МИНОБРНАУКИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«**САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО**»

Кафедра теории функций и стохастического анализа
наименование кафедры

Прогнозирование успешности обучения студентов методами машинного обучения на примере механико-математического факультета СГУ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ (ДИПЛОМНОЙ) РАБОТЫ

студентки 4 курса 451 группы

направления 38.03.05 - Бизнес -информатика
код и наименование направления

механико-математического факультета
наименование факультета, института, колледжа

Севостьяновой Ирины Ильиничны
фамилия, имя, отчество

Научный руководитель
доцент, к. ф-м. наук _____ Агафонова Н. Ю.
должность, уч. степень, уч. звание дата, подпись инициалы, фамилия

Заведующий кафедрой
доцент, д. ф-м. наук _____ Сидоров С. П.
должность, уч. степень, уч. звание дата, подпись инициалы, фамилия

Саратов 2020

Введение. Целью настоящей работы является исследование актуальных тенденций применения технологии машинного обучения в сфере образования.

Для достижения заявленной цели работы были поставлены следующие задачи:

— углубленное изучение теоретических основ интеллектуального анализа данных и машинного обучения;
— совершенствование навыков программирования на высокоуровневом языке программирования Python (Version 3.6);
— исследование реализаций алгоритмов машинного обучения на языке Python;
— сбор и анализ данных академической успеваемости студентов механико-математического факультета СГУ имени Н. Г. Чернышевского;
— разработка программного продукта и модели прогнозирования успешности обучения студентов на примере механико-математического факультета СГУ на языке программирования Python;
— анализ и сравнение построенных моделей, выделение закономерностей;
— анализ перспектив применения и развития разработанного программного продукта.

Объектом настоящего исследования является организация образовательного процесса в университете с использованием машинного обучения.

Предметом исследования служат возможность внедрения интеллектуальных технологий в учебный процесс в условиях высшего учебного заведения, эффективность их применения.

Актуальность работы обусловлена повсеместной информатизацией сферы образования путем повышения информационной культуры и грамотности молодых людей, создания комфортных условий для получения знаний и эффективного освоения навыков. Данная тенденция отражена и в национальном проекте «Образование», который в качестве одной из первостепенных задач предусматривает «создание современной и безопасной цифровой образовательной среды, обеспечивающей высокое качество и доступность образования всех видов и уровней». Однако стоит отметить, что чаще всего пилотные проекты разрабатываются, внедряются и исследуются именно на

базе учреждений высшего образования.

Эмпирической основой для проведения исследования являются наборы данных академической успеваемости студентов бакалавриата механико-математического факультета СГУ имени Н. Г. Чернышевского, предоставленные руководством платформы дистанционного обучения IpsilonUni.

Работа имеет следующую структуру:

1. первый раздел настоящей работы посвящен теоретическим основам интеллектуального анализа данных и машинного обучения, анализу наиболее распространенных алгоритмов машинного обучения;
2. второй раздел содержит исследование основных направлений применения современных информационных технологий в области образования, а также обзор некоторых существующих программных продуктов и моделей, разработанных на базе методов машинного обучения;
3. в третьем разделе представлены этапы и результаты разработки программного продукта для прогнозирования успешности обучения студентов методами машинного обучения на примере механико - математического факультета СГУ.

Основное содержание работы. В первом разделе рассматриваются теоретические основы машинного обучения, ключевые понятия области интеллектуального анализа данных; сформулирована задача бакалаврской работы в терминах машинного обучения; приведены методы, используемые для построения моделей.

Машинное обучение является частью научной области, называемой Data Mining, то есть интеллектуальным анализом данных. В основу современной технологии Data Mining положена концепция шаблонов (паттернов), отражающих фрагменты многоаспектных взаимоотношений в данных. Поиск шаблонов производится методами, не ограниченными рамками априорных предположений о структуре выборки и виде распределений значений анализируемых показателей [1].

Машинное обучение исследует методы построения алгоритмов, способных самостоятельно обучаться. Алгоритмы машинного обучения превращают набор данных в модель. Оптимальный алгоритм зависит от типа решаемой задачи, доступных вычислительных ресурсов и характера данных.

В разделе **«Информатизация сферы образования»** рассматривается феномен информатизации сферы образования в масштабе мирового сообщества. Обобщенно приводятся ключевые тенденции применения современных информационных технологий в образовании, а также перспективы и основные направления использования методов машинного обучения в данной области. В качестве теоретического обзора существующих решений и литературы рассматриваются характеристики некоторых наиболее известных программных продуктов, компьютерных платформ, прогнозных моделей, разработанных на базе алгоритмов машинного обучения.

Информатизация образования как процесс интеллектуализации деятельности обучающего и обучаемого, развивающийся на основе реализации возможностей средств новых информационных технологий, поддерживает интеграционные тенденции процесса познания закономерностей предметных областей и окружающей среды, сочетая их с преимуществами индивидуализации и дифференциации обучения.

Основные задачи информатизации сферы образования [2], [3]:

1. повышение качества подготовки специалистов на основе использования в учебном процессе современных информационных технологий;
2. применение активных методов обучения, повышение творческой и интеллектуальной составляющих учебной деятельности;
3. интеграция различных видов образовательной деятельности (учебной, исследовательской и т.д.);
4. адаптация информационных технологий обучения к индивидуальным особенностям обучаемого;
5. разработка информационных технологий дистанционного обучения;
6. совершенствование программно-методического обеспечения учебного процесса.

Раздел **«Прогнозирование успешности обучения студентов методами машинного обучения на примере механико - математического факультета СГУ»** посвящен описанию этапов разработки модели и программного продукта для прогнозирования успешности обучения студентов: сбору, обработке и предварительному анализу данных, построению и сравнительному анализу моделей, а также рассмотрению возможных пер-

спектив развития поставленной задачи.

Основное отличие от приведенных в литературном обзоре работ и решений [4], [5], [6] заключается в ограниченности имеющегося набора данных и признаков, которые выделены на основе обезличенных данных об академической успеваемости вследствие невозможности использования демографических данных о студентах. Также был рассмотрен более широкий круг методов машинного обучения для построения эффективной модели.

Построенные модели обеспечивают приемлемую обобщающую способность для различных метрик, на основе наиболее эффективных были оценены вероятности отчисления студентов механико-математического факультета. Основные выводы по проведенным экспериментам можно сформулировать следующим образом:

1. наиболее точно вычислить вероятность отчисления студента можно после получения итоговой отметки по дисциплине «математический анализ» в третьем или четвертом семестре (в зависимости от направления), таким образом, после второго курса определенного студента уже следует относить к группе риска отчисления;
2. для имеющихся данных лучший результат продемонстрировали модели, построенные на базе методов логистической регрессии и деревьев решений.

Стоит отметить, что улучшение качества моделей впоследствии может быть достигнуто за счет решения проблем ограниченности выборки и рассматриваемых признаков.

Сбор, анализ и предварительная обработка данных для обучения и тестирования моделей. В качестве обучающей выборки рассматривались различные наборы данных академической успеваемости студентов бакалавриата механико-математического факультета, предоставленные руководством платформы дистанционного обучения IpsilonUni. В качестве ключевых параметров построения моделей было решено выбрать дисциплины, присутствующие в учебной программе студентов практически всех направлений механико-математического факультета СГУ, а именно математический анализ, информатика, алгебра и геометрия, также были использованы сведения о баллах ЕГЭ и итоге обучения (студент отчислен/студент успешно завер-

шил обучение). Стоит отметить, что задолженности по вышеперечисленным дисциплинам являются наиболее частыми причинами отчисления студентов.

Общие сведения о собранных данных. Анализ данных был проведен с помощью библиотеки «pandas», визуализация с помощью инструмента «matplotlib» на высокоуровневом языке программирования общего назначения Python (Version 3.6). Общий объем выборки – 138 человек. Форма входных данных и диапазоны значений приведены в соответствии с таблицей 1.

Таблица 1 – Форма входных данных, диапазоны значений

ID	Идентификационный номер студента
exam	Суммарный балл ЕГЭ
it	Отметка по дисциплине «Информатика» в первом семестре
math	Отметка по дисциплине «Математический анализ» в первом семестре
algem	Отметка по дисциплине «Алгебра и геометрия» в первом семестре
per_2	Статус перевода во второй семестр (0 – отчислен/1 – переведен)
it_2	Отметка по дисциплине «Информатика» во втором семестре
math_2	Отметка по дисциплине «Математический анализ» во втором семестре
per_3	Статус перевода на второй курс (0 – отчислен/1 – переведен)
math_final	Итоговая отметка по дисциплине «Математический анализ»
per_4	Статус перевода в пятый семестр (0 – отчислен/1 – переведен)
per_5	Статус перевода на третий курс (0 – отчислен/1 – переведен)
Final	Итоговый статус (0 – отчислен, 1 – успешно закончил)

Основные выводы по анализу исходных данных:

— крайне высокая общая доля отчислений (более $\frac{1}{3}$ студентов);

— пик отчислений приходится на второй курс;

— самая низкая доля отчислений – после первого курса;

— доля отметок по дисциплине «математический анализ» относительно постоянна, однако наблюдается рост значений «неудовлетворительно» в итоговом семестре;

— отмечается обратная тенденция для дисциплины «информатика»: количество значений «неудовлетворительно» существенно снижается во втором (итоговом для данной дисциплины) семестре.

Формальная постановка задачи. Программный продукт должен решать задачу прогнозирования успешности обучения, а также оценивать вероятность отчисления студентов на примере набора данных студентов механико-математического факультета 2019 года выпуска.

Задачу прогнозирования успешности обучения студентов в статистической теории принятия решений можно рассматривать как задачу бинарной классификации результата обучения обучающихся на основе информации об их успеваемости.

Пусть имеется множество студентов Z_i, $i = 1, \ldots, n$, каждый из которых характеризуется t-мерным вектором признаков $X_i = (x_{i1}, \ldots, x_{it})^T$, пусть далее известна принадлежность каждого студента Y_i к одному из двух классов:

$$Y = \begin{cases} y = 1, & \text{студент успешно завершил обучение,} \\ y = -1, & \text{студент был отчислен.} \end{cases} \tag{1}$$

На основе данной выборки необходимо описать процедуры, с помощью которых можно было бы с наибольшей точностью отнести студентов, находящихся в процессе обучения S_j, $j = 1, \ldots, m$ к одному из классов (в данном случае $k = 2$), имея в качестве входной информации наборы признаков $X_j = (x_{j1}, \ldots, x_{jt})^T$, характеризующих студентов. Данные студентов могут содержать как дискретные и непрерывные количественные признаки, так и качественными признаки, поэтому следует рассматривать поставленную задачу классификации как задачу в пространстве разнотипных признаков. В качестве выходной информации используется принадлежность студента к одному из заявленных классов, а также апостериорное распределение.

Для решения поставленной задачи были выделены следующие этапы:

1. предварительная обработка данных;
2. выбор соответствующих алгоритмов машинного обучения;
3. определение метрик классификации;

4. построение моделей, подбор оптимальных параметров;
5. сравнение и анализ полученных результатов.

В качестве основного набора используемых алгоритмов были выбраны следующие:

— метод k-ближайших соседей;
— ядерный метод опорных векторов;
— логистическая регрессия;
— деревья решений.

Оценка значимости признаков проводилась с помощью метода сверхслучайных деревьев. В качестве метрик классификации рассматривались:

1. точность модели на тестовой и обучающей выборках (accuracy) для выявления признаков переобучения и недообучения модели;
2. матрица неточностей (confusion matrix), на основе которой были рассчитаны precision (точность) и recall (полнота).

Первая серия экспериментов была проведена с учетом следующих условий: не рассматривалась итоговая оценка по математическому анализу в третьем или четвертом семестре; после первых тестов были исключены признаки, содержащие категориальную информацию о статусе перевода в следующий семестр из-за высокой корреляции с итогом обучения.

Выводы по результатам первой серии экспериментов:

— лучший результат на тестовом наборе демонстрируют модели логистической регрессии и деревьев принятия решений (точность около 90%);
— наибольшая точность модели на основе логистической регрессии достигается при значении параметра $C = 100.0$, т. е. модель требует меньшей степени регуляризации, что подчеркивает важность правильной классификации каждой точки. При выборе высокого значения параметра регуляризации возникает вероятность переобучения модели, однако построенные модели признаки переобучения не демонстрируют;
— модели чувствительны к выбору соотношения между количеством примеров тестового и обучающего наборов, а также примерам, попавшим в тот или иной набор;
— модель, построенная на основе метода опорных векторов, демонстрирует признаки переобучения: наблюдается тенденция к точному опре-

делению студентов, которые получат диплом, в то же время модель практически не предсказывает отчисление;

— метод k-ближайших соседей позволяет построить модель, демонстрирующую низкую точность классификации при любых наборах параметров, что доказывает необходимость исключить метод из рассмотрения;

— наиболее значимый признак в данной серии экспериментов – результат ЕГЭ (около $0,31$);

— суммарная значимость отметок по математическому анализу за первый и второй семестры составляет – $0,308$, а по информатике – $0,276$, что позволяет сделать заключение о примерно одинаковой степени влияния дисциплин на вопрос об отчислении.

Цель второй серии экспериментов – выявление зависимости качества модели от выбора определенного набора признаков. Была выдвинута следующая гипотеза: модель, построенная на основе данных академической успеваемости по математическому анализу, способна более точно предсказывать отчисление студентов по сравнению с моделью, реализованной на основе данных об успеваемости по дисциплине «информатика».

Выводы по результатам второй серии экспериментов:

— вторая серия экспериментов позволила значительно улучшить точность моделей по сравнению с первой;

— лучшие показатели качества демонстрируют модели с набором признаков exam + math + algem + math_2 + math_final и math + algem + math_2 + math_final;

— модели, построенные на базе академической успеваемости по дисциплине «информатика», значительно менее эффективны по сравнению с моделями, основанными на использовании отметок по математическому анализу, что подтверждает выдвинутую гипотезу;

— учет итоговой отметки по математическому анализу существенно смещает значимость результатов ЕГЭ, при исключении данного показателя (exam) эффективность модели снижается незначительно, также стоит отметить падение значимости отметок по математическому анализу за первый и второй семестры;

— модель прогнозирования отчисления студента демонстрирует высокую

обобщающую способность после добавления итоговой отметки по дисциплине «математический анализ»;

— существенным недостатком построенных моделей является сохранившаяся чувствительность к выбору параметра random_state;

— сохраняется тенденция к превалированию методов логистической регрессии и деревьев решений;

— проведение предварительной обработки позволило значительно повысить эффективность моделей, в первую очередь, построенных ядерным методом опорных векторов.

Заключительный этап решения поставленной задачи включал в себя оптимизацию оценки качества построенных моделей, прогнозирование вероятности отчисления студента.

По результатам перекрестной проверки можно сделать следующие выводы:

— для построенных моделей существует относительно высокий разброс значений правильности, вычисленных для блоков. Подобный результат может означать, что модель сильно зависит от конкретных блоков, использованных для обучения, а также это может быть обусловлено небольшим размером набора данных;

— использование в качестве параметра cv (количество блоков перекрестной проверки) генератора разбиений позволило повысить эффективности перекрестной проверки;

— наиболее точно можно предсказать вероятность отчисления студента после получения итоговой отметки по дисциплине «математический анализ» в третьем или четвертом семестре (в зависимости от направления), следовательно после второго курса можно с вероятностью 0,92 отнести определенного студента к группе риска отчисления;

— по сравнению с оценками второй серии экспериментов использование перекрестной проверки позволило получить наиболее достоверное представление о качестве моделей.

В таблице 2 представлены примеры вероятности отчисления для студентов с различными данными академической успеваемости, предсказанных наиболее эффективными из полученных моделей.

Таблица 2 – Вероятности отчисления для студентов

мат. анализ 1 сем.	алгебра и геометрия	мат. анализ 2 сем.	мат. анализ итоговый сем.	logistic regression	decision tree
4	3	3	3	0.058	0.0
4	3	5	2	0.709	1.0
5	3	3	3	0.06	0.0
3	3	3	2	0.963	1.0
5	4	4	5	0.0	0.0

В качестве основного направления развития разработанного программного продукта выступает возможность создания проекта интеллектуальной надстройки для системы дистанционного образования университета на базе существующих платформ для сбора и учета данных по обучению (MOODLE или IpsilonUni) с использованием методов машинного обучения, где прогнозирование успешности обучения студента будет внедрено в роли одного из ключевых модулей.

Заключение. В работе были исследованы перспективы применения современных информационных технологий, в частности машинного обучения и интеллектуального анализа данных, в качестве основы развития сферы образования.

Были решены следующие задачи:

— углубленное изучение теоретических основ интеллектуального анализа данных и машинного обучения;
— совершенствование навыков программирования на высокоуровневом языке программирования Python (Version 3.6);
— исследование реализаций алгоритмов машинного обучения на языке Python;
— сбор и анализ данных академической успеваемости студентов механико-математического факультета СГУ имени Н. Г. Чернышевского;
— разработка программного продукта для прогнозирования успешности обучения студентов на примере механико-математического факультета СГУ на языке программирования Python;
— анализ и сравнение построенных моделей, выделение закономерностей;
— анализ перспектив применения и развития разработанного программ-

ного продукта.

Рассмотрение проблемы прогнозирования успешности обучения студентов позволило выявить скрытые зависимости между результатами освоения образовательных дисциплин и вероятностью отчисления обучающихся. Очевидно, вопросы отчисления непосредственно значимы как для государственного сектора вследствие снижения количества потенциальных квалифицированных специалистов, а также потери части ресурсов, выделенных на обучение, так и для самих обучающихся в связи с затратой времени и усилий. Таким образом, оптимизация решения вопросов об отчислении является насущным вопросом, который может быть решен путем превентивной аналитики и принятия соответствующих мер на уровне образовательного учреждения.

Разработанный продукт позволяет выявлять студентов в группе риска уже после третьего или четвертого семестра с точностью 92% на основе данных об академической успеваемости, накапливаемых на базе университета, что характеризует высокое качество построенного алгоритма обработки данных. Кроме стандартного регрессионного анализа были также использованы методы деревьев решений, ядерный метод опорных векторов, что позволило путем сравнительного анализа выявить оптимальные алгоритмы для решения поставленной задачи. Однако качество моделей значительно зависит от количества анализируемых семестров, за которые имеются данные об успеваемости, а также от распределения примеров по тестовым блокам, что свидетельствует о необходимости дальнейших исследований на более широких выборках.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 *Brownlee, J.* Basic concepts in machine learning / J. Brownlee // *Start Machine Learning.* — 2015.

2 *Зверева, Ю. С.* Информатизация образования / Ю. С. Зверева // *Молодой ученый.* — 2016. — № 6.3. — С. 23 – 26.

3 *Григорьев, С. Г.* Информатизация образования. Фундаментальные основы / С. Г. Григорьев, А. Д. Гришкун. — М., 2005.

4 *Wu, Q.* Predicting academic perfomance via machine learning methods / Q. Wu // *Submitted to the Undergraduate Research Scholars program at Texas A&M University in partial fulfillment of the requirements for the designation as an UNDERGRADUATE RESEARCH SCHOLAR.* — 2017.

5 *Berens, J.* Early detection of students at risk - predicting student dropouts using administrative student data and machine learning methods / J. Berens, K. Schneider, S. Oster, J. Burghoff // *SCHUMPETER DISCUSSION PAPERS. SCHUMPETER school of Business and Economics, University of Wuppertal, Germany.* — 2018.

6 Knewton: адаптивное обучение в действии [Электронный ресурс]. — URL: https://newtonew.com/tech/knewton-adaptivnoe-obuchenie-v-dejstvii (Дата обращения 10.02.2020). - Загл. с экрана. - Яз. рус.