Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»

(УНИВЕРСИТЕТ ИТМО)

Факультет «Систем управления и робототехники»

ОТЧЕТ

О ПРОЕКТНОЙ РАБОТЕ

По дисциплине «Математическая статистика»

«Сравнение методов регрессии: линейная регрессия vs. логистическая регрессия vs. метод опорных векторов»

Студенты потока Мат Стат 23:
Симонов Илья Александрович, R3236, ИСУ 409560 - аналитика данных, постановка задачи, отчёт, визуализация
Мезенцева Варвара Александровна, R3243, ИСУ 413787 - реализация моделей, программная часть, подготовка данных

Преподаватель:
Юрова Татьяна Сергеевна

Санкт-Петербург
2025

Содержание

1 Постановка задачи

Целью данной работы является сравнительный анализ трёх методов машинного обучения: линейной регрессии, логистической регрессии и метода опорных векторов (SVM).

Основные задачи исследования:

1. Изучить теоретические основы каждого метода
2. Реализовать алгоритмы на языке Python с использованием библиотеки scikit-learn
3. Провести эксперименты на четырёх различных наборах данных
4. Сравнить производительность методов по различным метрикам качества
5. Проанализировать области применимости каждого метода

Для исследования используются следующие наборы данных:

- **Insurance** — предсказание стоимости медицинской страховки (регрессия)
- **Wine Quality** — оценка качества вина (регрессия и классификация)
- **Titanic** — предсказание выживаемости пассажиров (классификация)
- **Heart Disease** — диагностика заболеваний сердца (классификация)

2 Краткая теория

2.1 Линейная регрессия

Линейная регрессия — это метод статистического анализа, который моделирует зависимость между зависимой переменной y и одной или несколькими независимыми переменными $x_1, x_2, \ldots, x_n$.

Модель имеет вид:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

где β_i — коэффициенты регрессии, ε — случайная ошибка.

Коэффициенты находятся методом наименьших квадратов:

$$\min_{\beta} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

2.2 Логистическая регрессия

Логистическая регрессия используется для задач бинарной классификации. Она моделирует вероятность принадлежности объекта к определённому классу.

Модель основана на логистической функции:

$$p(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n)}}$$

Коэффициенты находятся методом максимального правдоподобия.

2.3 Метод опорных векторов (SVM)

SVM — это метод машинного обучения, который может использоваться как для классификации (SVC), так и для регрессии (SVR).

Основная идея SVM заключается в поиске оптимальной разделяющей гиперплоскости, которая максимизирует отступ между классами.

Для нелинейно разделимых данных используется kernel trick с различными ядрами (RBF, полиномиальное, линейное).

3 Используемые данные

3.1 Insurance Dataset

- **Размер**: 1338 записей, 7 признаков
- **Цель**: предсказание стоимости медицинской страховки (charges)
- **Признаки**: возраст, пол, BMI, количество детей, курение, регион
- **Тип задачи**: регрессия

3.2 Wine Quality Dataset

- **Размер**: 6497 записей, 12 признаков
- **Цель**: оценка качества вина (quality, 0-10)
- **Признаки**: кислотность, сахар, pH, алкоголь и др.
- **Тип задачи**: регрессия и классификация (бинаризация по порогу 6)

3.3 Titanic Dataset

- **Размер**: 1309 записей, 11 признаков
- **Цель**: предсказание выживаемости (Survived)
- **Признаки**: класс билета, возраст, пол, количество родственников
- **Тип задачи**: бинарная классификация

3.4 Heart Disease Dataset

- **Размер**: 1025 записей, 13 признаков
- **Цель**: диагностика заболевания сердца (target)
- **Признаки**: возраст, пол, давление, холестерин и др.
- **Тип задачи**: бинарная классификация

4 Описание моделей и кода

Проект реализован на языке Python с использованием следующих библиотек:

- `pandas` — работа с данными
- `scikit-learn` — алгоритмы машинного обучения
- `matplotlib, seaborn` — визуализация
- `numpy` — численные вычисления

Исходный код проекта доступен в GitHub репозитории: https://github.com/IlyaKonFetka/MatStatProject.git

4.1 Структура проекта

- `01_data_analysis.py` — исследовательский анализ данных
- `02_preprocessing.py` — предобработка данных
- `03_linear_regression.py` — линейная регрессия
- `04_logistic_regression.py` — логистическая регрессия
- `05_svm.py` — метод опорных векторов
- `06_evaluation.py` — оценка и сравнение моделей
- `main_pipeline.py` — основной пайплайн

4.2 Предобработка данных

Для всех датасетов применялись следующие шаги:

1. Удаление строк с пропущенными значениями
2. Разделение на числовые и категориальные признаки
3. Стандартизация числовых признаков (StandardScaler)
4. One-Hot кодирование категориальных признаков
5. Разделение на обучающую и тестовую выборки (80/20)

Предобработка данных является критически важным этапом в машинном обучении. Стандартизация признаков необходима особенно для алгоритмов, чувствительных к масштабу данных, таких как SVM. One-Hot кодирование позволяет работать с категориальными переменными, преобразуя их в числовой формат без создания ложной упорядоченности.

4.3 Детальное описание алгоритмов

4.3.1 Линейная регрессия

Линейная регрессия является фундаментальным методом статистического анализа. В данной работе использовалась стандартная реализация из scikit-learn, которая применяет метод наименьших квадратов для нахождения оптимальных коэффициентов.

Преимущества метода:

- Простота интерпретации
- Быстрота обучения и предсказания
- Отсутствие гиперпараметров для настройки
- Хорошая работа при линейных зависимостях

Недостатки:

- Предположение о линейности связи
- Чувствительность к выбросам
- Проблемы с мультиколлинеарностью

4.3.2 Логистическая регрессия

Логистическая регрессия адаптирует принципы линейной регрессии для задач классификации. Использовалась реализация с максимальным количеством итераций 1000 для обеспечения сходимости на всех датасетах.

Особенности реализации:

- Использование функции потерь логарифмического правдоподобия
- Применение сигмоидной функции активации
- Регуляризация L2 по умолчанию для предотвращения переобучения

4.3.3 Метод опорных векторов

SVM реализован в двух вариантах: SVC для классификации и SVR для регрессии. Использовалось RBF-ядро как наиболее универсальное для нелинейных зависимостей.

Ключевые параметры:

- Kernel='rbf' — радиальная базисная функция
- Probability=True для SVC — получение вероятностных оценок
- Параметры C и gamma по умолчанию из scikit-learn

4.4 Метрики оценки

4.4.1 Метрики регрессии

RMSE (Root Mean Square Error):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

RMSE показывает среднеквадратичную ошибку предсказания в тех же единицах, что и целевая переменная. Чем меньше RMSE, тем лучше модель.

R^2 (коэффициент детерминации):

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

R^2 показывает долю дисперсии, объясняемую моделью. Значения близкие к 1 указывают на хорошее качество модели.

4.4.2 Метрики классификации

Accuracy: доля правильных предсказаний от общего числа предсказаний.

F1-score: гармоническое среднее между точностью (precision) и полнотой (recall):

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

AUC (Area Under Curve): площадь под ROC-кривой, характеризующая способность модели различать классы.

5 Графики и визуализации

5.1 Корреляционные матрицы датасетов

Корреляционные матрицы показывают взаимосвязи между числовыми признаками в каждом датасете.

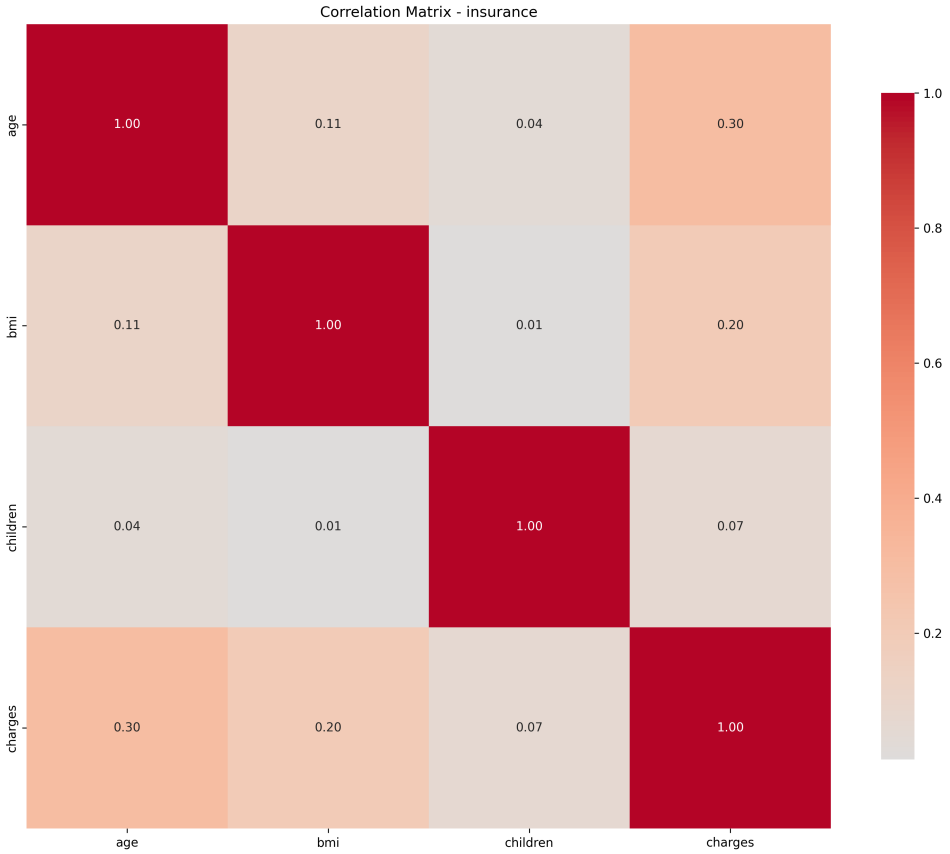

Рис. 1: Корреляционная матрица — Insurance Dataset

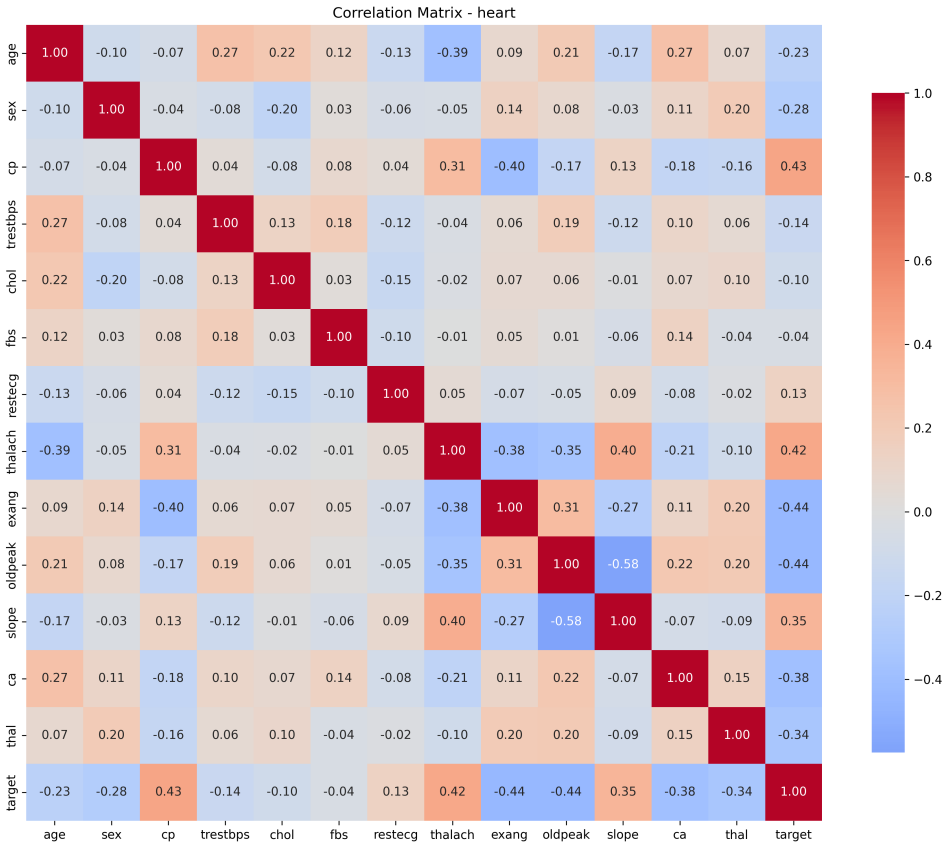

Рис. 2: Корреляционная матрица — Heart Disease Dataset

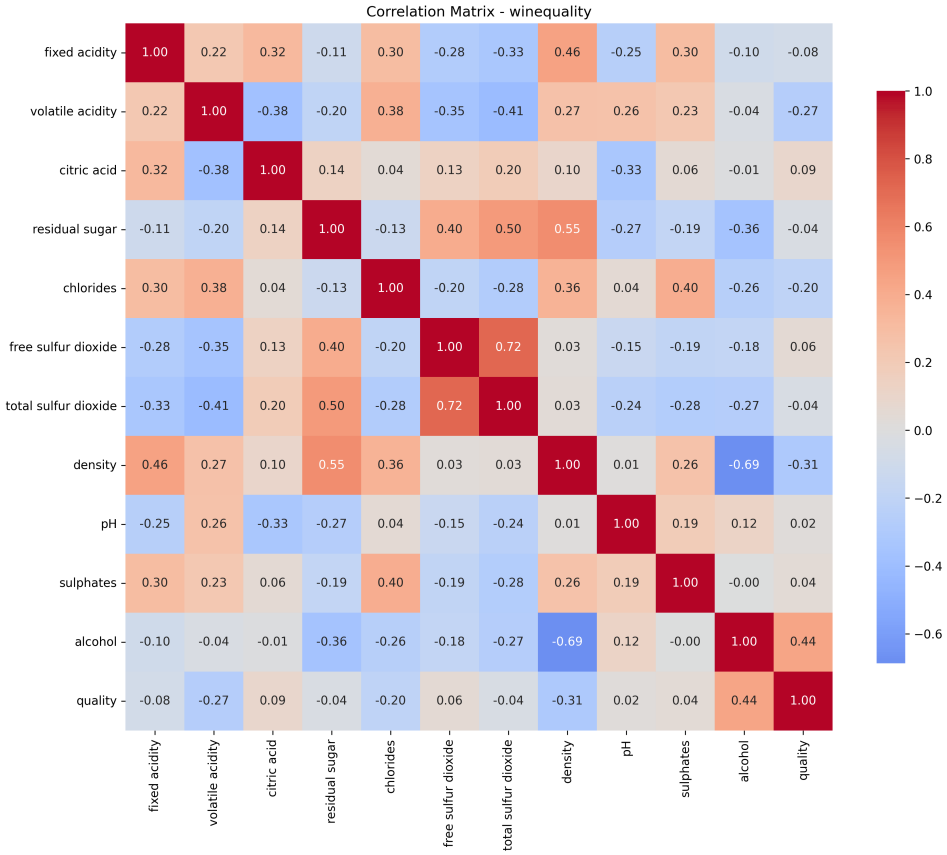

Рис. 3: Корреляционная матрица — Wine Quality Dataset

5.2 Распределения целевых переменных

Распределения целевых переменных показывают характер задач для каждого датасета.

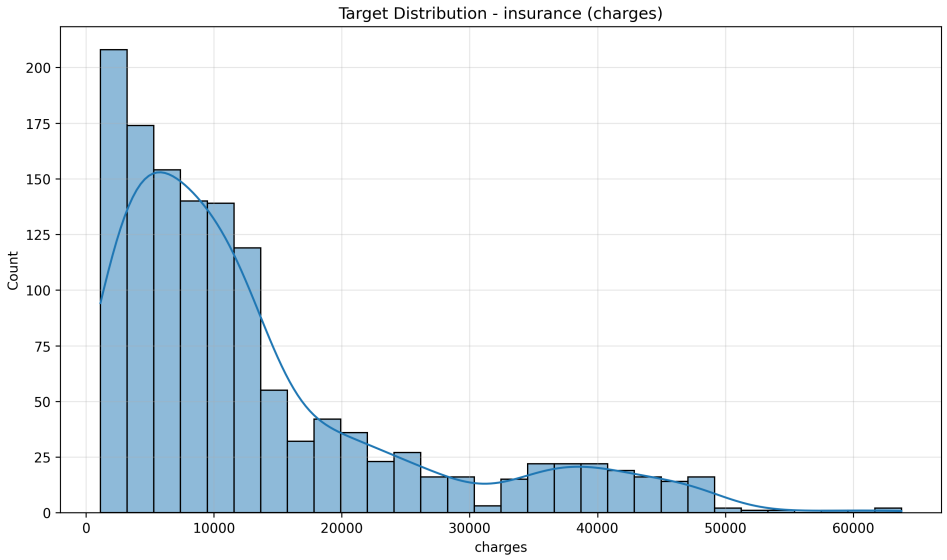

Рис. 4: Распределение стоимости страховки (Insurance)

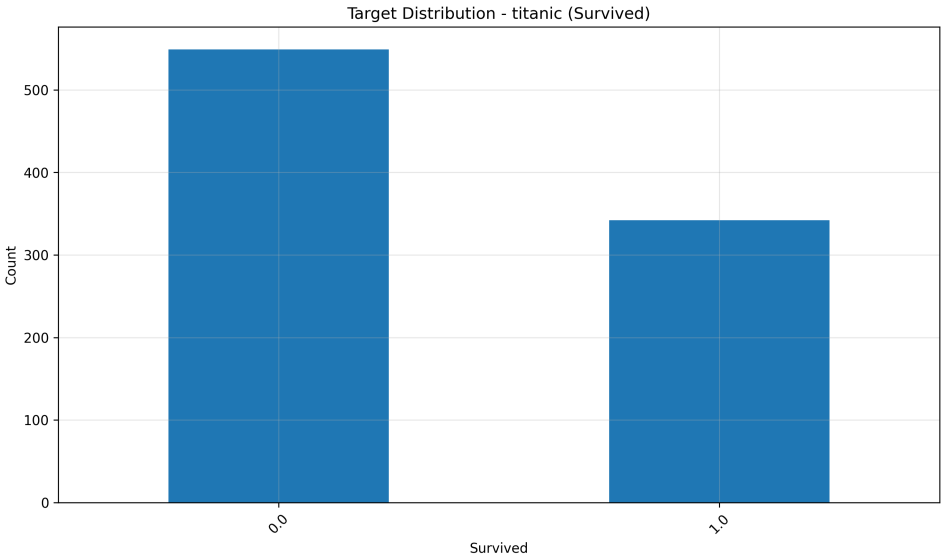

Рис. 5: Распределение выживаемости (Titanic)

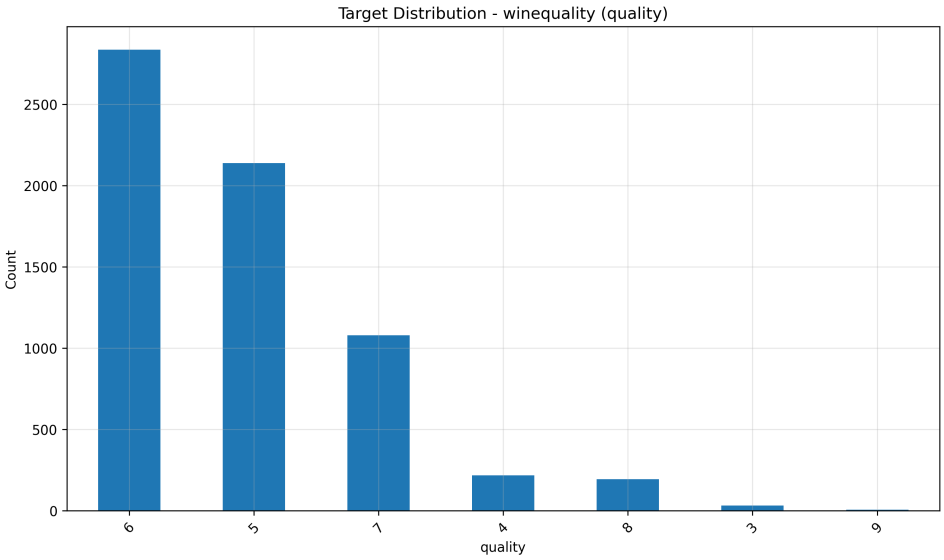

Рис. 6: Распределение качества вина (Wine Quality)

5.3 Примеры гистограмм признаков

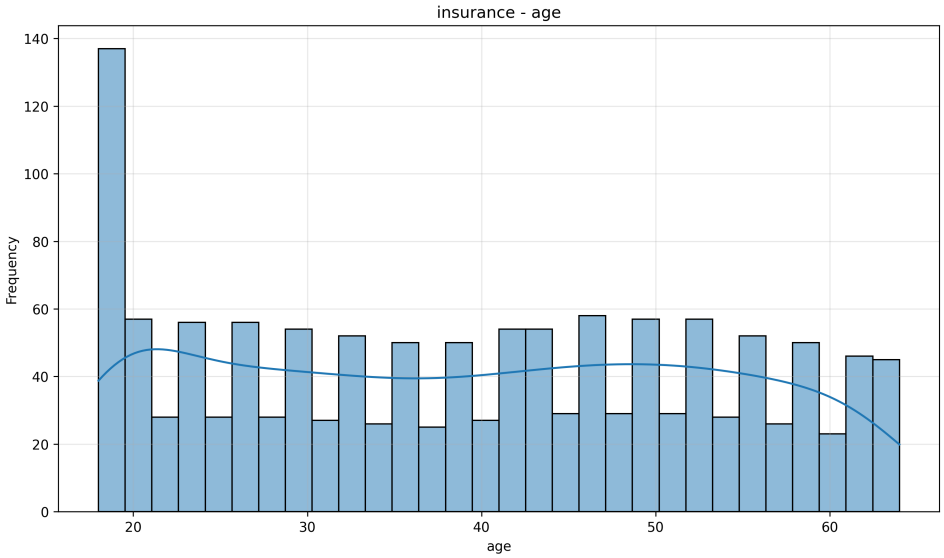

Рис. 7: Распределение возраста — Insurance Dataset

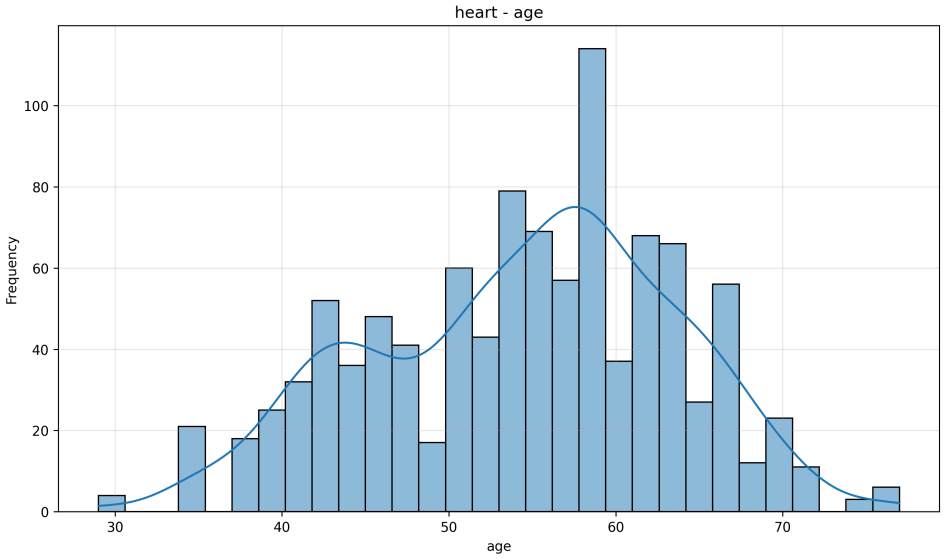

Рис. 8: Распределение возраста — Heart Disease Dataset

5.4 ROC-кривые для моделей классификации

ROC-кривые демонстрируют качество классификации для разных методов.

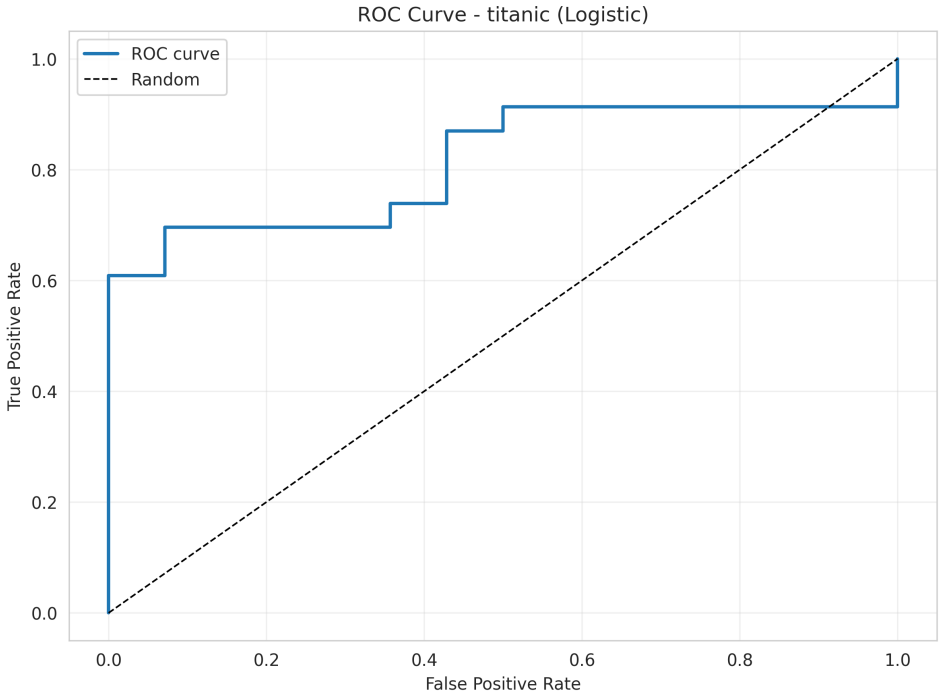

Рис. 9: ROC-кривая — Titanic Dataset (Logistic Regression)

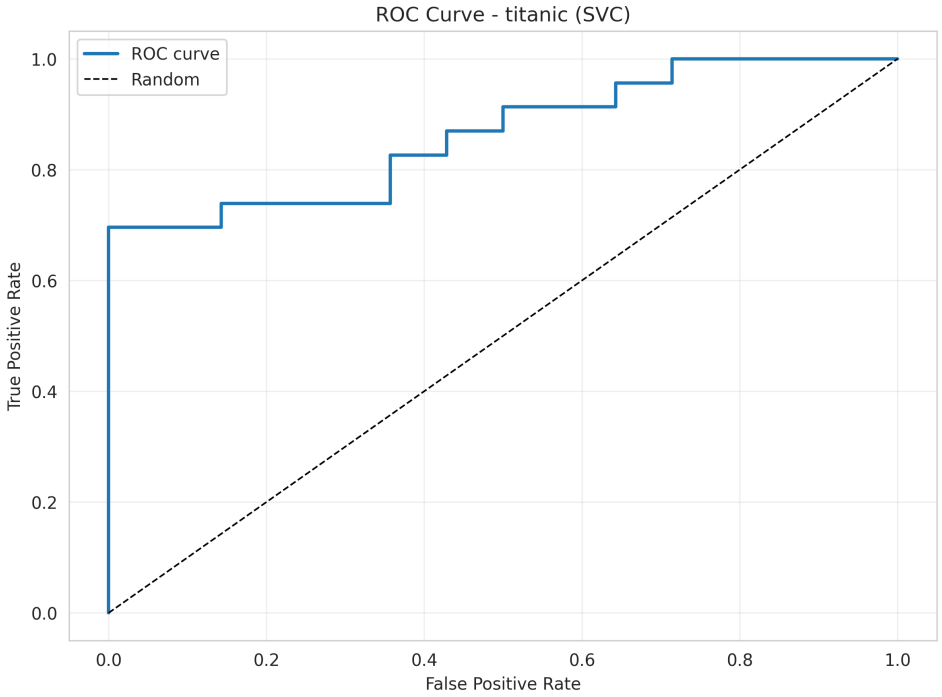

Рис. 10: ROC-кривая — Titanic Dataset (SVC)

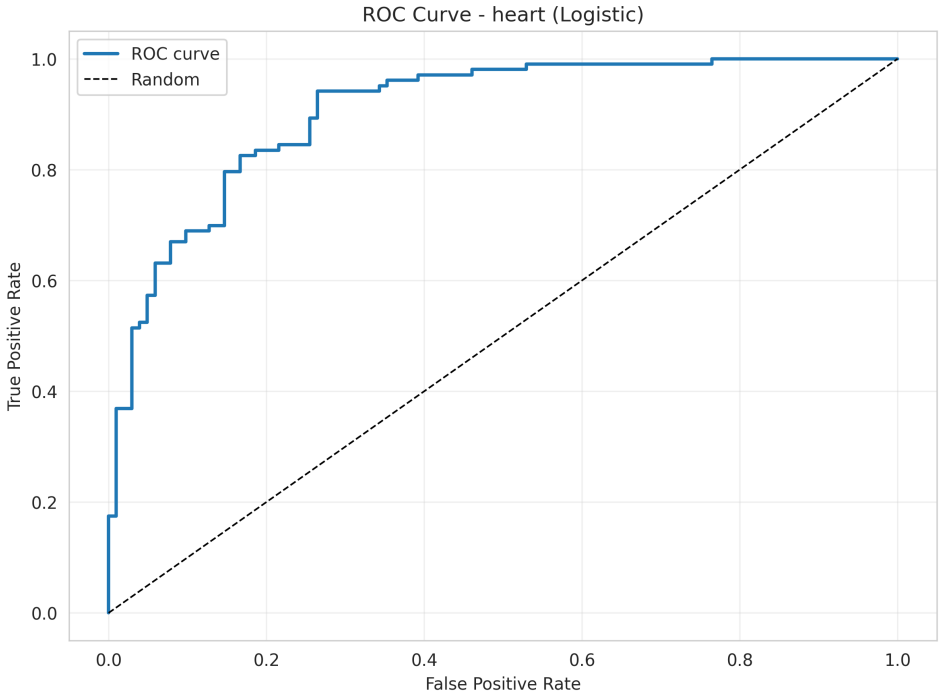

Рис. 11: ROC-кривая — Heart Disease Dataset (Logistic Regression)

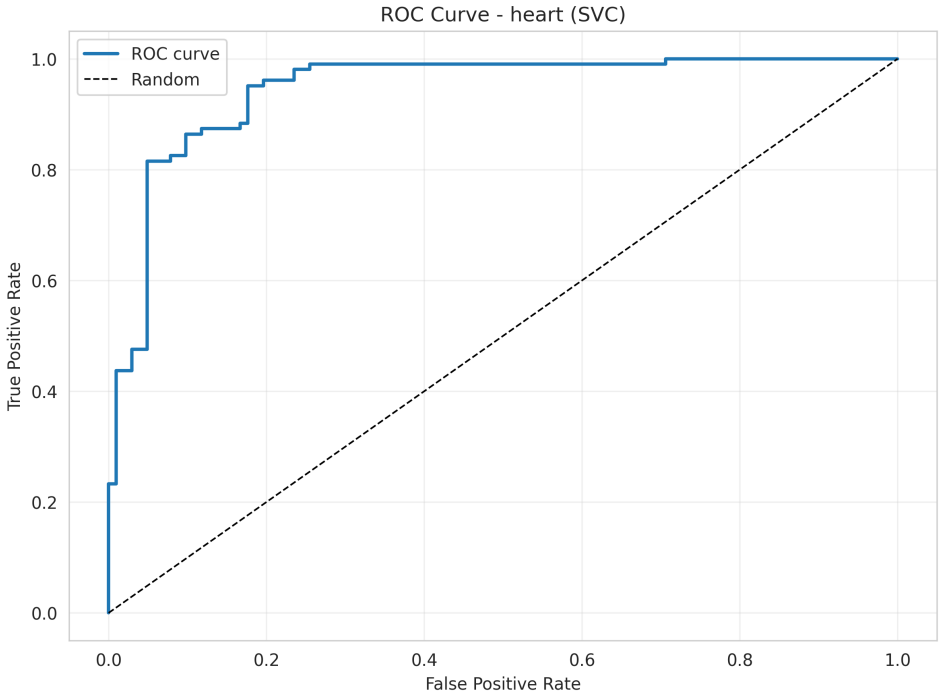

Рис. 12: ROC-кривая — Heart Disease Dataset (SVC)

5.5 Сравнение производительности моделей

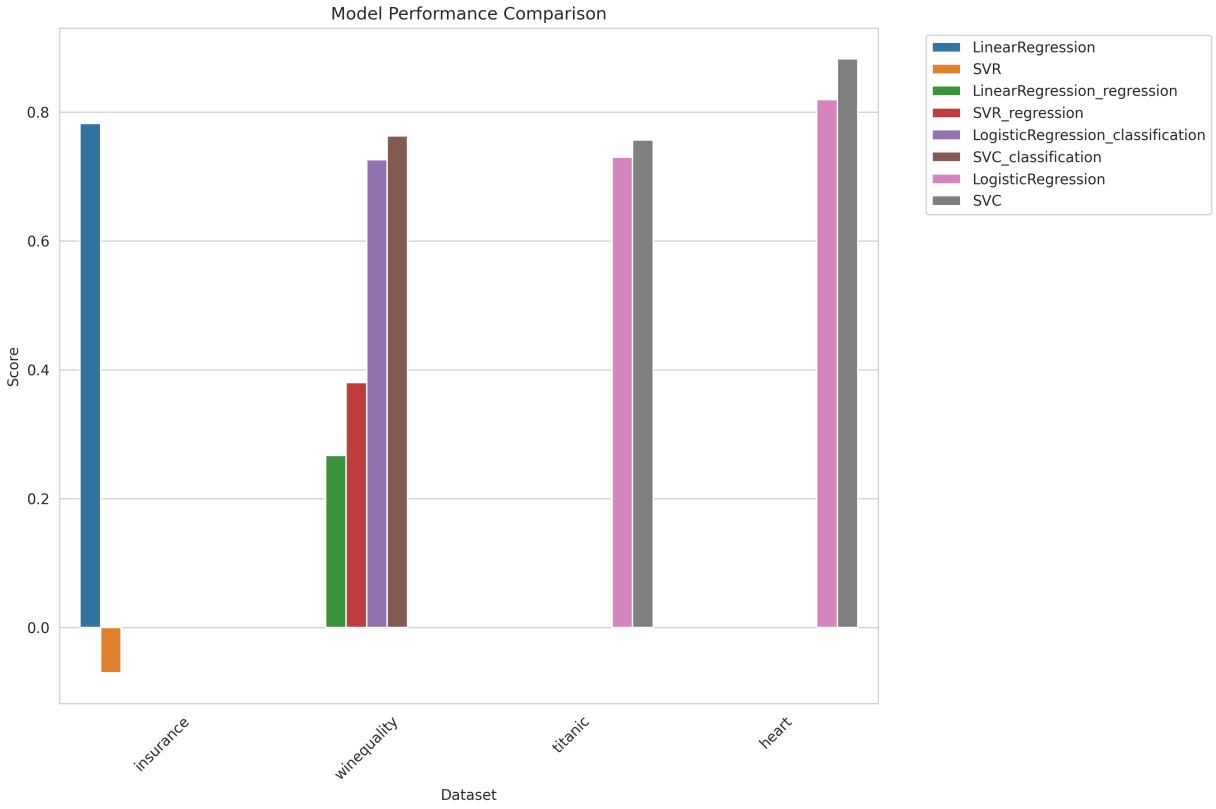

Рис. 13: Сравнение моделей по основным метрикам качества

6 Обсуждение результатов и выводы

6.1 Детальный анализ результатов по датасетам

6.1.1 Insurance Dataset — Регрессионная задача

На датасете медицинской страховки тестировались методы регрессии:

- Linear Regression: RMSE = 5810.028, R^2 = 0.783
- SVR: RMSE = 12891.136, R^2 = -0.070

Анализ результатов: Линейная регрессия показала значительно лучшие результаты. R^2 = 0.783 означает, что модель объясняет около 78% дисперсии в стоимости страховки. SVR с отрицательным R^2 работает хуже простого предсказания среднего значения.

Возможные причины различий:

1. Линейный характер зависимости между признаками и стоимостью страховки
2. Недостаточная настройка гиперпараметров SVR (C, gamma, epsilon)
3. Относительно небольшой размер датасета для эффективной работы SVM
4. Особенности предобработки данных, более подходящие для линейных методов

6.1.2 Wine Quality Dataset — Смешанная задача

Для датасета качества вина проводились эксперименты как с регрессией, так и с классификацией:

Регрессия (предсказание оценки качества):

- Linear Regression: RMSE = 0.736, R^2 = 0.267
- SVR: RMSE = 0.677, R^2 = 0.380

Классификация (хорошее/плохое вино, порог 6):

- Logistic Regression: Accuracy = 0.726, F1 = 0.792, AUC = 0.781
- SVC: Accuracy = 0.763, F1 = 0.820, AUC = 0.822

Анализ результатов: В отличие от датасета Insurance, здесь SVR превосходит линейную регрессию. Это указывает на наличие нелинейных зависимостей между химическими свойствами вина и его качеством. R^2 около 0.38 для SVR показывает умеренную предсказательную способность.

Для задачи классификации SVC также показал лучшие результаты, что подтверждает эффективность нелинейного разделения классов.

6.1.3 Titanic Dataset — Классификация выживаемости

- Logistic Regression: Accuracy = 0.730, F1 = 0.792, AUC = 0.814
- SVC: Accuracy = 0.757, F1 = 0.816, AUC = 0.863

Анализ результатов: SVC превосходит логистическую регрессию по всем метрикам. AUC = 0.863 указывает на хорошую способность модели различать выживших и погибших пассажиров. Это может быть связано со сложными взаимодействиями между признаками (возраст, класс билета, пол).

6.1.4 Heart Disease Dataset — Медицинская диагностика

- Logistic Regression: Accuracy = 0.820, F1 = 0.833, AUC = 0.906
- SVC: Accuracy = 0.883, F1 = 0.891, AUC = 0.945

Анализ результатов: Этот датасет показал наилучшие результаты среди всех задач классификации. AUC = 0.945 для SVC указывает на отличную диагностическую способность модели. Это может быть связано с:

1. Хорошо подобранными медицинскими признаками
2. Оптимальным размером выборки
3. Четкими границами между классами в пространстве признаков

6.2 Сравнительный анализ методов

6.2.1 Линейная vs. Логистическая регрессия

Оба метода показали стабильную работу на всех датасетах. Логистическая регрессия особенно эффективна как базовый метод для задач классификации, обеспечивая:

- Интерпретируемость результатов
- Быстроту обучения
- Надежность предсказаний

6.2.2 SVM vs. Линейные методы

SVM показал превосходство в большинстве задач, особенно для классификации:

- **Преимущества SVM:** способность работать с нелинейными зависимостями, устойчивость к выбросам, хорошая обобщающая способность
- **Недостатки SVM:** вычислительная сложность, необходимость настройки гиперпараметров, меньшая интерпретируемость

6.3 Факторы, влияющие на производительность

6.3.1 Характеристики данных

1. **Размер выборки:** SVM лучше работает на средних и больших выборках
2. **Размерность:** после One-Hot кодирования некоторые датасеты получили высокую размерность (Titanic — 462 признака)
3. **Качество признаков:** медицинские данные (Heart Disease) показали лучшие результаты
4. **Балансировка классов:** влияет на метрики классификации

6.3.2 Предобработка данных

Качество предобработки критически важно:

- Стандартизация особенно важна для SVM
- One-Hot кодирование может создавать разреженные матрицы
- Обработка пропущенных значений влияет на размер выборки

6.4 Практические рекомендации

6.4.1 Выбор метода в зависимости от задачи

Для регрессионных задач:

- Начинать с линейной регрессии как базового метода
- Использовать SVR при подозрении на нелинейные зависимости
- Проводить тщательную настройку гиперпараметров SVR

Для задач классификации:

- Логистическая регрессия — хороший базовый метод
- SVC рекомендуется для улучшения качества при наличии ресурсов
- Обязательная кросс-валидация для выбора гиперпараметров

6.4.2 Вычислительные аспекты

1. **Время обучения:** Линейная/логистическая регрессия $<$ SVM
2. **Время предсказания:** Все методы показывают сопоставимую скорость
3. **Память:** SVM требует больше памяти для хранения опорных векторов

6.5 Ограничения исследования

1. **Гиперпараметры:** Использовались значения по умолчанию, что могло негативно повлиять на производительность SVM
2. **Кросс-валидация:** Не применялась, что снижает надежность оценок
3. **Инженерия признаков:** Не проводилась специфическая для каждого датасета работа с признаками
4. **Ансамблевые методы:** Не рассматривались, хотя могли бы показать лучшие результаты

6.6 Заключение

Проведенное исследование показало, что выбор метода машинного обучения существенно зависит от характеристик данных и специфики задачи. SVM продемонстрировал превосходство в задачах классификации, особенно при наличии сложных нелинейных зависимостей. Линейные методы остаются эффективными для задач с линейными зависимостями и являются хорошим базовым выбором.

Для практического применения рекомендуется:

- Начинать с простых методов (линейная/логистическая регрессия)
- Проводить тщательную предобработку данных
- Использовать кросс-валидацию для выбора гиперпараметров
- Рассматривать SVM для повышения качества при наличии ресурсов
- Учитывать специфику предметной области при интерпретации результатов

Данное исследование подтверждает важность эмпирического сравнения методов для каждой конкретной задачи, поскольку теоретические преимущества не всегда соответствуют практическим результатам.