# Классификация пассажиров Титаника: модель SVM

Практическое применение метода опорных векторов для предсказания выживаемости пассажиров на основе исторических данных

# Постановка задачи классификации

## Что такое классификация?

Задача машинного обучения с учителем, где модель определяет принадлежность объекта к одному из заранее известных классов на основе его признаков.

## Целевая переменная

**Survived** — бинарный класс {0, 1}, где 1 означает выживание пассажира, 0 — гибель.

## Датасет Titanic

Исторические данные о 891 пассажире легендарного лайнера с информацией о социальном статусе, возрасте, поле и других характеристиках.

## Почему SVM?

- Строгая разделяющая гиперплоскость
- Эффективность на небольших выборках
- Устойчивость к переобучению

# Описание признаков датасета

### Исходные признаки

- **Age** — возраст пассажира
- **Sex** — пол
- **Pclass** — класс билета (1, 2, 3)
- **SibSp** — число братьев/сестер
- **Parch** — число родителей/детей
- **Fare** — стоимость билета
- **Embarked** — порт посадки

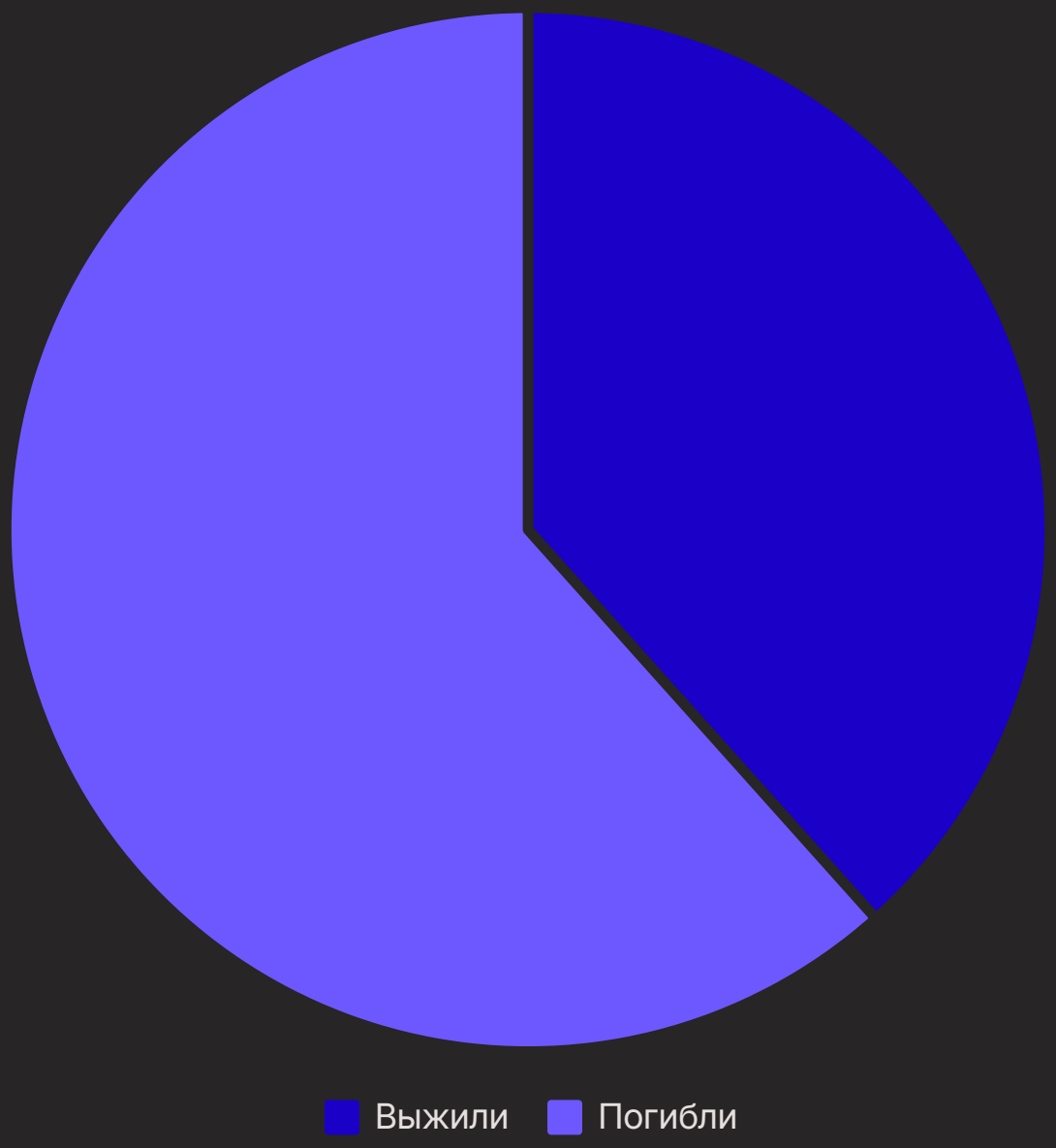
### Удаленные признаки

Текстовые поля, не несущие структурированной информации:

- **Name** — полное имя
- **Ticket** — номер билета
- **Cabin** — номер каюты (много пропусков)

### Созданные признаки

- **FamilySize** — размер семьи
- **IsAlone** — путешествует один
- **Title** — обращение из имени
- **Deck** — палуба из Cabin



# Предобработка данных

01

## Обработка пропусков

**Age** — заполнение медианным значением по группам Pclass и Sex для сохранения демографических закономерностей.

**Embarked** — заполнение модой (наиболее частым портом посадки).

02

## Кодирование категорий

**OneHotEncoder** для категориальных признаков: Sex, Embarked, Title. Преобразование в бинарные векторы для корректной работы SVM.

03

## Масштабирование

**StandardScaler** для числовых признаков: приведение к нулевому среднему и единичной дисперсии. Критично для SVM, чувствительного к масштабу.

04

## Разделение выборки

Train/Test split 80/20 с **stratify** по целевой переменной для сохранения пропорций классов в обеих выборках.

# Feature Engineering: создание новых признаков

1

### FamilySize

FamilySize = SibSp + Parch + 1

Общий размер семьи включая самого пассажира. Гипотеза: большие семьи имели разные шансы на спасение.

2

### IsAlone

IsAlone = (FamilySize == 1)

Бинарный признак одиночества. Одинокие путешественники могли действовать иначе в критической ситуации.

3

### Title

Извлечение из Name: Mr, Mrs, Miss, Master, др.

Социальный статус и пол, влиявшие на приоритет эвакуации.

4

### Age×Pclass

Интерактивный признак

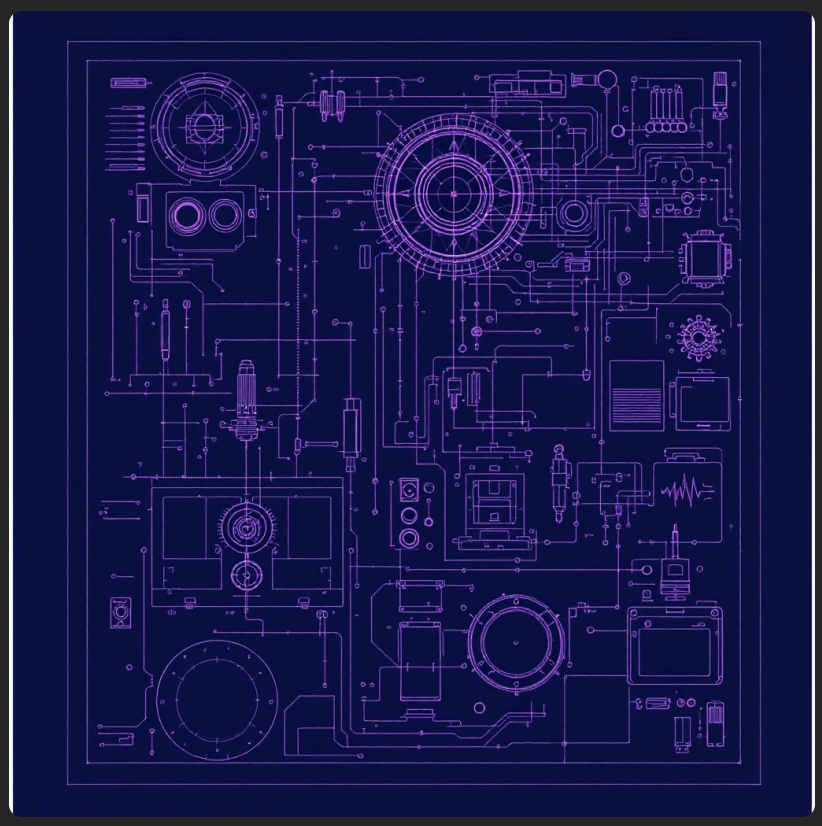
Взаимодействие возраста и класса билета для выявления нелинейных зависимостей.

5

### FarePerPerson

Fare / FamilySize

Нормализованная стоимость на человека для корректного сравнения семей разного размера.



**Почему это важно для SVM?**

Метод опорных векторов чувствителен к информативности признаков. Создание новых комбинаций помогает модели находить более четкую разделяющую гиперплоскость между классами.

# SVM: теоретические основы

## Разделяющая гиперплоскость

SVM строит оптимальную линейную границу между классами в многомерном пространстве признаков, максимизируя расстояние до ближайших точек обоих классов.

## Максимизация margin

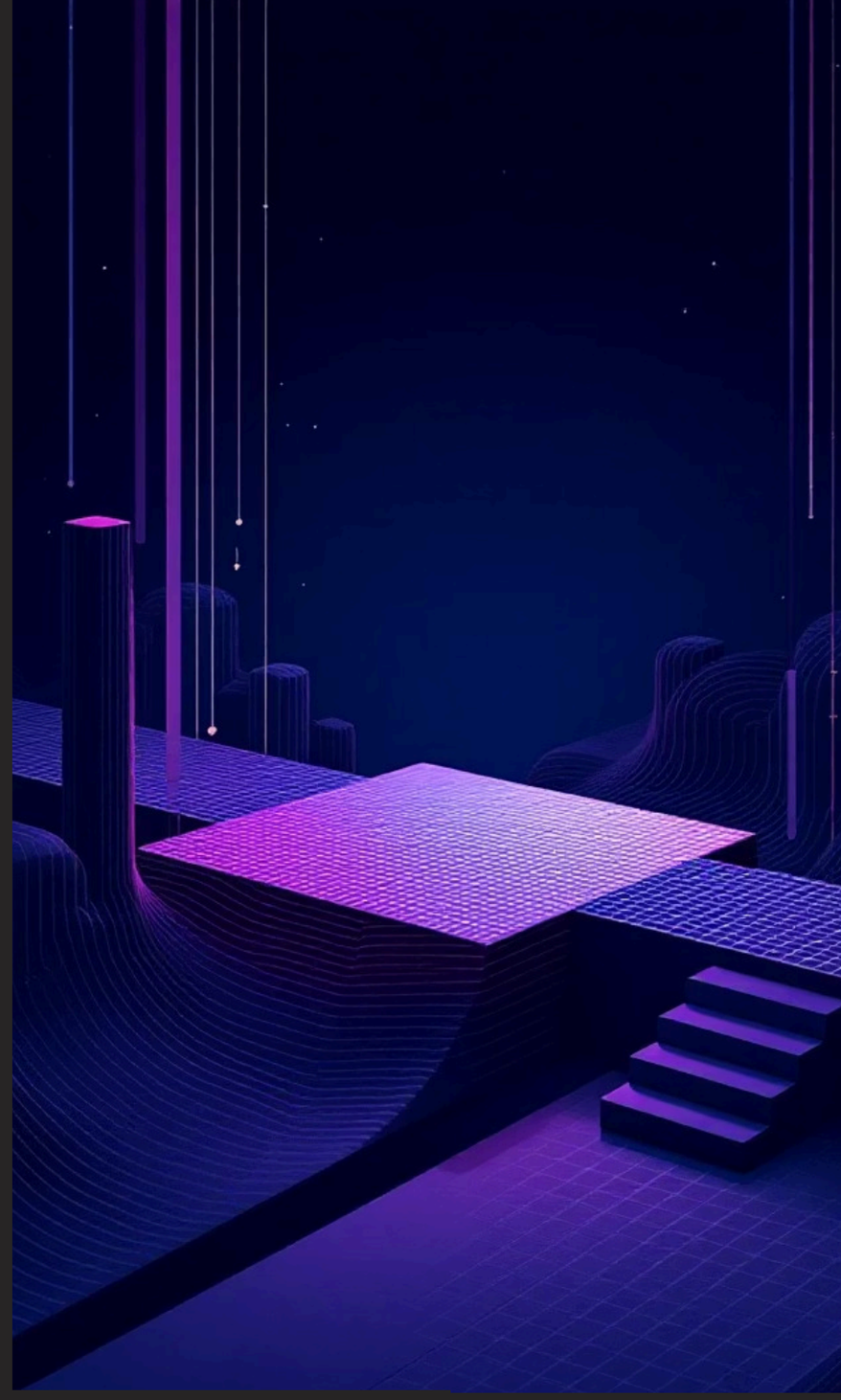
Ключевая идея: выбор такой гиперплоскости, которая максимально удалена от опорных векторов обоих классов. Это обеспечивает лучшую обобщающую способность модели.

## Параметр C

Коэффициент штрафа за неправильную классификацию. Большое C → жесткая граница, малое C → более мягкая граница с допустимыми ошибками.

## Функция ядра (kernel)

Преобразование исходного пространства в пространство более высокой размерности. **Linear** — для линейно разделимых данных, **RBF** — для сложных нелинейных зависимостей.



# Настройка гиперпараметров SVM

## Варьируемые параметры

### Параметр C

Диапазон: [0.01, 0.1, 1, 10, 100]

Контролирует баланс между максимизацией margin и минимизацией ошибок классификации на обучающей выборке.

### Тип ядра (kernel)

- **linear** — линейное разделение
- **rbf** — радиальная базисная функция
- **poly** — полиномиальное ядро

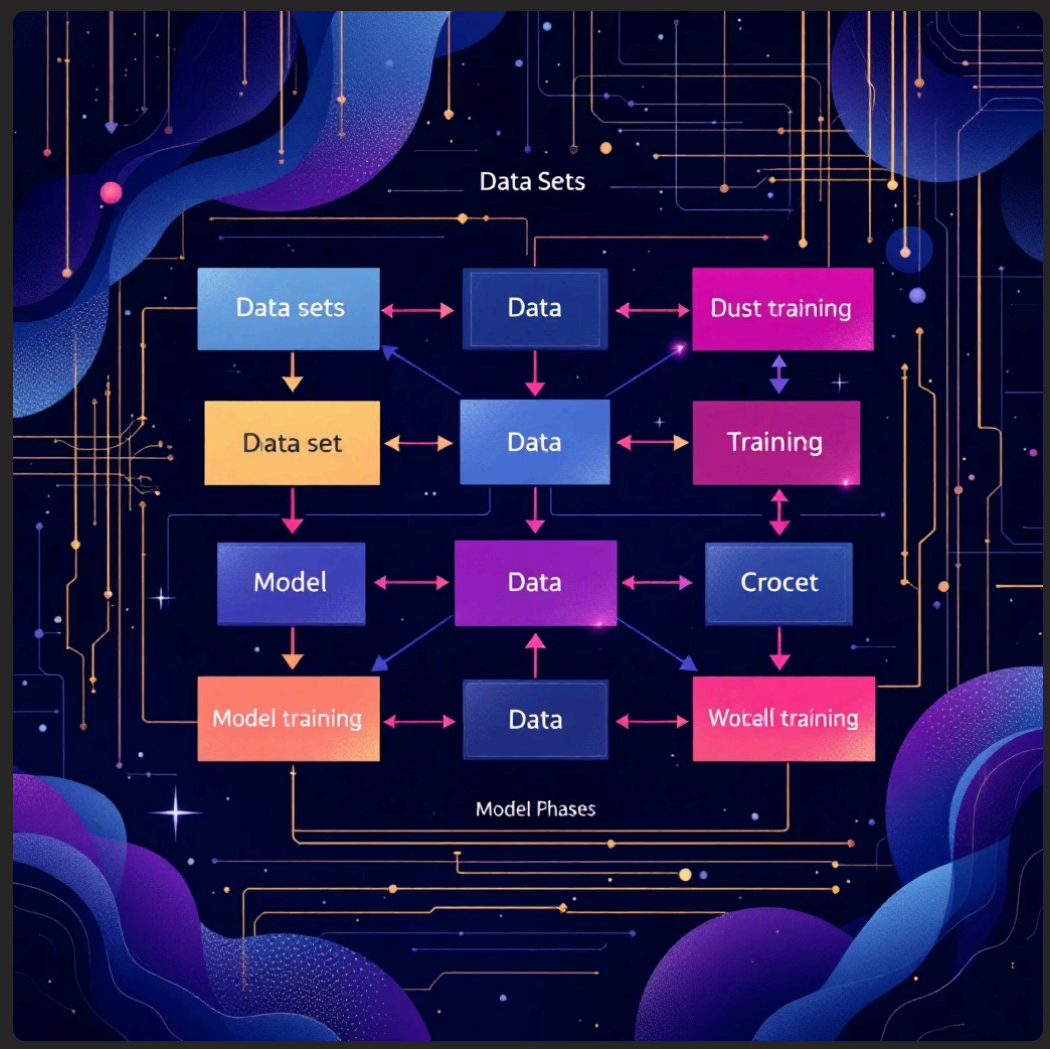
### Параметр gamma

Диапазон: [0.001, 0.01, 0.1, 1]

Для RBF-ядра: определяет радиус влияния одного обучающего примера. Малое gamma → широкое влияние, большое → узкое.

## Кросс-валидация

Применялся метод **5-fold cross-validation** для объективной оценки качества каждой комбинации параметров и предотвращения переобучения.




Сетка параметров перебиралась с помощью GridSearchCV для поиска оптимальной конфигурации по метрике F1-score.

# Результаты модели SVM

## Финальные метрики качества

**83.2%**

**Accuracy**

Общая точность предсказаний на тестовой выборке

**81.7%**

**Precision**

Доля корректно предсказанных выживших среди всех предсказанных как выжившие

**79.4%**

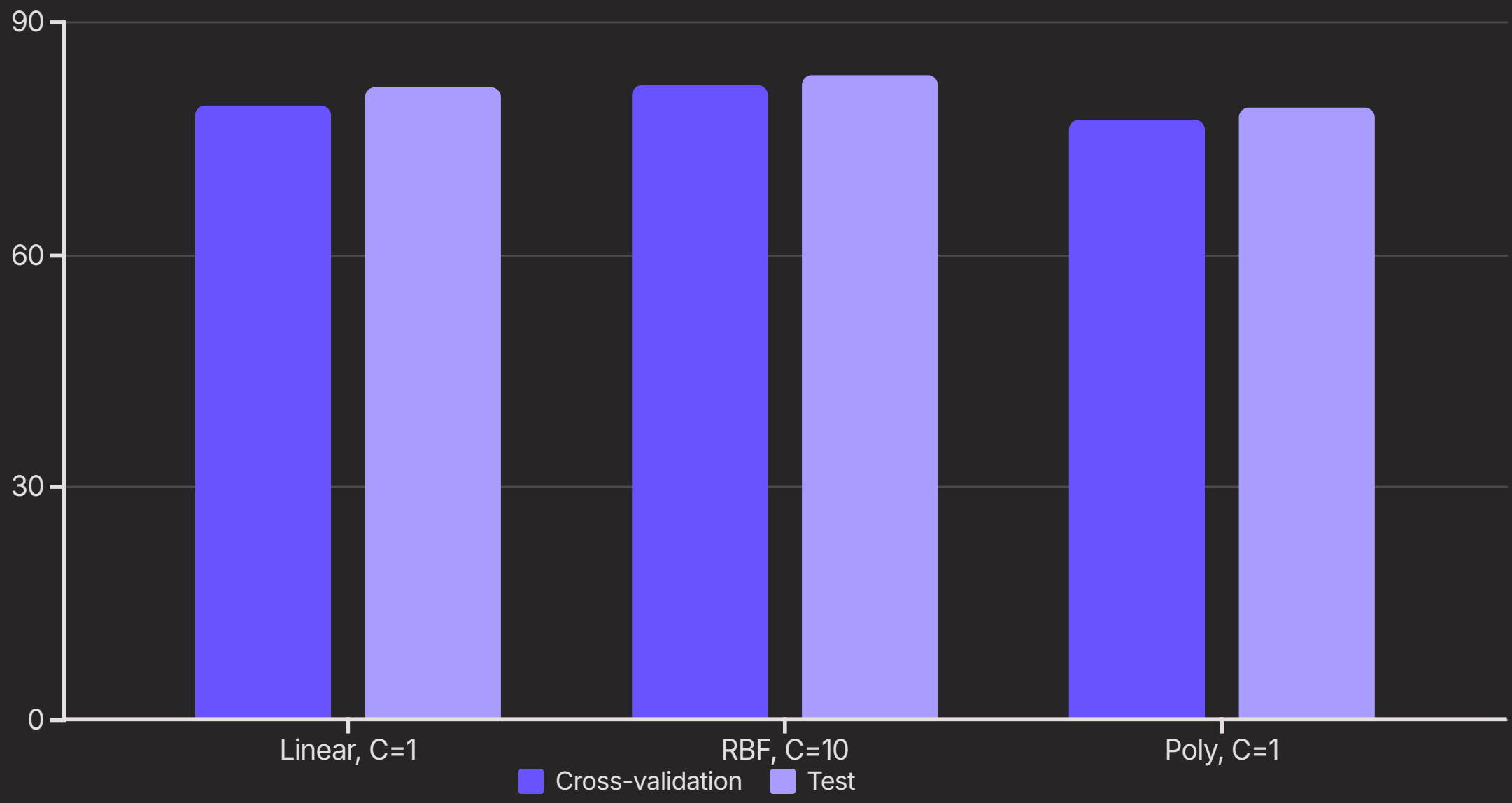
**Recall**

Доля найденных выживших среди всех реально выживших пассажиров

**80.5%**

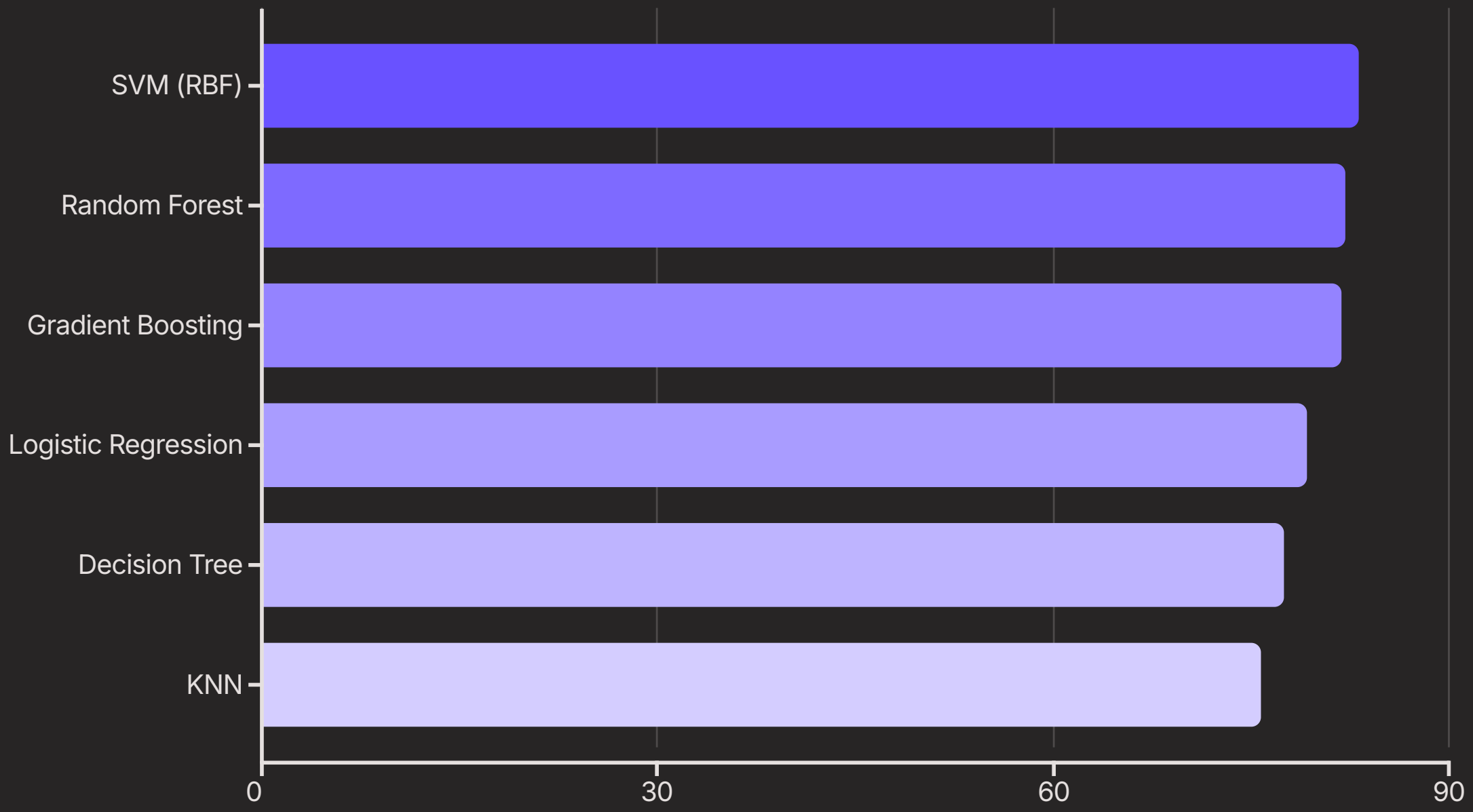
**F1-score**

Гармоническое среднее между Precision и Recall




Лучшие результаты показала конфигурация с RBF-ядром и параметром C=10, демонстрирующая стабильность на кросс-валидации и тестовой выборке.

# Сравнение SVM с другими моделями




## Преимущества SVM

- Наивысшая точность среди протестированных моделей
- Стабильность результатов на кросс-валидации
- Устойчивость к переобучению при небольшом объеме данных
- Эффективная работа в пространстве высокой размерности

## Наблюдения

Random Forest и Gradient Boosting показали сопоставимые результаты, но требовали больше времени на обучение. Логистическая регрессия уступает из-за предположения о линейности. KNN чувствителен к масштабу и зашумленности данных.

# Выводы и перспективы развития

- **Ключевые достижения**

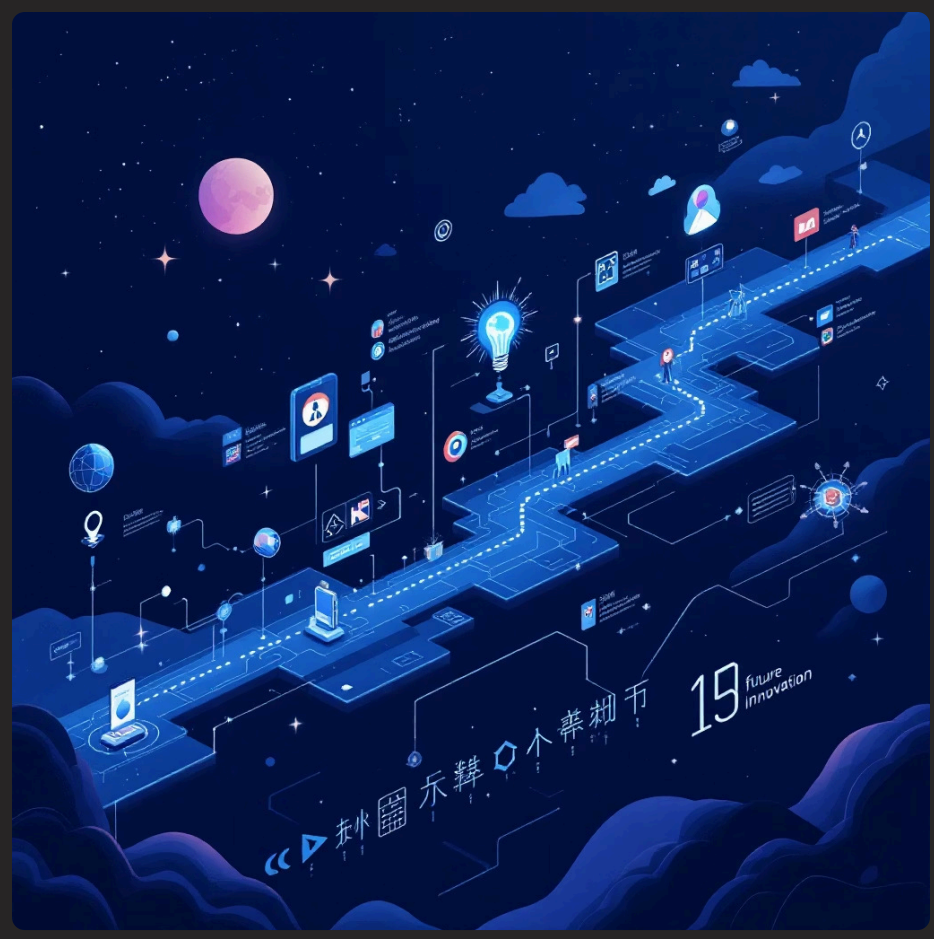
  Модель SVM с RBF-ядром достигла **83.2% точности**, демонстрируя стабильные результаты и превосходя альтернативные алгоритмы на данной задаче.

- **Критические преобразования**

  Нормализация данных, OneHotEncoding категориальных признаков и создание новых признаков (Title, FamilySize, IsAlone) существенно улучшили качество модели.

- **Устойчивость к малым данным**

  SVM показал эффективность на относительно небольшой выборке благодаря механизму максимизации margin и опорным векторам.




## Направления для улучшения

### Оптимизация ядер

Тестирование кастомных функций ядра и более детальная настройка параметров gamma

### Новые признаки

Извлечение информации из каюты, анализ взаимодействий между признаками высших порядков

### Ансамбли

Комбинирование SVM с другими моделями для повышения обобщающей способности

### Deep Learning

Сравнение с нейронными сетями при расширении датасета