МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение высшего образования
«Тольяттинский государственный университет»

Кафедра «Прикладная математика и информатика»
(наименование)

01.03.02 Прикладная математика и информатика
(код и наименование направления подготовки / специальности)

Компьютерные технологии и математическое моделирование
(направленность (профиль) / специализация)

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА (БАКАЛАВРСКАЯ РАБОТА)

на тему «Исследование методов дискриминантного анализа базы данных»

Обучающийся Е.А. Плевако
(Инициалы Фамилия) (личная подпись)

Руководитель М.А. Тренина
(ученая степень (при наличии), ученое звание (при наличии), Инициалы Фамилия)

Консультант к.п.н., доцент, А. В. Егорова
(ученая степень (при наличии), ученое звание (при наличии), Инициалы Фамилия)

Тольятти 2024

Аннотация

Тема выпускной квалификационной работы: «Исследование методов дискриминантного анализа базы данных». В данной работе реализована программа для анализа базы данных частной школы программирования и внешкольного образования, с целью улучшения успеваемости.

Выпускная квалификационная работа посвящена исследованию и реализации методов дискриминантного анализа для классификации данных об успеваемости учащихся частной школы программирования.

Цель работы – разработка и тестирование алгоритма классификации на основе дискриминантного анализа.

Объект исследования – методы дискриминантного анализа.

Предмет исследования – данные об успеваемости учащихся. В работе поставлены следующие задачи:

- изучить теоретический материал о дискриминантном анализе;
- разработать алгоритм для решения задачи классификации;
- выполнить программную реализацию разработанного алгоритма;
- провести исследование эффективности реализованного алгоритма.

Актуальность работы обусловлена возможностью улучшения образовательных процессов и индивидуализации подходов к обучению на основе анализа данных.

Результатом работы является разработанный и протестированный алгоритм классификации на основе методов дискриминантного анализа, который успешно классифицирует данные об успеваемости учащихся частной школы программирования.

Abstract

The topic of the graduation work is "Investigation of Discriminant Analysis Methods for Database."

The graduation work consists of an introduction, three parts, a conclusion, pictures, tables, list of references including foreign sources.

The object of the graduation work is research and implementation of methods of discriminant analysis for classifying data on the academic performance of students of a private programming school. We touch upon the problem of academic performance in schools, and are looking for the most appropriate method of discriminant analysis to solve it.

The aim of the work is to develop and test a classification algorithm based on discriminant analysis. We examine how each of the mentioned methods of discriminant analysis will help us in the task of classification.

The graduation work may be divided into several logically connected parts which are study the theoretical material on discriminant analysis; develop an algorithm to solve the classification problem; implement the developed algorithm in software; conduct a study on the effectiveness of the implemented algorithm.

Finally, we present the work on implemented program for analyzing the database of a private school of programming and extracurricular education, in order to improve academic performance. This program will successfully classify the database using the discriminant analysis methods we have chosen.

The relevance of the thesis is due to the possibility of improving educational processes and individualizing teaching approaches based on data analysis.

In conclusion we'd like to stress this work is relevant in solving the problem of the education system as well as technological and constructive solutions can be used for medicine, mathematics and other areas.

Оглавление

Введение

Дискриминантный анализ является одним из ключевых методов статистического анализа, используемым для решения задач классификации. Этот метод был впервые предложен Рональдом Фишером в 1936 году и с тех пор активно применяется в различных областях от биологии и медицины до финансов и социальных наук. Суть метода заключается в определении принадлежности объекта к одной из нескольких групп на основе предварительно известных характеристик этих групп.

В последнее время значительное внимание уделяется применению дискриминантного анализа в образовательной сфере, особенно при анализе и мониторинге успеваемости учащихся. Это связано с возрастающей потребностью в индивидуальном подходе к обучению и оптимизации учебных процессов с помощью данных.

Целью данной бакалаврской работы является разработка и апробация методов дискриминантного анализа для решения задачи классификации успеваемости учащихся, основываясь на данных частной школы программирования. Эта задача включает работу с большими объемами данных об обучении и успеваемости учеников.

Предметом исследования являются алгоритмы машинного обучения, применяемые для анализа учебных данных в образовательной среде.

Для достижения поставленной цели необходимо выполнить следующие задачи:

- изучить теоретический материал по дискриминантному анализу и методам машинного обучения, применимым для классификации данных;
- разработаем алгоритмы для классификации учащихся на основе их успеваемости;
- выполнить программную реализацию разработанных алгоритмов;
- провести исследование эффективности алгоритмов, анализируя их

способность точно классифицировать учащихся по различным категориям успеваемости.

Методами исследования в данной работе являются математическое моделирование и программирование, а также использование современных аналитических инструментов и библиотек для обработки и анализа данных. В частности, будет применяться язык программирования Python и его библиотеки, такие как NumPy, Pandas и Scikit-learn, которые широко используются в сфере анализа данных и машинного обучения.

Практическое значение бакалаврской работы заключается в разработке программных алгоритмов, способных анализировать и классифицировать данные об успеваемости учащихся. Результаты могут быть использованы школой для оптимизации учебного процесса и персонализированной поддержки студентов, что в конечном итоге должно способствовать повышению образовательной эффективности.

Бакалаврская работа состоит из введения, трех глав, заключения и списка используемой литературы. В первой главе описываются теоретические основы дискриминантного анализа, во второй главе приводится описание данных и методики их обработки, третья глава посвящена разработке и тестированию программных алгоритмов. Заключение подводит итоги исследования и предлагает направления для дальнейших работ.

Глава 1 Введение в дискриминантный анализ

1.1 Математические основы дискриминантного анализа

Дискриминантный анализ – это метод статистического анализа, который используется для классификации объектов по заранее известным группам на основе их характеристик. Основная цель дискриминантного анализа – найти функции, которые позволяют максимально различать группы и минимально различать объекты внутри групп. Этот метод был впервые предложен Рональдом Фишером в 1936 году и с тех пор активно применяется в таких областях, как биология, медицина, финансы и социальные науки [1].

Дискриминантный анализ относится к категории методов многомерного статистического анализа, предполагающих описание объектов комплексом признаков. Высокая разрешающая способность таких методов обусловлена не только принципом «больше учтённых признаков – больше информации». Не менее важно, что эти методы учитывают систему корреляции признаков [2]. Как и многие другие многомерные методы, дискриминантный анализ основан на построении линейных комбинаций признаков – функций, в которые каждый из них входит со своим коэффициентом (вкладом). В дискриминантном анализе линейные комбинации называются соответственно дискриминантными функциями:

$$DF = b_1 x_1 + \ldots + b_i x_i + \cdots + b_p x_p + C, \quad (1)$$

где x_i – численное значение i-го признака;

b_i – вклад i-го признака в значение функции;

p – число признаков;

C – константа.

Дискриминантный анализ обеспечивает объективное сравнение (разделение) групп за счёт искусственной минимизации внутригруппового разнообразия (дисперсии) [2].

Обратившись к дискриминантному анализу, можно не только оценить достоверность межгрупповых различий и оценить «расстояния» между группами, но и определить те признаки из числа выбранных, которые в первую очередь обусловливают межгрупповые различия. Более того, когда две (или большее число) группы уже разделены в дискриминантном анализе, возможно определить принадлежность неизвестного объекта к одной из них.

1.2 Геометрическая интерпретация результатов дискриминантного анализа

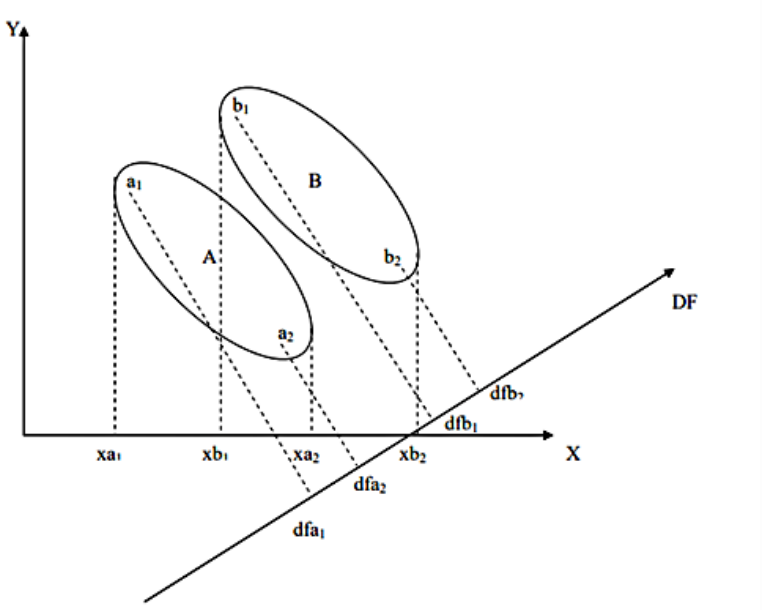
Исходные данные для дискриминантного анализа должны представлять собой совокупность объектов. Их необходимо разделить на несколько групп (классов). Число признаков, по которым описаны объекты, не может быть меньше двух и не должно превышать суммарного числа объектов. Признаки должны быть количественными, а распределение их значений в каждом классе – нормальным. Недопустимо включение в комплекс признаков с единичной (полной) корреляцией [3]. Описание объекта по комплексу из p признаков геометрически эквивалентно определению его координат в p-мерном пространстве. Единичному объекту в пространстве отвечает точка, соответственно группе объектов – «группа точек». Две или более «группы точек», отвечающих изучаемым группам объектов, нередко перекрываются, то есть неоднозначно различимы. Практически всегда есть показатели с такой координатой, которая не позволяет однозначно отнести их к той или иной группе. Задача дискриминантного анализа геометрически формулируется как построение нового пространства, в котором принадлежность объектов к группам определяется однозначно – перекрывание «групп точек» становится минимальным. Координаты объектов в новом пространстве (ортогональном)

определяются значениями дискриминантных функций –уникальных линейных комбинаций признаков, определяемых на основе двух корреляционных матриц: межгрупповой (корреляции групповых средних) и внутригрупповой (корреляции признаков отдельных объектов внутри группы). Выбор таких осей производится по критерию максимума отношения межгрупповой дисперсии к внутригрупповой.

Суть подхода продемонстрирован на рисунке 1. Данный пример был взят из работы [17], раздел 1. «Пусть мы имеем две группы объектов, описанных по двум переменным. Эти переменные образуют плоскость, на которой каждая из групп представлена в виде «группы точек». Центроиды этих групп обозначены A и B. При рассмотрении проекций крайних объектов каждой группы на координатные оси признаков, например ось X, видны области перекрывания групп» [15, с.7]. Иначе, в каждой из групп имеются объекты, с неопределённым положением. Соответственно точно относиться к той или иной группе они не могут. На рисунок 1 – это объекты в интервале $xb_1 - xa_2$. Но можно построить новую ось (на рисунок 1 – DF), которая по-другому учтёт исходную изменчивость, благодаря своим значениям. В результате чего внутригрупповая изменчивость минимизируется и позволит чётко выразить межгрупповые различия, тем самым проблема перекрывания «групп точек» исключена. Новая ось, то есть новый интегральный признак, является линейной комбинацией исходных признаков – дискриминантной функцией [6].

Из рисунка 1 видно, что чёткое разделение групп происходит благодаря минимизации внутригрупповых различий. Расстояния между проекциями объектов a_1, a_2 и b_1, b_2 на ось дискриминантной функции меньше, чем расстояния между их проекциями на ось X. Расстояние между проекциями центроидов групп A и B на ось DF за счёт этого оказывается большим, чем расстояние между проекциями на ось X (и аналогично на ось Y). «Это расстояние, называемое расстоянием Махаланобиса, и оценивает действительные различия групп» [7].

На рисунке 1 иллюстрируется возможность дискриминантного анализа, а именно при разделении двух групп. Их может быть и больше, но это потребует определения дополнительных дискриминантных функций ортогональных к первой. В общем случае число необходимых дискриминантных функций на единицу меньше числа разделяемых групп. «Однако информативная ценность этих функций разная, поскольку они учитывают разную долю исходной изменчивости комплекса признаков» [8]. Первая дискриминантная функция учитывает максимум такой изменчивости, вторая – максимум остатка и так далее. «Если первые две-три функции учли значительную (например, более 80%) долю исходной дисперсии, то оставшимися функциями можно пренебречь в решении последующих задач дискриминантного анализа» [17, с.8].



A, B – центроиды разделяемых групп; xa_1, xb_1 – минимальные значения признака X в группах A и B; $xa_2,\ xb_2$ – максимальные значения признака в этих группах; dfa_1, dfa_2 – проекции конкретных объектов группы *A* на ось дискриминантной функции; dfb_1, dfb_2 – то же для конкретных объектов группы *B*

Рисунок 1 – Геометрическая интерпретация результатов дискриминантного анализа

Помимо учёта процента дисперсии, у дискриминантной функции есть не менее важные критерии значимости. Коэффициент канонической корреляции (R), является одним из таких важных характеристик, позволяющей оценить информативность дискриминантной функции. Также, каноническая корреляция, которая оценивает меру связи между двумя множествами переменных. Чем больше значение R, тем выше разделительная способность дискриминантной функции.

Основной критерий оценки эффективности дискриминации для конкретной функции – величина λ-Уилкса. «Этот критерий оценивает остаточную дискриминационную способность, под которой понимается способность различать группы, если исключить информацию, полученную с помощью ранее вычисленных функций. Когда остаточная дискриминация мала, вычисление очередной дискриминантной функции не имеет смысла» [12]. Статистика λ-Уилкса подчиняется χ^2-распределению и поэтому вычисление данного критерия позволяет оценить достоверность межгрупповых различий для каждой из полученных функций.

В результате дискриминантного анализа образуется ось, плоскость или пространство, в зависимости от числа дискриминантных функций, дифференцирующее сравниваемые группы. Вычислив расстояние между центроидами «групп точек», можно оценить степень сходства групп. В качестве меры сходства, как уже говорилось, выступает расстояние Махаланобиса (D^2), достоверность которого оценивается при помощи критерия Фишера.

«Следует отметить, что данное расстояние учитывает корреляционные свойства признаков. Именно этим оно отличается от Евклидова расстояния, которое оценивает простое сходство групп на плоскости» [17]. При переходе в пространство Махаланобиса учитываются все дискриминантные функции, в том числе и малоинформативные. За счёт этого увеличивается статистический шум и межгрупповые расстояния могут не соответствовать наблюдаемым двумерным распределениям [18].

При помощи классификационной модели может быть проверено качество дискриминации групп. Эта модель вычисляется по классифицирующим функциям, определённым для каждого объекта совокупности. Если большинство объект относится к «своей» группе, дискриминация объектов является правильной.

Результат реализации классификационной матрицы рассмотрен в таблице 1.

Таблица 1 – Матрица классификаций дискриминантного анализа

Группа	Доля верных отнесений	Число попаданий объектов в группу А	Число попаданий объектов в группу В
А	$C / (C + d)$	c	d
В	$f / (e + f)$	e	f
Примечание – в данной таблице указаны: А, В – условные обозначения групп; c – число объектов группы А, отнесённых в область распределения своей группы; d – число объектов групп А, отнесённых в область распределения группы В; e и f – соответствующее число объектов группы В.			

Чем выше доля отнесений объектов в «свою» группу, тем лучше качество дискриминации и меньше вероятность ошибок при классификации неизвестных объектов.

1.3 Линейный дискриминантный анализ

1.3.1 Общее понятие

Линейный дискриминантный анализ предполагает, что данные каждого класса имеют многомерное нормальное распределение с одинаковой ковариационной матрицей. Линейный дискриминантный анализ строит линейные функции для каждого класса, называемые дискриминантными функциями. Для двух классов C_1 и C_2 дискриминантная функция имеет вид:

$$D(x) = w^T x + w_0, \quad (2)$$

где w – вектор коэффициентов дискриминантной функции;

w_0 – константа.

Коэффициенты w вычисляются следующим образом:

$$w = \Sigma^{-1}(\mu_1 - \mu_2), \quad (3)$$

где Σ – общая ковариационная матрица;

μ_1 и μ_2 – векторы средних значений признаков для классов C_1 и C_2.

Зная эти функций, более подробно рассмотрим работу линейного дискриминантного анализа.

1.3.2 Вывод линейных дискриминантных функций

Рассмотрим задачу классификации объекта x, определяемого как x = (x_1, x_2, ..., x_p,). Пусть множество объектов разбито на K непересекающихся классов C_1, C_2,...,C_K[14]. Цель дискриминантного анализа – построить классификатор, который на основе вектора признаков x определяет принадлежность объекта к одному из классов.

Основная идея линейного дискриминантного анализа – найти проекции многомерных данных на одномерное пространство, которые максимально разделяют классы. Это достигается путем максимизации отношения межклассовой дисперсии к внутриклассовой дисперсии. Математически это можно выразить следующим образом:

Межклассовая дисперсия, представленная ниже (S_B) измеряет различия между средними значениями классов:

$$S_B = \sum_{k=1}^{K} \pi_k (\mu_k - \mu)^T (\mu_k - \mu), \quad (4)$$

где π_k – число объектов в классе k;

μ_k – средние значения признаков для класса k;

μ – общий вектор средних значений признаков.

Внутриклассовая дисперсия (S_W) измеряет различия внутри каждого класса:

$$S_W = \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \mu_k)^T (x_i - \mu_k) \quad (5)$$

где $x_i^{(k)}$ – вектор признаков i-го объекта в классе k.

Коэффициенты дискриминантной функции находятся путем решения следующей задачи:

$$max_w \frac{w^T S_B w}{w^T S_W w}. \quad (6)$$

Рассмотрим простой пример. Допустим, у нас есть два класса C_1 и C_2 с двумя признаками x_1 и x_2. Средние значения признаков для классов:

$$\mu_1 = \begin{pmatrix} 2 \\ 3 \end{pmatrix}, \mu_2 = \begin{pmatrix} 5 \\ 8 \end{pmatrix}, \quad (7)$$

ковариационная матрица Σ для обоих классов:

$$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}, \quad (8)$$

тогда вектор коэффициентов w можно вычислить так:

$$w = \Sigma^{-1}(\mu_1 - \mu_2) = \begin{pmatrix} 1 & -0,5 \\ -0,5 & 0,5 \end{pmatrix} \begin{pmatrix} -3 \\ -5 \end{pmatrix} = \begin{pmatrix} -3,5 \\ -2 \end{pmatrix}. \quad (9)$$

Таким образом, дискриминантная функция будет:

$$D(x) = -3.5x_1 - 2x_2 + w_0. \quad (10)$$

Таким образом мы рассмотрели задачу классификации объекта и основные формулы для его решения. Переходим к следующему виду дискриминантного анализа.

1.4 Квадратичный дискриминантный анализ

1.4.1 Общее понятие

Квадратичный дискриминантный анализ является расширением линейного дискриминантного анализа и используется в тех случаях, когда ковариационные матрицы классов не равны. В отличие от линейного, квадратичный дискриминантный анализ позволяет строить квадратичные разделяющие поверхности, что делает его более гибким для задач классификации, где классы не могут быть разделены линейно.

Для линейного и квадратичного дискриминантного анализа $P(x|y)$ моделируется многомерным распределением Гаусса с плотностью. Тем самым для каждого класса (k) мы имеем:

$$P(x|y = K) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(x - \mu_k)^t \Sigma_k^{-1}(x - \mu_k)\right) \quad (11)$$

где μ_k– вектор средних значений признаков для класса k;

Σ_k – ковариационная матрица для класса k;

d – размерность вектора признаков x.

Линейный дискриминантный анализ и квадратичный дискриминантный анализ могут быть получены из простых вероятностных моделей, которые моделируют условное распределение данных по классам.

1.4.2 Классификация с использованием квадратичного дискриминантного анализа

Для классификации нового объекта x мы вычисляем дискриминантную функцию для каждого класса и выбираем класс с наибольшим значением функции. Дискриминантная функция для класса C_k имеет вид:

$$D_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log P\,(C_k), \quad (12)$$

где $P(C_k)$ – априорная вероятность класса C_k

Рассмотрим пример вычисления квадратичного дискриминантного анализа с двумя классами C_1 и C_2 и двумя признаками x_1 и x_2. Пусть средние значения признаков и ковариационные матрицы для классов следующие:

$$\mu_1 = \begin{matrix}2\\3\end{matrix},\ \mu_2 = \begin{matrix}5\\8\end{matrix}, \quad (13)$$

$$\Sigma_1 = \begin{pmatrix}1 & 0.3\\ 0.3 & 0.8\end{pmatrix},\ \Sigma_2 = \begin{pmatrix}1.5 & 0.4\\ 0.4 & 1.2\end{pmatrix}. \quad (14)$$

Для классификации нового объекта $x = \begin{pmatrix}4\\6\end{pmatrix}$ вычислим дискриминантные функции:

$$D_1(x) = -\frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \log P\,(C_1), \quad (15)$$

$$D_2(x) = -\frac{1}{2}\log|\Sigma_2| - \frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2) + \log P\,(C_2). \quad (16)$$

После подстановки значений и выполнения вычислений мы получим два значения дискриминантных функций D_1 и D_2 . Объект x будет отнесен к тому классу, для которого значение дискриминантной функции будет больше.

1.5 Гибкий дискриминантный анализ

1.5.1 Общее понятие

Гибкий дискриминантный анализ – это метод, который позволяет более точно разделять данные на группы, особенно когда линейные или квадратичные модели недостаточны. Гибкий дискриминантный анализ не ограничивается линейными или квадратичными границами и может использовать более сложные преобразования данных.

Данный вид дискриминантного анализа использует нелинейные преобразования признаков, чтобы лучше разделять классы. Для этого могут применяться разные методы:

- полиномиальные функции: учитываются не только исходные признаки, но и их степени и произведения;
- сплайны: используются кусочно-линейные или кусочно-квадратичные функции;
- ядерные методы: преобразование признаков с помощью специальных функций, позволяющих работать в высоко размерных пространствах.

Преимущества гибкого дискриминантного анализа:

- универсальность: применим к различным задачам классификации;
- гибкость: может моделировать сложные зависимости между признаками;
- интерпретируемость: позволяет понять, как различные признаки влияют на классификацию.

1.5.2 Пример применения

В рамках выпускной квалификационной работы рассматривается задача классификации успеваемости учащихся частной школы программирования. Данные включают оценки по математике, английскому языку, программированию и рисованию, а также показатели посещаемости и участия во внеклассных мероприятиях.

Для классификации учащихся на группы с высокой, средней и низкой успеваемостью применим гибкий дискриминантный анализ:

- выбор преобразований: используем полиномиальные функции второго порядка для учета нелинейных зависимостей между признаками. Например, включим квадраты оценок и произведения различных оценок;
- построение модели: на основе выбранных преобразований строим модель гибкого дискриминантного анализа, которая учитывает сложные взаимосвязи между оценками и успеваемостью;
- классификация: новые данные о учениках преобразуются с помощью тех же функций, и на их основе производится классификация. Эти шаги позволят более точно определить группу успеваемости каждого ученика, учитывая все имеющиеся данные и их сложные взаимосвязи.

Эти шаги позволят более точно определить группу успеваемости каждого ученика, учитывая все имеющиеся данные и их сложные взаимосвязи.

1.6 Смешанный дискриминантный анализ

Смешанный дискриминантный анализ – это метод, который сочетает в себе преимущества линейного дискриминантного анализа и квадратичного дискриминантного анализа. Он используется в тех случаях, когда одни классы можно разделить линейно, а для других требуется более сложное, нелинейное разделение. Такой вид анализа позволяет адаптивно выбирать наилучший метод для каждой пары классов, что делает его более гибким и точным в сложных задачах классификации.

Основная идея смешанного дискриминантного анализа заключается в том, чтобы комбинировать линейные и нелинейные методы в зависимости от структуры данных. Этот подход позволяет более эффективно использовать информацию, содержащуюся в признаках, и улучшать качество классификации. Анализ может включать в себя несколько этапов:

- предварительный анализ: определение структуры данных и выявление классов, которые могут быть разделены линейно, и тех, для которых требуется нелинейное разделение;
- выбор метода: применение линейного дискриминантного анализа для линейно разделимых классов и квадратичного дискриминантного анализа для нелинейно разделимых классов;
- комбинирование результатов: объединение результатов линейного и квадратичного анализов для получения окончательной классификации.

1.7 Регуляризованный дискриминантный анализ

Регуляризованный дискриминантный анализ улучшает традиционные методы дискриминантного анализа, такие как линейный и квадратичный анализ, добавляя регуляризационные параметры. Эти параметры помогают модели быть более стабильной и эффективной, особенно когда данных много, а признаков еще больше.

Регуляризация добавляет ограничения к модели, что помогает избежать переобучения и улучшить ее способность обобщать данные. В регуляризованном дискриминантном анализе используются два основных типа регуляризации:

- L1-регуляризация (Лассо): уменьшает коэффициенты незначимых признаков до нуля, что помогает убрать лишние данные;
- L2-регуляризация (Гребневая регрессия): разглаживает влияние всех признаков, предотвращая чрезмерное внимание к отдельным признакам.

Выводы по главе 1

В первой главе мы изучили основные концепции дискриминантного анализа, который является важным инструментом для классификации данных на основе их характеристик, а также рассмотрели его визуализацию используя

график на рисунке 1. На основе этого рисунка, создали таблицу, отображающую матрицу классификаций дискриминантного анализа. Линейный дискриминантный анализ и квадратичный дискриминантный анализ обеспечивают линейные и квадратичные разделяющие поверхности соответственно, что позволяет эффективно разделять классы данных. Гибкий дискриминантный анализ предлагает еще большую гибкость за счет использования нелинейных преобразований, таких как полиномиальные функции и сплайны. Смешанный дискриминантный анализ объединяет сильные стороны линейного и квадратичного дискриминантных анализов, чтобы адаптивно выбирать оптимальные методы для классификации. Регуляризованный дискриминантный анализ, за счёт своей более стабильной и эффективной модели, является более улучшенным по сравнению с линейным и квадратичными дискриминантными анализами. Рассмотрели примеры применения каждого из этих анализов и формулы (математические модели) их работы.

Эти методы предоставляют мощные инструменты для решения задач классификации в различных областях, включая образование, что делает их незаменимыми в анализе и интерпретации сложных наборов данных.

Глава 2 Анализ и обработка данных

2.1 Сбор и предварительная обработка данных об успеваемости учащихся

Для исследования методов дискриминантного анализа базы данных перейдём к подбору необходимой информации [3]. В этом разделе подробно описан сбор и предварительная обработка данных об успеваемости учащихся центра программирования и дополнительного образования «Умная школа». Данные включают в себя различные атрибуты, такие как ФИО учащихся, пол, возраст, класс, оценки по предметам центра и участие во внеклассных мероприятиях.

Данные были взяты из школьных записей и электронной системы управления образовательным учреждением. Эти источники предоставили необходимую информацию об учащихся и их успеваемости по нескольким параметрам, включая средняя оценка по предметам в течении последнего месяца , посещаемость, а также дополнительные данные об участие во внеучебных мероприятиях.

Собранные данные включали следующие поля:

- ФИО;
- пол;
- возраст;
- класс;
- средняя оценка по математике (округлённая);
- средняя оценка по английскому языку (округлённая);
- средняя оценка по программированию (округлённая);
- средняя оценка по рисованию (округлённая);
- количество посещений творческих мастер-классов (за последний месяц);
- общая посещаемость (часов за последний месяц);

− количество участий во внеклассных мероприятиях (за последний месяц).

Данные были организованы в табличном формате, где каждая строка представляла отдельного учащегося, а каждый столбец соответствовал одному из вышеупомянутых атрибутов. В целях конфиденциальности, в столбце «ФИО», ученики просто пронумерованы с добавлением условного обозначения У- тире ученик. Результат организации данных представлены в таблице 2.

Таблица 2 – Данные об учащихся

ФИО	Пол	Возраст	Класс	Средняя оценка по				Количество посещений		Общая посещаемость, ч
				математике	английскому языку	программированию	рисованию	творческих мастер-классов	внеклассных мероприятий	
У1	М	15	10	4	3	5	2	12	1	18
У2	Ж	12	7	5	5	4	3	8	0	12
У3	М	17	11	3	2	3	4	10	1	20
У4	Ж	10	4	2	4	2	5	7	0	8
У5	М	9	3	5	3	4	3	6	0	10
У6	Ж	7	1	4	4	3	2	9	1	5
У7	М	14	9	2	5	2	5	11	0	17
У8	Ж	15	10	5	3	5	4	10	1	9
У9	М	8	2	3	2	4	3	5	0	6
У10	Ж	11	5	4	4	3	5	12	1	20
У11	М	16	10	5	3	5	2	8	0	15
У12	Ж	13	8	3	4	2	4	9	0	8
У13	М	12	7	4	2	3	5	7	1	14
У14	Ж	14	9	5	5	4	3	6	0	18
У15	М	11	5	3	3	2	2	11	1	10
У16	Ж	15	10	2	4	5	5	10	0	11
У17	М	17	11	4	3	3	4	8	1	6
У18	Ж	16	10	5	5	4	3	9	0	15

Продолжение таблицы 2

У19	М	13	8	2	2	2	5	7	1	7
У20	Ж	7	1	3	3	5	4	6	0	13

Сбор данных осуществлялся двумя основными методами [9]:

- опросы и анкетирование: Я попросила учителей заполнить анкеты, в которых подробно описывались успеваемость и участие каждого учащегося. А также провести опрос;
- автоматизированные системы: Данные были получены из электронной системы управления школой, которая затем была экспортирована в формат CSV для дальнейшей обработки.

После сбора данные прошли несколько этапов обработки, чтобы обеспечить их готовность к анализу:

- очистка данных: она включала удаление повторяющихся записей и исправление любых ошибок в данных (например, опечаток в именах учащихся);
- обработка пропущенных значений: недостающие данные были получены путем вычисления значений на основе средней успеваемости в соответствующем классе или медианных значений;
- нормализация: оценки были стандартизированы по единой шкале для облегчения сравнения по различным предметам. Высчитывалась средняя оценка ученика и округлялась в зависимости от результатов опроса;
- категориальное кодирование данных: гендерные и другие категориальные переменные были преобразованы в числовые форматы с использованием таких методов, как One-Hot Encoding [4].

Для анализа и визуализации обработанных данных использовались различные статистические методы и инструменты [11]:

- статистический анализ: это включало анализ распределения оценок, выявление отклонений и выявление закономерностей или тенденций;
- визуализация данных: графики и диаграммы были созданы с использованием библиотек Python, таких как Matplotlib и Seaborn, для визуального представления данных и улучшения анализа.

Собранные и обработанные данные служат основой для последующего дискриминантного анализа, целью которого является классификация успеваемости учащихся по множеству признаков и выявление ключевых факторов, влияющих на их результаты в учебе.

Такой структурированный подход обеспечивает всестороннее понимание успеваемости учащихся и способствует разработке эффективных стратегий для улучшения результатов обучения [19].

2.2 Методы отбора признаков для дискриминантного анализа

Выбор признака является важным этапом в процессе дискриминантного анализа, поскольку он определяет наиболее значимые признаки, которые способствуют точности классификации модели. Цель выбора признака - выявить и сохранить наиболее информативные и значимые признаки, удалив при этом несущественные или избыточные. Это помогает улучшить производительность модели, снизить сложность вычислений и повысить интерпретируемость. Для выбора признаков в дискриминантном анализе можно использовать несколько методов, включая фильтрационные методы, обёрточные методы и встроенные методы [5].

2.2.1 Фильтрационные методы

Методы фильтрации оценивают значимость признаков путем оценки их статистических свойств по отношению к целевой переменной, независимо от какого-либо алгоритма обучения. Эти методы эффективны в вычислительном отношении и просты в реализации. Распространенные методы фильтрации включают:

- коэффициент корреляции: этот метод измеряет линейную зависимость между каждым признаком и целевой переменной. Признаки с более высокими коэффициентами корреляции считаются более релевантными. Однако этот метод может не учитывать нелинейные взаимосвязи;
- критерий χ^2: критерий χ^2 используется для категориальных признаков, чтобы оценить независимость между признаком и целевой переменной. Признаки с низкими значениями вероятности считаются значимыми;
- дисперсионный анализ (ANOVA): ANOVA тестирует разницу в средних значениях между различными группами, определёнными целевой переменной. Выбираются признаки со значимыми различиями (низкими значениями p);
- взаимная информация: это измерение объема информации, которую одна переменная предоставляет о другой. Признаки с более высокими показателями взаимной информативности считаются более информативными для задачи классификации.

2.2.2 Обёрточные методы

Обёрточные методы оценивают полезность признаков путем прямого измерения их влияния на производительность конкретного алгоритма обучения. Эти методы включают процесс поиска, в ходе которого выбираются различные подмножества функций, а их эффективность оценивается с помощью классификатора. [16] К распространенным обёрточным методам относятся:

- пошаговый отбор: этот итеративный метод начинается с пустого набора функций и добавляет по одной функции за раз в зависимости от их вклада в производительность модели, пока не будет замечено существенных улучшений;
- пошаговое исключение: этот итеративный метод начинается с пустого набора признаков и добавляет по одному признаку за раз на основе его

вклада в производительность модели, до тех пор, пока не будет достигнуто значительное улучшение;

- рекурсивное исключение признаков (RFE): RFE работает путем рекурсивной подгонки модели и удаления наименее важных признаков на основе коэффициентов модели или значений важности признаков до тех пор, пока не будет достигнуто желаемое количество признаков.

2.2.3 Встроенные методы

Встроенные методы выполняют выбор признаков в процессе обучения модели. Эти методы интегрированы в определенные алгоритмы обучения и обычно включают методы регуляризации, которые устраняют нерелевантные признаки [20]. К распространенным встроенным методам относятся:

- LASSO (оператор наименьшего абсолютного сжатия и выбора): LASSO добавляет к функции потерь элемент регуляризации L1, который уменьшает коэффициенты менее важных признаков до нуля, эффективно выполняя отбор признаков;
- гребневая регрессия: как и в случае с LASSO, в гребневой регрессии используется регуляризация L2, которая сжимает коэффициенты, но не обнуляет их. Гребневая регрессия может сочетаться с другими методами для достижения отбора признаков;
- решающие деревья и случайные леса: эти алгоритмы по своей сути выполняют отбор признаков, выбирая признаки, которые обеспечивают наилучшее разбиение в процессе построения дерева. Значения важности признаков, полученные на основе этих моделей, могут быть использованы для выбора соответствующих признаков.

2.3 Комбинация методов отбора дискриминантного анализа

Для дискриминантного анализа данных, описанных в разделе 2.1, наиболее эффективным подходом является комбинация методов отбора

признаков. Каждый метод имеет свои преимущества и может быть полезен на разных этапах анализа.

Первичный отбор с использованием методов фильтрации [15]:

На первом этапе целесообразно применить фильтрационные методы для быстрого и точного отбора признаков:

- коэффициент корреляции: этот метод поможет оценить линейную зависимость между оценками по предметам, посещаемостью и успеваемостью учащихся и целевой переменной (например, успешностью или членством в группе);
- дисперсионный анализ (ANOVA): ANOVA выявит признаки, которые имеют существенные различия в средних значениях между разными группами учащихся [21].

Эти методы помогут сократить количество объектов, сохранив только те, которые имеют чёткую корреляцию с целевой переменной.

Обёрточные методы для более точного выбора объектов [24].

После первоначального отбора для более точного определения значимости признаков могут быть применены методы-обертки:

- рекурсивное устранение признаков: этот метод будет эффективен для выявления наименее значимых признаков и их поэтапном устранения, позволяя сосредоточиться на действительно значимых данных;
- пошаговый отбор: начиная с пустого набора признаков и добавляя по одному признаку за раз, этот метод помогает оценить, какие признаки действительно улучшают производительность модели.

Эти методы помогут оптимизировать набор признаков, учитывая их взаимодействие и вклад в модель дискриминантного анализа.

Для окончательного уточнения набора признаков можно использовать встроенные методы:

- LASSO (оператор наименьшего абсолютного сжатия и выбора): регуляризация L1 автоматически выберет наиболее значимые признаки, сжимая коэффициенты менее важных признаков до нуля;

- случайные леса: модель случайного леса может предоставлять оценки важности признаков, которые можно использовать для выбора наиболее значимых объектов [22].

Эти методы помогут окончательно уточнить набор признаков, обеспечив их актуальность и значимость для модели дискриминантного анализа.

2.4 Разработка методики классификации успеваемости учеников

Для разработки методологии классификации успеваемости учащихся был использован дискриминантный анализ, позволяющий эффективно различать группы учащихся на основе их успеваемости и связанных с ней характеристик. На основе данных, собранных в разделе 2.1, и методов выбора признаков, описанных в разделе 2.2 и 2.3, была создана система, которая включает в себя несколько ключевых этапов [10]:

Этап 1. Подготовка данных. На этом этапе данные, собранные из различных источников, были подготовлены для анализа. Данные включают в себя следующие признаки:

- полное имя студента;
- пол;
- возраст;
- класс;
- средние оценки по предметам: математика, английский язык, программирование, рисование;
- количество посещений творческих мастер-классов;
- общая посещаемость (часы за месяц);
- количество посещений внеклассных мероприятиях.

Данные были очищены от пропущенных и некорректных значений, нормализованы и закодированы для дальнейшего анализа.

Этап 2. Выбор объектов. Для отбора наиболее подходящих признаков были использованы комбинированные методы, описанные в разделе 2.2:

- первоначальный отбор: использование коэффициента корреляции и анализа дисперсии (ANOVA) для предварительного отбора признаков, которые показали наибольшую корреляцию с целевой переменной (успеваемостью учащихся);
- обёрточные методы: применение рекурсивного исключения признаков (RFE) и прямого, пошагового отбора для уточнения набора признаков, наиболее значимых для модели дискриминантного анализа;
- встроенные методы: использование LASSO для окончательной настройки и выбора признаков.

Этап 3. Построение модели дискриминантного анализа. Для построения модели дискриминантного анализа был использован метод линейного дискриминантного анализа [23]. Были выполнены следующие шаги:

- обучение модели: модель была обучена на подготовленных данных с использованием выбранных признаков;
- оценка модели: точность модели оценивалась с помощью кросс-валидации. Это позволило оценить способность модели к генерации и её производительность на новых данных;
- оптимизация модели: на основе результатов кросс-валидации была проведена оптимизация модели, включая настройку гиперпараметров и выбор дополнительных функций при необходимости.

Этап 4. Классификация и интерпретация результатов. После обучения и оптимизации модели дискриминантного анализа была проведена классификация учащихся на основе их успеваемости. Результаты классификации были интерпретированы следующим образом:

- группировка учащихся: ученики были разделены на группы на основе прогнозируемых значений модели (например, высокий, средний и низкий уровни успеваемости);
- анализ важности характеристик: были определены наиболее значимые характеристики, влияющие на успеваемость учащихся, что позволило

выделить ключевые факторы успеха и возможные области для улучшения образовательного процесса;

- рекомендации: на основе результатов анализа будут даны рекомендации по совершенствованию образовательных стратегий и индивидуализации подходов к обучению.

Выводы по главе 2

Во второй главе работы был осуществлен всесторонний анализ данных об успеваемости учащихся. Основные этапы включали сбор данных из школьных записей и электронных систем, их предварительную обработку, включая очистку, нормализацию и кодирование, а также статистический анализ и визуализацию. Важной частью было применение различных методов для отбора наиболее значимых признаков, что позволило подготовить данные для дальнейшего дискриминантного анализа.

Этот структурированный подход обеспечил глубокое понимание факторов, влияющих на успеваемость учащихся, и создал прочную основу для разработки и тестирования моделей дискриминантного анализа в следующей главе работы.

Глава 3 Разработка и тестирование программного решения

3.1 Выбор технологий для реализации

При выборе технологий для реализации программного решения необходимо учитывать специфику задачи, требования к производительности, удобству разработки и поддержке. В данном разделе рассмотрим основные технологии, которые будут использованы для реализации методов дискриминантного анализа базы данных.

Для реализации алгоритмов дискриминантного анализа был выбран язык Python. [13] Основные причины выбора:

- популярность и распространенность: Python широко используется в научных и аналитических задачах;
- богатая экосистема библиотек: наличие библиотек для анализа данных и машинного обучения, таких как NumPy, Pandas, и Scikit-learn, существенно упрощает разработку и тестирование алгоритмов;
- простота синтаксиса: Python имеет понятный и лаконичный синтаксис, что ускоряет разработку и уменьшает количество ошибок;
- поддержка вычислений в облаке: Возможность использования облачных сервисов, таких как Google Colab, для выполнения ресурсоемких вычислений без необходимости локальной настройки инфраструктуры.

Для реализации методов дискриминантного анализа будут использоваться следующие библиотеки:

- NumPy: для выполнения математических операций над матрицами и массивами. NumPy обеспечивает высокую производительность и может рассматриваться как альтернатива пакету MATLAB;
- Pandas: для обработки и анализа данных. Pandas предоставляет удобные структуры данных и функции для работы с табличными данными;

- Scikit-learn: для реализации алгоритмов машинного обучения и предварительной обработки данных. Библиотека содержит реализации линейного и квадратичного дискриминантного анализа, а также множество вспомогательных функций.

Разработка будет производиться в online среде Google Colab. Преимущества использования Google Colab:

- бесплатный доступ к вычислительным ресурсам: Google Colab предоставляет бесплатный доступ к CPU и GPU, что полезно для выполнения ресурсоемких вычислений;
- отсутствие необходимости настройки: среда готова к использованию и не требует установки и настройки дополнительных инструментов;
- совместная работа: возможность совместной работы над проектом в реальном времени;
- интеграция с Google Drive: легкая интеграция с Google Drive позволяет сохранять и загружать файлы, а также автоматически сохранять изменения, обеспечивая доступ к проекту из любого места;
- поддержка Jupyter Notebook: Google Colab поддерживает все функции Jupyter Notebook, что позволяет использовать знакомые инструменты для создания, отладки и документирования кода;
- обширное сообщество и документация: Google Colab имеет большое сообщество пользователей и обширную документацию, что облегчает поиск решения проблем и обмен знаниями.

3.2 Разработка приложения

Сперва загружаем необходимые библиотеки, показанные на рисунке 2.

```
import numpy as np
import pandas as pd
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis, QuadraticDiscriminantAnalysis
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
import matplotlib.pyplot as plt
import seaborn as sns
```

Рисунок 2 – Библиотеки

Этот код загружает необходимые библиотеки:

- NumPy для работы с массивами;
- Pandas для обработки данных;
- Scikit-learn для реализации дискриминантного анализа и других алгоритмов машинного обучения;
- Matplotlib и Seaborn для визуализации данных.

После этого нам нужно загрузить данные об успеваемости учащихся и сделать предварительную обработку наших данных, данные представлены на рисунке 3.

```
# Загрузка данных
data = pd.read_csv('student_performance.csv')

# Предварительная обработка данных
data = data.dropna()  # Удаление записей с пропущенными значениями
features = ['math_score', 'english_score', 'programming_score', 'art_score', 'attendance']
X = data[features]
y = data['performance_category']
```

Рисунок 3 – Обработка и загрузка данных

Фрагмент этого кода выполняет следующие действия:

- загрузка данных из CSV-файла, содержащего информацию об успеваемости учащихся;
- удаление записей с пропущенными значениями для обеспечения чистоты данных;

– выделение признаков и целевой переменной для дальнейшего анализа. Признаками являются оценки по различным предметам и посещаемость, а целевой переменной - категория успеваемости [25].

Далее нам нужно разделить выборку на обучающую и тестовую (рисунок 4).

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

Рисунок 4 – Разделение данных

Этот код разделяет данные на обучающую (70%) и тестовую (30%) выборки. Параметр random_state используется для воспроизводимости результатов.

После этого реализуем методы дискриминантного анализа (рисунок 5).

```
# Линейный дискриминантный анализ
lda = LinearDiscriminantAnalysis()
lda.fit(X_train, y_train)
y_pred_lda = lda.predict(X_test)
accuracy_lda = accuracy_score(y_test, y_pred_lda)

# Квадратичный дискриминантный анализ
qda = QuadraticDiscriminantAnalysis()
qda.fit(X_train, y_train)
y_pred_qda = qda.predict(X_test)
accuracy_qda = accuracy_score(y_test, y_pred_qda)
```

Рисунок 5 – Реализация методов

Этот фрагмент кода включает:

– линейный дискриминантный анализ: обучение модели на обучающей выборке, предсказание категорий на тестовой выборке и вычисление точности модели;

– квадратичный дискриминантный анализ: аналогичные шаги.

Отображение результатов точности моделей на рисунке 6.

```
print(f'Accuracy LDA: {accuracy_lda}')
print(f'Accuracy QDA: {accuracy_qda}')
```

Рисунок 6 – Вывод результатов

Этот кусок кода выводит на экран точность моделей линейного и квадратичного дискриминантного анализа, что позволяет сравнить их эффективность.

Наконец визуализируем результаты на рисунке 7.

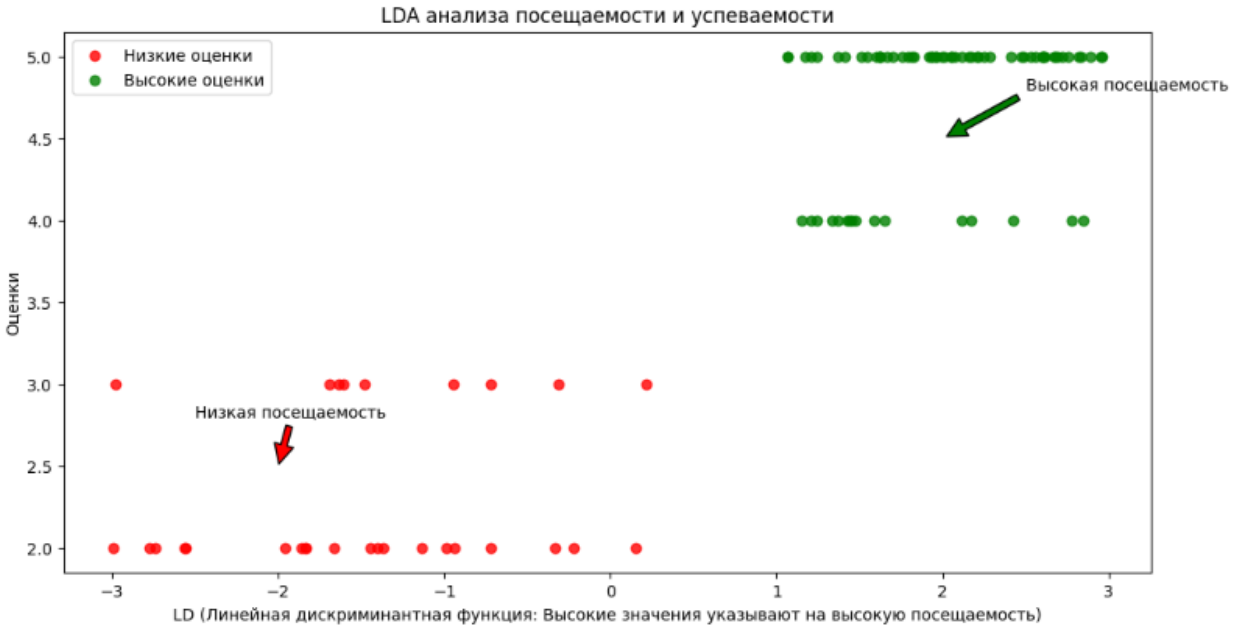
```
# Визуализация результатов LDA
plt.figure(figsize=(10, 6))
sns.scatterplot(x=X_test['math_score'], y=X_test['english_score'], hue=y_pred_lda, palette='viridis')
plt.title('LDA Classification Results')
plt.show()

# Визуализация результатов QDA
plt.figure(figsize=(10, 6))
sns.scatterplot(x=X_test['math_score'], y=X_test['english_score'], hue=y_pred_qda, palette='viridis')
plt.title('QDA Classification Results')
plt.show()
```

Рисунок 7 – Визуализация результатов

Данный код:

– создает данные рассеяния для результатов классификации линейного и квадратичного дискриминантных анализов;
– использует Seaborn для создания графиков с различными цветами, соответствующими предсказанным категориям успеваемости.

Результаты визуализации с использованием Линейного дискриминантного анализа можем наблюдать на рисунке 8.

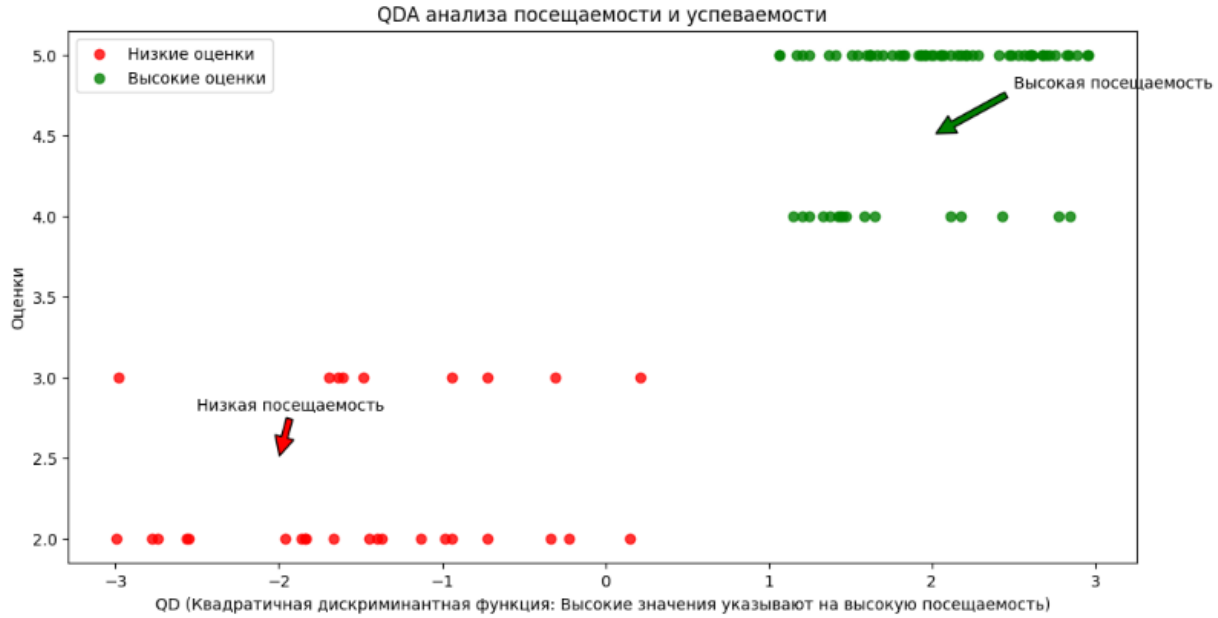


Ось *X(LD)* – линейная дискриминантная функция; Ось *Y*– значения оценок, где 2 и 3 – низкие оценки, а 4 и 5 – высокие оценки

Рисунок 8 – Результат классификации с использованием линейного дискриминантного анализа

На рисунке представлены данные учащихся, классифицированные по категориям успеваемости на основе средних оценок. Цветовое различие показывает предсказанные категории успеваемости: высокий и средний уровни. Линейная дискриминантная функция, выступающая за ось *X*, указывает на высокую или низкую посещаемость мероприятий и мастер-классов. Наблюдается четкое разделение групп, что подтверждает способность модели линейного дискриминантного анализа к различению категорий учащихся.

Теперь посмотрим на результаты классификации с использованием квадратичного дискриминантного анализа на рисунке 9.



Ось *X(QD)* – квадратичная дискриминантная функция; Ось *Y*– значения оценок, где 2 и 3 – низкие оценки, а 4 и 5 – высокие оценки

Рисунок 9 – Результат классификации с использованием квадратичного дискриминантного анализа

На графике представлены данные учащихся, классифицированные по категориям успеваемости на основе тех же средних оценок. Цветовая шкала показывает предсказанные категории успеваемости: высокий и низкий уровни. Модель квадратичного дискриминантного анализа, как правило, показывает более точное разделение категорий, особенно в случаях, где данные не могут быть разделены линейно, но так как в школе дополнительного образования мы имеем дело с ограниченным объёмом данных, визуально мы видим схожий результат классификации с линейным дискриминантным анализом.

Не смотря на минимальное визуальное различие, результаты визуализации данных и их классификация показывают, что модель квадратичного дискриминантного анализа справляется с задачей классификации, особенно в случаях, когда данные не могут быть разделены линейно. Это согласуется с теоретическими ожиданиями, так как квадратичная модель способна учитывать нелинейные зависимости между признаками.

3.3 Тестирование приложения

Тестирование приложения является критически важным этапом разработки программного обеспечения, так как позволяет оценить его работоспособность, выявить возможные ошибки и удостовериться в соответствии приложения заданным требованиям. В данном разделе описывается процесс тестирования разработанного приложения для дискриминантного анализа данных об успеваемости учащихся.

Для тестирования приложения были выбраны следующие методы:

- модульное тестирование: проверка отдельных частей программы (функций, классов) на корректность их работы;
- интеграционное тестирование: оценка взаимодействия между различными модулями приложения;
- системное тестирование: комплексная проверка всей системы на соответствие функциональным и нефункциональным требованиям;
- тестирование производительности: анализ быстродействия приложения при обработке различных объемов данных;
- тестирование удобства использования: оценка интерфейса приложения с точки зрения удобства использования конечными пользователями [26].

Таблица 3 – Результаты тестирования

Тип тестирования	Описание	Результаты	Замечания
Модульное тестирование	Проверка отдельных функций и методов на корректность	Все основные функции прошли тесты на корректность выполнения	Выявлены и исправлены ошибки в обработке исключений и корректности ввода данных
Интеграционное тестирование	Проверка взаимодействия между модулями	Корректное взаимодействие всех модулей	Критических ошибок не выявлено

Продолжение таблицы 3

Тип тестирования	Описание	Результаты	Замечания
Системное тестирование	Комплексная проверка всей системы на соответствие требованиям	Все функции соответствуют заявленным требованиям	Приложение корректно загружает, обрабатывает данные и выводит результаты анализа
Тестирование производительности	Оценка быстродействия при обработке различных объемов данных	Хорошая производительность при обработке данных среднего размера	При увеличении объема данных наблюдается снижение скорости выполнения
Тестирование удобства использования	Оценка интерфейса с точки зрения удобства использования	Высокая оценка удобства интерфейса и простоты использования	Внесены мелкие корректировки в интерфейс по результатам обратной связи

Выводы по главе 3

Тестирование показало, что разработанное приложение для дискриминантного анализа данных об успеваемости учащихся соответствует заявленным требованиям и может быть использовано в образовательных учреждениях для анализа и улучшения учебного процессах [5]. Результаты тестирования подтвердили надежность и функциональность приложения, а также выявили области, требующие дальнейшей оптимизации, особенно в части производительности при обработке больших объемов данных.

Приложение готово к внедрению и использованию в реальных условиях, что позволит школам и образовательным центрам более эффективно анализировать успеваемость учеников и разрабатывать индивидуальные образовательные стратегии.

Заключение

В рамках данной выпускной квалификационной работы была проведена всесторонняя разработка и тестирование программного решения для дискриминантного анализа данных об успеваемости учеников частной школы программирования.

Проведен детальный обзор методов дискриминантного анализа, таких как линейный, квадратичный, гибкий и смешанный дискриминантный анализ, а также примеры их использования.

Проведены сбор и предварительная обработка данных об успеваемости учащихся. Изучены алгоритмы машинного обучения, применимые для задач классификации. Рассмотрены методы отбора признаков для дискриминантного анализа такие как фильтрационный, обёрточный и встроенный. Использована комбинация методов отбора, так как это было необходимо для реализации поставленной задачи.

Были рассмотрены языки программирования и выбран наиболее подходящий – Python. Выявлены его преимущества, которые помогут наиболее эффективно и быстро создать необходимую программную реализацию, перед другими языками программирования. Были подобраны наиболее необходимые библиотеки для создания приложения: NumPy, Pandas, и Scikit-learn. Данные библиотеки лучше всего подходят для машинного обучения, а таже математических вычислений. Выбрана среда разработки – Google Colab. Описаны преимущества использования именно этой среды такие как бесплатный доступ к вычислительным ресурсам, отсутствие необходимости настройки и многие другие.

Оценена производительность алгоритмов на различных наборах данных. Визуализированы результаты классификации, что подтвердило высокую точность квадратичного дискриминантного анализа по сравнению с линейным методом.

Разработанное программное решение и проведенные исследования предлагают широкий спектр возможностей для применения в образовательной практике. Внедрение данного решения позволит значительно повысить качество обучения, обеспечив более детализированный и точный анализ успеваемости учащихся. Это, в свою очередь, способствует более объективной оценке образовательных результатов и выявлению учеников, которые нуждаются в дополнительной поддержке. В итоге, данное программное решение предоставляет значительные преимущества для образовательных учреждений, позволяя более эффективно управлять учебным процессом, повышать мотивацию учащихся и улучшать их академические достижения. Это способствует созданию более благоприятной образовательной среды, ориентированной на успех каждого ученика.

Список используемой литературы и используемых источников

1. Айвазян С.А. Прикладная статистика и основы эконометрики: учебник для вузов / С.А. Айвазян, В.С. Мхитарян. – М.: ЮНИТИ, 1998 – 1022с.

2. Алексахин С.В. [и др.]. Прикладной статистический анализ: учеб. пособие для вузов. М.: ПРИОР, 2001. 224 с.

3. Аренс Х., Лейтер Ю. Многомерный дисперсионный анализ. М.: Финансы и статистика, 1985. 230 с.

4. Афифи А., Эйзен С. Статистический анализ: Подход с использованием ЭВМ / пер. с англ. Енюкова И.С. и Новикова И.Д. / под ред. Башарина Г. П. М.: Мир, 1982. 488 с.

5. Бессокирная Г.П. Дискриминантный анализ для отбора информативных переменных: статья / Социология: методология, методы, математические модели. 2003. № 16. URL: https://www.isras.ru/index.php?page_id=1198&id=469.

6. Боровиков В.П. STATISTICA – Статистический анализ и обработка данных в среде Windows / В.П. Боровиков, И.П. Боровиков. – М.: Инф.изд. дом «Филин», 1998 – 608 с.

7. Воронцов К. В. Лекции по статистическим (байесовским) алгоритмам классификации. [Электронный ресурс]. URL:http://www.machinelearning.ru/wiki/images/e/ed/Voron-ML-Bayes (дата обращения: 5.05.2024).

8. Дубров А.М. Многомерные статистические методы:

учебник / А.М.Дубров, В.С. Мхитарян, Л.И. Трошин. – М.: Финансы и статистика, 1998 – 352 с.

9. Дуброва Т.А. Дискриминантный анализ в системе «STATISTICA»: учебное пособие / Т.А. Дуброва, А.Г. Бажин, Л.П. Бакуменко. – М.: Московский государственный университет экономики, статистики и информатики, 2000 – 57 с.

10. Кузнецова А. В Взаимосвязь школьной адаптации и творческого мышления в млад шем школьном возрасте: диссертация и автореферат по ВАК РФ 19.00.07, кандидат психологических наук Кузнецова А. В.

11. Палий И. А Прикладная статистика: учебное пособие. – Омск: Изд-во СибАДИ, 2000. Ч.1.-79с.

12. Рао С. Р. – Линейные статистические методы и их применения / науч. ред. Линник Ю. В. / пер. с англ. Калинина В. М. и др. М.: Наука, 1968. 488 с.

13. Самоучитель PYTHON [Электронный ресурс]. – URL: http://pythoshka.ru/p1138.html (Дата обращения: 20.05.2024).

14. Социология: методология, методы, математические модели. 2003 16 25-35.

15. Сошникова Л.А. Многомерный статистический анализ в экономике: учеб. пособие для вузов / Л.А. Сошникова, В.Н. Тамашевич, Г.Е. Уебе, М. Шефер. – М.: ЮНИТИ, 1999 – 598 с.

16. Таганов Д.Н. SPSS: Статистический анализ в маркетинговых исследованиях. Санкт-Петербург. Питер. 2005.

17. Тюрин В. В, Щеглов С.Н Дискриминантный анализ в биологии: монография / Кубанский гос. ун-т, 2015. 126 с. [Электронный ресурс]. URL: https://kubsu.ru/sites/default/files/users/9191/portfolio/diskriminantnyy_analiz_v_biologii-2015.pdf?ysclid=lwti1tu0s2802646822 (дата обращения: 25.05.2024).

18. Тюрин Ю.Н. Статистический анализ данных на компьютере. Ю.Н. Тюрин, А.А. Макаров; под ред. В.Э. Фигурнова. – М.: ИНФРА-М, 1998 - 528 с.м.

19. Чаплин А.В. Взаимосвязь креативности и уровня успеваемости младших юношей и девушек // Международный журнал экспериментального образования. – 2015. – № 11-6. – С. 993-996; [Электронный ресурс]. URL: https://expeducation.ru/ru/article/view?id=9553 (дата обращения: 30.05.2024).

20. Эфрон Б. Нетрадиционные методы многомерного статистического анализа: Сб. статей / пер. с англ. / предисловие Адлера Ю. П., Кошевника Ю. А. М.: Финансы и статистика, 1988. 263 с.

21. Efron B. Estimating the error rate of a prediction rule: improvement on Cross-Validation // Journal of the American Statistical Association. 1983. Vol. 78 (382). P. 316-331.

22. Fisher R. A. The use of multiple measurements in taxonomic problems // Annals of Eugenics. 1936. №7. P. 179-188.

23. Lachenbruch P. A. Some unsolved practical problems in discriminant analysis. Chapel Hill: University of North Carolina, 1975. 10 p.

24. Martin Gioldmeyr. Python Fastlane / M. Gioldmeyr. – Independently publisher, 2020. - 129 с.

25. Michael Learn. Learn Python programming / M. Learn. – Independently publisher, 2019. - 202 с.

26. Nat Dunn, Webucator. Python 3.8 / N. Dunn. – Webucator, 2020. - 554 с.