

АҚПАРАТТЫҚ-КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР  
ИНФОРМАЦИОННО-КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ  
INFORMATION AND COMMUNICATION TECHNOLOGIES

DOI 10.51885/1561-4212\_2024\_2\_187  
MFTAA 28.23.13

**А.С. Тлебалдинова<sup>1</sup>, М.А. Карменова<sup>2</sup>**

<sup>1</sup>Дәулет Серікбаев атындағы Шығыс Қазақстан техникалық университеті,  
Өскемен қ., Қазақстан  
E-mail: a\_tlebaldinova@mail.ru\*

<sup>2</sup>Сәрсен Аманжолов атындағы Шығыс Қазақстан университеті, Өскемен қ., Қазақстан  
E-mail: mmm\_0582@mail.ru

**БІЛІМ АЛУШЫЛАРДЫҢ ОҚУ ҮЛГЕРІМДЕРІН  
МАШИНАЛЫҚ ОҚЫТУ ӘДІСІ НЕГІЗІНДЕ ТАЛДАУ ЖӘНЕ БОЛЖАУ**  
**АНАЛИЗ И ПРОГНОЗИРОВАНИЕ УСПЕВАЕМОСТИ ОБУЧАЮЩИХСЯ  
НА ОСНОВЕ МЕТОДА МАШИННОГО ОБУЧЕНИЯ**  
**ANALYSIS AND PREDICTION OF STUDENTS' PERFORMANCE BASED  
ON THE MACHINE LEARNING METHOD**

**Аңдатпа.** Қазіргі таңда әрбір салада деректер жиынын талдау мен машиналық оқыту әдістерін қолдану өзекті мәселелердің бірі болып отыр. Себебі, жыл сайын жинақталып отырған деректер жиынын талдау ғылыми зерттеу салаларында жаңа білімдерді, заңдылықтарды анықтауға көп көлемде өсерін тигізіп отыр. Сол сияқты, білім беру саласында да, деректер жиынын талдау мен машиналық оқыту әдістері кең мүмкіндіктерді және шешімдерді ұсынады, атап айтсақ, ол білім алушылардың оқу үлгерімдерін болжау. Білім алушылардың оқу үлгерімдерін болжауда деректер жиынындағы әртүрлі атрибуттар бұрыннан белгілі және белгісіз байланыстарды анықтап беруде. Ол өз кезегінде, білім беру үрдісін тиімді ұйымдастыру мен білім сапасын жақсарту мәселелерін де шешуге көп үлес тигізеді. Қарастырылып отырған мақалада, білім алушылардың оқу үлгерімін болжау үшін деректер жиыны талданып, машиналық оқыту алгоритмі, яғни сызықтық регрессия әдісі қолданылған. Деректерді жиынын талдау екі кезеңде жүзеге асырылған, бірінші кезеңде деректер жиынына статистикалық талдау жүргізіліп, екінші кезеңде сызықтық регрессия үлгісі құрылған. Білім алушылардың оқу үлгерімін болжауда сызықтық регрессия алгоритмі негізінде релевантты атрибуттар анықталды. Олар, оқушылардың тоқсан бойынша бағалары G1-G3, білім алушылардың ата-анасының білімі жайлы мәліметтер және білім алушының қаншалықты сабаққа қатысуы, яғни сабақты босату туралы деректері. Сызықтық регрессияны қолдану негізінде алынған үлгіні бағалау 90 % жоғары нәтижені көрсетті.

**Түйін сөздер.** деректерді талдау; білім алушы үлгерімі; болжау, статистикалық талдау; машиналық оқыту; сызықтық регрессия; визуализация; веб-қосымша; дашборд.

**Аннотация.** В настоящее время применение методов анализа данных и машинного обучения является актуальным вопросом для каждой отрасли. Это связано с тем, что анализ ежегодно собираемых наборов данных оказывает большое влияние на выявление новых знаний, закономерностей в областях научных исследований. Точно так же, как в сфере образования, методы анализа наборов данных и машинного обучения предоставляют широкие возможности и решения, а именно прогнозирование успеваемости обучающихся. При прогнозировании успеваемости обучающихся различные атрибуты в наборах данных выявляют давно известные и неизвестные связи. Это, в свою очередь, вносит большой вклад в решение вопросов эффективной организации образовательного процесса и улучшения качества образования. В рассматриваемой статье были проанализированы наборы данных для прогнозирования успеваемости обучающихся и применен алгоритм машинного обучения, а именно метод линейной регрессии. Анализ набора данных осуществлялся в два этапа, на первом этапе проводился статистический анализ набора данных, а на втором этапе строилась модель линейной регрессии. В прогнозировании успеваемости обучающихся определены релевантные атрибуты на основе алгоритма линейной регрессии. Такими атрибутами являются оценки G1-

G3 обучающихся за четверть, данные об образовании родителей обучающихся, а также пропуски от занятий обучающегося. Оценка выборки, полученная на основе использования линейной регрессии, показала результат выше 90 %.

**Ключевые слова.** анализ данных; успеваемость обучающихся; прогнозирование; статистический анализ; машинное обучение; линейная регрессия; визуализация; веб-приложение; дашборд.

**Abstract.** Currently, the application of data analysis and machine learning methods is a pressing issue for every industry. This is due to the fact that the analysis of annually collected data sets has a great impact on identifying new knowledge and patterns in areas of scientific research. Just like in the field of education, methods of analyzing data sets and machine learning provide powerful opportunities and solutions, namely, predicting student performance. When predicting student performance, various attributes in data sets reveal long-known and unknown relationships. This, in turn, makes a great contribution to resolving issues of effective organization of the educational process and improving the quality of education. In this article, data sets were analyzed to predict student performance and a machine learning algorithm was applied, namely the linear regression method. The analysis of the data set was carried out in two stages, in the first stage a statistical analysis of the data set was carried out, and in the second stage a linear regression model was built. In predicting student performance, relevant attributes are identified based on a linear regression algorithm. Such attributes are grades G1-G3 of students for the quarter, data on the education of the students' parents, as well as absences from classes of the student. The sample estimate obtained using linear regression showed a result above 90%.

**Keywords.** data analysis; student performance; forecasting; statistical analysis; machine learning; linear regression; visualization; web application; dashboard.

*Кіріспе.* Жалпы білім беру саласында білім алушылардың үлгерімін талдау мен болжау мәселелері немесе білім алушының үлгеріміне қатысты шешім қабылдау жүйелерін жобалау – ол көптеген мүдделі тұлғалар қатысатын күрделі, көп өлшемді үрдіс. Тиімді шешім қабылдау үшін оқу үрдісіне қатысушылардан оның әртүрлі кезеңдерінде алынған ақпаратты талдау маңызды болып табылады. Жылдар бойы оқу орындарының ақпараттық жүйелерінде білім беру үрдісінің әртүрлі аспектілері туралы ақпарат жинақталып отыр. Мысалы, оқушылар және олардың үлгерімі туралы, мұғалімдер және олардың ғылыми-білім беру жұмыстары туралы т.с.с.

Қазіргі кезеңде қашықтықтан оқыту курстары, білім беру форумдары, оқушыларды тестілеу және сауалнама жүргізу жүйелері және тағы басқа жүйелер құрылу мен қолдану үстінде. Осылайша, соңғы жылдары білім беру үрдісіне қатысты белгілі бір дәрежеде көптеген мәліметтер жиыны жинақталды және жинақталуда деуге болады. Білім алушылардың үлгерімін болжау мен олардың құзыреттіліктерінің қалыптасуын бағалау көп жағдайда білім алушының моделіне байланысты болып келеді. Мұндай білім алушы модельдері білім беру деректерін талдау негізі болып келеді. Білім беру деректерін талдауды қолдану негізінде айқындалған мақсат-міндеттер мен олар арқылы шешілетін мәселелер көп жағдайда техникалық әдістер арқылы шешім тауып отырады. Осындай техникалық әдістердің көбі машиналық оқыту саласынан алынған. Осы әдістермен [1, 2, 3, 4] жұмыстарынан танысуға болады. Білім беру деректерін талдау немесе білім беру деректерін интеллектуалды талдау - бұл қазіргі кезде дамып келе жатқан пәндердің бірі болып табылады, әрі оның шеңберінде білім алушылардың қалай оқитынын және оны қандай жағдайда жасайтынын түсіну үшін есептеу және педагогикалық әдістер мен тәсілдер қолданылады [5]. Осы тұста білім беру деректеріне білім алушылардың педагогикалық бағдарламалық жүйелермен өзара іс-қимылының әр түрлі түрлері ғана емес, сондай-ақ әкімшілік деректер, демографиялық деректер, жеке бас қасиеттері туралы деректер мен білім алушылардың өзара іс-қимылы туралы деректері де жатады [6].

Білім беру саласында деректерді талдау салыстырмалы түрде жаңа зерттеу бағыты болып табылатыны барлығымызға мәлім. EDM-ды қолдану негізінде оқу мекемелері білім беру үрдісін ұйымдастыруда жоғары жетістіктерге жету, тиімді шешімдерді қабылдау және білім алушылардың оқу үлгерімдерін жоғарылату мақсаттарында қазіргі

уақытта жиі қолданып келеді. EDM – білім беру және оқыту арқылы жиналған деректерді талдайтын, содан кейін бағалау, білім, жетістік, мотивация және оқушылардың көзқарасы сияқты егжей-тегжейлі ақпаратты зерттеу арқылы білім алушылардың болашақ мінез-құлқын болжау үшін машиналық оқыту және деректерді өндіру әдістерін қолданатын жаңа құралдардың бірі [7]. Нақты атап айтатын болсақ, EDM – бұл білім беру деректерінің үлкен жинақтарындағы заңдылықтарды анықтау үшін деректерді өндіруді, машиналық оқытуды және статистикалық әдістерді қолдануға бағытталған зерттеу саласы [8]. Мысалы, [9] жұмысында EDM-ды қолдану арқылы білім алушылардың оқу үлгерімдерін болжайды. Тәжірибе нәтижелерінде көпқабатты персептронның жоғары өнімділігі анықталды. EDM-ның әр түрлі әдістерін қолдану, соның ішінде, мысалы жіктеу сияқты әдістері негізіндегі зерттеулер әдістердің тиімділіктері мен дәлдіктерін дәлелдеп отыр [10]. Жалпы алғанда, data mining (деректерді іздеу) термині әртүрлі қызмет салаларында шешім қабылдау үшін қажет жаңа, тривиалды емес және іс жүзінде пайдалы білімді анықтауды білдіреді [11]. Сонымен, білім беру деректерін интеллектуалды талдауға қатысты анықтаманы түйіндесек, білім беру деректерін іздеу – ол шешім қабылдауды қолдау мақсатында білім беру үрдісі және оның қатысушылары туралы бұрын белгісіз, іс жүзінде пайдалы және түсіндірілген білімді анықтау әдістерінің жиынтығы.

*Әдеби көздерге шолу.* Білім беру саласында білім алушылардың үлгерімін талдау мен болжау, білім беру үрдістері негізінде жинақталған деректер жиынына байланысты зерттеулер әлемдік деңгейде қарқынды даму үстінде. Білім беру деректер жиынына талдау жасау, жоғарыда атап кеткен, деректерді интеллектуалды талдау технологиясы мен машиналық оқыту әдістері қолдану мысалдары өте көп. Мысалы, келесі [12] зерттеуде, авторлар білім алушылардың оқу үлгерімдерін болжау үшін және оларды жіктеуде екі деректер жиынын қарастырып, машиналық оқытудың 4 алгоритмдерін қолданған. Деректер жиынын талдау барысында он сегіз рет зерттеу тәжірибесін жасаған. Білім алушылардың оқу үлгерімін болжау үшін машиналық оқытудың кері таралу (BP, Backpropagation), тірек векторлары регрессиясы (SVR) және LSTM (Long-Short Term Memory) алгоритмдері қолданылған. Деректер жиынын зерттеу бойынша жіктеу есебін шешуде қосымша тағы BP and SVM алгоритмдері қолданылған да, жіктеу кезеңінде тағы да Gradient Boosting Classifier (GBC) әдісі қолданған. Зерттеу нәтижелерінде әрбір машиналық оқыту әдістерінің жұмыс өнімділігі салыстырылып, қолданылған алгоритмдер ішінде болжамда ең төменгі орташа квадрат мәні және ең жоғары R2 және EV Scores SVR арқылы алынғанын көрсетеді. BP алгоритмі ең төменгі болжам көрсеткіштерін бейнедегенмен, ол жіктеу есебін шешу тіжрибелерінде басқа жіктеу алгоритмдерінен 87,78% жоғары болғанын көрсеткені туралы қорытынды шығарады. Сол сияқты [13] зерттеуде де, авторлар гибридік тәсілдемеге біріктіру арқылы машиналық оқыту әдістері мен алгоритмдерін SVM тірек векторлар әдісі, Naive Bayes, Decision Tree ағаштар шешімі, Neural Network жасанды нейрондық желі қолдану негізінде білім алушылардың оқу үлгерімдерін болжайды. Болжау нәтижесінде кластеризация мен жіктеу алгоритмдерін біріктірген гибридік тәсілдеме білім алушының мінез-құлқы мен үлгерімі арасында тығыз байланыс бар екенін айқындалады. Қарастырып отырған [14] зерттеуде, мысалы кластеризация әдістері тиімді қолданылады. Мұнда рекурсивті кластеризация білім алушыларды олардың оқу үлгерімдеріне сәйкес білгілі бір курстар бойынша топтстырады. Анықталған әрбір топ сол топқа сай өздерінің оқу бағдарламаларымен мен нұсқаулықтарын автоматты түрде ала алады. Осындай тәсілдемені қолдану мақсаты төменгі топ білім алушыларын жоғары топқа ауыстыруында болып табылады. Осыған ұқсас, келесі [15] зерттеуде де, әр білім алушының оқу үлгерімі ескерілетіндей, өзіндік ерекшеліктеріне байланысты жеке болжам әдісі жүзеге асырылады. Мұндай зерттеу тәсілдемесін авторлар жекелендірілген болжау әдісі деп атап ұсынған. Келесі [16]

зерттеуде, білім алушылардың оқу нәтижелерін болжау мен үлгілеуде нейрондық желі машиналық оқыту әдісі жүзеге асырылған.

*Деректер жиынына сипаттама.* Деректер жиыны Шығыс Қазақстан облысы білім басқармасы Өскемен қаласы бойынша Білім бөлімінің «Шоқан Уәлиханов атындағы №3 мектеп-лицейі» коммуналдық мемлекеттік мекемесінде оқитын білім алушылардың оқу үлгеріміне қатысты жинақталған. Деректер жиыны білім алушылардың үлгерімі, демографиялық сипаттамалары, әлеуметтік және мектеп сипаттамалары атрибуттарынан тұрады, сонымен қатар, мектеп есептері мен сауалнамаларын қолдана отырып жинақталған. Деректер жиыны екі түрлі пән бойынша жинақталған деректер жиынтығын ұсынады, олар Информатика (инф) және Математика (мат) пәндері. Білім алушылардың оқу үлгерімі мәліметтерін талдау мен болжау бойынша деректер жиынтығын алдын ала өңдеуге байланысты жұмыстар жүргізілді. Деректер жиынында сандық емес мәліметтер болғандықтан білім алушылардың оқу үлгерімін талдау мен болжауды, модельдеуді жасамас бұрын деректерді тазарту жүргізілді. Деректер жиынын алдын ала өңдеуге қатысты негізінен өңделмеген деректерді алып тастау немесе жетіспейтін мәндерді өңдеу, сонымен қатар деректерді талдауда жетіспейтін мәндердің бар болуына немесе жоқ болуына қарай зерттеуші ретінде назар аудару қажет деп айтып кеткен жөн. Деректер жиыны сипаттамасына келер болсақ, деректер жиынтығында әрбір кесте бағанасына сәйкес 35 атрибут және де кесте жолдарына байланысты 426 жазба бар. Деректер жиыны атрибуттарын шартты түрде білім алушылардың жеке бас мәліметтерінен бастайық. Атрибуттар атаулары: School (орта мектеп)- орта мектеп оқушысы, Sex («F» - қыз; - «M» - ұл) – оқушы жынысы, Age (10-нан 18-ге дейінгі сандар) – оқушы жасы, Address («U» - қала; - «R» - ауыл) – мекен-жайы. Білім алушылардың отбасы жағдайына байланысты мәліметтер 1-кестеде келтірілген.

**1-кесте.** Білім алушылардың отбасылары бойынша мәліметтері

№	Атрибут атауы	Деректер жиынындағы шартты белгіленуі	Түсініктемесі
1	Famsize	- «LE3» - 3 тең немесе одан артық емес; - «GT3» - 3-тен артық	Отбасы мүшелерінің саны
2	Pstatus	- «T» - бірге тұру; - «A» - бөлек тұру	Отбасы мүшелерімен бірге тұру мәртебесі
3	Medu	- 0 - жоқ; - 1 - бастауыш білім (төрт сынып); - 2- 5-9 сынып; - 3 - орта білім; - 4 - жоғары білім	Анасының білімі жайлы мәлімет
4	Fedu	- 0 - жоқ; - 1 - бастауыш білім (төрт сынып); - 2- 5-9 сынып; - 3 - орта білім; - 4 - жоғары білім	Әкесінің білімі жайлы мәлімет
5	Mjob	«мұғалім», «денсаулық», азаматтық «қызмет» (мысалы, әкімшілік немесе полиция қызметкері), «үй шаруасында» немесе «басқалар»)	Анасының жұмысы
6	Fjob	«мұғалім», «денсаулық», азаматтық «қызмет» (мысалы, әкімшілік немесе полиция қызметкері), «үй шаруасында» немесе «басқалар»)	Әкесінің жұмысы

Деректер жиыны атрибуттары сипаттамалары бойынша төмендегі 3-кестеге сәйкес білім алушылардың мектепке байланысты оқу үрдісіне қатысты тағы да қосымша мәліметтері қолданылады.

**2-кесте.** Білім алушылардың оқу үрдісіне қатысты қосымша деректері

№	Атрибут атауы	Деректер жиынындағы шартты белгіленуі	Түсініктемесі
1	Reason	үйге, мектепке жақын, беделі, бағыты, өз қалауы және т.б.	Осы мектепті таңдау себебі
2	Guardian	Анасы, әкесі немесе т.б.	Оқушының қамқоршысы
3	Treaveltime	- 1 - <15 минут, - 2 - 15-30 минут - 3 - 30 минут - 1 сағат - 4 - > 1 сағат	Мектепке дейінгі уақыт
4	Studytime	- 1 - 2 сағат - 2 - 2 сағат - 3 - 5-10 сағат - 4 - 10 сағат	Бір аптада оқу уақыты
5	Failures	егер $1 \leq n$ болса, онда n, әйтпесе 4	Босатылған сабақтар саны
6	Schoolsup	иә немесе жоқ	Қосымша білім
7	Famsup	иә немесе жоқ	Білім алуын отбасылық қолдау
8	Paid	иә немесе жоқ	Информатика және математика бойынша қосымша ақылы курс
10	Nursery	иә немесе жоқ	Жеке бөлмесінің болуы
11	Higher	иә немесе жоқ	Жоғары білім алғыңыз келеді ме
12	Internet	иә немесе жоқ	Үйде Интернет бар ма?
13	Romantic	иә немесе жоқ	Романтикалық қарым-қатынас бар
14	Famrel	1 саны - өте төмен - 5-ке дейін - өте жоғары	Отбасылық қатынастардың сапасы
15	Freetime	1 саны - өте төмен - 5-ке дейін - өте жоғары	Мектептен кейінгі бос уақыт
16	Goout	1 саны - өте төмен - 5-ке дейін - өте жоғары	Достармен демалыс
17	Dalc	1 саны - өте төмен - 5-ке дейін - өте жоғары	Жұмыс күндері алкоғольді тұтыну
18	Walc	1 саны - өте төмен - 5-ке дейін - өте жоғары	Демалыс күндері алкоғольді тұтыну
19	Health	1-ден өте нашар - 5-ке дейін - өте жақсы	Қазіргі уақытта денсаулық жағдайы
20	Absences	0-ден 93-ке дейінгі сан	Босатқан күндері бойынша саны

Деректер жиыны атрибуттары бойынша кесте 4-ке сәйкес білім алушылардың информатика және математика пәндері бойынша үлгерім бағалары туралы мәліметтері ұсынылады. G1 (бірінші тоқсан бағалары), G2 (екінші тоқсан бағалары), G3 (қорытынды бағалар) атрибуттарының мәндері дейінгі сандар аралығында шартты түрде белгіленді. Нақты 0-ден 20 аралығындағы сандарға қатысты бағалаудың мәндеріне келетін болсақ, олар келесідей шарттар негізінде алынып отыр: 16-20 А – өте жақсы деңгей; 14-15 В – жақсы деңгей; 12-13 С – қанағаттанарлық деңгей; 10-11 D – жеткілікті деңгей; 0-9 F – нашар деңгей. Жалпы бұл шартты деңгейлер 5 деңгейлі жіктеу – қайта есептеу жүйесі Erasmus бағалау жүйесіне негізделеді.

*Зерттеу әдісі.* Кез келген математикалық модельдерді шешудің объективтілігі негізінен бастапқы ақпараттың сенімділігіне байланысты екенін білеміз. Мұнда осы ақпарат ретінде білім алушының бағалау қолданылады. Бұл мұғалімнің өзі анықтайтын ең объективті бағалау. Барлық белгісіздік үшін ол әлі де оқу сапасының көрсеткіші және өлшемі болып табылады. Оқу пәндері бойынша оқыту сапасын талдау және басқару мақсатында регрессиялық теңдеулер мен болжамды бағаларды алу тәсілі қарастырылған. Көрсетілген теңдеулер білім алушылардың үлгерімі туралы нақты ақпаратты өңдеу және регрессиялық талдауды қолдану негізінде алынуы мүмкін. Сызықтық регрессия (LR - Linear regression) оқытушымен оқыту есептерін шешеді, яғни белгілі бір алдын ала белгілеулері бар деректер жиынтығымен жұмыс жасап, содан кейін біз әлі қарастырылмаған жаңа деректерге туралы болжам жасалынады. Сызықтық регрессия жағдайында белгілі бір нақты айнымалы болжанады және сызықтық тәуелділікті атрибуттарға байланысты модельденіледі, яғни қандай да бір атрибуттар бар делік және олар басқа атрибуттарға байланысты болады. Мақсатымыз сол тәуелділікті табу болып табылады. Сызықтық регрессия дегеніміз не деп қысқаша талдау жасап өтейік. Айталық, мақсатты айнымалы бар. Атрибуттар салмақтарын негізінде деректерді анағұрлым жақсы сипаттайтын атрибуттар салмақтарын табу қажет.

$$y = a_0 + a_1x_1 + a_2x_2 + \dots + a_kx_k, \quad (1)$$

мұндағы  $y$  – мақсатты айнымалы,  $x$  – белгілері,  $a$  – үлгі салмақтары.

Мұны әртүрлі белгілердің өлшенген қосындысы ретінде қарастыруға болады және бұл салмақтарды тек модельде таңдалынады. Мысалы, кейін осы салмақтарды атрибуттардың қайсысы маңызды екенін түсіну үшін қолдануға болады.

Сызықтық регрессия – ол оқытушымен оқыту әдістер тобына жатады. Зерттеу барысына байланысты таңдап алынған деректер жиыны бар, оларды талдау арқылы сипаттама жасаймыз да, содан кейін жаңа деректер туралы болжам жасауға болатынына көз жеткізуімізге болады. Деректермен жұмыс істегенде «деректерді жақсы сапада сипаттау» туралы алдымызға сұрақ қойылады. Бұл жағдайда үйретуші таңдалымды қалай жақсы сипаттай аламыз деген де сұрақ туындайды. Сол себептен, үйретуші таңдалымды жақсы оқыту үшін қате функциясы қажет, яғни сол арқылы оны оңтайландырамыз деген сөз. Қате функциясы ретінде, мысалы, Mean Absolute Error немесе орташа абсолютті қатеі қолдана аламыз, бұл біздің орташа болжамдарымыздың дұрыс жауаптардан қаншалықты алыс екенін көрсетеді.

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|, \quad (2)$$

мұндағы  $N$  – таңдалымның мөлшері,  $y_i$  – дұрыс жауап,  $\hat{y}_i$  – біздің болжамымыз. Бұл өте қисынды және жақсы түсіндірілген жоғалту функциясы, бірақ ол сараланбайды, сондықтан оны, мысалы, градиентті түсіру немесе градиентті оңтайландыру әдістері жағдайында қолдануға болмайды. Сондықтан әдетте күрделі және сәл өзгеше функциялар қолданылады. Мысалы, Mean Squared Error, яғни дәл осылай жасайтын орташа квадраттық қате біздің болжамдарымыздың нақты жауаптардан қашықтықтарының квадраттарын ғана анықтайды.

$$MAE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2, \quad (3)$$

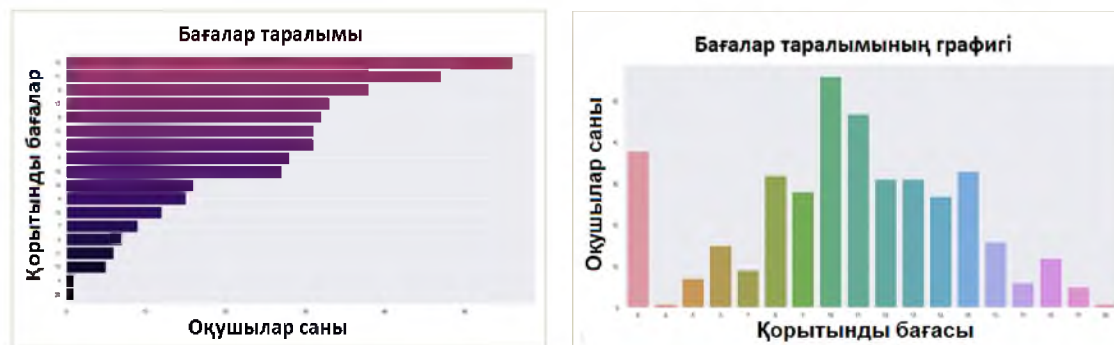
мұндағы  $N$  – таңдалымның мөлшері,  $y_i$  – дұрыс жауап,  $\hat{y}_i$  – біздің болжамымыз. Осыған байланысты оны қазірдің өзінде оңтайландыруға болады, сонымен қатар, оны саралауға

да болады және де ол градиентті түсіру әдістерінде өте жиі кездеседі. Бірақ та, мұндай жоғалту функциясы бар сызықтық регрессия жағдайында мәселені аналитикалық түрде шеше аламыз, градиент түсіруді қолданудың қажеті бұл жағдайда болмайды. Осыған қарамастан, ол үшін үлкен өлшемді кеңістіктерде күрделі матрицаларды айналдыру керек, және бұл әрдайым жұмыс істемейді, кейде әрдайым дерлік қолданылатын градиент әдістерін қолдану оңайырақ болады. Сонымен, жоғалту функциясын біз төмен түскіміз келетін бет сияқты елестетуге болады, яғни өз функциямызды азайтып, оның минимумын тапқымыз келеді. Дәл сол жерде біздің модель жақсы жұмыс істейтін болады. Ол үшін өздеріңіз білетіндей, градиент функцияның ең жылдам өсу бағытын көрсетеді, ал логикалық тұрғыдан антиградиент функцияның ең аз төмендеу бағытын көрсетеді де, оны осы тұста қолдануға болады.

*Тәжірибе жүзінде жүзеге асырылуы.* Достигнутые в исследовании авторов результаты могут быть использованы и в других подобных информационно-образовательных средах, а также полезны для продолжения проводимой работы, обеспечив комплексный подход к политике ИБ в масштабах всей организации. Можно рекомендовать результаты и в качестве учебных ресурсов при преподавании в вузе курсов по защите информации. Білім алушылардың оқу үлігерімі деректер жиынына талдау мен болжау жасау мәселесін шешудегі алғашқы қадам арнайы пәндер таңдалған. Ол белгілі бір мамандықтың оқу жоспарының құрылымдық-логикалық схемаларының логикалық ойларына және талаптарына бағынуы керек. Мұнда мысал ретінде екі арнайы пән алынған. Енді солардың атаулары мен шартты белгілерін берелік:

- арнайы пәндер бойынша болжам багалары –  $y$ ;
- пән атауы – «Математика» –  $x_1$ ;
- пән атауы – «Информатика» –  $x_2$ ;

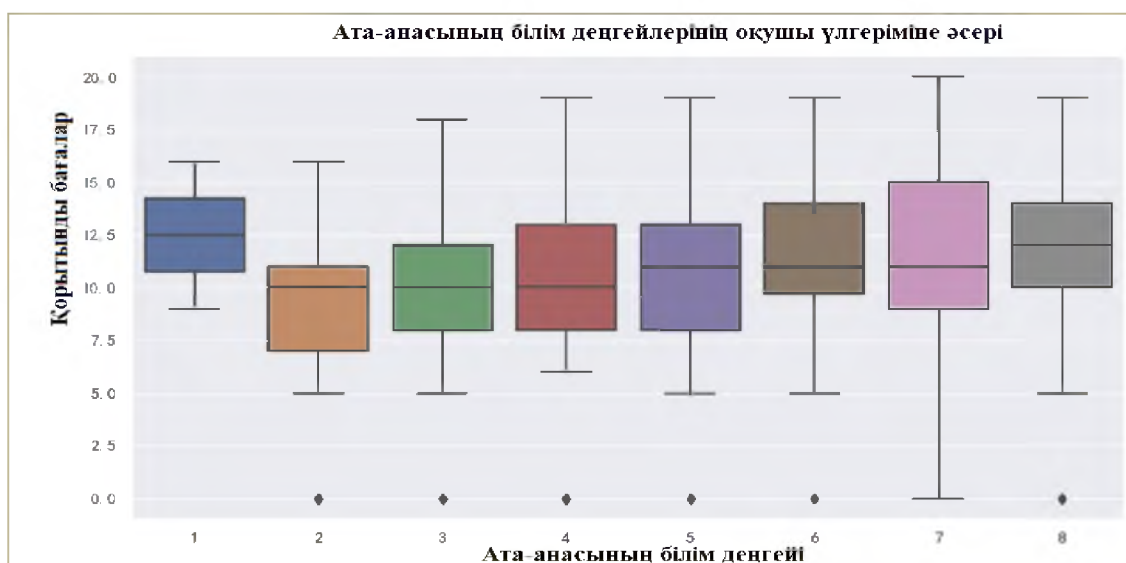
– оқушының жеке бас мәліметтері (School, Sex, Age, Address), шыққан отбасына қатысты мәліметтер (Famsize, Pstatus, Medu, Fedu, Mjob, Fjob), оқушылардың оқу үрдісіне қатысты қосымша деректері (Reason, Guardian, Treaveltime, Studytime, Failures, Schoolsup, Famsup, Paid, Nursery, Higher, Internet, Romantic, Famrel, Freetime, Goout, Dalc, Walc, Health, Absences), білім алушылардың оқу пәндері бойынша мәліметтеріне кіретін атрибуттар –  $x_k, \dots, x_{k+n}$ . Бастапқы ақпарат ретінде оқушылардың 1-ші және 2-ші тоқсан нәтижелерінің багалары және осы бағаларға тікелей әсері бар атрибуттар алынды. Білім алушылардың оқу үлігерімі деректер жиынына қатысты зерттеу жүргізудің бастапқы қадамы, деректерді визуализациялау әдістері арқылы жүргізілді де, екінші кезекте сызықтық регрессия моделі құрылды. Білім алушылардың оқу үлігерімі деректер жиынына визуализациялау әдістерін қолдану кеосі түрде жүргізілді. Білім алушылардың қорытынды багаларына талдау жасау негізінде жалпы бағалардың таралымын байқадық (1-сурет).



1-сурет. Білім алушылардың бағалары бойынша таралымының графигі

Білім алушылардың оқу үлгерімі деректер жиынына қатысты сабақ үлгеріміне әсер ететін факторларды анықтау және оларға жан-жақты талдауды жалғастыра отырып, атрибуттар арасындағы корреляциялық талдау жүзеге асырылды. Алдымен әр атрибут пен G3 қорытынды баға атрибутының арасындағы қатынас есептелді. Нәтижесінде анағұрлым корреляцияланушы атрибуттар қатарына G1 мен G2-ден басқа ешбір атрибут кірген жоқ. Білім алушылардың оқу үлгерімі деректер жиынын терең барлау үшін релеванттылығы жоғары атрибуттардың оқу үлгеріміне әсерін зерттедік. Сондай атрибуттардың бірі Failures (Өткен сыныптағы босатылған сабақтар саны) атрибутын қарастырдық.

Көріп тұрғанымыздай, нәтиже айқын. Сабақты көп босатқан білім алушылардың оқу үлгерімі де жоғары емес. Олай болса, Failures атрибутының оқу үлгеріміне әсері маңызды екендігіне көз жеткіздік. Осы тұста, сонымен қатар, Medu және Fedu (Ата-анасының білім деңгейлері жайлы мәлімет) атрибуттарының оқу үлгеріміне деген әсері де қарастырылып, келесі 2-суретте келтірілген.



**2-сурет.** Medu және Fedu атрибуттарының оқу үлгеріміне әсері

Отбасылық білім деңгейінің жоғарылауымен белгілі бір диапазонда білім алушылардың орташа бағалары да жақсарғанын көруге болады, бірақ графиктің қарама-қарсы ұштарында жүйелілік байқалмайды. Ең төменгі білім деңгейі бар білім алушылар басқа деңгейдегі көптеген отбасыларға қарағанда жоғары балл алған, ал ең жоғары білім деңгейі бар отбасылардың балаларында бағалары төмен. Мұның себебі әзірше белгісіз және деректерде кателіктер кету мүмкіндігін де жоққа шығаруға болмайды. Осы зерттеуде, деректер жиынын талдауға байланысты сызықтық регрессия әдісінен басқа да, машиналық оқыту әдістері қолданылып, білім алушылардың оқу үлгерімдерін болжауға байланысты әдістердің ішінен тиімдісі анықталды. Жалпы зерттеуде жіктеу әдістерінің ішінен ағаштар шешімдері (Decision Tree), кездейсоқ ағаштар шешімдері (Random Forest) және Support Vector Classification (SVC) әдістері қолданылды. Аталған әдістерді қолдану негізінде олардың тиімділігі Cross Validation Score әдісі арқылы бағаланып, дәлдік көрсеткіші ең жоғарғысы ретінде сызықтық регрессия әдісі анықталды, кесте 3-ке сәйкес. Сызықтық регрессия әдісі үлгісінің тиімділігін бағалау нәтижесінде дәлдік көрсеткіші – 90,6 %-ға тең болды.



### 3-кесте. Салыстырмалы түрде қолданылған әдістердің тиімділігін бағалау нәтижелері

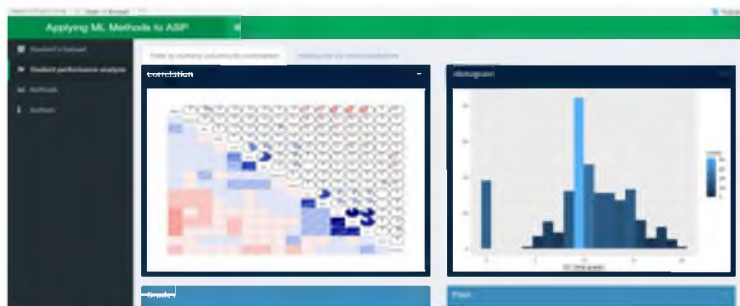
№	Бағалау үлгісі	Decision Tree	Random Forest	SVC	LR (сызықтық регрессия)
1	Model Score	0.887	0.963	0.929	0.898
2	Cross Validation Score	0.893	0.875	0.867	0.906

Білім алушылардың оқу үлгерімі деректер жиынына талдау мен болжау жасау мәселесін қарастырудың екінші кезеңінде тікелей сызықтық регрессия әдісі қолданылды. Білім алушылардың үлгерімін машиналық оқыту әдістерімен болжау, соның ішінде сызықтық регрессиялық модельмен бағалауды жүзеге асыру үшін R бағдарламалау тілінің Shiny пакеті арқылы веб-қосымша да әзірленді. Веб-қосымша зерттеудің бірінші кезеңі мен екінші кезеңі нәтижелері мен визуализацияларын өзіне кіріктіріп отыр. Веб-қосымша R бағдарламалау тілінің Shiny пакетін қолдану негізінде орындалды.

Shiny пакетімен жұмыс жасағанда бастапқыда жеңіл болғанымен, бірақ оның қалай жұмыс істейтініне қаншалықты терең үңілсеңіз, соғұрлым ол бағдарламалық жасақтаманы әзірлеудің күшті принциптері бар жалпы құрылыс блоктарынан жасалғанын түсінесіз. Shiny пакеті көмегімен қолданушы мен аналитиктер үшін тартымды және заманауи тұрғыда интерфейс дайындалды. Веб-қосымша құруда Shiny пакеті мүмкіндіктері өте тиімді болып келеді, сонымен қатар серверлік мүмкіндіктері де тиімді түрде жұмысты ұйымдастыруға көмектеседі.

Веб-қосымшаның навигациясы ретінде веб-қосымшаның сол жақ бөлігіндегі вертикаль мәзірге назар аудару қажет. Вертикаль мәзірде келесі атаулармен қажетті бөлімдер бар, олар: Student's Dataset; Student performance analysis; Methods; Authors. Веб-қосымшаның құрылымдық сызбасына сәйкес келесі бөлімдердің визуализацияларын келтіре кетейік. Веб-қосымшада білім алушылардың үлгерімін болжау үшін зерттеуде сипатталып кеткен деректер жиыны қолданылды. Мұнда «Student's Dataset» мәзірінде деректер жиыны жазбаларын қолданушы қажетінше парақ нөміріне сәйкес толығымен көре алады немесе іздеу функциясы арқылы да қажетті жазбаны тауып ала алады. Сонымен қоса, қажетінше 10 немесе 15, 20 т.с.с. жазбалар көлемін өз қалауыңша шығарып көру мүмкіндігіне ие бола алады.

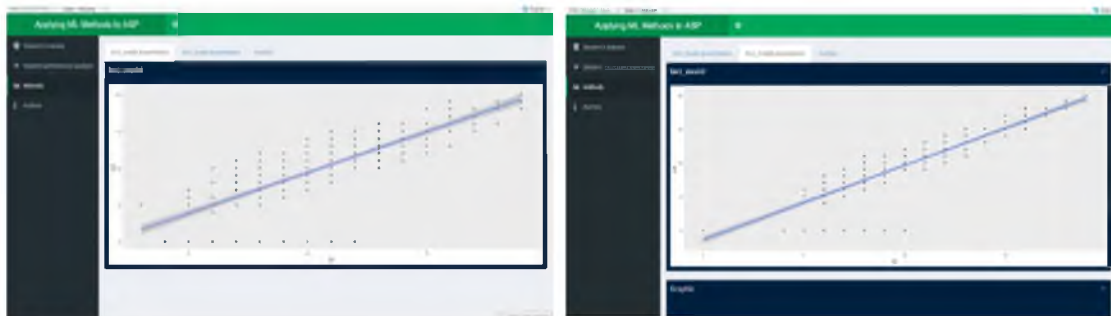
Веб-қосымшаның келесі «Student performance analysis» мәзірі бойынша екі қосымша ұйымдастырылған, олар: 1) «Filter to numeric columns for correlation», 2) «Analysis by G3 versus studytime». Осы құрылымға байланысты «Filter to numeric columns for correlation» қосымшасына шерту арқылы сурет 3-ке қатысты деректер жиыны негізінде атрибуттар арасындағы корреляциялық талдамалар графикалық түрде визуализацияланып көрсетіледі.



3-сурет. Деректер жиыны атрибуттары арасындағы корреляциялық талдама

Білім алушылардың үлгерімін болжауда тек қана бағалардың маңыздылығы бар екеніне ғана көзімізді жеткізіп қана қоймай, сонымен бірге, білім алушылардың мінез-құлқы, бос уақытын өткізуі мен отбасы және тұрмыстық жағдайларының да маңыздылығы бар екенін анықтап алдық. Сол себептен, білім алушылардың мінез-құлқына байланысты да талдамалар жасалып, олардың қорытындысы графикалық түрде алынған болатын. Бұл графикалық түрдегі талдамалар веб-қосымшаның арнайы бөлімдерінде визуализацияланып шығарылған болатын.

Келесі кезекте, веб-қосымшада зерттеудің мақсат-міндеттеріне байланысты таңдалып алынған зерттеу әдісі, яғни сызықтық регрессия моделіне қатысты визуализациялау келтірілген. Мұнда деректер жиынына байланысты білім алушылардың үлгерімін болжауға байланысты сызықтық регрессия модельдері графиктерін 4-ші суреттерден көруге болады.



**4-сурет.** Білім алушылар үлгерімін болжаудағы сызықтық регрессияның 1-ші және 2-ші моделдері

Білім алушылардың оқу үлгерімі деректер жиынына талдау мен болжау үшін машиналық оқыту әдісі ретінде сызықтық регрессияны қолдану нәтижесінде оқушылардың үлгерімін болжауда келесідей релевантты атрибуттар анықталды, яғни ол деген сөз: оқушылардың білімін болжау үшін келесі деректерге мұқият назар аудару қажет. Олар біріншіден, оқушылардың тоқсан бойынша бағалары кестеде олар G1/G2/G3 белгісімен сипатталып тұр, сонымен қатар, оқушылардың ата-анасының білімі жайлы мәліметтер маңызды болып табылады және соңғы ең маңызды дерек – ол білім алушының қаншалықты сабаққа қатысуы, яғни сабақты босату туралы деректер. Осылардың негізінде оқушылардың оқу үлгеріміне болашақта болжау жасауға болады. Сонымен қатар, құрылған модельді бағалау үшін арнайы метрика Ассигасу метрикасы қолданылып, бағалау нәтижесі жоғары пайызды көрсетті (90 %).

*Қорытынды.* Қарастырылып отырған зерттеу мақаласында білім алушылардың деректер жиынына талдау жасауда және оқу үлгерімін болжауда статистикалық талдау мен сызықтық регрессия моделі құрылып қолданылды. Сонымен қатар, корреляциялық әдісті қолдану нәтижесінде релевантты атрибуттар тізімі де анықталған. Деректер жиынтығы оқушылардың үлгерімі, демографиялық сипаттамалары, әлеуметтік және мектеп сипаттамалары атрибуттарын қамтыды. Деректер жиыны екі түрлі Информатика (инф) және Математика (мат) пәндері бойынша жинақталған. Машиналық оқыту әдісі сызықтық регрессияны қолдану нәтижесінде білім алушылардың үлгерімін болжауда релевантты атрибуттар анықталды. Олар, оқушылардың тоқсан бойынша бағалары G1-G3, білім алушылардың ата-анасының білімі жайлы мәліметтер, білім алушының қаншалықты сабаққа қатысуы, яғни сабақты босату туралы деректері маңызды болып табылатыны анықталды. Сызықтық регрессияны қолдану негізінде алынған модельді

бағалау нәтижесі жоғары пайызды көрсетті (90,6 %). Зерттеу жұмыстарының қорытынды нәтижелері мен визуализациясы Shiny Dashboard құралы негізінде білім алушылардың оқу үлгерімдерін талдау мен болжау интерактивті қосымшасы әзірленді.

#### References

1. Yağcı, M. Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learn. Environ.* 9, 11 (2022). <https://doi.org/10.1186/s40561-022-00192-z>.
  2. Alam, A., Mohanty, A. (2023). Predicting Students' Performance Employing Educational Data Mining Techniques, Machine Learning, and Learning Analytics. In: Tomar, R.S., et al. *Communication, Networks and Computing. CNC 2022. Communications in Computer and Information Science*, vol 1893. Springer, Cham. [https://doi.org/10.1007/978-3-031-43140-1\\_15](https://doi.org/10.1007/978-3-031-43140-1_15).
  3. H. Pallathadka, A. Wenda, E. Ramirez-Asís, M. Asís-López, J. Flores-Albornoz, K. Phasinam. Classification and prediction of student performance data using various machine learning algorithms. *Volume 80, Part 3, 2023, pp. 3782-3785, ISSN 2214-7853*, <https://doi.org/10.1016/j.matpr.2021.07.382>.
  4. Fahd, K., Venkatraman, S., Miah, S.J. et al. Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. *Educ Inf Technol* 27, 3743–3775 (2022). <https://doi.org/10.1007/s10639-021-10741-7>.
  5. Munir, H.; Vogel, B.; Jacobsson, A. Artificial Intelligence and Machine Learning Approaches in Digital Education: A Systematic Revision. *Information* 2022, 13, 203. <https://doi.org/10.3390/info13040203>.
  6. G. Feng, M. Fan and Y. Chen, "Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining," in *IEEE Access*, vol. 10, pp. 19558-19571, 2022, doi: 10.1109/ACCESS.2022.3151652.
  7. Ahuja, R., Jha, A., Maurya, R., Srivastava, R. (2019). Analysis of Educational Data Mining. In: Yadav, N., Yadav, A., Bansal, J., Deep, K., Kim, J. (eds) *Harmony Search and Nature Inspired Optimization Algorithms. Advances in Intelligent Systems and Computing*, vol 741. Springer, Singapore. [https://doi.org/10.1007/978-981-13-0761-4\\_85](https://doi.org/10.1007/978-981-13-0761-4_85).
  8. A. Hernández-Blanco, B. Herrera-Flores, D. Tomás, and B. Navarro-Colorado, "A Systematic Review of Deep Learning Approaches to Educational Data Mining," *Complexity*, vol. 2019, pp. 1–22, May 2019, doi: 10.1155/2019/1306039.
  9. C. Jalota and R. Agrawal, "Analysis of Educational Data Mining using Classification," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, pp. 243-247, doi: 10.1109/COMITCon.2019.8862214.
  10. Alshareef, F., Alhakami, H., Alsubait, T., & Baz, A. (2020). Educational Data Mining Applications and Techniques. *International Journal of Advanced Computer Science and Applications*, 11.
  11. Batool, S., Rashid, J., Nisar, M.W. et al. Educational data mining to predict students' academic performance: A survey study. *Educ Inf Technol* 28, 905–971 (2023). <https://doi.org/10.1007/s10639-022-11152-y>.
  12. Sekeroglu B., Dimililer K., Tuncal K. Student Performance Prediction and Classification Using Machine Learning Algorithms. *ICEIT 2019: Proceedings of the 2019 8th International Conference on Educational and Information Technology*. March 2019, pp.7–11. <https://doi.org/10.1145/3318396.3318419>.
  13. Francis, B.K., Babu, S.S. Predicting Academic Performance of Students Using a Hybrid Data Mining Approach. *J Med Syst* 43, 162 (2019). <https://doi.org/10.1007/s10916-019-1295-4>.
  14. Trakunphutthirak, R., & Lee, V. C. S. (2022). Application of Educational Data Mining Approach for Student Academic Performance Prediction Using Progressive Temporal Data. *Journal of Educational Computing Research*, 60(3), 742-776. <https://doi.org/10.1177/07356331211048777>.
  15. D. Aggarwal, S. Mittal, and V. Bali, "Prediction model for classifying students based on performance using machine learning techniques," *International Journal of Recent Technology and Engineering*, vol. 8, no. 2 Special Issue 7, pp. 496–503, Jul. 2019, doi: <https://doi.org/10.35940/ijrte.B1093.0782S719>.
  16. Lau, E.T., Sun, L. & Yang, Q. Modelling, prediction and classification of student academic performance using artificial neural networks. *SN Appl. Sci.* 1, 982 (2019). <https://doi.org/10.1007/s42452-019-0884-7>.
- 
-