doi.org/10.4324/9780203052754. URL https://eric.ed.gov/?id=EJ862360.

[6] M.L.E. Mapesela M.P. Jama and A.A. Beylefeld. Theoretical perspectives on factors affecting the academic performance of students. South African Journal of Higher Education, 22(5), 1 Jan 2008. URL https://hdl.handle.net/10520/EJC37488.

[7] Jianwu Wang Athanasios V. Vasilakos Lina Zhou, Shimei Pan. Machine learning on big data: Opportunities and challenges. Journal of Asynchronous Learning Networks, 237(3):350–361, 10 May 2017. doi: https://doi.org/10.1016/j.neucom.2017.01.026. URL https://www.sciencedirect.com/science/article/abs/pii/S0925231217300577.

[8] Ethem Alpaydin. Machine learning: The new ai. page 224, 07 October 2016. URL https://dl.acm.org/doi/book/10.5555/3017658.

[9] K. R. Srinath. Python – the fastest growing programming language. 4(12): 354–357, Dec-2017. URL https://www.irjet.net/.

[10] Sarah Guido Andreas C. Müller. Introduction to machine learning with python. page 224, September 2016. URL https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/.

[11] Christopher Frederick W. B., Watson and Li. Failure rates in introductory programming revisited. Annual Conference on Innovation and Technology in Computer Science Education, pages 39–44, 2014. URL https://dro.dur.ac.uk/19223/.

[12] Erkan Er. Identifying at-risk students using machine learning techniques: A case study with is 100. International Journal of Machine Learning and Computing , 2(4):476–480, 08 August 2012. doi: 10.7763/IJMLC.2012.V2.171. URL http://www.ijmlc.org/show-32-132-1.html.

[13] V. Nikolopoulos G. Mpardis I. Lykourentzou, I. Giannoukos and V. Loumos. Dropout prediction in e-learning courses through the combination of machine learning techniques. Education Resources Information Center, 53(3):950–965, 11 Nov 2009. URL https://eric.ed.gov/?id=EJ848773.

[14] C. Pierrakeas S. Kotsiantis and P. Pintelas. Preventing student dropout in distance learning systems using machine learning techniques. Springer-Verlag, 2774:267–274, 2003. URL https://link.springer.com/article/10.1007/s10462-022-10155-y.

[15] Huzefa Rangwala Yujing Chan, Aditya Johri. Running out of stem: A comparative study across stem majors of college students at-risk of dropping

out early. Proceedings of the 8th International Conference on Learning Analytics and Knowledge, March 2018. doi: 10.1145/3170358.3170410. URL https://www.researchgate.net/publication/323860348_Running_out_of_STEM_a_comparative_study_across_STEM_majors_of_college_students_at-risk_of_dropping_out_early.

[16] Student final grade prediction based on linear regression. 'failure rates in introductory programming revisited. Indian Journal of Computer Science and Engineering (IJCSE), 8(3):274–279, Jun-Jul 2017. URL http://www.ijcse.com/docs/INDJCSE17-08-03-080.pdf.

[17] Georgios Bekas Christos Troussas Cleo Sgouropoulou Zoe Kanetaki, Constantinos Stergiou. A hybrid machine learning model for grade predictionin online engineering education. International Jounal of Engineering Pedagogy, 12(3), 2022-05-31. doi: https://doi.org/10.3991/ijep.v12i3.23873. URL https://online-journals.org/index.php/i-jep/article/view/23873.

[18] A. Shami A. Moubayed, M. Injadat and H. Lutfiyya. Student engagement level in an e-learning environment: Clustering using k- means. American Journal of Distance Education , 34(2):137–156, March 2020. URL https://www.irrodl.org/index.php/irrodl/article/view/6453.

[19] T. Anderson. Applications of machine learning to student grade prediction in quantitative business courses. 1(3), 2017. URL https://www.igbr.org/wp-content/uploads/articles/GJBP_Vol_1_No_3_2017-pgs-13-22.pdf.

[20] D. Bairamova. Data collection to identify students at risk of not completing a course using machine learning. Natural and Technical Sciences, 2:151–166, March 2023. URL https://journals.sdu.edu.kz/index.php/nts/article/view/931.

[21] D. Bairamova. Predicting course grades of students academic performance using the lightgbm regressor. Natural and Technical Sciences, 1:34–

48, March 2023. URL https://journals.sdu.edu.kz/index.php/nts/article/view/952.

Appendix A

Outliers identification

```
def impute outliers IQR(df):
q1=df.quantile(0.25)
q3=df.quantile(0.75)
IQR=q3-q1
upper = df[ (df>(q3+1.5*IQR))].max()
lower = df[ (df<(q1-1.5*IQR))].min()
df = np.where(df > upper,
df.mean(),
np.where(
df < lower,
df.mean(),
df ) ) return df
```

Appendix B

Students clustering with Birch Algorithm

```
from numpy import unique
  from numpy import where
  from sklearn.datasets import make classification
  from sklearn.cluster import Birch
  from matplotlib import pyplot
  model = Birch(threshold=0.01, n clusters=4)
  model.fit(df)
yhat = model.predict(df)
clusters = unique(yhat)
for cluster in clusters:
row ix = where(yhat == cluster)
plot clusters1(df)
pyplot.show()
```

Appendix C

K Elbow and Silhouette Visualizer with

```
from sklearn.cluster import KMeans
from yellowbrick.cluster import KElbowVisualizer
SilhouetteVisualizer
model = KMeans(random state=42)
elb visualizer = KElbowVisualizer(model, k=(2,11))
elb visualizer.fit(df)
elb visualizer.show()
sil visualizer = SilhouetteVisualizer(model)
sil visualizer.fit(df)
sil visualizer.show()
```