Ministry of Education and Science of the Republic of Kazakhstan
Suleyman Demirel University

Akerke Abilkassym, Begayim Kamar, Aruzhan Omarbekova, Nagima Zhailau

Grade Prediction System

A thesis submitted for the degree of
Bachelor in Computer Systems and Software and Bachelor in Information Systems
(degree code: 5B070400, 5B070300)

Kaskelen, 2020

Ministry of Education and Science of the Republic of Kazakhstan
Suleyman Demirel University
Faculty of Engineering and Natural Sciences

Grade Prediction System

A thesis submitted for the degree of
Bachelor in Computer Systems and Software and Bachelor in Information Systems
(degree code: 5B070400, 5B070300)

Author: **Akerke Abilkassym, Begayim Kamar, Aruzhan Omarbekova, Nagima Zhailau**

Supervisor: **Ardak Shalkarbay-uly**

Dean of the faculty:
Prof. Andrey Bogdanchikov

Kaskelen, 2020

Abstract

Most universities have a portal where students can see their academic performance. We developed the system that later can be integrated into the university portal. The aim of this project is to improve student achievement and the quality of education in university. This system gives more information about courses of the upcoming semester and will be used during the course registration. With the help of this system students can analyze the courses they want to choose and make timely decisions. Another part of this system is that with the help of machine learning algorithms we predict final exam grade using midterm exams that helps students to know earlier their predicted grade and will be useful also for teachers to know about students' performance.

Аңдатпа

Көптеген университеттерде студенттердің академиялық үлгерімдерін көретін портал бар. Біз кейінірек университеттің порталына қосыла алатын жүйені жасадық. Бұл жобаның мақсаты - студенттердің оқу жетістіктерін және университеттегі білім сапасын жақсарту. Бұл жүйе алдағы семестр курстары туралы көбірек ақпарат береді және курсты тіркеу кезінде қолданылады. Бұл жүйенің көмегімен студенттер өздері таңдағысы келетін курстарды талдап, уақытылы шешім қабылдай алады. Бұл жүйенің тағы бір бөлігі - машиналық оқыту алгоритмдерінің көмегімен аралық бақылаудың көмегімен қорытынды емтихан бағасын болжай аламыз, бұл студенттерге алдын-ала болжанған бағасын білуге және мұғалімдерге оқушылардың үлгерімі туралы білуге көмектеседі.

Аннотация

В большинстве университетов есть портал, где студенты могут увидеть свои академические результаты. Мы разработали систему, которая впоследствии может быть интегрирована в университетский портал. Целью данного проекта является повышение успеваемости студентов и качества образования в университете. Эта система предоставляет больше информации о курсах предстоящего семестра и будет использоваться при регистрации курса. С помощью этой системы студенты могут анализировать курсы, которые они хотят выбрать, и принимать своевременные решения. Другая часть этой системы заключается в том, что с помощью алгоритмов машинного обучения мы прогнозируем итоговую оценку экзамена, используя промежуточные экзамены, которые помогают учащимся раньше узнать свой прогнозируемый балл и будут полезны также учителям для оценки успеваемости учащихся.

Contents

Chapter 1

Introduction

1.1 Data Analysis and its importance in education

Higher educational institutes are considered as the first step to adulthood. It is an important time for everyone, it decides our future. So, university's tasks are to support students everywhere and provide everything they need. At the beginning of university life, everything will be new for 1st year students. So, we have an orientation week that helps students to feel confident in a new environment. This is done in order to get used to in general, but still there will be problems in university portal when they should register courses. They will need an explanation about the courses in general. Higher class students also face this problem when they should choose elective courses, because there is no support for them. Usually, they receive this kind of information from previous year students. To solve this problem we made a system using data analysis.

Data Analysis is an effective way to know about what is happening with the data in the background. Using data visualization, machine learning gives us the possibility to easily understand and analyze the processes and make decisions. Nowadays, using Data Analysis in universities and colleges is regarded as the modern tool to improve the education system. We choose the Microsoft Power BI as the best way to visualize the data. Using Power BI data visualization tools, machine learning algorithms, python libraries like Pandas, Matplotlib, Scikit-learn we analyze students' performances and make predictions. It will help students know about their courses and check their academic achievement.

1.2 Thesis outline

2. Data Collection and Preparation
3. Data Visualization
4. Correlation and Prediction

Chapter 2

Data Collection and Preparation

This chapter covers everything related to the dataset. In this section, we discuss details to collect data for prediction. We will prepare, clean, decompose and finally present appropriate data.

2.1 Data Collection

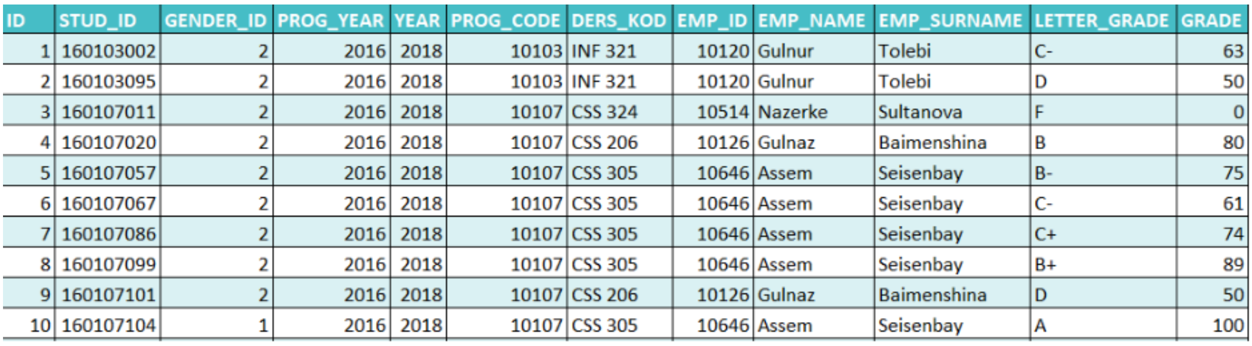
Since this project is related to the grading system of university, to solve our analysis questions, we began our activities by collecting data from the Suleyman Demirel University's portal. This dataset includes all passed courses of 2016,2017,2018,2019 year registered students of Information Systems (IS), Computer Systems and Software (CSS) (Engineering faculty, SDU). The number of students enrolled in the course for a given year ranged from 200 to 550, with the overall number of about 1600 students in the dataset. Each student record has the following attributes: student name, student ID, gender, final GPA, ENT scores and all the courses taken by the student with the teacher data who taught this course, and including the course grades. As for the grade system, we have given weight of the performance assessments in the Portal: 30% midterm 1, 30% midterm 2 and 40% final exam. However in this academic year (2019) SDU changed the grading policy to 60%(FA) that contain the weight of midterm exams and 40% final exam.

2.2 Data Preparation and Pre-Processing

Firstly we removed records with null values where students dropped the course as well as students who did not get a grade. We applied some pre-processing for the collected data to get appropriate data for various predictions. We prepared a dataset to predict GPA using ENT scores, predict final exam grades, make a correlation between similar courses and analyze elective courses related to Data Science. As we know that the algorithm is learning from past years, the predictions become more accurate when more data from previous years become available, then we added academic years exactly 2014,2015 for elective courses.

Additionally, with absence_count we wanted to predict the final grade of course, so we pre-processed data with absence and attendance count. But, there we encountered issues such as data deficiency. In analysis we find out that the electronic attendance system began to be used only in 2017 years, and changes the limit of absent counts 2 and more times. That's why we don't predict course grades by absence, but in the future it will be a good way to predict final grades.

After preprocessing we are left with 5011 records for 1st year courses visualization, 721 records for final grade prediction, 863 records for elective courses visualization, 3098 records to predict GPA and 589 records for correlation between similar courses. In total – 10282 records.



ID	STUD_ID	GENDER_ID	PROG_YEAR	YEAR	PROG_CODE	DERS_KOD	EMP_ID	EMP_NAME	EMP_SURNAME	LETTER_GRADE	GRADE
1	160103002	2	2016	2018	10103	INF 321	10120	Gulnur	Tolebi	C-	63
2	160103095	2	2016	2018	10103	INF 321	10120	Gulnur	Tolebi	D	50
3	160107011	2	2016	2018	10107	CSS 324	10514	Nazerke	Sultanova	F	0
4	160107020	2	2016	2018	10107	CSS 206	10126	Gulnaz	Baimenshina	B	80
5	160107057	2	2016	2018	10107	CSS 305	10646	Assem	Seisenbay	B-	75
6	160107067	2	2016	2018	10107	CSS 305	10646	Assem	Seisenbay	C-	61
7	160107086	2	2016	2018	10107	CSS 305	10646	Assem	Seisenbay	C+	74
8	160107099	2	2016	2018	10107	CSS 305	10646	Assem	Seisenbay	B+	89
9	160107101	2	2016	2018	10107	CSS 206	10126	Gulnaz	Baimenshina	D	50
10	160107104	1	2016	2018	10107	CSS 305	10646	Assem	Seisenbay	A	100

Figure 2.1: Sample of student dataset

Chapter 3

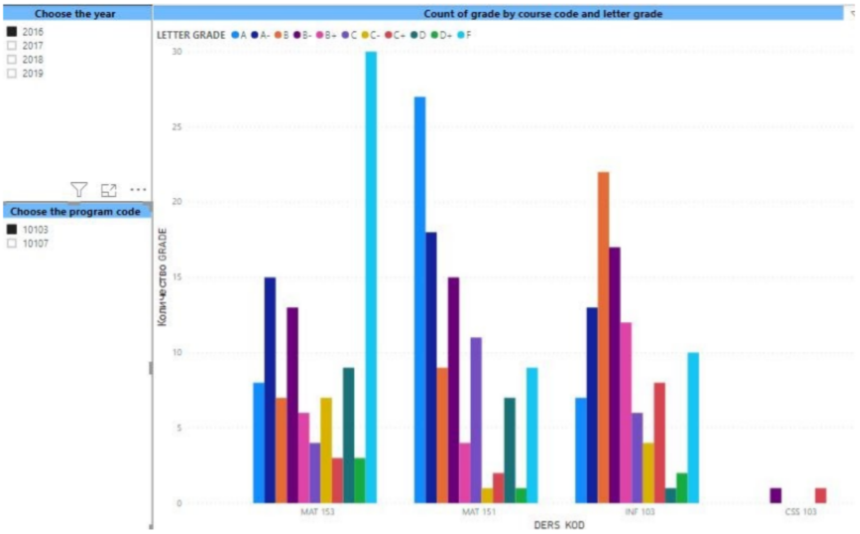
Data Visualization

Data visualization is the graphical representation of data using visual elements like Graphs, Maps and Charts with filters. In addition data visualization tools allow trends and patterns to be more easily seen. We used PowerBI to show our data analysis in the Dashboard. The dashboard consists of two parts that help students deeper understand about their courses.So, we divided this chapter into 2 parts:

3.1. 1st year courses dashboard
3.2. Elective courses dashboard

3.1 1st year courses dashboard

This is the first part of this project, it can be used at the beginning of the semester. It is like an “orientation dashboard” for 1st year students. The diagram below shows counts of each letter grade by year, by program code (10103 – IS, 10107 – CSS). In all cases, the Grades are represented ordinally from A to F.



Choose the year
■ 2016
□ 2017
□ 2018
□ 2019

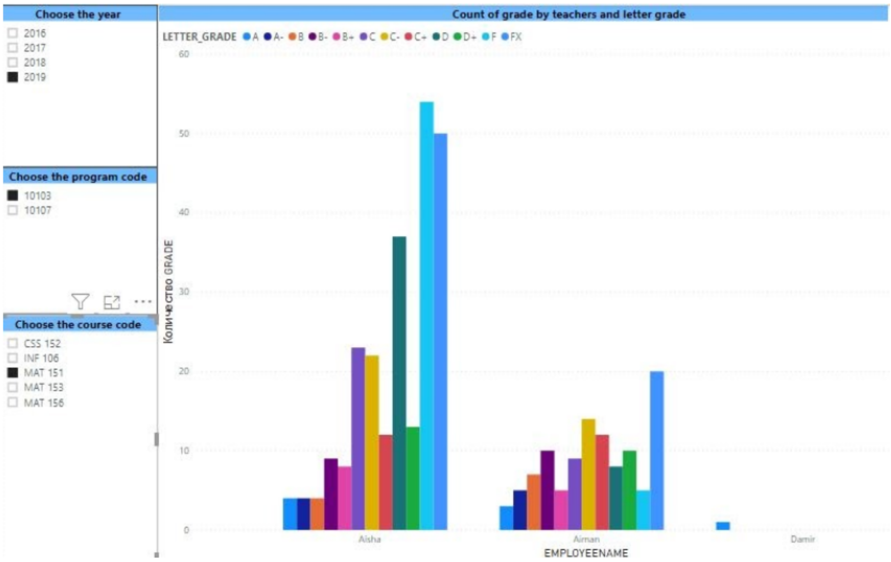
Choose the program code
■ 10103
□ 10107

Figure 3.1: Grade count of 2016 year students of IS

In the dashboard we have 2 types of widgets, Filter and Bar Chart. With the help of these filters in the third widget, exactly in Bar Chart data will change. In this case we selected year as 2016 and program code as 10103, it means we get IS students grade by 2016 year.

As you can see from these figures, we can probably see exactly which courses are difficult to learn and have more F-Fails. With this data students can get chances to give all strength and more time for these courses to get a good score.

In the second part we analyze and design a report by teachers. It can help students to choose the teachers according to their grade differences, because all teachers' grading rules are different from each other.



Choose the year
□ 2016
□ 2017
□ 2018
■ 2019

Choose the program code
■ 10103
□ 10107

Choose the course code
□ CSS 152
□ INF 106
■ MAT 151
□ MAT 153
□ MAT 156

Count of grade by teachers and letter grade
LETTER_GRADE ● A ● A- ● B ● B- ● B+ ● C ● C- ● C+ ● D ● D+ ● F ● FX
60
50
40
30
20
10
0
Количество GRADE
Aisha
Aiman
Damir
EMPLOYEENAME

Figure 3.2: Grade count of 2019 year students of IS by teachers

In Figure 3.2 we presented the list of teachers and students' progress by course. A majority of the students choose Aisha teacher, and we can see that more of them failed the course and barely achieved the D.

3.2 Elective courses dashboard

As we know some courses can be easier than others and some of them can be hard for all students even if the student's GPA is the highest. This kind of prediction gives information for students when they should choose the elective courses and even can help in mandatory courses and he will know for which course he must try better. So there we analyze data and design dashboard report for elective courses.

In this part, we got detailed data about elective courses related to Data Science. Such as Introduction to Data analysis, Introduction to Machine Learning, Database Management and etc (in Figure 3.3). Moreover, in this part we will also show the list of teachers who taught these courses (in Figure 3.4).

There we encountered issues such as data shortages. As mentioned earlier the predictions become more accurate when more data from previous years become available.

Due to the fact that we have a choice to choose any of the pre-taught lessons, here we can see that these courses were chosen by a minority of students and to solve this problem we added academic years datasets.

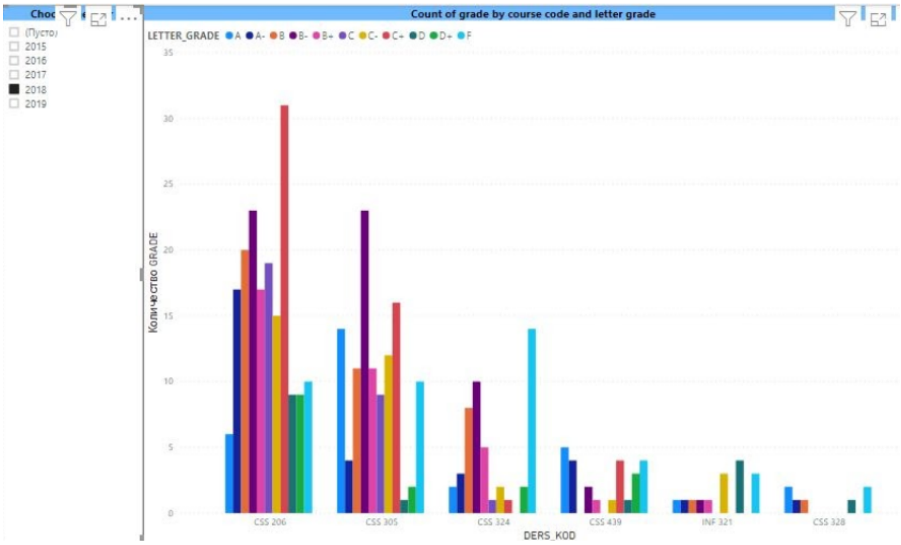


Figure 3.3: Grade count of 2018 year students of IS elective courses

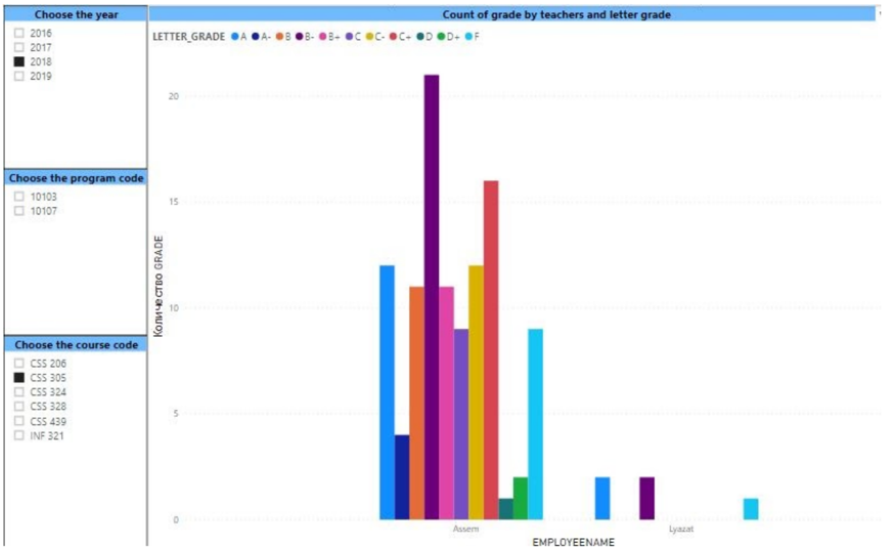


Figure 3.4: Grade count of 2018 year IS students' elective courses by teachers

Finally, we realized that the majority of students prefer another sphere of elective courses. This may be due to teacher choice and difficulty of the course. Generally, these reports help us to understand data easily by visual widgets and see all the flaws.

Chapter 4

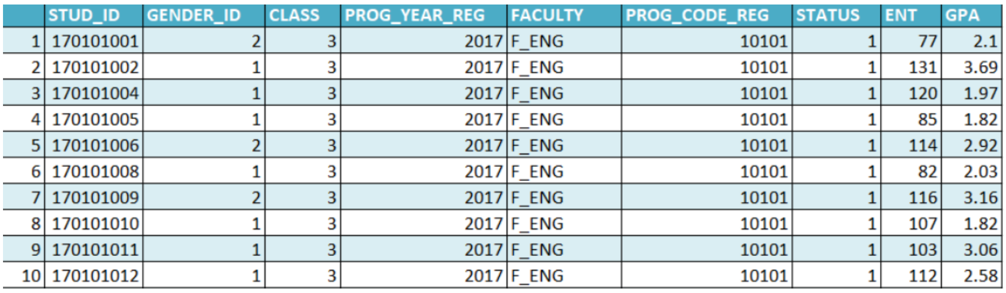
Correlations and predictions

This chapter includes making correlations and predictions using students' information. It describes which features have more effect on students' performance and how we can predict grades. With the help of this, students can improve their academic performances and also helps teachers analyze students' grades. We used Machine Learning algorithms to make correlations, predictions and Python Visual in PowerBI to visualize the results in the dashboard. We made three types of predictions, so divided this chapter into three parts:

4.1. Correlation between ENT score and GPA
4.2. Correlation between similar courses
4.3. Final grade prediction

4.1 Correlation between ENT score and GPA

At the beginning of our project, we had the idea to predict GPA using ENT score, in order to find out how the results of ENT affect the GPA and if the school performances of students are similar to university performances. We collected data from the SDU portal. It contains 2017,2018,2019 year students' main information. We removed the rows where GPA is less than 1.47 and ENT score less than 50. After preprocessing, we are left with 3098 records. Data includes Student ID, gender ID, class, year, faculty code, program code, status, ENT score and GPA (Figure 4.1). Firstly, we decided to know the relation between ENT



	STUD_ID	GENDER_ID	CLASS	PROG_YEAR_REG	FACULTY	PROG_CODE_REG	STATUS	ENT	GPA
1	170101001	2	3	2017	F_ENG	10101	1	77	2.1
2	170101002	1	3	2017	F_ENG	10101	1	131	3.69
3	170101004	1	3	2017	F_ENG	10101	1	120	1.97
4	170101005	1	3	2017	F_ENG	10101	1	85	1.82
5	170101006	2	3	2017	F_ENG	10101	1	114	2.92
6	170101008	1	3	2017	F_ENG	10101	1	82	2.03
7	170101009	2	3	2017	F_ENG	10101	1	116	3.16
8	170101010	1	3	2017	F_ENG	10101	1	107	1.82
9	170101011	1	3	2017	F_ENG	10101	1	103	3.06
10	170101012	1	3	2017	F_ENG	10101	1	112	2.58

Figure 4.1: Dataset with ENT score and GPA

score and GPA. We calculated Pearson correlation coefficient using the Python Pandas package. The result was low positive correlation with coefficient r = 0.185. Here we can say that the relationship between ENT score and GPA is weak and unimportant (Figure 4.2).

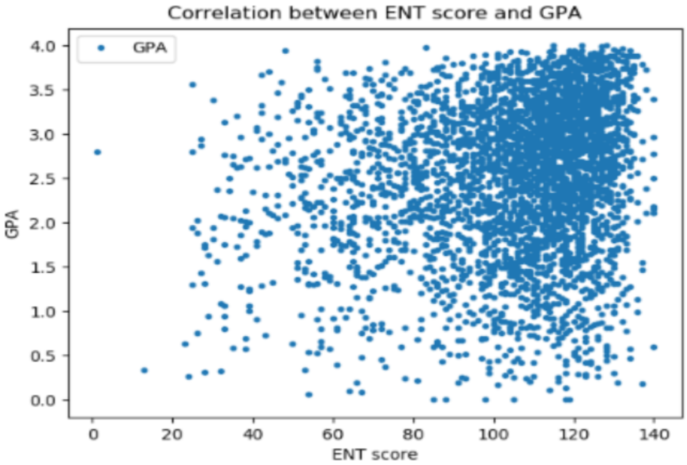


Figure 4.2: Correlation between ENT score and GPA

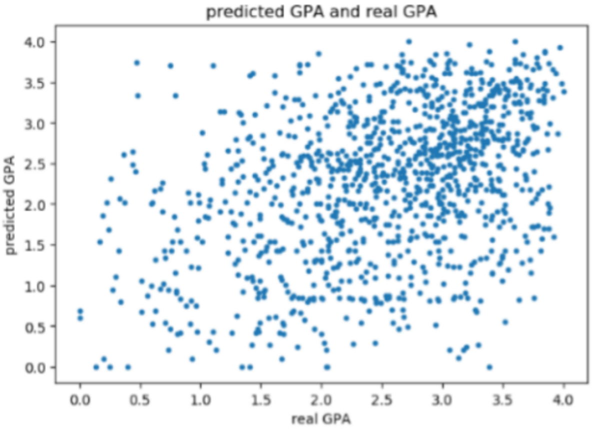
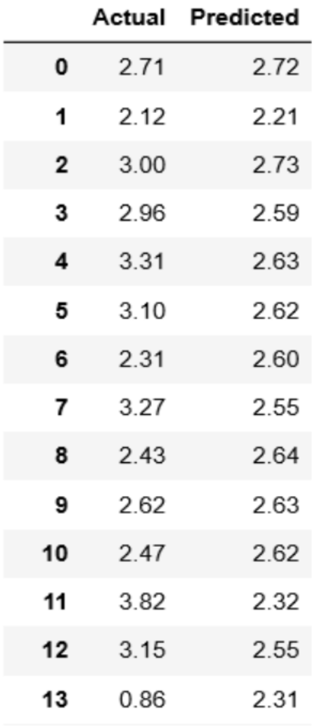


Figure 4.3: Predicting GPA using class,program code and ENT score

Using only an ENT score to predict GPA gives no result. So, we took the GPA as a target variable and in order to make a prediction with higher probability we added some features like class and program code. Because GPA changes every semester, depending on chosen courses and faculties of students also affect the change in the GPA . For example, mostly Engineering faculty students' GPA is lower than others.

We chose one of the well known machine learning algorithms - Decision Tree, because it's easy to understand, can work with any type of data and does not always require data preparation. We built it in Python library Scikit-learn using DecisionTreeClassifier() class. The probability is extremely low. It's impossible to predict GPA with +50% accuracy, because GPA is a very small number (Figure 4.3, 4.4) and from the results of correlation, we can say that the ENT score



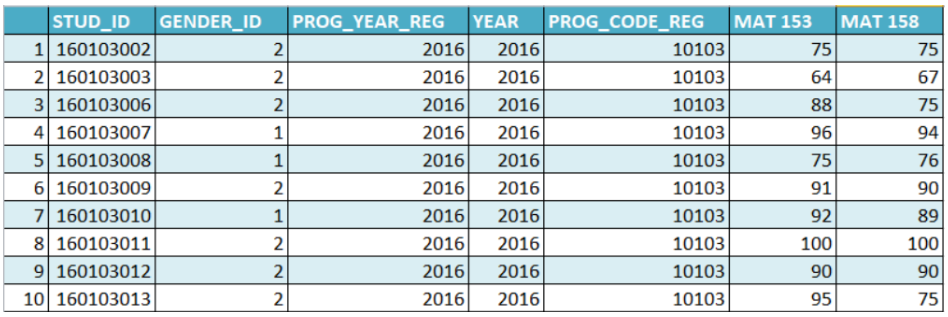
	Actual	Predicted
0	2.71	2.72
1	2.12	2.21
2	3.00	2.73
3	2.96	2.59
4	3.31	2.63
5	3.10	2.62
6	2.31	2.60
7	3.27	2.55
8	2.43	2.64
9	2.62	2.63
10	2.47	2.62
11	3.82	2.32
12	3.15	2.55
13	0.86	2.31

Figure 4.4: Predicting GPA using class,program code and ENT score

strongly does not affect the change in GPA. But if we have more data to test, the probability increases and it will predict more accurately.

4.2 Correlation between similar courses

Many students' grades from similar courses are often closer to each other. Because the same forces are leaving and the same skills will be needed. Predicting a course grade using similar courses gives an opportunity for students to analyze the grade and for teachers to know the difference between grades. We took MAT 153 and MAT 158 (Mathematics for Computer Science 1,2) as an example and collected students' data from the university portal. It includes Student ID, Gender ID, Registered year, Course passed year, Program code and final grade from MAT



	STUD_ID	GENDER_ID	PROG_YEAR_REG	YEAR	PROG_CODE_REG	MAT 153	MAT 158
1	160103002	2	2016	2016	10103	75	75
2	160103003	2	2016	2016	10103	64	67
3	160103006	2	2016	2016	10103	88	75
4	160103007	1	2016	2016	10103	96	94
5	160103008	1	2016	2016	10103	75	76
6	160103009	2	2016	2016	10103	91	90
7	160103010	1	2016	2016	10103	92	89
8	160103011	2	2016	2016	10103	100	100
9	160103012	2	2016	2016	10103	90	90
10	160103013	2	2016	2016	10103	95	75

Figure 4.5: Sample of dataset with final grades from MAT 153 and MAT 158

153 and MAT 158. We removed rows with null values, so overall we are left with 589 records (Figure 4.4).

We made a correlation between these courses using the Python Pandas package. The result was high positive correlation with Pearson correlation coefficient r=0.65. It is a very good result to predict. The diagram in Figure 4.5 shows the correlation between these courses. You can see that the relationship between these courses is very good, even if the teachers have changed.

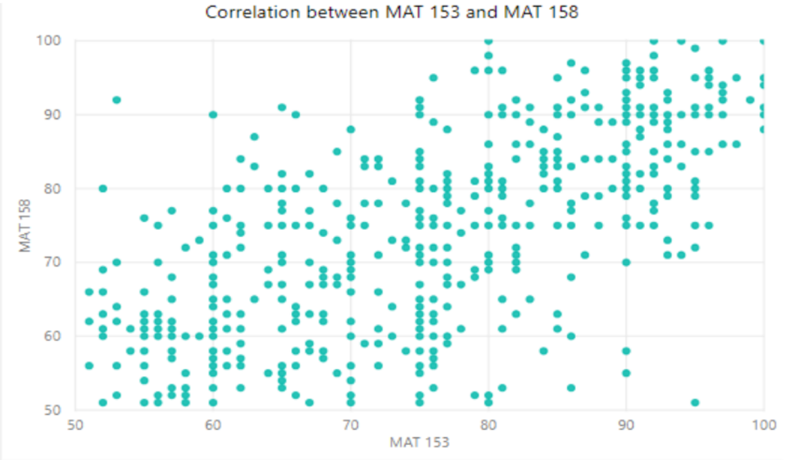


Figure 4.6: Correlation between MAT 153 and MAT 158

For prediction we use students' grades from MAT 158, which is considered as the second part of the course MAT 153. Students often pass this course in the second semester, so we collected previous semester grades from MAT 153. According to the results of correlation (r=0.65), we can say that these courses are highly dependent on each other, so it helps us to predict more accurately.

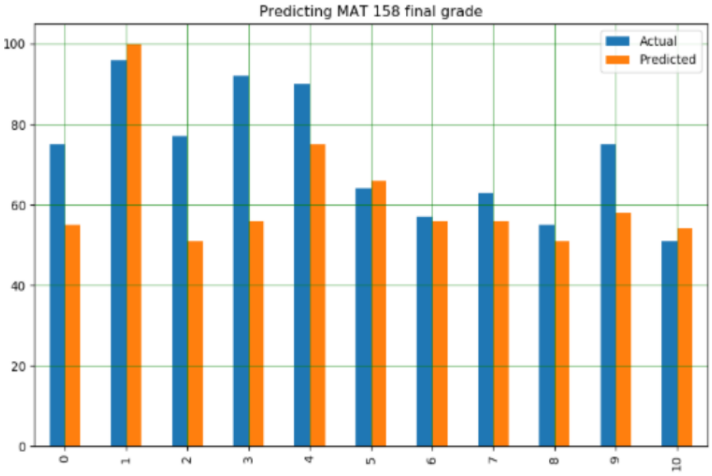
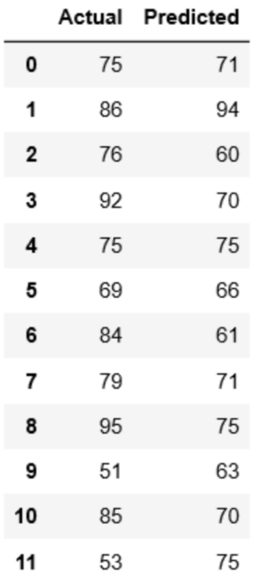


Figure 4.7: Predicting MAT 158 final grades

We predicted MAT 158 final grade, so took the MAT 158 as a target variable, and added features like course passed year, program code and MAT 153 final grades.

We built it using the DecisionTreeClassifier() class from Scikit-learn library. We predicted the final grade with the accuracy 40%. The results are shown in Figure 4.6,4.7. In Figure 4.7 shown the actual and predicted grade. Here we can see that we have a difference between them, but the main task is not to have perfect accuracy. Because the grade that we used to predict is a numeric value, so it is hard to calculate the exact result.



	Actual	Predicted
0	75	71
1	86	94
2	76	60
3	92	70
4	75	75
5	69	66
6	84	61
7	79	71
8	95	75
9	51	63
10	85	70
11	53	75

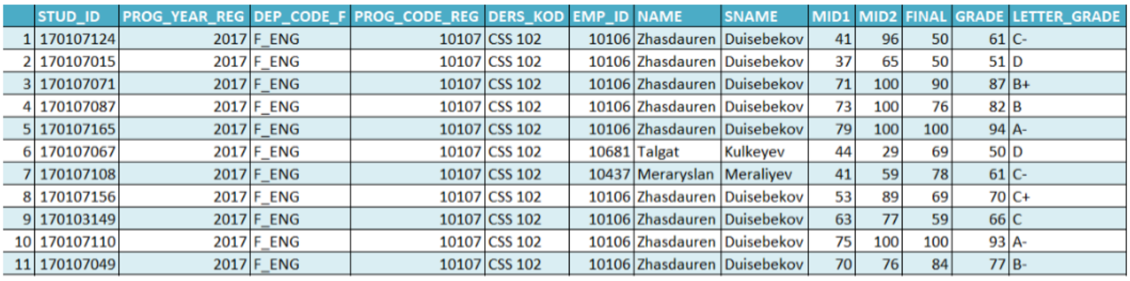
Figure 4.8: Sample of predicted and actual grade

4.3 Final grade prediction

In this section we will show the prediction of the final exam score using all last year students' grades of the same course. It helps students to know their predicted grade before the final exam to analyze their performances and prepare in time. We chose the course CSS102 (Programming Technologies) to predict the final exam score. We collected data from the university portal. Dataset is given in Figure 4.8, includes Student ID, year, Faculty code, Program code, course code, teacher ID, name, surname, Midterm1, Midterm 2, Final exam grades, Total grade and Total letter grade.

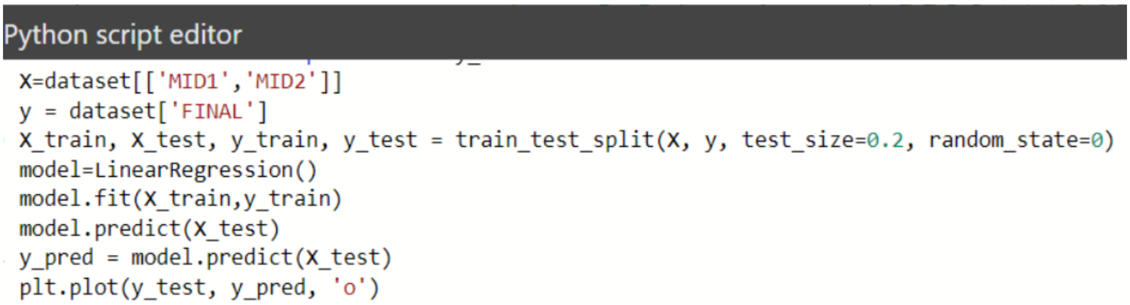
In general, all teachers' have their own grading policy. Some teachers' midterm exams are easy, but final exams can be very difficult. Some of them add midterm, final projects, which are evaluated differently. This kind of difference can make huge changes during the final grade prediction. So, we will predict the final exam score according to which teacher teaches the course.

We used Linear Regression class from Scikit-learn library, to predict final grade using midterm grades as a regression model. Made correlation between midterm grades and final grade. Midterm-1 grades' coefficient is 0.55, midterm-2 0.45. But this always changes depending on the chosen course.



	STUD_ID	PROG_YEAR_REG	DEP_CODE_F	PROG_CODE_REG	DERS_KOD	EMP_ID	NAME	SNAME	MID1	MID2	FINAL	GRADE	LETTER_GRADE
1	170107124	2017	F_ENG	10107	CSS 102	10106	Zhasdauren	Duisebekov	41	96	50	61	C-
2	170107015	2017	F_ENG	10107	CSS 102	10106	Zhasdauren	Duisebekov	37	65	50	51	D
3	170107071	2017	F_ENG	10107	CSS 102	10106	Zhasdauren	Duisebekov	71	100	90	87	B+
4	170107087	2017	F_ENG	10107	CSS 102	10106	Zhasdauren	Duisebekov	73	100	76	82	B
5	170107165	2017	F_ENG	10107	CSS 102	10106	Zhasdauren	Duisebekov	79	100	100	94	A-
6	170107067	2017	F_ENG	10107	CSS 102	10681	Talgat	Kulkeyev	44	29	69	50	D
7	170107108	2017	F_ENG	10107	CSS 102	10437	Meraryslan	Meraliyev	41	59	78	61	C-
8	170107156	2017	F_ENG	10107	CSS 102	10106	Zhasdauren	Duisebekov	53	89	69	70	C+
9	170103149	2017	F_ENG	10107	CSS 102	10106	Zhasdauren	Duisebekov	63	77	59	66	C
10	170107110	2017	F_ENG	10107	CSS 102	10106	Zhasdauren	Duisebekov	75	100	100	93	A-
11	170107049	2017	F_ENG	10107	CSS 102	10106	Zhasdauren	Duisebekov	70	76	84	77	B-

Figure 4.9: Dataset to predict final exam score



```
X=dataset[['MID1','MID2']]
y = dataset['FINAL']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
model=LinearRegression()
model.fit(X_train,y_train)
model.predict(X_test)
y_pred = model.predict(X_test)
plt.plot(y_test, y_pred, 'o')
```

Figure 4.10: Sample of the python code in PowerBI Python Visual

We must consider the fact that we will not have high accuracy during the final grade prediction, because the final exam weighs mostly 30-40% of final grade, so it's very high weight. The main goal is not to have an exact value, but that there should not be a big difference between actual and predicted grades.

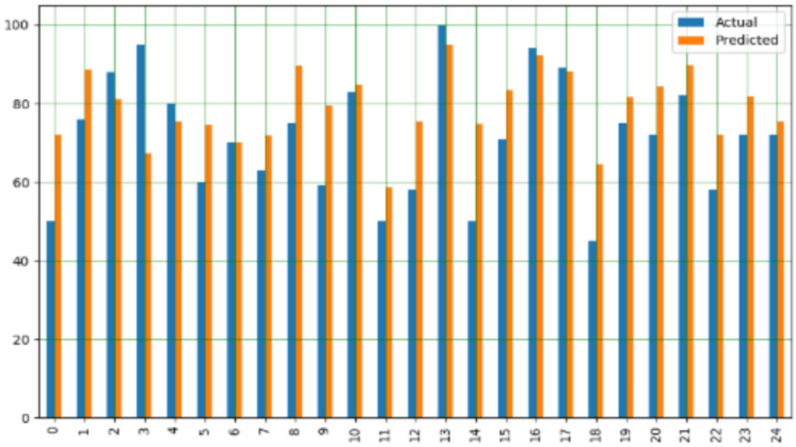
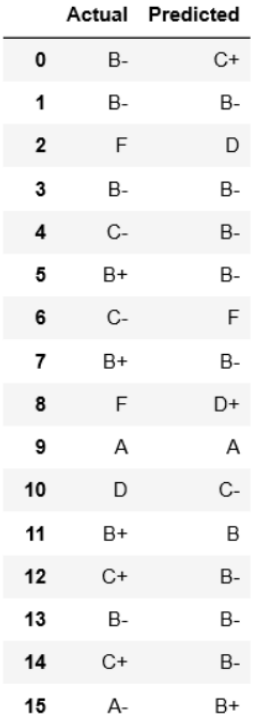


Figure 4.11: Predicted and actual final exam scores

We got a good result as shown in Figure 4.11. But here we can see that since grades are numeric values, it is hard to predict 100% accurately. So, to get better results we decided to predict grades using letter grades. We built it in the DecisionTreeClassifier() class in Scikit-learn library. Added features like year, term, teacher ID and used letter grade as a target variable. So, we predicted the letter grade using a classification method. The accuracy was 35%. These results are better than regression model results. Because predicting letter



	Actual	Predicted
0	B-	C+
1	B-	B-
2	F	D
3	B-	B-
4	C-	B-
5	B+	B-
6	C-	F
7	B+	B-
8	F	D+
9	A	A
10	D	C-
11	B+	B
12	C+	B-
13	B-	B-
14	C+	B-
15	A-	B+

Figure 4.12: Predicted and actual letter grades

grade shows approximately grades, when the regression method shows exact values (Figure 4.12). Also in classification we used many features that helped to get more accurate results. Since the results of each year are different from each other, some teachers are more stringent than others, and the term when a student takes the course has a big impact on the grade, the classification method is the best way to predict the final grade.

Chapter 5

Conclusion

In this paper we presented a case study in Data Analysis and its importance in education. We learned how Data Analysis and Machine learning algorithms help evaluate academic data and improve the educational system in university. We analyzed students' academic performances according to their course grades. We collected the dataset from the university portal and prepared the data to make correlations, removed outliers and null values. Using this dataset we showed a dashboard that helps students deeper understand their courses. It also helps students to choose the suitable elective courses according to their skills. We made a correlation between ENT and GPA, the result was weak and unimportant, so we decided that school performances are not similar to university performances. Also we made correlations between similar courses. With the help of this we showed whether and how strongly courses are related and how we can use them to predict grades. We used the Decision Tree Classifier() class in the Scikit-learn library to predict the final grade of students from each course. We discovered which features have a big impact on the students' final grade. These predictions help students know their predicted grades before the final exam and prepare in time. During this project, we encountered problems from the dataset. Since the university changed grading policy, we had to remove the data that don't match the old system, because they are considered as outliers and may indicate error. In the future, when the university portal will be systematized, we can collect large dataset. So we can predict more accurately and get good results.

[5] [4] [7] [2] [1] [3] [6]

References

[1] Mohammed M Abu Tair and Alaa M El-Halees. “Mining educational data to improve students’ performance: a case study”. In: *Mining educational data to improve students’ performance: a case study* 2.2 (2012).

[2] Mashael A Al-Barrak and Muna Al-Razgan. “Predicting students final GPA using decision trees: a case study”. In: *International Journal of Information and Education Technology* 6.7 (2016), p. 528.

[3] Prashant Gupta. “Decision trees in machine learning”. In: *Towards Data Science* (2017).

[4] Zafar Iqbal et al. “Machine learning based student grade prediction: A case study”. In: *arXiv preprint arXiv:1708.08744* (2017).

[5] Emaan Abdul Majeed and Khurun Nazir Junejo. “Grade prediction using supervised machine learning techniques”. In: *e-Proceedings of the 4th Global Summit on Education* (2016).

[6] Qasem A Al-Radaideh, Emad M Al-Shawakfa, and Mustafa I Al-Najjar. “Mining student data using decision trees”. In: *International Arab Conference on Information Technology (ACIT’2006), Yarmouk University, Jordan.* 2006.

[7] Linlin Zhang and Kin Fun Li. “Education analytics: Challenges and approaches”. In: *2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA).* IEEE. 2018, pp. 193–198.