



Московский государственный  
технический университет имени  
Н.Э. Баумана



Кафедра ИУ5  
«Системы обработки  
информации и управления»

# СИСТЕМА АНАЛИЗА АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧ КЛАССИФИКАЦИИ С ИСПОЛЬЗОВАНИЕМ PANDAS

---

Выпускная квалификационная работа бакалавра

ВЫПОЛНИЛА: Соловьева А.С. ИУ5Ц-103Б

НАУЧНЫЙ РУКОВОДИТЕЛЬ: Григорьев Ю.А.

# АКТУАЛЬНОСТЬ

---

- НЕОБХОДИМОСТЬ **СОЗДАНИЯ ЭФФЕКТИВНЫХ ИНСТРУМЕНТОВ ДЛЯ АНАЛИЗА** РАСТУЩИХ ОБЪЕМОВ ДАННЫХ
- **УПРОСТИТ ПРОЦЕСС АНАЛИЗА** ДАННЫХ, **УСКОРИТ ОБУЧЕНИЕ АЛГОРИТМОВ**, **ПОВЫСИТ ТОЧНОСТЬ ПРЕДСКАЗАНИЙ** И **УЛУЧШИТ КАЧЕСТВО РЕШЕНИЙ НА ИХ ОСНОВЕ**
- **СИСТЕМА БУДЕТ ПОЛЕЗНА** ДЛЯ ИССЛЕДОВАТЕЛЕЙ, СПЕЦИАЛИСТОВ ПО ДАННЫМ, ДЛЯ СТУДЕНТОВ И ПРЕПОДАВАТЕЛЕЙ, ЗАНИМАЮЩИХСЯ ИЗУЧЕНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

# ЦЕЛЬ РАБОТЫ

---

СОЗДАНИЕ СИСТЕМЫ АНАЛИЗА  
АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ  
РЕШЕНИЯ ЗАДАЧ КЛАССИФИКАЦИИ С  
ИСПОЛЬЗОВАНИЕМ БИБЛИОТЕКИ PANDAS

# ЗАДАЧИ

---

1

ИССЛЕДОВАНИЕ  
ПРЕДМЕТНОЙ ОБЛАСТИ

4

ОЦЕНКА КАЧЕСТВА  
КАЖДОЙ МОДЕЛИ

2

ВЫБОР АЛГОРИТМОВ  
КЛАССИФИКАЦИИ

5

СРАВНЕНИЕ РЕЗУЛЬТАТОВ И  
ОПРЕДЕЛЕНИЕ НАИЛУЧШЕЙ  
МОДЕЛИ

3

ОБУЧЕНИЕ МОДЕЛЕЙ

6

РЕАЛИЗАЦИЯ В ВИДЕ  
ВЕБ-ПРИЛОЖЕНИЯ

# PANDAS

---

PANDAS – **БИБЛИОТЕКА ДЛЯ АНАЛИЗА ДАННЫХ,**  
**ОСНОВАННАЯ НА ЯЗЫКЕ ПРОГРАММИРОВАНИЯ PYTHON**  
БИБЛИОТЕКА PANDAS ПРЕДСТАВЛЯЕТ СОБОЙ **НАБОР**  
**ИНСТРУМЕНТОВ ДЛЯ ОПЕРАЦИЙ С ДАННЫМИ:**  
СОРТИРОВКИ, ФИЛЬТРАЦИИ, ОЧИСТКИ, УДАЛЕНИЯ  
ДУБЛИКАТОВ И МНОГИЕ ДРУГИЕ.

# ФУНКЦИОНАЛЬНЫЕ ТРЕБОВАНИЯ

---

СИСТЕМА ВЫПОЛНЯЕТ СЛЕДУЮЩИЕ ФУНКЦИИ

- Обработка данных
- Масштабирование данных и корреляционный анализ
- Разделение данных на обучающую и тестовую выборку
- Обучение моделей
- Оценка каждой модели
- Построение графиков сравнения оценок моделей

# ИСПОЛЬЗУЕМЫЕ ТЕХНОЛОГИИ

---

- ЯЗЫК ПРОГРАММИРОВАНИЯ **PYTHON**
- БИБЛИОТЕКИ: **PANDAS, NUMPY, MATPLOTLIB, SEABORN, SCIKIT-LEARN**
- ФРЕЙМВОРК **STREAMLIT**
- СРЕДА РАЗРАБОТКИ: **JUPYTER NOTEBOOK**



# НАБОР ДАННЫХ

---

ПРЕДМЕТНАЯ ОБЛАСТЬ РАЗРАБОТКИ:

**НАБОР ДАННЫХ «СТАТИЧЕСКИЕ ДАННЫЕ О ЗАНЯТОСТИ И  
БЕЗРАБОТИЦЕ СРЕДИ НАСЕЛЕНИЯ РФ ПО ВОЗРАСТНЫМ ГРУППАМ»**

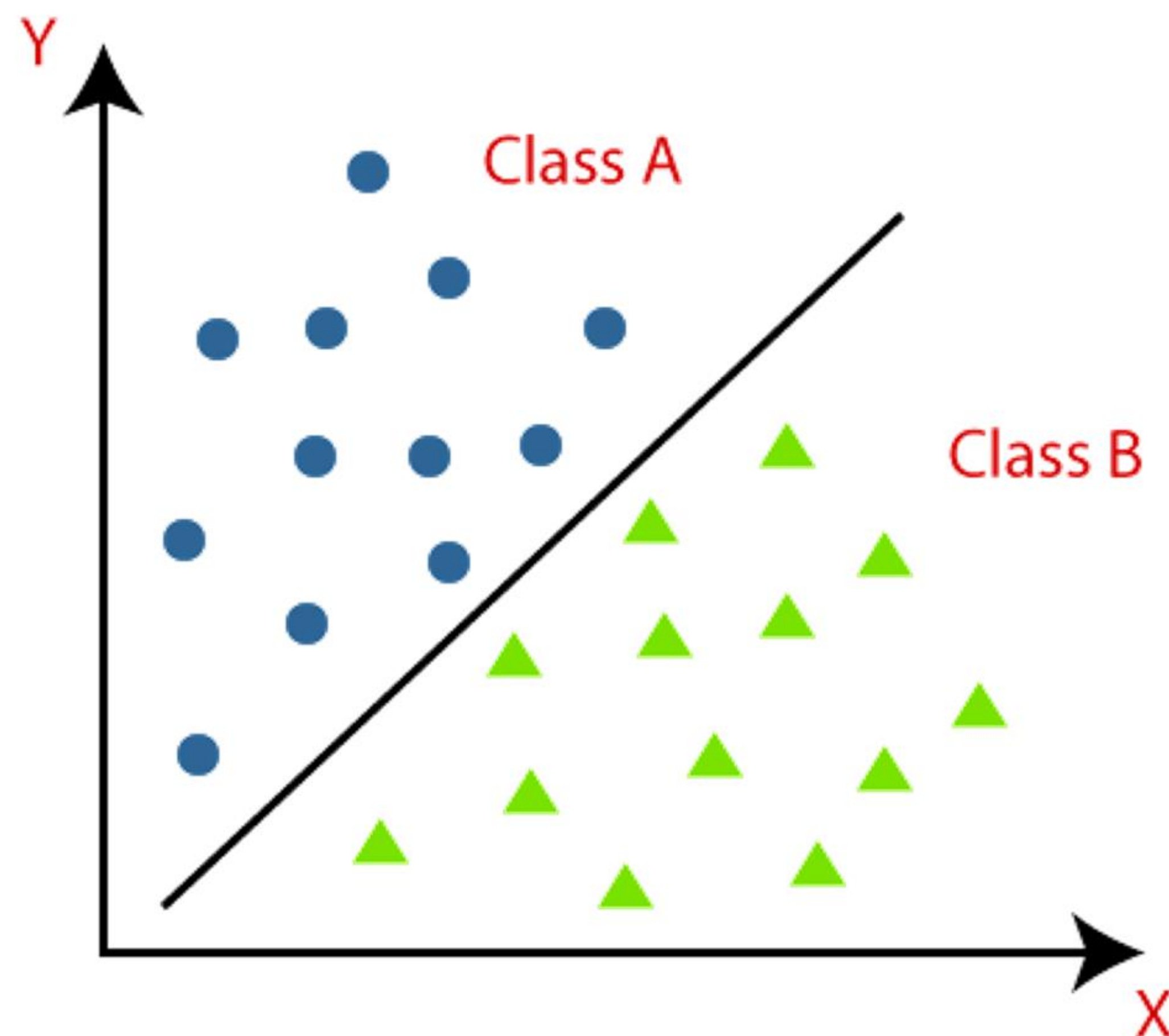
ДАННЫЕ ПРЕДСТАВЛЕНЫ ЗА ПЕРИОД 2001-2019 Г.

ДАТАСЕТ СОДЕРЖИТ 831 СТРОКИ И 32 СТОЛБЦА



# КЛАССИФИКАЦИЯ

---



— ЭТО ЗАДАЧА МАШИННОГО ОБУЧЕНИЯ, В КОТОРОЙ МОДЕЛЬ ПРОГНОЗИРУЕТ КАТЕГОРИЮ ИЛИ КЛАСС, К КОТОРОМУ ОТНОСИТСЯ НОВЫЙ НАБЛЮДАЕМЫЙ ОБЪЕКТ, НА ОСНОВЕ ЕГО ХАРАКТЕРИСТИК ИЛИ ПРИЗНАКОВ

В ОСНОВЕ КЛАССИФИКАЦИИ ЛЕЖИТ **ОБУЧЕНИЕ С УЧИТЕЛЕМ**, ГДЕ ДЛЯ КАЖДОГО ОБЪЕКТА ИМЕЕТСЯ ИЗВЕСТНАЯ МЕТКА КЛАССА

# МОДЕЛИ

---

1

## ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

ИСПОЛЬЗУЕТСЯ ДЛЯ ОЦЕНКИ ДИСКРЕТНЫХ ЗНАЧЕНИЙ, ТАКИХ КАК 0 И 1, ДА ИЛИ НЕТ. ПРЕДСКАЗЫВАЕТ ВЕРОЯТНОСТЬ ПРИНАДЛЕЖНОСТИ К ОДНОМУ ИЗ ДВУХ КЛАССОВ

2

## К – БЛИЖАЙШИХ СОСЕДЕЙ (KNN)

КЛАССИФИЦИРУЕТ ОБЪЕКТ НА ОСНОВЕ МЕТОК КЛАССОВ ЕГО БЛИЖАЙШИХ СОСЕДЕЙ В ПРОСТРАНСТВЕ ПРИЗНАКОВ

3

## СЛУЧАЙНЫЙ ЛЕС

СОЗДАЕТ МНОЖЕСТВО РЕШАЮЩИХ ДЕРЕВЬЕВ (НЕЗАВИСИМЫХ МОДЕЛЕЙ) И ИСПОЛЬЗУЕТ ИХ ДЛЯ ПРЕДСКАЗАНИЯ КЛАССОВ ОБЪЕКТОВ

# МЕТРИКИ ОЦЕНКИ КАЧЕСТВА

1

## ACCURACY

ДОЛЯ ПРАВИЛЬНО  
КЛАССИФИЦИРОВАННЫХ  
ОБЪЕКТОВ СРЕДИ ВСЕХ  
ОБЪЕКТОВ

2

## PRECISION

ДОЛЯ ИСТИННО  
ПОЛОЖИТЕЛЬНЫХ  
СРЕДИ ВСЕХ ОБЪЕКТОВ,  
КОТОРЫЕ МОДЕЛЬ  
КЛАССИФИЦИРУЕТ КАК  
ПОЛОЖИТЕЛЬНЫЕ

3

## RECALL

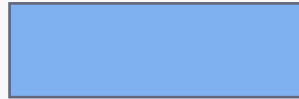


ДОЛЯ ИСТИННО  
ПОЛОЖИТЕЛЬНЫХ  
СРЕДИ ВСЕХ  
ДЕЙСТВИТЕЛЬНО  
ПОЛОЖИТЕЛЬНЫХ

4

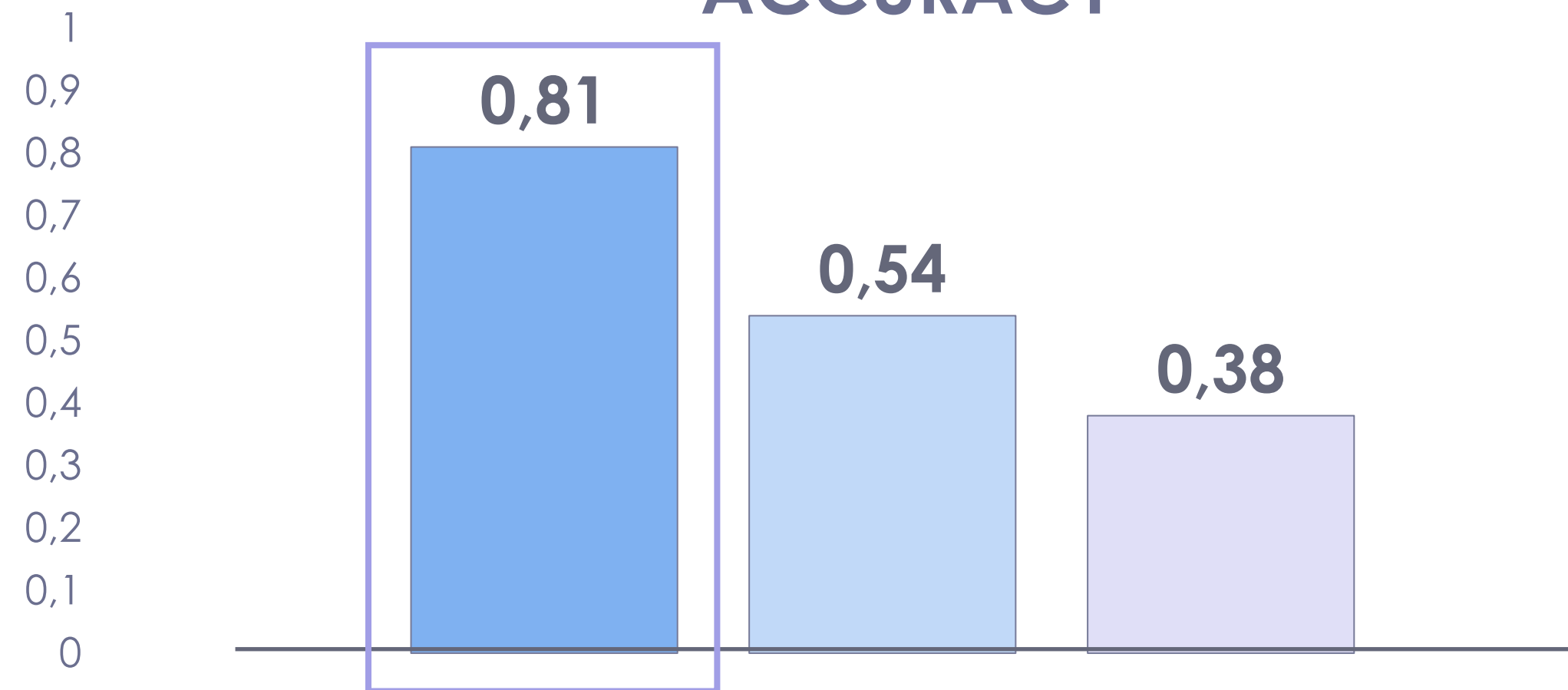
## F1-SCORE

ГАРМОНИЧЕСКОЕ  
СРЕДНЕЕ МЕЖДУ  
ТОЧНОСТЬЮ И  
ПОЛНОТОЙ

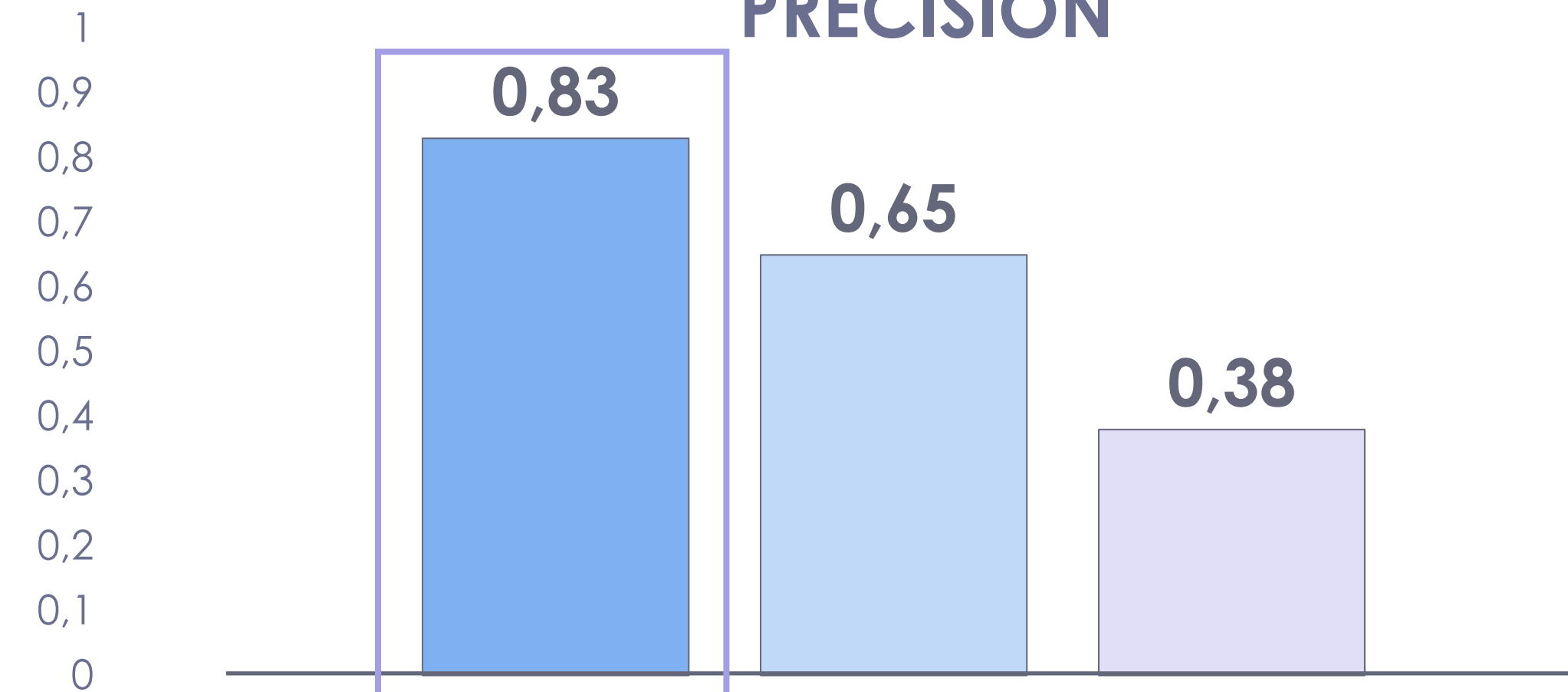
# РЕЗУЛЬТАТЫ

 - ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ  
 - KNN  
 - СЛУЧАЙНЫЙ ЛЕС

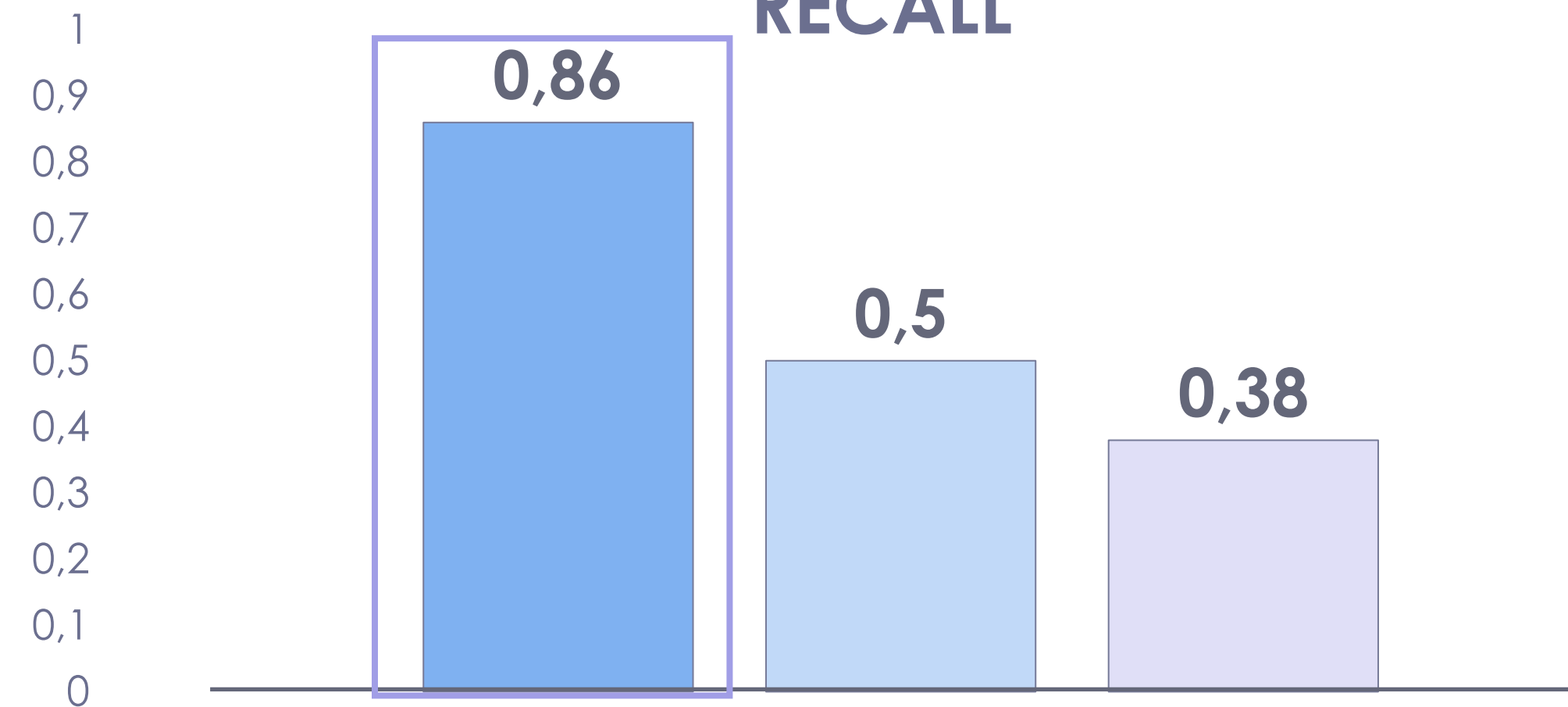
## ACCURACY



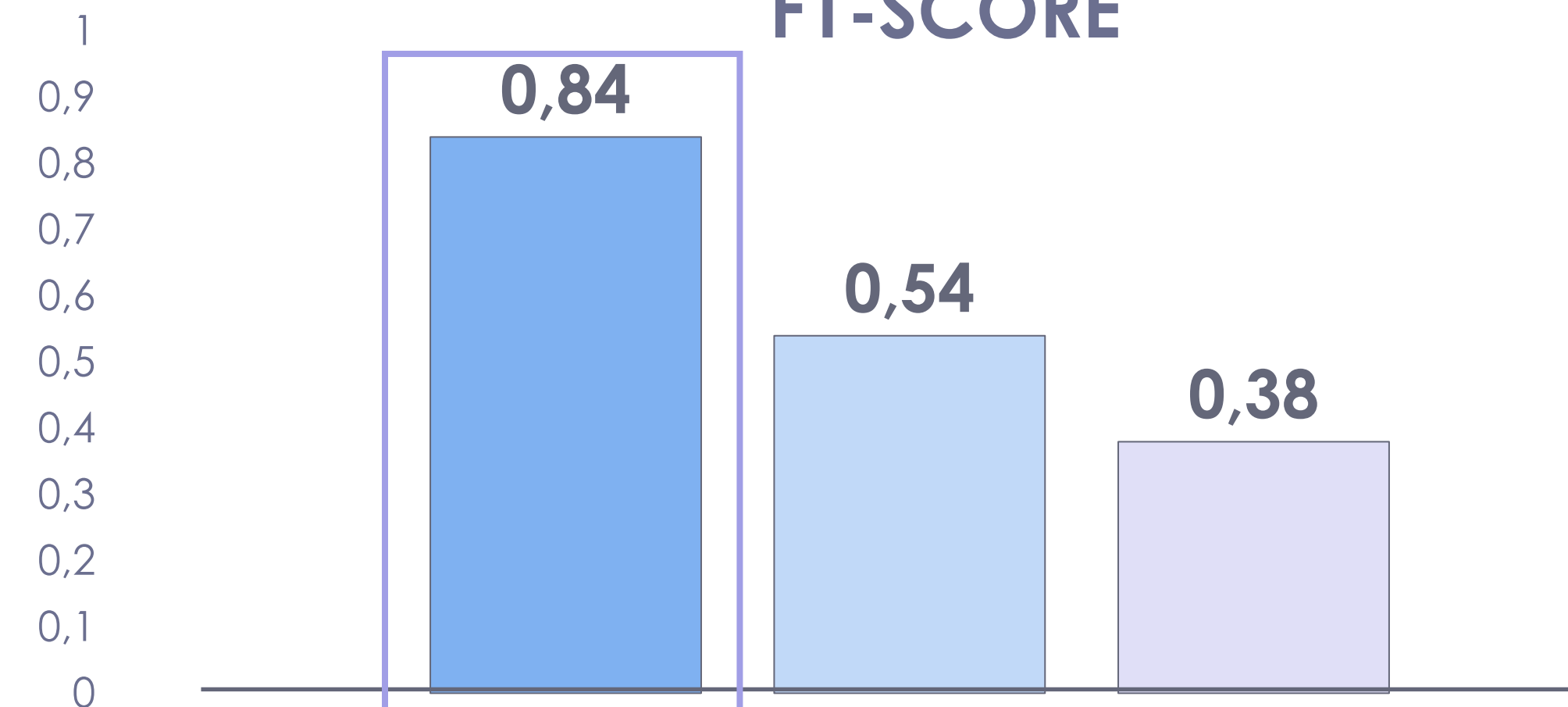
## PRECISION



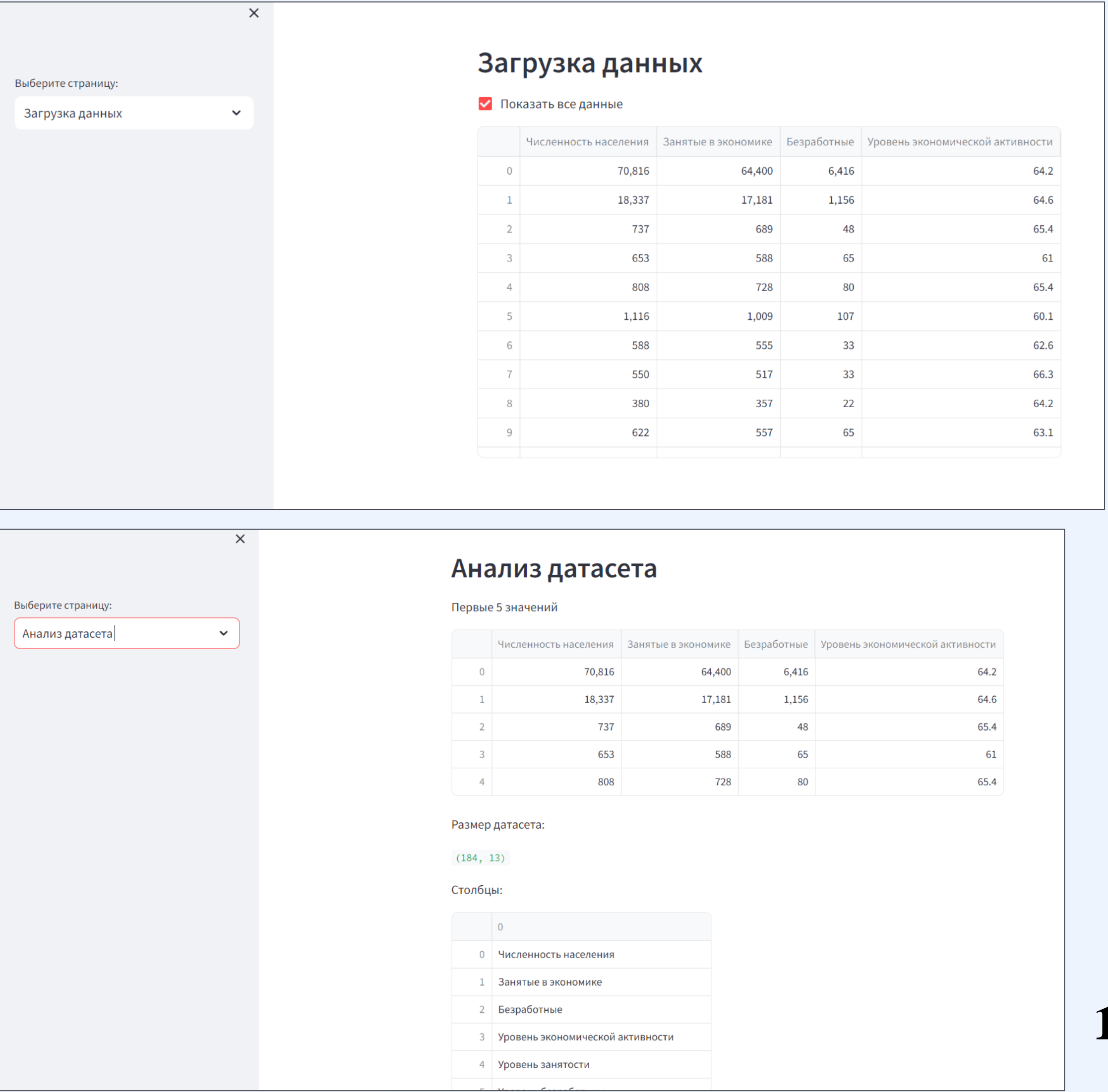
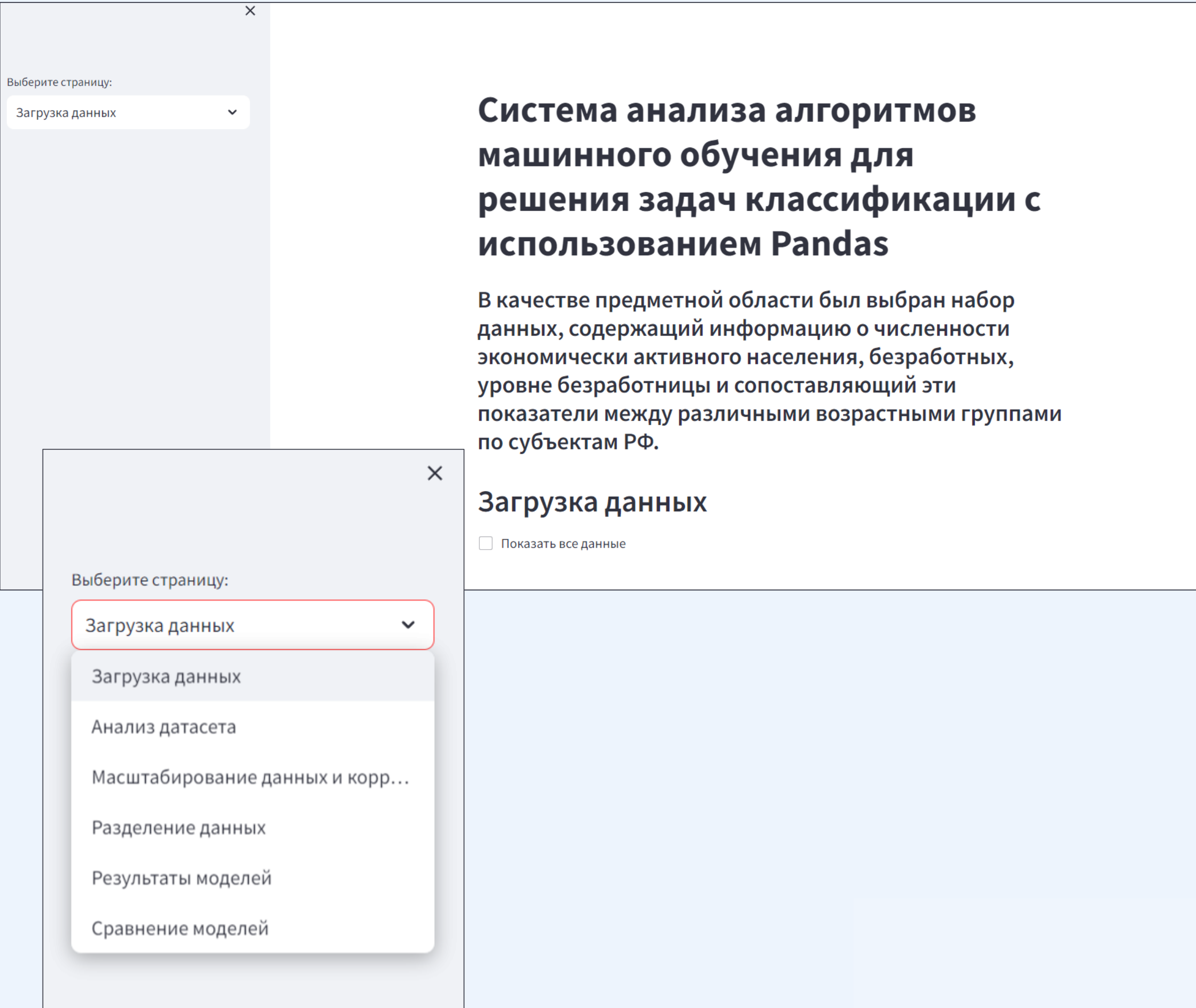
## RECALL



## F1-SCORE



# ИНТЕРФЕЙС ПОЛЬЗОВАТЕЛЯ





# ИНТЕРФЕЙС ПОЛЬЗОВАТЕЛЯ

Выберите страницу:

Масштабирование данных и к...

Масштабирование данных

☐ Показать данные

Корреляционный анализ данных

☐ Показать исходные данные (до масштабирования)

☐ Показать масштабированные данные

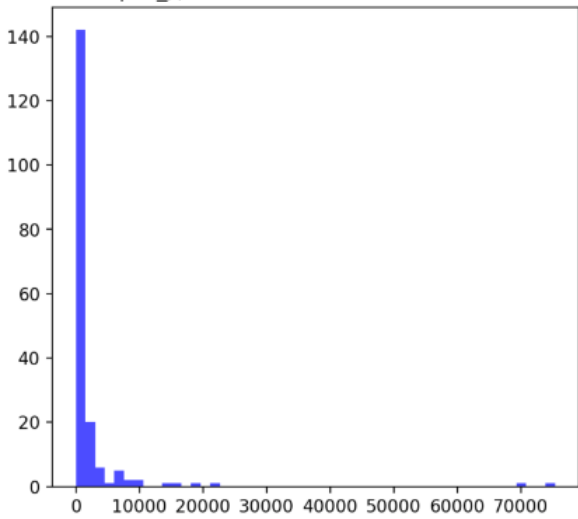
Выберите страницу:

Масштабирование данных и к...

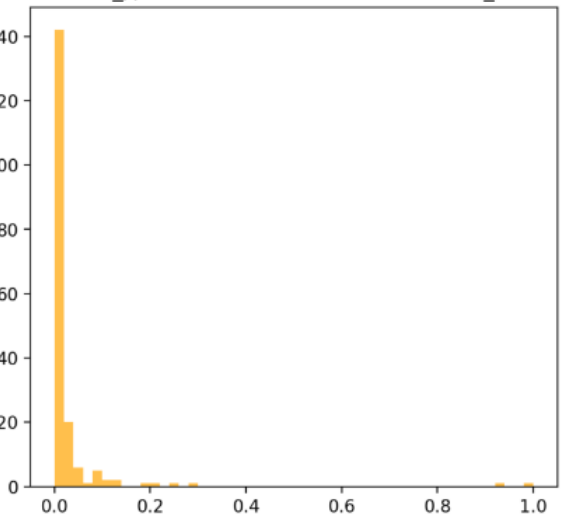
Масштабирование данных

☒ Показать данные


Ориг\_данные: Численность населения




Масшта\_данные: Численность населения\_scaled



Ориг\_данные: Занятые в экономике



Масшта\_данные: Занятые в экономике\_scaled



Выберите страницу:

Масштабирование данных и к...

☒ Показать исходные данные (до масштабирования)

Исходные данные (до масштабирования)

Численность населения	1.00	1.00	0.93	0.01	0.06	-0.08	0.80	0.94	0.94	0.89	0.99	0.98	0.00
Занятые в экономике	1.00	1.00	0.92	0.01	0.06	-0.09	0.78	0.94	0.93	0.88	0.98	0.97	0.01
Безработные	0.93	0.92	1.00	0.01	-0.00	0.02	0.96	1.00	1.00	0.99	0.97	0.98	-0.09
Уровень экономической активности	0.01	0.01	0.01	1.00	0.83	-0.09	0.04	0.01	0.01	0.02	-0.01	0.03	-0.27
Уровень занятости	0.06	0.06	-0.00	0.83	1.00	-0.62	-0.01	-0.00	-0.01	-0.00	0.01	0.05	0.07
Уровень безработицы	-0.08	-0.09	0.02	-0.09	-0.62	1.00	0.06	0.02	0.02	0.03	-0.03	-0.04	-0.51
Численность безработных (до 20 лет)	0.80	0.78	0.96	0.04	-0.01	0.06	1.00	0.94	0.95	0.98	0.87	0.89	-0.17
Численность безработных (от 20 до 29 лет)	0.94	0.94	1.00	0.01	-0.00	0.02	0.94	1.00	1.00	0.99	0.98	0.98	-0.08
Численность безработных (от 30 до 39 лет)	0.94	0.93	1.00	0.01	-0.01	0.02	0.95	1.00	1.00	0.99	0.97	0.98	-0.09
Численность безработных (от 40 до 49 лет)	0.89	0.88	0.99	0.02	-0.00	0.03	0.98	0.99	0.99	1.00	0.94	0.95	-0.12
Численность безработных (от 50 до 59 лет)	0.99	0.98	0.97	-0.01	0.01	-0.03	0.87	0.98	0.97	0.94	1.00	0.99	-0.03
Численность безработных (60 и более лет)	0.98	0.97	0.98	0.03	0.05	-0.04	0.89	0.98	0.98	0.95	0.99	1.00	-0.04
Год	0.00	0.01	-0.09	-0.27	0.07	-0.51	-0.17	-0.08	-0.09	-0.12	-0.03	-0.04	1.00

☒ Показать масштабированные данные

Масштабированные данные

Численность населения_scaled	1.00	1.00	0.93	0.01	0.06	-0.08	0.80	0.94	0.94	0.89	0.99	0.98	0.00
Занятые в экономике_scaled	1.00	1.00	0.92	0.01	0.06	-0.09	0.78	0.94	0.93	0.88	0.98	0.97	0.01
Безработные_scaled	0.93	0.92	1.00	0.01	-0.00	0.02	0.96	1.00	1.00	0.99	0.97	0.98	-0.09
Уровень экономической активности_scaled	0.01	0.01	0.01	1.00	0.83	-0.09	0.04	0.01	0.01	0.02	-0.01	0.03	-0.27

# ИНТЕРФЕЙС ПОЛЬЗОВАТЕЛЯ

Выберите страницу:

Разделение данных

### Разделение данных на обучающую и тестовую выборки

Размер обучающей выборки (147, 3) (147,)

Размер тестовой выборки (147, 3) (147,)

Выберите страницу:

Сравнение моделей

### Сравнение моделей

Accuracy Сравнение

Model	Accuracy
Logistic Regression	0.81
KNN Model	0.54
Random Forest	0.43

Precision Сравнение

Выберите страницу:

Оценка моделей

Выберите модель для оценки:

Logistic Regression

Logistic Regression

KNN

Random Forest

### Оценка обученных моделей

Модель: Logistic Regression

Accuracy: {0.8108108108108109}

Precision: {0.8260869565217391}

Recall: {0.8636363636363636}

F1 Score: {0.8444444444444444}



# ЗАКЛЮЧЕНИЕ

---

- ИССЛЕДОВАНА ПРЕДМЕТНАЯ ОБЛАСТЬ
- ВЫБРАНЫ АЛГОРИТМЫ КЛАССИФИКАЦИИ
- ОБУЧЕНЫ МОДЕЛИ
- ПРОВЕДЕНА ОЦЕНКА КАЖДОЙ МОДЕЛИ
- ПРОВЕДЕНО СРАВНЕНИЕ МОДЕЛЕЙ И  
ВЫЯСНЕНО, КАКАЯ ИЗ НИХ ЛУЧШАЯ
- СИСТЕМА РЕАЛИЗОВАНА В ВИДЕ  
ВЕБ-ПРИЛОЖЕНИЯ