



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**

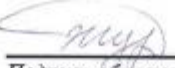
  
Подпись руководителя

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.

## **АННОТАЦИЯ**

Курсовая работа содержит 34 страницы, включает 16 рисунков, 2 таблицы и 1 приложение, представляющие код. В ходе работы были проведены проверки гипотез, описательная статистика, классификация методом опорных векторов, создание модели для прогнозирования принадлежности трека к жанру по различным аудио показателям. Были визуализированы все результаты, посчитаны различные метрики и сделаны выводы по курсовой работе. В ходе выполнения были использованы такие программы, как VSCode и Glarus BI.

Ключевые слова: машинное обучение, классификация, метод опорных векторов, музыка.

# СОДЕРЖАНИЕ

АННОТАЦИЯ .....	3
ВВЕДЕНИЕ.....	6
1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ.....	8
1.1. Описание предметной области .....	8
1.1.1. Значимость анализа данных в области музыкального бизнеса .....	8
1.1.2. Основные типы данных (числовые, категориальные, временные ряды) .....	9
1.1.3. Ключевые задачи статистического анализа (прогнозирование, выявление закономерностей, оптимизация процессов) .....	10
1.2. Основные методы статистической обработки данных .....	11
1.2.1. Описательная статистика (среднее, медиана, мода, дисперсия).....	11
1.2.2. Методы проверки гипотез (t-тест, ANOVA, хи-квадрат) .....	11
1.2.4. Классификация и метод опорных векторов (SVM).....	13
1.2.5. Метрики оценки качества классификации .....	13
2. ПРАКТИЧЕСКАЯ ЧАСТЬ .....	15
2.1. Описание источников данных .....	15
2.1.1. Источники данных .....	15
2.1.2. Структура и характеристики данных.....	15
2.1.3. Предварительная обработка данных .....	16
2.2. Исследовательский анализ данных .....	16
2.2.1. Визуализация распределений .....	16
2.2.2. Корреляционный анализ.....	18
2.2.3. Выявление выбросов и трендов.....	19
2.3. Применение методов статистического анализа .....	20
2.3.1. Описательная статистика .....	20
2.3.2. Проверка гипотез .....	20
2.4. Классификация методом опорных векторов .....	22
2.4.1. Классификация данных (метод опорных векторов).....	22
2.4.2. Результаты и оценка классификации .....	23
2.5. Визуализация данных .....	24

2.5.1. Визуализация в R (ggplot2, plotly) .....	24
2.5.2. Интерактивные дашборды в Glarus BI.....	27
2.5.3. Сравнение методов визуализации .....	28
3. АВТОМАТИЗАЦИЯ И ОТЧЁТНОСТЬ В АНАЛИЗЕ ДАННЫХ.....	30
3.1. Генерация отчётов в R .....	30
3.1.1. Обоснование необходимости автоматизации отчётов .....	30
3.1.2. Использование RMarkdown для создания отчётов .....	31
3.1.3. Экспорт отчётов в PDF, HTML, Word .....	31
3.2. Формирование интерактивных отчётов в Glarus BI .....	32
3.2.1. Различие между статичными и интерактивными отчётами .....	32
3.2.2. Создание дашбордов в Glarus BI .....	32
3.2.3. Экспорт отчётов в Glarus BI.....	33
3.3. Сравнение инструментов R и Glarus BI.....	33
3.3.1. Анализ сильных и слабых сторон инструментов .....	33
3.3.2. Возможности интеграции R и Glarus BI .....	34
3.3.3. Применимость инструментов для различных типов задач.....	35
ЗАКЛЮЧЕНИЕ .....	36
СПИСОК ИСТОЧНИКОВ .....	37
ПРИЛОЖЕНИЕ.....	38

# 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 VI.

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.

	BPM	RMS.Energy	Zero.Crossing.Rate	Инструментальность	Вокал	Код.Жанра
1	103	0.2015555	0.19095720	0.08	0.92	2
2	111	0.2583140	0.08387791	0.35	0.65	2
3	114	0.3284943	0.09978011	0.07	0.93	2
4	121	0.2282471	0.16713450	0.36	0.64	2
5	90	0.2768621	0.14672060	0.06	0.94	2
6	115	0.2244783	0.08520809	0.37	0.63	2

Рисунок – Обзор структуры данных

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

### 2.1.3. Предварительная обработка данных

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).
- Преобразование категориальных переменных: переменная `Код.Жанра` в фактор с помощью `as.factor()`, что необходимо для корректной работы SVM.

```
train$Код.Жанра <- factor(train$Код.Жанра)
test$Код.Жанра <- factor(test$Код.Жанра, levels = levels(train$Код.Жанра))
```

#### Рисунок – Преобразование в факторы

Эти шаги позволяют подготовить данные к корректной работе алгоритма классификации и минимизируют возможные ошибки при обучении модели.

## 2.2. Исследовательский анализ данных

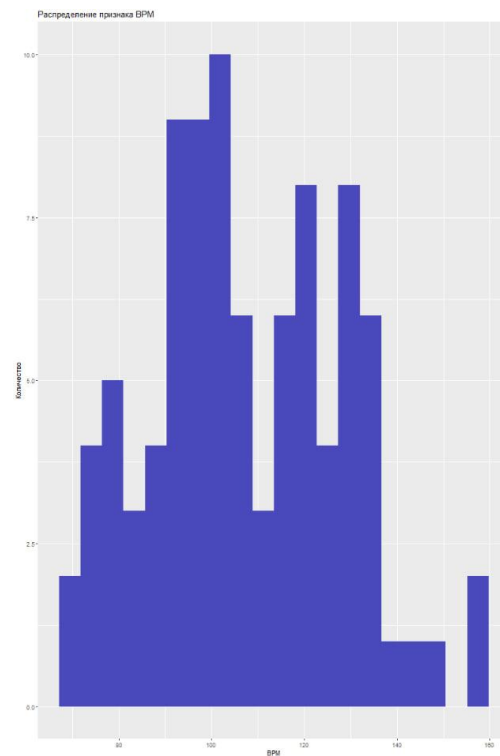
Исследовательский анализ данных (EDA) позволяет получить общее представление о распределении признаков, выявить возможные зависимости, аномалии и особенности структуры выборки. Он является важным предварительным этапом, особенно при построении интерпретируемых моделей, таких как метод опорных векторов.

### 2.2.1. Визуализация распределений

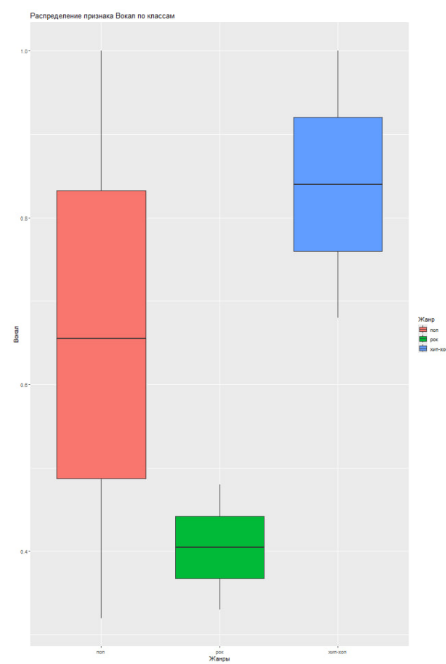
На первом этапе анализа была произведена визуализация распределения признаков, таких как BPM, Инструментальность, Zero Crossing Rate, RMS Energy. В частности, с помощью гистограмм и диаграмм рассеяния были выявлены:

- выраженная концентрация треков с BPM около 100;
- отсутствие выбросов по признаку Вокал в трех классах;
- различия в распределении инструментальности и BPM между треками в жанре рок и в жанре хип-хоп.

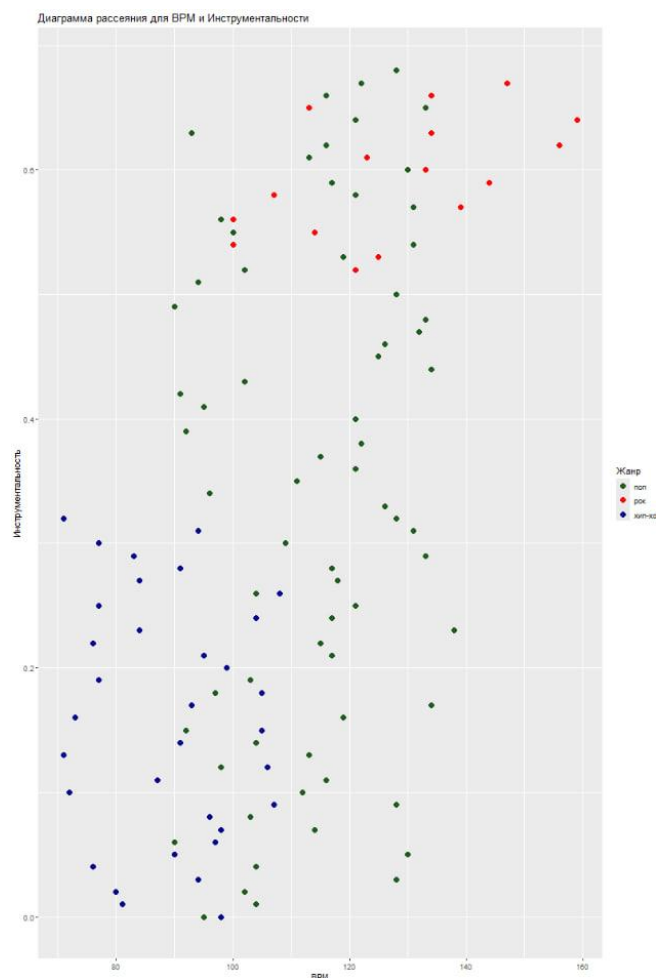




**Рисунок – Гистограмма распределения BPM**



**Рисунок – Боксплот распределений признака Вокал по классам**



**Рисунок – Диаграмма рассеяния для признаков BPM и Инструментальности**

### **2.2.2. Корреляционный анализ**

Далее был проведён корреляционный анализ между числовыми признаками: BPM, Инструментальностью, RMS Energy и Zero Crossing Rate. Матрица корреляции позволила выявить следующие закономерности:

- BPM (1.0) имеет умеренную корреляцию с RMS Energy (0.44) и слабую с Zero Crossing Rate (0.34).
- RMS Energy (1.0) достаточно сильно коррелирует с Инструментальностью (0.56) и Zero Crossing Rate (0.47), что может указывать на связь между энергией звука, его инструментальной природой и частотой пересечения нуля.

- Zero Crossing Rate (1.0) имеет не очень высокую корреляцию с Инструментальностью (0.4), что может означать, что частота пересечения нуля слабо связана с инструментальностью звука.

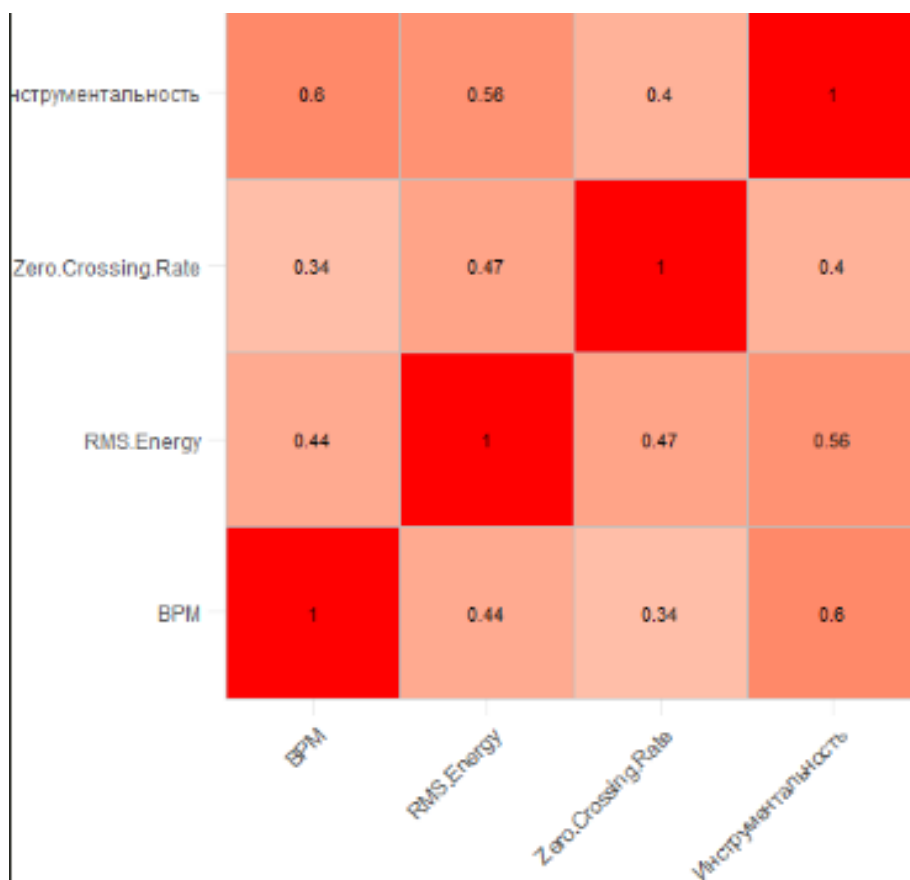


Рисунок — Корреляционная матрица

### 2.2.3. Выявление выбросов и трендов

С помощью визуальных инструментов, таких как боксплоты и агрегированные столбчатые диаграммы, обнаруженных потенциальных выбросов не было.

Несмотря на отсутствие временных рядов в датасете, были изучены тренды через упорядоченные признаки, что позволило выявить некоторые закономерности, полезные для дальнейшего моделирования.

## 2.3. Применение методов статистического анализа

### 2.3.1. Описательная статистика

На первом этапе были рассчитаны базовые статистические характеристики для ключевых признаков: BPM, RMS Energy, Zero Crossing Rate, Инструментальность, Вокал.

Результаты анализа показывают, что:

- средний темп песен составляет 107.4 удара в минуту (BPM);
- средний уровень значения RMS Energy варьируется в диапазоне от 0.1066 до 0.7761, с медианой 0.1253 – распределение относительно симметрично, но с несколькими выбросами в сторону высоких значений;
- показатель Инструментальность варьируется от 0 до 1, в среднем музыка в датасете имеет умеренную до высокой инструментальность, в большинстве значения сосредоточены выше 0.4.

BPM	RMS.Energy	Zero.Crossing.Rate	Инструментальность
Min. : 71.0	Min. : 0.1066	Min. : 0.05153	Min. : 0.0000
1st Qu.: 93.0	1st Qu.: 0.2244	1st Qu.: 0.09637	1st Qu.: 0.1375
Median : 104.5	Median : 0.2772	Median : 0.12528	Median : 0.3850
Mean : 107.4	Mean : 0.3300	Mean : 0.13059	Mean : 0.3551
3rd Qu.: 122.2	3rd Qu.: 0.4079	3rd Qu.: 0.14711	3rd Qu.: 0.5625
Max. : 159.0	Max. : 0.7761	Max. : 0.29485	Max. : 0.6800
Вокал	Код.Жанра	genre	InstrumentalityCategory
Min. : 0.3200	Min. : 0.000	рок : 16	Low : 42
1st Qu.: 0.4375	1st Qu.: 1.000	хип-хоп: 33	Medium: 47
Median : 0.6150	Median : 1.000	поп : 43	High : 3
Mean : 0.6449	Mean : 1.293		
3rd Qu.: 0.8625	3rd Qu.: 2.000		
Max. : 1.0000	Max. : 2.000		

Рисунок – Описательная статистика

### 2.3.2. Проверка гипотез

Для количественной оценки различий между жанрами (хип-хопа, рока и попа) были применены методы проверки статистических гипотез.

t-тест Стьюдента показал, что среднее значение темпа роковых треков статистически значимо отличается от среднего темпа песен в жанре поп, причем тем

рок-песен выше. ( $p < 0.01$ ).

```
6 rock <- ds$`BPM`[ds$`Код.Жанра` == 0]
7 pop <- ds$`BPM`[ds$`Код.Жанра` == 2]
8 t_test_result <- t.test(rock, pop, var.equal = TRUE)
9
```

PROBLEMS 11 OUTPUT DEBUG CONSOLE TERMINAL PORTS GITLENS

Two Sample t-test

data: rock and pop  
t = 3.1333, df = 57, p-value = 0.002729  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
5.150804 23.392801  
sample estimates:  
mean of x mean of y  
128.0625 113.7907

Рисунок – t-test

ANOVA (дисперсионный анализ) показывает, что средние темпы песен (BPM) статистически значимо различаются между всеми жанрами, а это подтверждает гипотезу о том, что жанр оказывает влияние на темп (исходя из данных).

```
6 hip-hop <- ds$`BPM`[ds$`Код.Жанра` == 1]
7 rock <- ds$`BPM`[ds$`Код.Жанра` == 0]
8 pop <- ds$`BPM`[ds$`Код.Жанра` == 2]
9 ds$genre <- factor(ds$Код.Жанра, levels = c(0, 1, 2),
10 labels = c("рок", "хип-хоп", "поп"))
11 anova_result <- aov(BPM ~ genre, data = ds)
12 t_test_result <- t.test(rock, pop, var.equal = TRUE)
13
```

PROBLEMS 12 OUTPUT DEBUG CONSOLE TERMINAL PORTS GITLENS

Terms:

	genre	Residuals
Sum of Squares	19649.52	18184.78
Deg. of Freedom	2	89

Residual standard error: 14.29417  
Estimated effects may be unbalanced

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
genre	2	19650	9825	48.08	6.94e-15 ***
Residuals	89	18185	204		

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Рисунок – ANOVA-test

## 2.4. Классификация методом опорных векторов

### 2.4.1. Классификация данных (метод опорных векторов)

В рамках данной работы была построена модель методом опорных векторов на основе обучающей выборки. Для обучения использовалась функция `svm()` из пакета `e1071`.

Целевая переменная: Код Жанра (0 — рок, 1 — хип-хоп, 2 — поп).

```
# 3. Разделение на обучающую и тестовую выборки
set.seed(123) # для воспроизводимости
n <- nrow(data_music)
train_indices <- sample(1:n, size = round(0.8 * n))
train <- data_music[train_indices, ]
test <- data_music[-train_indices, ]

# ПРЕОБРАЗУЕМ В ФАКТОРЫ ПОСЛЕ РАЗДЕЛЕНИЯ
train$Код.Жанра <- factor(train$Код.Жанра)
test$Код.Жанра <- factor(test$Код.Жанра, levels = levels(train$Код.Жанра))

# 4. Обучение модели SVM
svm_model <- svm(Код.Жанра ~ BPM + RMS.Energy + Zero.Crossing.Rate
+ Инструментальность + Вокал,
data = train,
kernel = "radial", # Радиальное ядро
scale = TRUE,      # Масштабирование признаков
probability = TRUE) # Для получения вероятностей

# Предсказание
predictions <- predict(svm_model, test)

# Преобразуем predictions в фактор (уровни берем из train)
predictions <- factor(predictions, levels = levels(train$Код.Жанра))
```

Рисунок – SVM-модель

## 2.4.2. Результаты и оценка классификации

После обучения модель была применена к тестовой выборке, и предсказания были сопоставлены с истинными метками. Также была выполнена количественная оценка качества классификации с помощью стандартных метрик.

Результаты оценки модели:

- Точность (Accuracy): 0.833 — модель правильно классифицирует почти 83% песен;
- Средняя полнота (Recall): 0.785 — 79% песен были корректно определены моделью;
- Средняя точность (Precision): 0.875 — доля всех правильно положительных предсказанных песен среди всех положительных.
- F1-мера: 0.8 — сбалансированный показатель между полнотой и точностью.

```
Overall Statistics

Accuracy : 0.8333
95% CI : (0.5858, 0.9642)
No Information Rate : 0.3889
P-Value [Acc > NIR] : 0.0001479

Kappa : 0.7353

McNemar's Test P-Value : NA

Statistics by Class:

                Class: 0 Class: 1 Class: 2
Sensitivity      0.5000   1.0000   0.8571
Specificity      1.0000   0.9091   0.8182
Pos Pred Value   1.0000   0.8750   0.7500
Neg Pred Value   0.8750   1.0000   0.9000
Prevalence       0.2222   0.3889   0.3889
Detection Rate   0.1111   0.3889   0.3333
Detection Prevalence 0.1111 0.4444 0.4444
Balanced Accuracy 0.7500   0.9545   0.8377
Precision:
Class: 0 Class: 1 Class: 2
    1.000   0.875   0.750

Recall:
Class: 0 Class: 1 Class: 2
0.5000000 1.0000000 0.8571429

F1-Score:
Class: 0 Class: 1 Class: 2
0.6666667 0.9333333 0.8000000

Средний Precision: 0.875
Средний Recall: 0.7857143
Средний F1-Score: 0.8
```

Рисунок – Оценочные метрики классификации

## 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)

В рамках данной работы активно использовалась библиотека ggplot2 для построения визуализаций на языке R. Этот инструмент позволяет создавать профессиональные и гибко настраиваемые графики для анализа и представления данных.

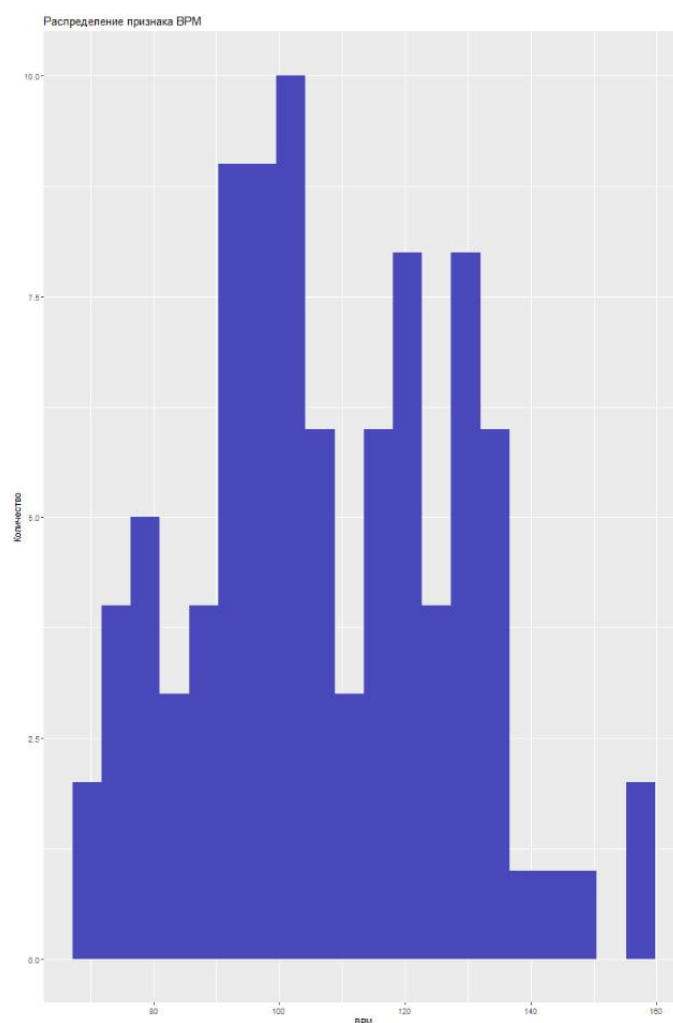
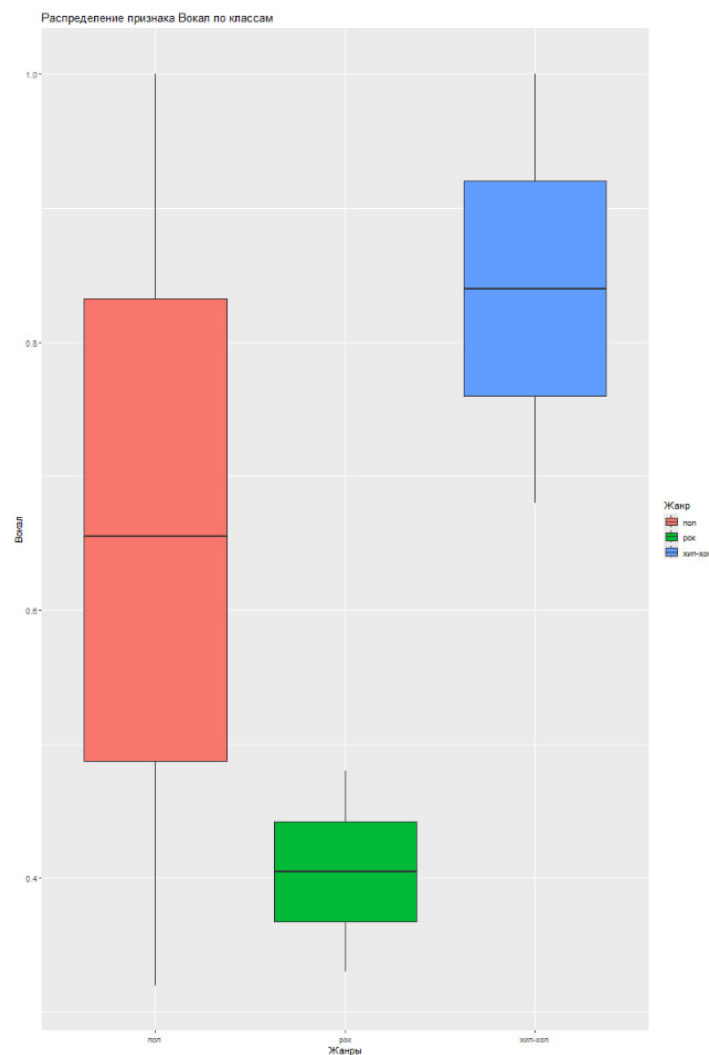


Рисунок – Гистограмма распределения BPM

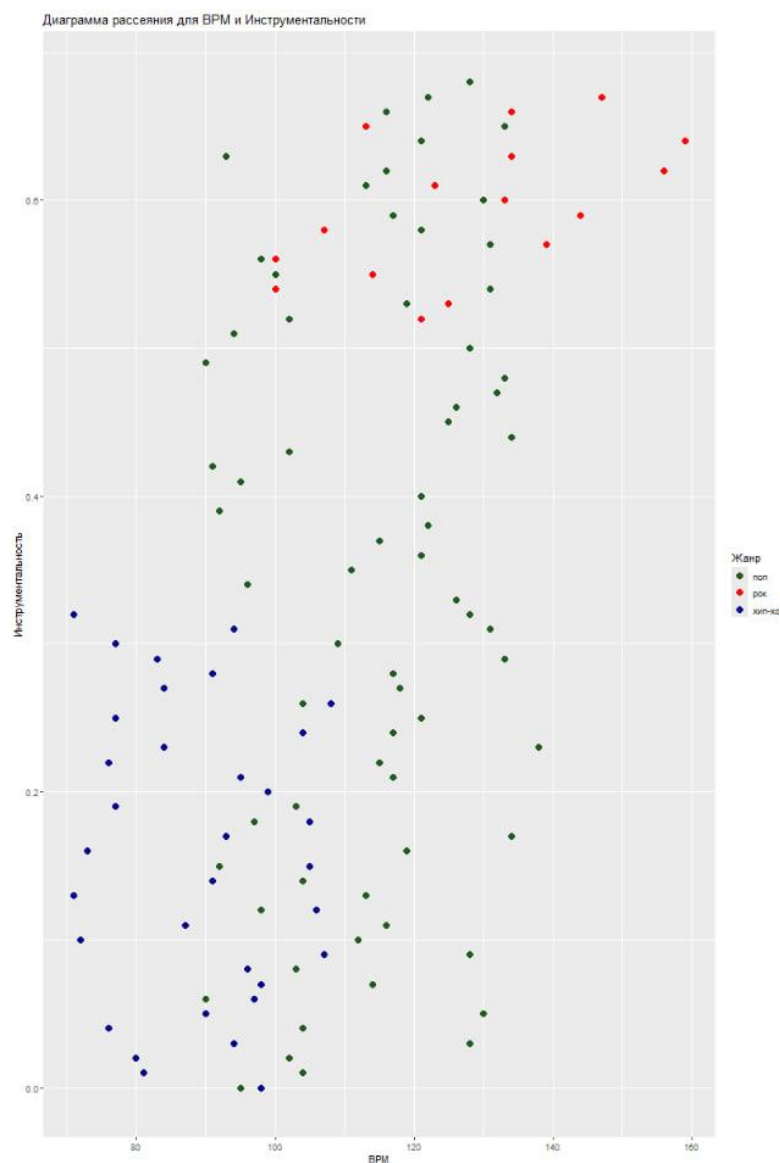


Гистограмма распределения показателя BPM показала, что большинство значений сосредоточены в диапазоне от 90 до 100.



**Рисунок – Боксплот распределений признака Вокал по классам**

Боксплот распределений признака Вокал по группам продемонстрировал отсутствие значительных выбросов по жанрам, а также значения в разных классах имеют различие в медианах.



**Рисунок – Диаграмма рассеяния для признаков BPM и Инструментальности**

Диаграмма рассеяния между BPM и Инструментальности показывает тенденцию увеличению признака ударов в минуту (BPM) с возрастания уровня инструментальности музыкального произведения.

### 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.

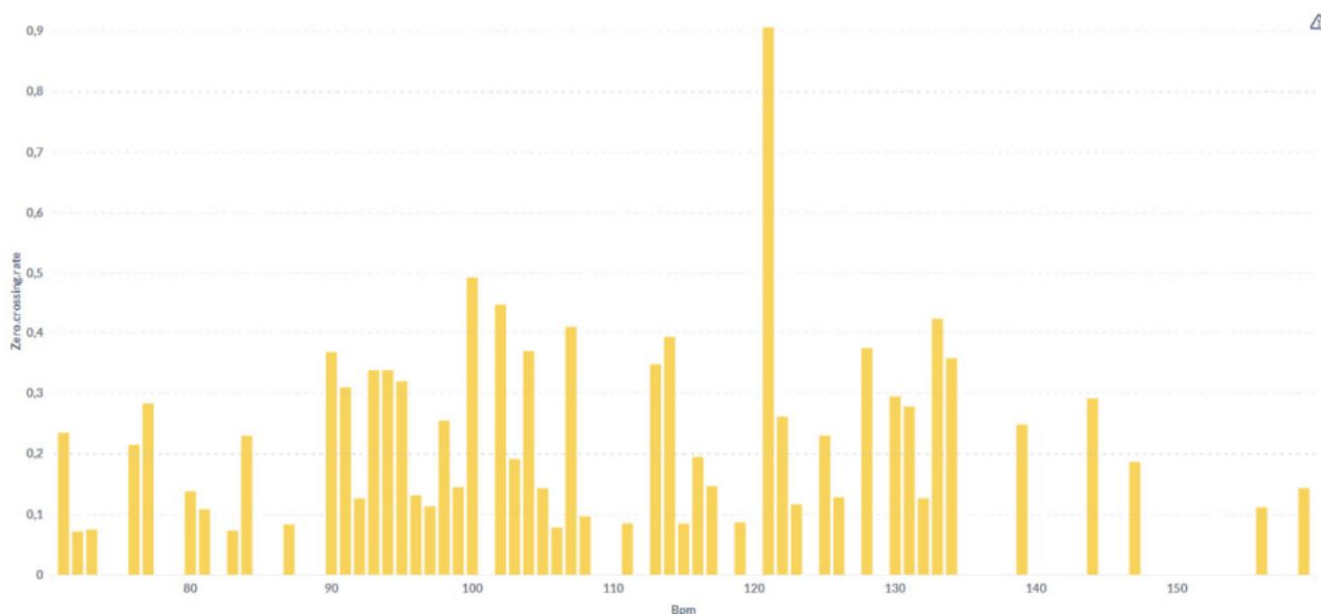
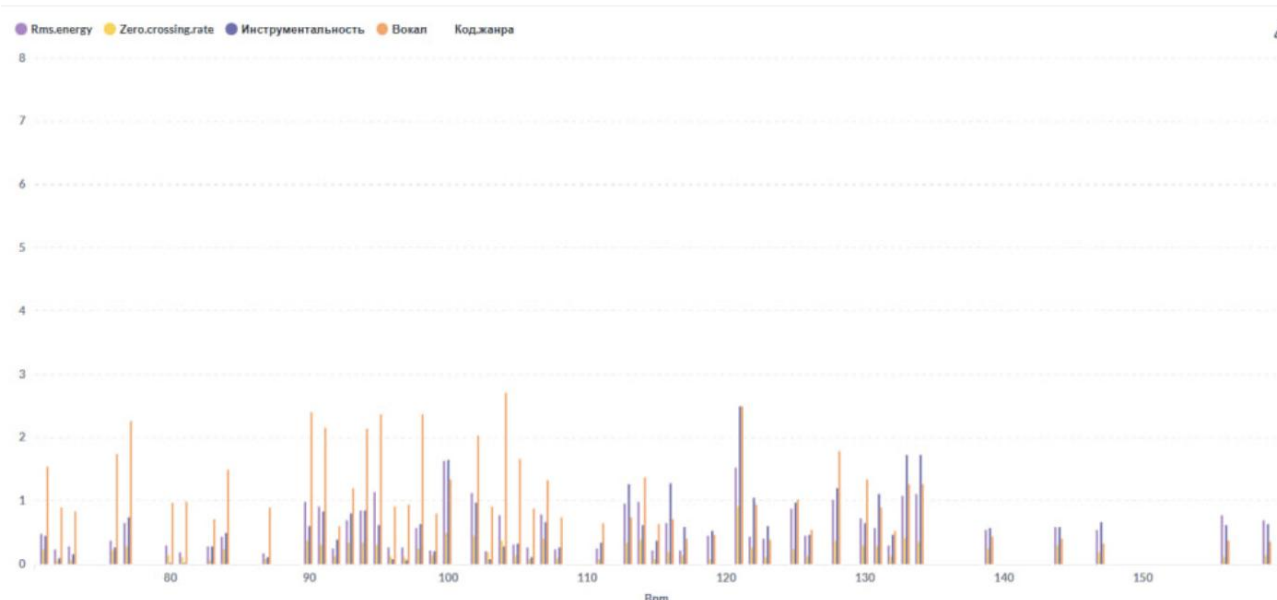


Рисунок — Дашборд Glarus BI



**Рисунок — Дашборд Glarus BI**

### 2.5.3. Сравнение методов визуализации

Ниже представлено сравнение визуализации с помощью R (ggplot2) и Glarus BI.

*Таблица — Сравнение методов визуализации*

Критерий	R (ggplot2)	Glarus BI
Гибкость настройки	Крайне высокая (настраивается каждая деталь)	Средняя (готовые шаблоны, ограниченная кастомизация)
Интерактивность	Отсутствует	Высокая (интерактивные фильтры, выпадающие меню)
Автоматизация	Полная (сценарии, отчёты, RMarkdown)	Частичная (автообновление, расписание)
Удобство для конечного пользователя	Требует навыков программирования	Не требует — ориентирован на конечного пользователя
Поддержка сложных графиков	Любые типы визуализаций	Ограничена простыми диаграммами и таблицами

Язык R обладает удобным инструментарием с библиотеками визуализации, которые подойдут специалистам в области анализа данных, ведь они обладают гибкостью и предельной точностью в построении графиков. Glarus BI, в свою очередь, удобен для создания отчётов и дашбордов, которые предназначены для быстрого принятия решений и наглядного представления информации без необходимости писать код. Совместное использование этих инструментов позволяет объединить аналитическую мощь с доступной визуализацией.

### **3. АВТОМАТИЗАЦИЯ И ОТЧЁТНОСТЬ В АНАЛИЗЕ ДАННЫХ**

Автоматизация отчётности помогает эффективно работать с данными в условиях регулярной аналитики. В музыкальной сфере, где важны точность, актуальность и прозрачность данных, автоматизация позволяет формировать отчёты быстро, последовательно и без риска. Благодаря современным инструментам, таким как R и Glarus BI, можно создавать как статические, так и интерактивные отчёты, обновляемые при изменении данных.

#### **3.1. Генерация отчётов в R**

##### **3.1.1. Обоснование необходимости автоматизации отчётов**

Во многих учреждениях аналитика по-прежнему сопровождается ручной подготовкой графиков, таблиц и отчётных документов. Такой подход не только затратен по времени, но и подвержен риску ошибок и несогласованностей. Автоматизация, особенно в задачах повторяющейся аналитики позволяет минимизировать трудозатраты и обеспечить воспроизводимость результатов.

Преимущества автоматизации отчётности:

- сокращение времени на подготовку документов;
- исключение человеческого фактора;
- возможность многократного воспроизведения с обновлёнными данными;
- интеграция текста, графиков и расчётов в одном документе.

### 3.1.2. Использование RMarkdown для создания отчётов

Одним из мощных инструментов автоматизации отчётности в R является RMarkdown — формат, объединяющий текст, код и визуализацию в едином документе. Он позволяет:

- выполнять R-код прямо в структуре отчёта;
- автоматически добавлять в отчёт таблицы, графики и результаты анализа;
- поддерживать стилизованный текст, заголовки, списки, формулы и изображения;
- легко обновлять отчёт при изменении данных — просто перезапустив рендеринг.

Файл.Rmd можно запускать через интерфейс RStudio, и он будет автоматически превращён в готовый отчёт с актуальными результатами.

### 3.1.3. Экспорт отчётов в PDF, HTML, Word

Одно из главных преимуществ RMarkdown — это поддержка различных форматов экспорта. В зависимости от цели, отчёт можно сгенерировать в:

- HTML — интерактивный отчёт, удобный для веб-публикации и навигации;
- PDF — надёжный формат для печати, архивирования и официальных документов;
- Microsoft Word (.docx) — привычный формат для офисной среды.

Экспорт можно выполнять как через меню RStudio, так и программно.

Таким образом, автоматизация отчётности с использованием R и RMarkdown позволяет не только ускорить подготовку аналитических документов, но и обеспечить их консистентность, прозрачность и удобство для обмена между участниками процесса.

## **3.2. Формирование интерактивных отчётов в Glarus BI**

### **3.2.1. Различие между статичными и интерактивными отчётами**

Статичные отчёты представляют собой неизменяемые документы, обычно сохранённые в форматах PDF или Word. Они удобны для печати и архивирования, но не позволяют пользователю взаимодействовать с данными напрямую. В отличие от них, интерактивные отчёты дают возможность фильтровать, сортировать, настраивать отображение информации без изменения исходных данных. Это особенно важно при работе с большим объёмом показателей, когда необходимо быстро находить нужные зависимости или формировать представление под конкретную задачу.

### **3.2.2. Создание дашбордов в Glarus BI**

Glarus BI предоставляет удобную визуальную среду для построения интерактивных отчётов и дашбордов. Создание отчёта начинается с импорта данных (из Excel, CSV или SQL), после чего аналитик может добавить на рабочее пространство графики, таблицы, фильтры и другие элементы визуализации.

Функции, доступные при создании дашбордов:

- выбор визуализаций: линейные и столбчатые диаграммы, круговые графики, таблицы;
- создание фильтров по признакам: BPM, RMS Energy и т.д.;
- группировка и агрегация данных;
- добавление вычисляемых показателей (например, доля треков в жанре рок);
- автоматическое обновление данных.



### 3.2.3. Экспорт отчётов в Glarus BI

Созданные отчёты в Glarus BI можно:

- сохранять в системе для внутреннего использования;
- делиться ими с людьми через общие ссылки или корпоративную BI-среду;
- экспортировать в PDF или Excel — для отправки по электронной почте или печати;
- встраивать в веб-интерфейсы или отчёты других систем через iframe или API (если включено).

Таким образом, Glarus BI обеспечивает не только визуализацию и исследование данных, но и полноценную платформу для построения интерактивной отчётности, которая может быть адаптирована под потребности конкретных отделов, специалистов или управленцев.

## 3.3. Сравнение инструментов R и Glarus BI

### 3.3.1. Анализ сильных и слабых сторон инструментов

Каждый из инструментов — R и Glarus BI — обладает своими преимуществами и ограничениями, которые определяют их применимость в различных аналитических сценариях.

Сильные стороны R:

- высокая гибкость и масштабируемость при работе с данными;
- богатая экосистема пакетов для анализа, моделирования и визуализации (ggplot2, caret, MASS и др.);
- полная автоматизация аналитических процессов и отчётности (через RMarkdown);
- возможность обучения и тестирования сложных моделей машинного обучения.

Слабые стороны R:

- требует уверенных навыков программирования и знания статистики;
- менее удобен для пользователей без технической подготовки;
- визуализация требует написания кода и тонкой настройки.

Сильные стороны Glarus BI:

- доступность для бизнес-пользователей без навыков программирования;
- быстрый запуск и визуальное построение отчётов через интерфейс;
- возможности фильтрации, агрегации и совместной работы;
- публикация интерактивных дашбордов и обновление по расписанию.

Слабые стороны Glarus BI:

- ограниченность в построении сложных моделей анализа данных;
- зависимость от готовых визуальных компонентов и шаблонов;
- ограниченная расширяемость и поддержка кастомной логики.

### **3.3.2. Возможности интеграции R и Glarus BI**

Несмотря на различия в подходе и целевой аудитории, инструменты R и Glarus BI можно использовать совместно. Интеграция может быть реализована следующим образом:

- подготовка и агрегация данных в R, экспорт в формат CSV/Excel и последующая загрузка в Glarus BI для визуализации;
- генерация отчётов в RMarkdown и добавление ссылок на дашборды Glarus BI как внешние компоненты;
- использование одного и того же набора данных как в R (для моделирования), так и в Glarus BI (для презентации);
- автоматическое формирование экспортных файлов в R с дальнейшим импортом в BI-платформу.

Такой подход позволяет объединить аналитическую мощь R с наглядностью и доступностью Glarus BI.

### 3.3.3. Применимость инструментов для различных типов задач

Для наглядности сравнения возможностей R и Glarus BI в зависимости от специфики аналитических задач, приведём сводную таблицу.

*Таблица — Сравнение применимости R и Glarus BI*

Тип задачи	Подходит R	Подходит Glarus BI
Статистический анализ	Да	Ограничено
Машинное обучение	Да	Нет
Визуализация данных	Через ggplot2, plotly	Через интерфейс
Интерактивные дашборды	Ограничено (через Shiny)	Да
Подготовка отчётов	RMarkdown	Встроенный экспорт
Работа без программирования	Нет	Да

Таким образом, R идеально подходит для проведения глубокого анализа, построения моделей и статистических расчётов, в то время как Glarus BI — для создания интерактивной отчётности, презентации результатов и бизнес-визуализации. Их совместное использование позволяет получить как точные аналитические выводы, так и удобную форму их представления конечному пользователю.

## 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.

## СПИСОК ИСТОЧНИКОВ

1. Хомченко И. Г. Язык программирования R. Анализ и визуализация данных. — М.: ДМК Пресс, 2020. — 416 с.
2. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning. — New York: Springer, 2021. — 426 p.
3. Kuznetsova E. R Markdown на практике. — СПб.: Питер, 2021. — 256 с.
4. Официальная документация R. URL: <https://www.r-project.org/> (дата обращения: 27.05.2025).
5. Glarus BI. Официальный сайт. URL: <https://glarusbi.ru/> (дата обращения: 26.05.2025).

## 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