Анализ кардиологического датасета и классификация с помощью PCA и t-SNE
心电图数据集分析与降维分类

Ёркинжон Валиев — ИУ1-41М

1 Введение

Электрокардиограмма (ЭКГ) представляет собой графическое отображение электрической активности сердца. Анализ таких сигналов позволяет выявлять отклонения в работе сердечно-сосудистой системы. В данной работе проводится исследование кардиологического датасета, содержащего параметры ЭКГ, с целью изучения структуры данных и выявления закономерностей с помощью методов уменьшения размерности.

2 Обработка и очистка данных

Перед анализом необходимо обеспечить высокое качество данных. Пропущенные значения могут негативно повлиять на результат моделирования, особенно при работе с медицинскими данными. Были удалены признаки, содержащие более 50% пропущенных значений, так как они несут недостаточно информации для анализа.

Кроме того, из выборки были удалены выбросы — аномальные значения,

引言

心电图(Electrocardiogram, 简称ECG)是对心脏电流活动的图形表示，应用于心脏健康状态的分析与评估。本研究分析包含ECG参数的数据集，通过降维技术探索数据结构与隐含的模式。

数据预处理

在分析之前，必须确保数据质量。缺失值可能会显著影响模型性能，特别是在处理医疗数据时。因此，我们删除了缺失超过50%的特征，因为它们提供的信息不足。

此外，还清除了异常值，即远离正常分布的观测。使用四分位距（IQR）方法定义异常值如下：

$$\mathrm{IQR} = Q_3 - Q_1$$

сильно отклоняющиеся от распределения большинства данных. Для этого использовался метод межквартильного размаха (IQR):

$$\text{IQR} = Q_3 - Q_1$$

Границы выбросов рассчитываются по формуле:

$$[Q_1 - 1.5 \cdot IQR,\ Q_3 + 1.5 \cdot IQR]$$

Удаление выбросов помогает улучшить устойчивость моделей и точность выводов.

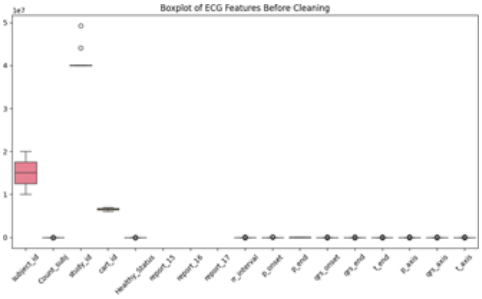


Рис. 1: Боксплот признаков до очистки

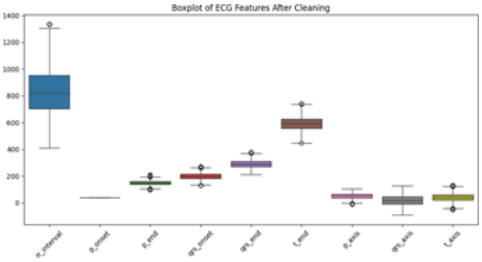
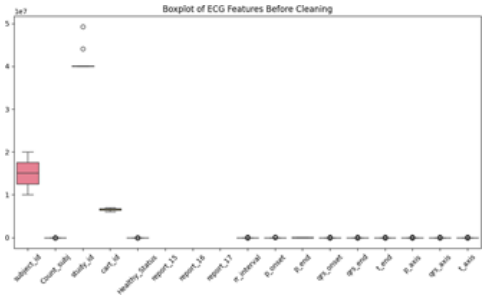
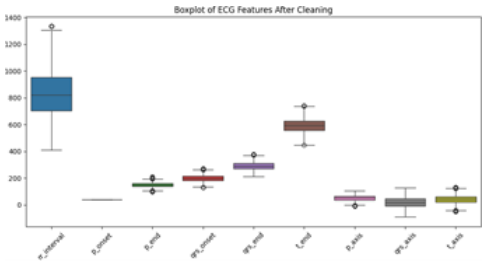


Рис. 2: Боксплот признаков после очистки

异常范围为：

$$[Q_1 - 1.5 \cdot IQR,\ Q_3 + 1.5 \cdot IQR]$$





去除异常值有助于提高模型的鲁棒性和预测精度。

3 Анализ корреляций и визуализация

Корреляционная матрица

Для оценки взаимосвязей между признаками рассчитывался коэффициент корреляции Пирсона:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}}$$

Визуализация представлена в виде тепловой карты.

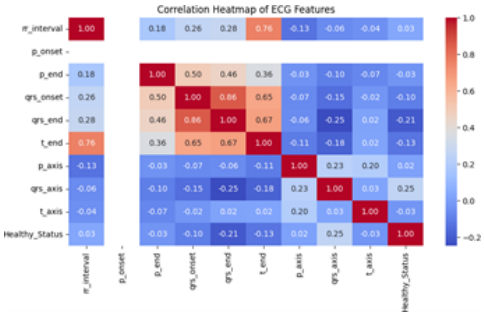


Рис. 3: Тепловая карта корреляций

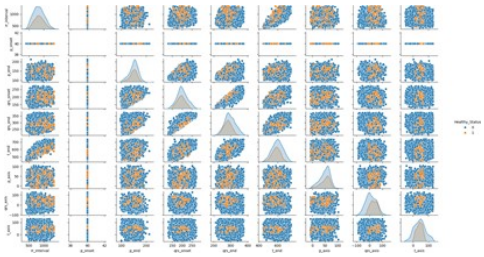


Рис. 4: Pairplot — распределение по классам

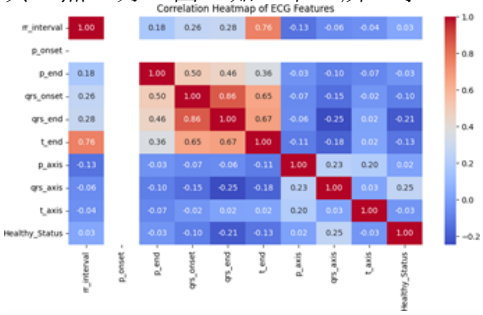
特征相关分析

皮尔逊相关系数

为了评估特征之间的相关性，使用了皮尔逊系数：

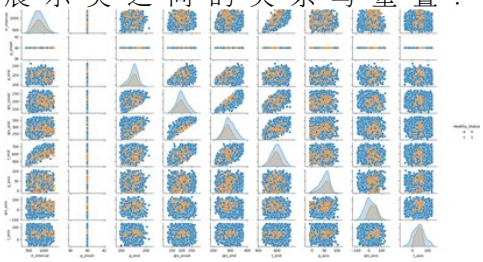
$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}}$$

其热力图如下所示：



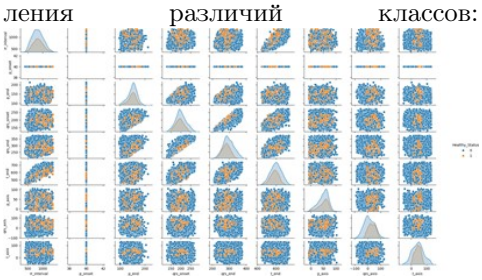
Pairplot 可视化

展示类之间的关系与重叠：



Парная визуализация (Pairplot)

Используется для выявления различий классов:



4 Снижение размерности: PCA

Метод главных компонент (PCA) применяется для преобразования данных в новое пространство, где максимизируется дисперсия. Это позволяет выявить основные направления изменения данных.

$$Z = XW$$

где:

- X — стандартизированные признаки,
- W — матрица собственных векторов ковариации,
- Z — новые проекции данных.

Результат визуализирован ниже:

PCA 降维

PCA方法通过寻找方差最大的方向来重构数据空间：

$$Z = XW$$

其中：

- X — 标准化后的数据；
- W — 协方差矩阵的特征向量；
- Z — 主成分投影结果。

注： 可视化表明，线性变换未能实现类别的清晰分离，说明数据结构可能具有非线性特征。

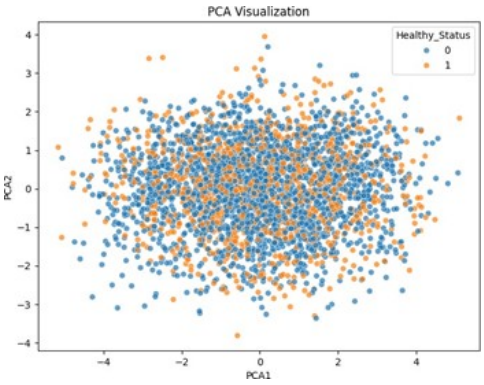


Рис. 5: PCA — визуализация

Замечание: несмотря на трансформацию, четкого разделения классов не достигнуто, что указывает на нелинейную природу данных.

5 Снижение размерности: t-SNE

Метод t-SNE (t-distributed stochastic neighbor embedding) проецирует данные в пространство меньшей размерности, сохраняя вероятностную близость точек.

$$P_{ij} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2/2\sigma^2)}$$

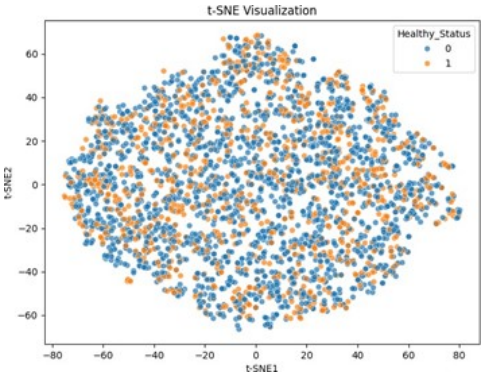
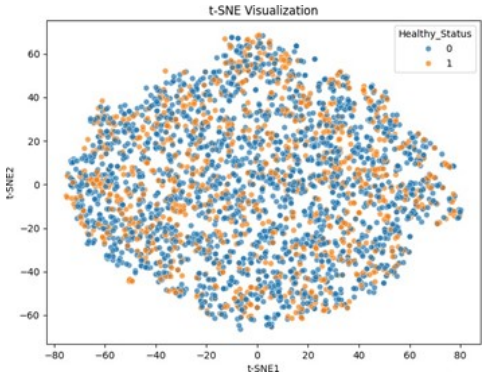


Рис. 6: t-SNE — визуализация

t-SNE 非线性降维

t-SNE将高维数据投影到二维/三维空间，保持邻域结构：

$$P_{ij} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2/2\sigma^2)}$$



小结： t-SNE揭示了数据的潜在结构，尽管类别边界模糊，但提供了对数据分布的重要洞察。

Вывод: t-SNE позволяет выявить скрытую структуру данных, несмотря на отсутствие чётких границ между классами.

6 Заключение

- Выполнена полная очистка и предварительная обработка данных;
- Методы визуализации и корреляции выявили взаимосвязи;
- PCA показал ограниченную эффективность в разделении классов;
- t-SNE успешно выявил скрытые структуры;
- Для повышения точности необходимы более сложные модели (например, нейросети).

总结

- 成功进行了数据清洗与标准化处理;
- 可视化和相关性分析揭示特征之间的关系;
- PCA线性降维效果有限;
- t-SNE展现了数据的隐藏模式;
- 可尝试使用神经网络等复杂模型提高分类效果。

Список литературы / 参考文献

Список литературы

[1] AI-is-out-there. *Data2Lab Repository.* https://github.com/AI-is-out-there/data2lab.git