1. Титульный слайд

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«МИРЭА - Российский технологический университет»
Институт информационных технологий (ИИТ)
Кафедра прикладной математики
Направление «Прикладная математика»
профиль «Анализ данных»

Курсовая работа по дисциплине «Языки программирования для статистической обработки данных» по теме:
«XXX»

Автор: студент группы ИМБО-11-23
Журавлев Ф. А.
Руководитель: к.ф.-м.н., доцент каф. ПМ
Царькова Е. Г.

Москва 2025

2. Введение

- 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.

3. Актуальность исследования

- Классификация музыки играет критически важную роль. Она позволяет эффективно организовывать, искать и рекомендовать музыку, формируя пользовательский опыт и принося значительную выгоду различным участникам рынка.

4. Цель работы

- 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.

```r
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))

# Матрица ошибок
conf_matrix <- confusionMatrix(predictions, test$Код.Жанра)
print("Матрица ошибок:")
print(conf_matrix)

# Извлечение Precision, Recall и F1-Score
precision <- conf_matrix$byClass[, "Precision"]
recall <- conf_matrix$byClass[, "Recall"]
f1_score <- conf_matrix$byClass[, "F1"]
```

5. Язык R и его возможности

- Язык R — мощный и гибкий инструмент для статистических вычислений и визуализации данных. Он широко используется в анализе данных, машинном обучении, биоинформатике и других научных областях. Основные возможности R включают обработку и очистку данных, проведение статистических тестов, визуализацию результатов, создание интерактивных графиков, и работу с регрессионными моделями.

6. Описание данных

- В работе используется синтетический набор данных.
- Датасет состоит из 94 записей, также включает в себя 5 признаков и одну целевую переменную (Код Жанра).
- Признаки разделяются на числовые (BPM, RMS Energy, Zero Crossing Rate, Инструментальность и Вокал) и категориальные (Жанр).

	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

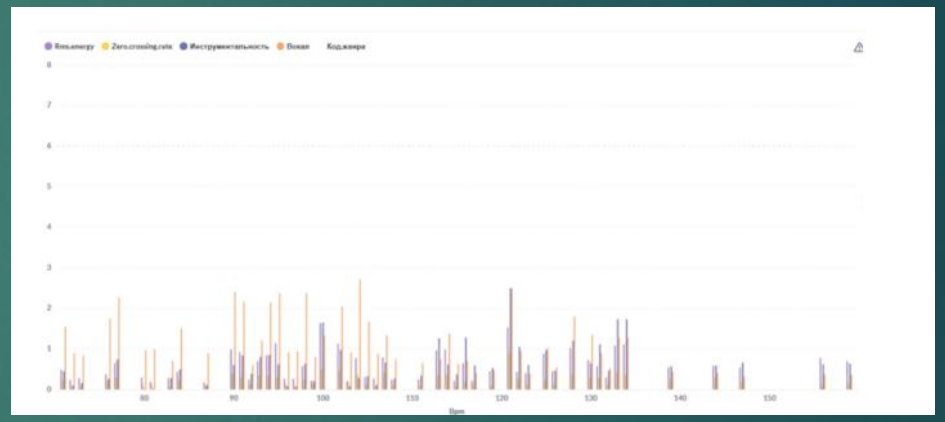
7. Методы статистической обработки данных

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

```
      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
```

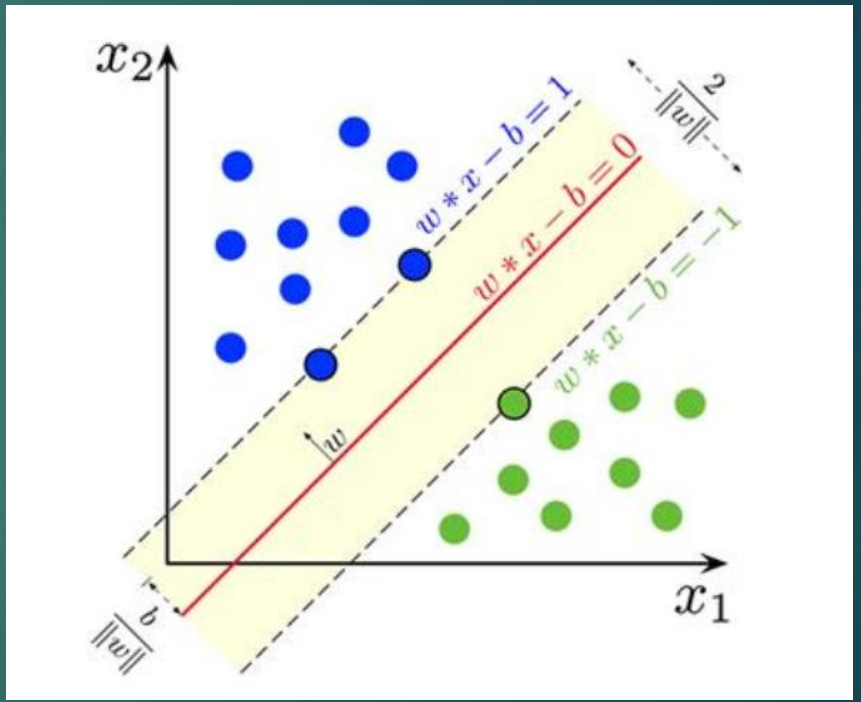
8. Задача классификации

- Задачу классификации музыки можно определить как процесс определения жанра трека на основе его аудиоданных.
- Существуют различные методы классификации, такие как логистическая регрессия, байесовский классификатор, однако в работе используется метод опорных векторов.



9. Метод опорных векторов

- Метод опорных векторов (SVM) ищет гиперплоскость, которая максимально разделяет данные разных классов.
- Опорные векторы – это объекты обучения, лежащие ближе всего к разделяющей гиперплоскости.
- Цель – максимизировать расстояние между гиперплоскостью и ближайшими точками каждого класса.



10. Реализация метода опорных векторов

```r
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))

# Матрица ошибок
conf_matrix <- confusionMatrix(predictions, test$Код.Жанра)
print("Матрица ошибок:")
print(conf_matrix)

# Извлечение Precision, Recall и F1-Score
precision <- conf_matrix$byClass[, "Precision"]
recall <- conf_matrix$byClass[, "Recall"]
f1_score <- conf_matrix$byClass[, "F1"]
```

11. Критерии качества классификации

- Критериями качества классификация являются такие меры, как precision, accuracy, recall, f1-score.

- Они помогают определить насколько модель классификации эффективно справляется со своей поставленной задачей.

```
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
```

12. Визуализация

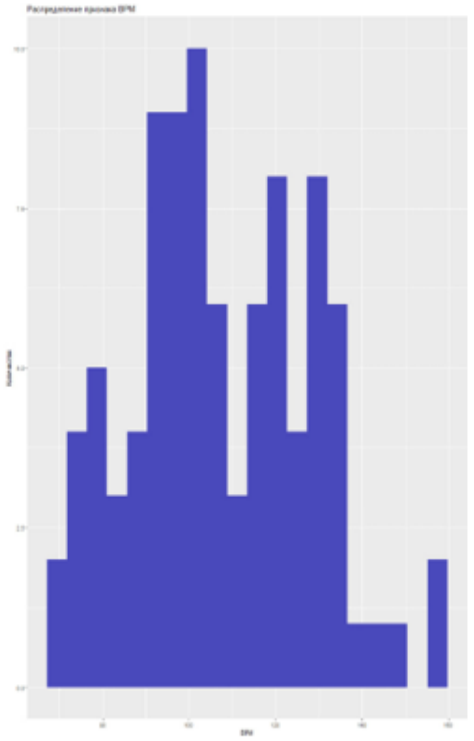


Рисунок – Гистограмма распределения BPM

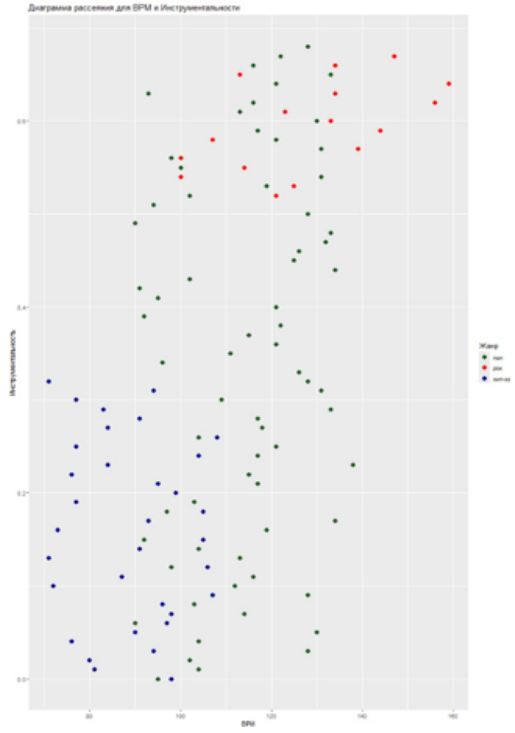


Рисунок – Диаграмма рассеяния для признаков BPM и Инструментальности

13. 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.

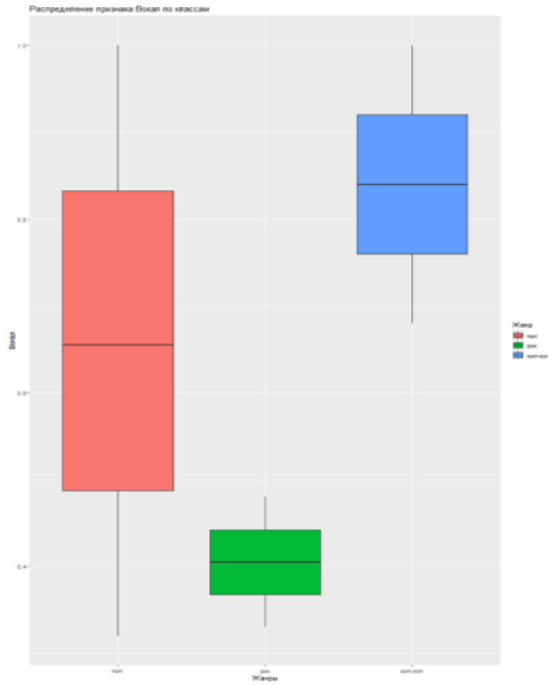


Рисунок – Боксплот распределений признака Вокал по классам

Спасибо за внимание!