



МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«МИРЭА – Российский технологический университет»
РТУ МИРЭА

Институт информационных технологий (ИИТ)
Кафедра прикладной математики (ПМ)

КУРСОВАЯ РАБОТА

по дисциплине

«Языки программирования для статистической обработки данных»

Тема курсовой работы: «Разработка программы классификации на базе алгоритмов метода опорных векторов по категорированию жанров музыки»

Студент группы ИМБО-11-23 Журавлев Фёдор Андреевич


(подпись)

Руководитель
курсовой работы

к.ф.-м.н., доцент каф. ПМ Царькова
Е.Г.


(подпись)

Работа представлена к защите «__»_____2025 г.


Допущен к защите «__»_____2025 г.

Москва 2025 г.



МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«МИРЭА – Российский технологический университет»
РТУ МИРЭА

Институт информационных технологий (ИИТ)
Кафедра прикладной математики (ПМ)

Утверждаю
Заведующий кафедрой ПМ

(подпись) Смоленцева Т.Е.
«22» февраля 2025 г.

ЗАДАНИЕ

на выполнение курсовой работы

по дисциплине «Языки программирования для статистической обработки данных»

Студент Журавлев Фёдор Андреевич

Группа ИМБО-11-23

Тема «Разработка программы классификации на базе алгоритмов метода опорных векторов по категорированию жанров музыки»

Исходные данные: собранный студентом набор данных по теме работы

Перечень вопросов, подлежащих разработке, и обязательного графического материала:

Характеристика изучаемой предметной области, алгоритма, набора данных (описание текущего состояния исследуемой предметной области, выделение перспективных направлений исследований, применимость алгоритмов анализа и обработки данных, описание полей набора данных)

Математическая формулировка предлагаемого метода анализа и обработки данных (классическая постановка задачи, формулировка задачи статистической обработки данных, описание параметров, описание критерия качества решения конечной задачи)

Анализ полученной выборки данных с использованием предложенных методов анализа и обработки данных (описание последовательности действий или сценария обработки данных)

Построение визуализаций и качественных выводов по проделанной работе

Срок представления к защите курсовой работы:

до «23» мая 2025 г.

Задание на курсовую работу выдал

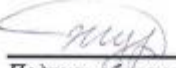

Подпись руководителя

Царькова Е.Г.

(ФИО руководителя)

«22» февраля 2025 г.

Задание на курсовую работу получил


Подпись обучающегося

Журавлев Ф.А.

(ФИО обучающегося)

«22» февраля 2025 г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ.....	4
1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ.....	6
1.1. Описание предметной области	6
1.1.1. Значимость анализа данных в области музыкального бизнеса	6
1.1.2. Основные типы данных (числовые, категориальные, временные ряды)	7
1.1.3. Ключевые задачи статистического анализа (прогнозирование, выявление закономерностей, оптимизация процессов)	8
1.2. Основные методы статистической обработки данных	8
1.2.1. Описательная статистика (среднее, медиана, мода, дисперсия)	8
1.2.2. Методы проверки гипотез (t-тест, ANOVA, хи-квадрат)	9
1.2.3. Корреляционный анализ	10
1.2.4. Классификация и метод опорных векторов (SVM)	10
1.3. Преимущества использования языка R для анализа данных	13
2. ПРАКТИЧЕСКАЯ ЧАСТЬ	14
2.1. Описание источников данных	14
2.1.1. Источники данных	14
2.1.2. Структура и характеристики данных	14
2.1.3. Предварительная обработка данных (заполнение пропусков, фильтрация, преобразование)	14
2.5. Визуализация данных	15
2.5.1. Визуализация в R (ggplot2, plotly)	15
2.5.2. Интерактивные дашборды в Glarus BI	17
ЗАКЛЮЧЕНИЕ	19
ПРИЛОЖЕНИЕ	20

ВВЕДЕНИЕ

Музыка является неотъемлемой частью культуры и играет важную роль в нашей жизни. С развитием цифровых технологий и стриминговых сервисов объемы доступной музыки растут многократно. Это создает потребность в эффективных методах автоматической организации и анализа музыкальных данных. Одним из ключевых аспектов является классификация музыкальных произведений по жанрам.

Автоматическая классификация музыкальных жанров – это задача машинного обучения, заключающаяся в определении жанра музыкального произведения на основе его аудио характеристик. Традиционные методы ручной классификации являются трудоемкими, субъективными и не масштабируемыми для обработки больших объемов данных. В связи с этим разработка эффективных и точных автоматизированных методов классификации музыкальных жанров представляет собой актуальную и важную задачу.

Настоящая курсовая работа посвящена разработке программы классификации музыкальных произведений по жанрам с использованием алгоритмов метода опорных векторов (SVM). Метод опорных векторов является мощным инструментом машинного обучения, который обладает высокой обобщающей способностью и эффективен в задачах классификации, в том числе для нелинейно разделимых данных.

Цель данной работы - разработка программы классификации музыкальных произведений по жанрам на основе алгоритмов метода опорных векторов (SVM), способной автоматически определять жанр музыкального произведения по его аудио характеристикам.

Для достижения поставленной цели необходимо решить следующие задачи:

- Изучить предметную область и провести анализ существующих методов классификации музыкальных жанров, выявив их достоинства и недостатки.
- Выполнить предварительную обработку данных, получив общее представление о распределении признаков, а также выявить возможные

зависимости, аномалии и особенности структуры выборки.

- Провести обзор алгоритмов метода опорных векторов (SVM), изучив теоретические основы, различные типы ядер, методы оптимизации и параметры регуляризации.
- Разработать модель классификации, определив набор аудио характеристик (признаков), используемых для классификации, разработав архитектуру модели SVM и обучив ее на размеченном наборе данных.
- Реализовать программное обеспечение, реализующее разработанную модель классификации музыкальных жанров на языке R.
- Оценить эффективность разработанной программы, проведя тестирование на контрольном наборе данных и оценив ее точность, полноту и другие метрики качества.
- Визуализировать данные с помощью графики, диаграммы, дашборды, используя язык программирования R и возможности аналитической системы Glarus BI.

Объектом исследования являются синтетические данные музыкальных характеристик, описывающих уровни шума, ритма и других параметров для жанров музыки (рок, хип-хоп и поп). Предметом - метод опорных векторов (SVM) и его применение для классификации музыкальных жанров.

Практическая реализация выполняется в программной среде R с использованием различных библиотеки и инструментов для анализа и визуализации данных. Также уделяется внимание автоматизации анализа и формированию отчётов с помощью R Markdown, что позволяет создать удобный и расширяемый аналитический процесс.

1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1. Описание предметной области

1.1.1. Значимость анализа данных в области музыкального бизнеса

В современном цифровом мире, где объем музыкального контента растет экспоненциально, классификация музыки играет критически важную роль. Она позволяет эффективно организовывать, искать и рекомендовать музыку, формируя пользовательский опыт и принося значительную выгоду различным участникам рынка.

Методы машинного обучения используются различными компаниями в сфере музыки. Стриминговые сервисы полагаются на классификацию музыки для организации огромных библиотек, улучшения поиска и, самое главное, для создания персонализированных рекомендаций и плейлистов. Чем точнее и релевантнее рекомендации, тем выше вовлеченность пользователей, дольше время, проведенное на платформе, и, как следствие, выше выручка от подписок и рекламы.

Также классификация музыкального контента помогает решать задачи, такие как:

- **повышение удержания пользователей:** персонализированные плейлисты, допустим, напрямую влияют на удержание пользователей, снижая отток и стимулируя продление подписок;
- **оптимизация лицензионных соглашений:** классификация музыки позволяет более точно определять жанры и стили, что важно при заключении лицензионных соглашений с правообладателями;
- **монетизация контента:** для независимых музыкантов и лейблов классификация предоставляет возможность эффективнее монетизировать свою музыку. Наиболее продвинутая модель классификации позволяет их трекам быть обнаруженными целевой аудиторией, попадать в релевантные плейлисты и, как следствие, генерировать больше прослушиваний и доходов.

- экономия ресурсов: модель классификации значительно экономит время и ресурсы по сравнению с ручной категоризацией. Это особенно важно для платформ с огромными музыкальными библиотеками, где ручная классификация получилась бы значительно дороже.

1.1.2. Основные типы данных (числовые, категориальные, временные ряды)

Типы данных в контексте музыкальных данных:

- **Аудиоданные (временные ряды)** представляют собой изменение амплитуды звука во времени. Это сырой, необработанный формат. Однако в данной курсовой работе не используется).

- **Числовые данные (аудио признаки)** извлекаются из аудио данных с помощью различных алгоритмов. В данной работе применяются следующие характеристики:

- Темп (BPM - Количество ударов в минуту) отражает скорость музыки.
- RMS Energy (Root Mean Square - Среднеквадратичная энергия) показывает общую громкость сигнала. Анализ должен определять среднее значение RMS Energy для трека и учитывать его влияние на восприятие музыки. Например, треки с высокой RMS Energy часто кажутся более энергичными и агрессивными, а треки с низкой RMS Energy - более тихими и спокойными.

- Zero Crossing Rate (Сигнальный шум) демонстрирует частоту, с которой сигнал пересекает нулевую линию. Более высокие значения ZCR обычно указывают на более "шумный" или высокочастотный сигнал, который может быть связан с определенными инструментами (например, перкуссия) или жанрами (например, электронная музыка).

- Инструментальность показывает вероятность того, что в треке отсутствует вокал. Значение 1.0 означает, что трек полностью инструментальный, а значение 0.0 - что в треке преобладает вокал.

- Фактически, параметр "Вокал" обратно пропорционален

"Инструментальности". Анализ может быть таким же, как и у "Инструментальности", или можно использовать этот параметр для уточнения классификации, если есть сомнения между жанрами, где вокал является определяющим (например, поп против инструментального хип-хопа).

- **Категориальные данные (жанры):** Метки, присваиваемые музыкальным произведениям (рок, поп, хип-хоп). Преобразуются в кодировку от 0 до 2 для определения класса прогнозирования.

1.1.3. Ключевые задачи статистического анализа (прогнозирование, выявление закономерностей, оптимизация процессов)

К основным задачам статистического анализа в анализе музыкальных данных относятся:

- **Прогнозирование:** в данном контексте – прогнозирование жанра музыкального произведения на основе его аудио признаков. SVM (Support vector machine), как алгоритм машинного обучения, является прогностическим инструментом.
- **Выявление закономерностей:** цель - найти статистические закономерности, связывающие аудио признаки и жанры. Это позволит понять, какие характеристики звука типичны для каждого жанра.
- **Оптимизации извлечения признаков:** выбор наиболее информативных признаков и оптимизация параметров SVM: настройка параметров ядра и регуляризации для достижения максимальной точности классификации.

1.2. Основные методы статистической обработки данных

1.2.1. Описательная статистика (среднее, медиана, мода, дисперсия)

Описательная статистики – это методы, используемые для обобщения и представления данных в понятной форме.

Среднее (Mean): среднее значение аудио признака (например, среднего темпа всех песен в жанре рок). Позволяет оценить типичное значение признака для данного жанра.

Медиана (Median): значение, разделяющее упорядоченный набор данных пополам. Менее чувствительна к выбросам, чем среднее.

Мода (Mode): наиболее часто встречающееся значение в наборе данных. Показывает наиболее типичное значение признака.

Дисперсия (Variance): мера разброса значений вокруг среднего. Показывает, насколько сильно значения признака варьируются в данном жанре.

Стандартное отклонение (Standard Deviation): квадратный корень из дисперсии. Более интерпретируемая мера разброса, чем дисперсия, т. к. выражается в тех же единицах измерения, что и сам признак.

1.2.2. Методы проверки гипотез (t-тест, ANOVA, хи-квадрат)

Методы проверки гипотез – это статистические процедуры, используемые для проверки утверждений о генеральной совокупности на основе выборочных данных.

t-тест (Student's t-test) используется для сравнения средних значений двух групп. Например, для проверки гипотезы о том, что средний темп песен в жанре рок статистически значимо отличается от среднего темпа песен в жанре поп.

ANOVA (Analysis of Variance) применяется для сравнения средних значений трех и более групп. Например, для проверки гипотезы о том, что средний темп песен статистически значимо различается между жанрами рок, поп и джаз.

Хи-квадрат (Chi-square test) нужен для анализа категориальных данных. Например, для проверки гипотезы о том, что наличие определенного инструмента (например, гитары) статистически связано с определенным жанром.

1.2.3. Корреляционный анализ

В статистике корреляционный анализ определяется как методы, используемые для измерения степени линейной взаимосвязи между двумя или более переменными.

Например, может существовать корреляция между яркостью звука и уровнем энергии. Это может быть полезно при отборе признаков для классификации.

Он может быть полезен для выявления избыточных признаков (признаков, которые сильно коррелируют друг с другом). В этом случае следует исключить один из коррелирующих признаков, чтобы упростить модель и избежать мультиколлинеарности.

Корреляция является вспомогательным инструментом для последующего анализа, потому что может позволить понять какие существуют связи, или же может помочь определить, насколько приведенные данные качественные.

Перед построением моделей классификации важно исследовать взаимосвязи между признаками. Для этого применяются различные известные формулы для подсчета коэффициентов корреляции:

- **Коэффициент корреляции Пирсона** - измеряет линейную зависимость между количественными переменными.
- **Коэффициент корреляции Спирмена** - оценивает монотонную зависимость, устойчив к нелинейностям и выбросам.

Высокая корреляция между признаками может указывать на мультиколлинеарность, что важно учитывать при выборе модели.

1.2.4. Классификация и метод опорных векторов (SVM)

Классификация – это задача машинного обучения с учителем, которая в своей задаче относит объекты к одному из заранее определённых классов. В данном случае – к жанрам музыки.

Метод опорных векторов (SVM, Support Vector Machine) – это мощный алгоритм, который строит гиперплоскость, максимально разделяющую классы в признаковом пространстве. SVM эффективен в задачах с высокой размерностью и хорошо справляется с нелинейными границами решений за счёт использования ядерных функций (например, радиальной базисной, RBF).

1.2.5. Метрики оценки качества классификации

Для оценки моделей классификации в машинном обучении используются следующие метрики:

- **Precision (точность)** – доля верно предсказанных положительных классов среди всех предсказанных. Следует обращать внимание на нее, когда необходимо минимизировать количество ложноположительных результатов.
- **Accuracy** - доля всех правильно предсказанных классов (как положительных, так и отрицательных) среди всех классов. Это общая мера производительности модели. Accuracy полезна, когда классы сбалансированы, но может быть вводящей в заблуждение, если классы сильно несбалансированы.
- **Recall (полнота)** – доля верно предсказанных положительных классов среди всех истинных и можно использовать для понимания, когда необходимо минимизировать количество ложноотрицательных результатов.
- **F1-score** – гармоническое среднее precision и recall, балансирует между ними. Эта метрика используется, когда необходимо найти баланс между точностью и полнотой.
- **AUC-ROC (Area Under the ROC Curve)** – оценивает способность модели различать классы, учитывая соотношение True Positive Rate (TPR) и False Positive Rate (FPR). Чем ближе AUC к 1, тем лучше модель.

Эти метрики позволяют глубже понять производительность модели классификации и выбрать наиболее подходящую в зависимости от конкретной задачи. Важно учитывать, что выбор метрики зависит от контекста задачи и

требований к модели, поэтому рекомендуется использовать несколько метрик для комплексной оценки.

1.3. Преимущества использования языка R для анализа данных

Язык **R** является одним из ведущих инструментов в анализе данных и машинном обучении благодаря:

- Богатым библиотекам (caret, e1071, pROC, tidyverse) для обработки данных, визуализации и построения моделей.
- Удобной работе со статистическими методами и матричными операциями.
- Гибкости в создании пользовательских функций и скриптов.
- Поддержке современных методов машинного обучения и визуализации результатов (ggplot2, ROC-кривые).

Таким образом, применение R в данной работе позволяет эффективно провести анализ данных, построить модели классификации и оценить их качество.

2. ПРАКТИЧЕСКАЯ ЧАСТЬ

2.1. Описание источников данных

2.1.1. Источники данных

В данной работе используется синтетический набор данных, сгенерированный нейросетью, т. к. не удалось найти достойный набор данных в открытых источниках.

2.1.2. Структура и характеристики данных

Датасет состоит из 94 записей, каждая из которых соответствует отдельному музыкальному произведению. Структура данных включает 5 признаков и одну целевую переменную (Код Жанра). Признаки можно разделить следующим образом:

- Числовые признаки: BPM, RMS Energy, Zero Crossing Rate, Инструментальность, Вокал

- Категориальные признаки: Жанр.

Целевая переменная: Код Жанра, где 1 — рок, 2 — хип-хоп, 3 — поп.

Данные не включают временные ряды — каждая запись представляет собой статический набор музыкальных показателей.

2.1.3. Предварительная обработка данных (заполнение пропусков, фильтрация, преобразование)

Предварительная обработка данных включает следующие этапы:

- Обнаружение и удаление пропусков: используется `summary()`, `is.na()`, и затем `na.omit()` для удаления неполных строк (пропусков не обнаружено).

ОСТАЛЬНЫЕ ПУНКТЫ С EDA ДОДЕЛАТЬ

2.5. Визуализация данных

Визуализация играет ключевую роль в интерпретации результатов анализа данных. Графическое представление позволяет выявлять закономерности, объяснять модели и демонстрировать результаты широкому кругу специалистов, включая тех, кто не обладает навыками работы с кодом.

2.5.1. Визуализация в R (ggplot2, plotly)

Доделать

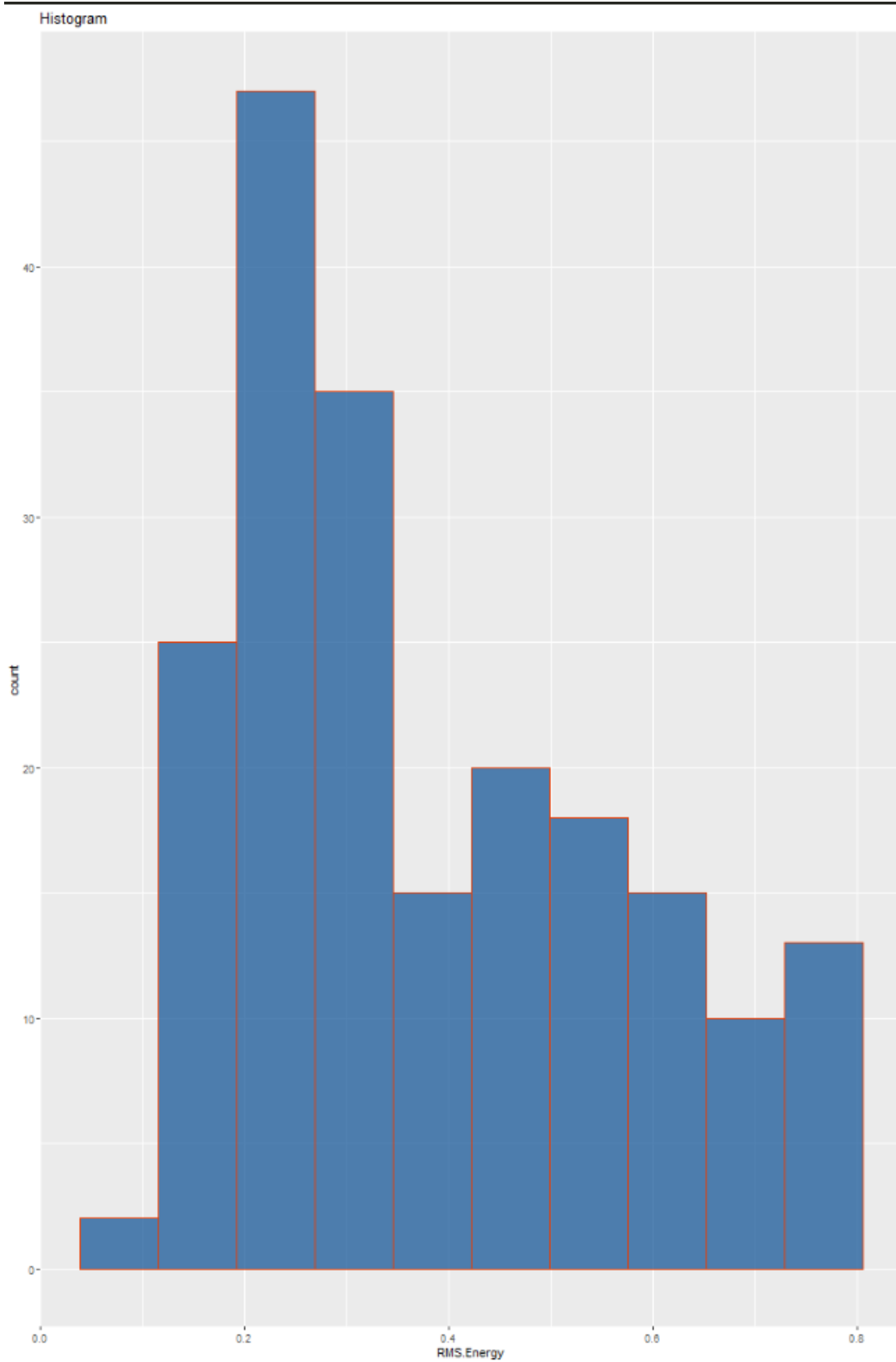


Рисунок – Гистограмма распределений RMS Energy

2.5.2. Интерактивные дашборды в Glarus BI

Для визуального представления результатов анализа и построения отчётов был использован инструмент Glarus BI — отечественная BI-платформа для создания интерактивных дашбордов. Благодаря простому графическому интерфейсу Glarus BI позволяет быстро загружать данные, строить визуализации и предоставлять интерактивные отчёты.

- Гистограмма распределений всех характеристик относительно параметра BPM;
- Гистограмма значений по характеристикам анализа музыки Zero Crossing Rate и BPM.

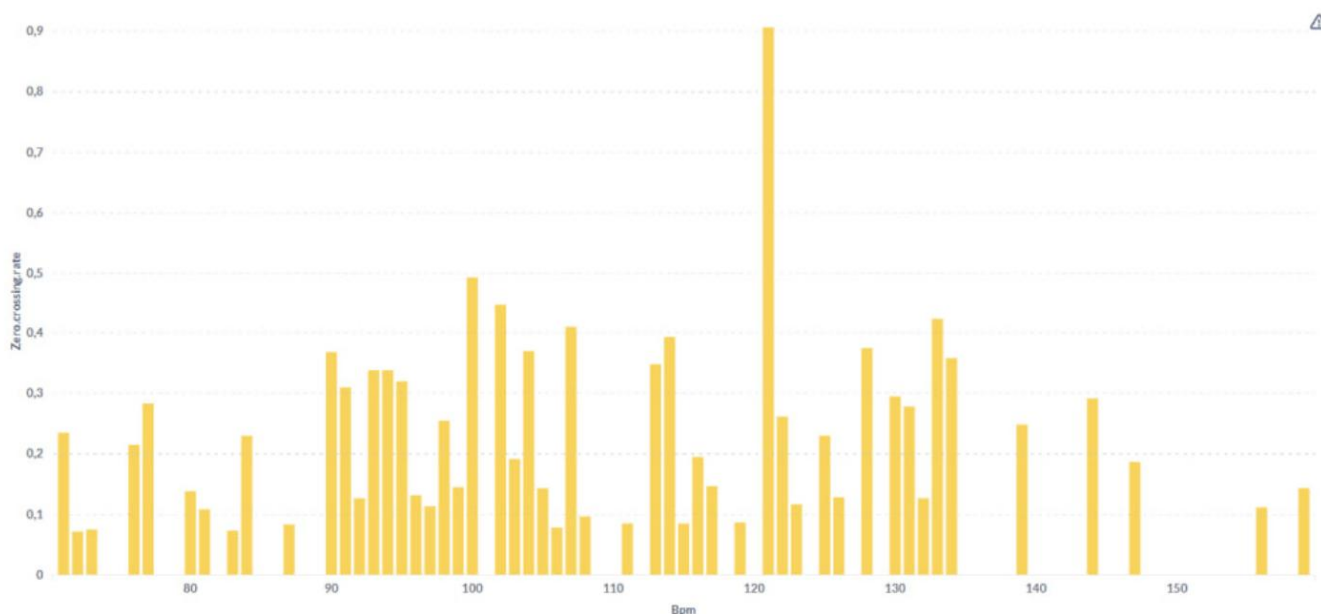


Рисунок — Дашборд Glarus BI

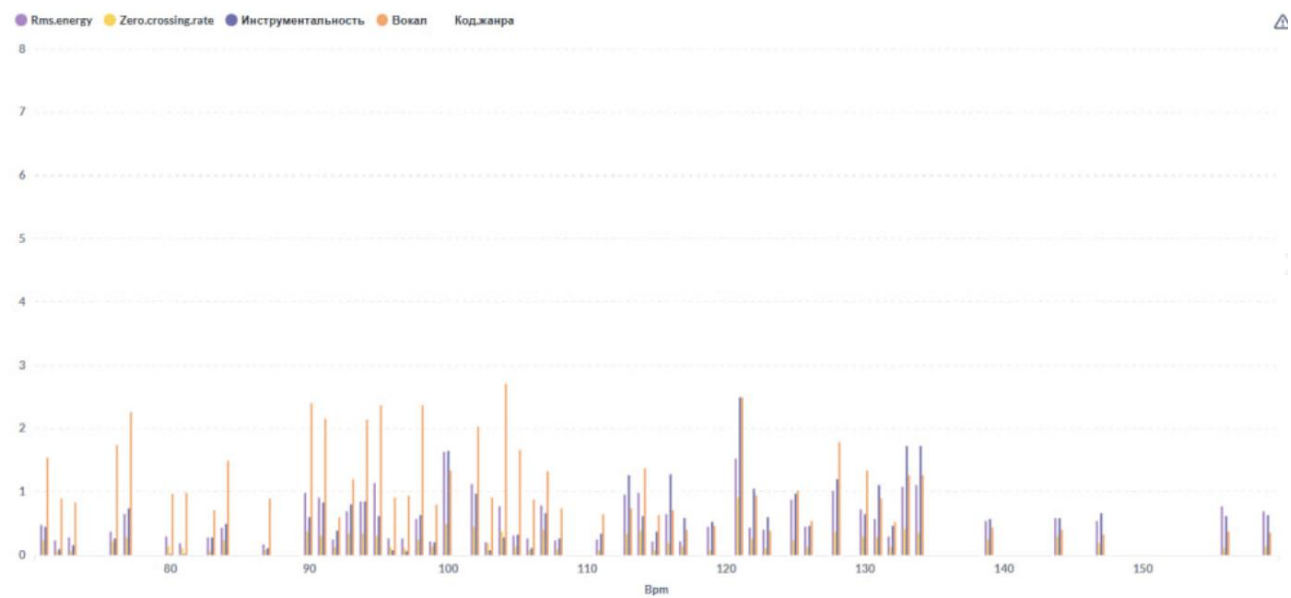


Рисунок — Дашборд Glarus VI

ЗАКЛЮЧЕНИЕ

Классификация музыкальных жанров с использованием метода опорных векторов (SVM) представляет собой мощный инструмент в области машинного обучения и анализа данных. Метод опорных векторов позволяет эффективно разделять данные на классы, находя оптимальную гиперплоскость, которая минимизирует ошибку классификации.

В ходе работы было показано, что SVM хорошо справляется с задачами классификации, особенно в условиях высокой размерности данных, что является характерным для аудио признаков. Использование различных аудио характеристик, таких как темп, уровень энергии, инструментальность и вокал, позволяет создать информативные векторы признаков, которые значительно повышают точность классификации.

Кроме того, метод опорных векторов обладает высокой устойчивостью к переобучению, что делает его особенно подходящим для работы с ограниченными объемами данных, как это часто бывает в музыкальных коллекциях. Применение SVM в сочетании с методами предварительной обработки данных, такими как нормализация и отбор признаков, позволяет добиться еще более высоких результатов.

В заключение классификация музыкальных жанров с использованием метода опорных векторов открывает новые горизонты для автоматизации музыкальных рекомендаций, создания плейлистов и улучшения пользовательского опыта в музыкальных сервисах. Дальнейшие исследования могут быть направлены на оптимизацию параметров модели и интеграцию SVM с другими методами машинного обучения для достижения еще более точных результатов.

ПРИЛОЖЕНИЕ

В приложении представлены фрагменты выполненного кода на языке R.

```
1 library(e1071)      You, 5 days ago • svm
2 library(caret)
3 library(rgl)
4
5 data_music <- read.csv("D:/Documents/Learning/3/R/ff.csv")
6
7 # 3. Разделение на обучающую и тестовую выборки
8 set.seed(123) # для воспроизводимости
9 n <- nrow(data_music)
10 train_indices <- sample(1:n, size = round(0.8 * n))
11 train <- data_music[train_indices, ]
12 test <- data_music[-train_indices, ]
13
14 # ПРЕОБРАЗУЕМ В ФАКТОРЫ ПОСЛЕ РАЗДЕЛЕНИЯ
15 train$Код.Жанра <- factor(train$Код.Жанра)
16 test$Код.Жанра <- factor(test$Код.Жанра, levels = levels(train$Код.Жанра)) # Важно сохранить уровни
17
18 # 4. Обучение модели SVM
19 svm_model <- svm(Код.Жанра ~ BPM + RMS.Energy + Zero.Crossing.Rate
20                 + Инструментальность + Вокал,
21                 data = train,
22                 kernel = "radial", # Радиальное ядро
23                 scale = TRUE,      # Масштабирование признаков
24                 probability = TRUE) # Для получения вероятностей
25
26 # Предсказание
27 predictions <- predict(svm_model, test)
28
29 # Преобразуем predictions в фактор (уровни берем из train)
30 predictions <- factor(predictions, levels = levels(train$Код.Жанра))
31
```

Рисунок — Код на R

```

32 # Матрица ошибок
33 conf_matrix <- confusionMatrix(predictions, test$Код.Жанра)
34 print("Матрица ошибок:")
35 print(conf_matrix)
36
37 # Извлечение Precision, Recall и F1-Score
38 precision <- conf_matrix$byClass[, "Precision"]
39 recall <- conf_matrix$byClass[, "Recall"]
40 f1_score <- conf_matrix$byClass[, "F1"]
41
42 # Вывод результатов
43 cat("\nPrecision:\n")
44 print(precision)
45 cat("\nRecall:\n")
46 print(recall)
47 cat("\nF1-Score:\n")
48 print(f1_score)
49
50 # Вывод средних значений
51 cat("\nСредний Precision:", mean(precision), "\n")
52 cat("Средний Recall:", mean(recall), "\n")
53 cat("Средний F1-Score:", mean(f1_score), "\n")

```

Рисунок — Код на R

```

1 library(ggplot2)
2 library(readr)
3 df <- read.csv("music_genre_dataset.csv")
4 p <- ggplot(df, aes(x = RMS.Energy)) +
5   geom_histogram(bins = 10, fill = '#094e92', color = '#e43f08', alpha = 0.7) +
6   labs(title = "Histogram")
7 print(p)

```

Рисунок — Код на R

```

1  library(readr)
2  ds <- read.csv("D:/Documents/Learning/3/R/ff.csv")
3  summary(ds$BPM)
4  library(modeest)
5  mode_value <- mfv(ds$BPM)
6  print(paste("Мода: ", mode_value))
7  variance <- var(ds$BPM)
8  std_dev <- sd(ds$BPM)
9  print(paste("Дисперсия", variance))
10 print(paste("Отклонение", std_dev))
11
12 ttest <- t.test(ds$`Код.Жанра` ~ ds$`BPM`)
13 print(ttest)
14
15 wtest <- wilcox.test(ds$`Код.Жанра` ~ ds$`BPM`)
16 print(wtest)
17
18 ctable <- table(ds$`Код.Жанра`, ds$`Вокал`)
19 ctable1 <- chisq.test(ctable)
20 print(ctable1)

```

Рисунок — Код на R