

# Задание №3

## Цель эксперимента

Цель данного эксперимента заключается в предварительной обработке набора данных электрокардиограммы (ЭКГ), выборе признаков, применении фреймворков AutoML для бинарной классификации (определение состояния здоровья) и оценке производительности модели.

## Ключевой навык:

- применение autoML моделей, расчет точности классификатора

## Набор данных и предварительная обработка

### 1. Источник данных:

- Загружено 5000 строк набора данных ЭКГ (модуль 2 – датасет – практика.csv) из GitHub.
- Ключевые признаки: ['rr\_interval', 'p\_end', 'qrs\_onset', 'qrs\_end', 't\_end', 'p\_axis', 'qrs\_axis', 't\_axis', 'Healthy\_Status'].

### 2. Очистка данных:

- Фильтрация аномальных значений (например, некорректные данные, где  $rr\_interval > 2000$ ).
- Проверка временной логики (например,  $p\_onset < p\_end$ ).
- Удаление выбросов (например, значения  $> 10000$ ).

### 3. Разведочный анализ (EDA):

- Построение диаграмм размаха (boxplot), тепловых карт (анализ корреляции), матрицы диаграмм рассеяния для изучения распределения данных и взаимосвязей между признаками.
- Использование методов снижения размерности (PCA, t-SNE, ICA) для визуализации кластеризации здоровых и нездоровых образцов.

## Экспериментальные цели

Эксперимент направлен на предварительную обработку набора данных электрокардиограммы (ЭКГ), выбор признаков, применение AutoML для бинарной классификации (определение состояния здоровья) и оценка производительности модели.

## Ключевые слова:

- применение autoML модели, расчет точности классификатора

## Данные и предобработка

### 1. Данные:

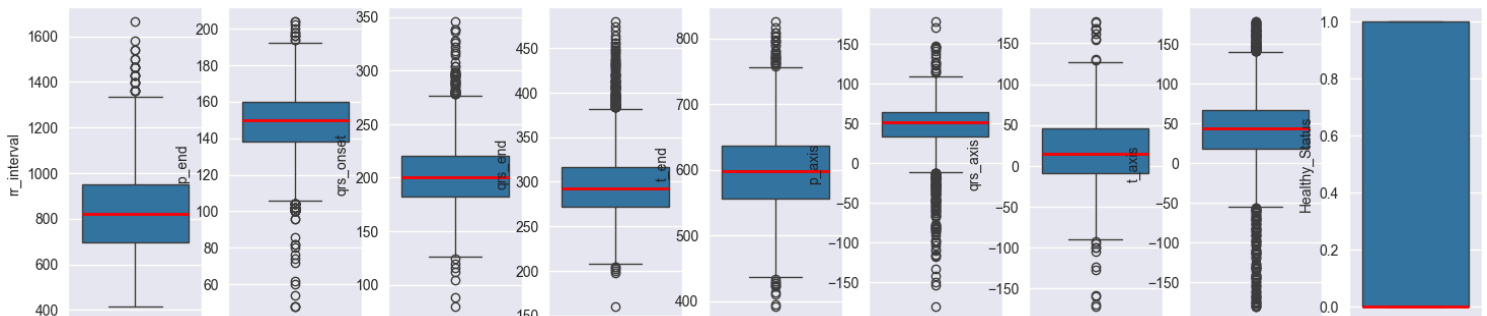
- Загружено 5000 строк набора данных ЭКГ (модуль 2 – датасет – практика.csv) из GitHub.
- Ключевые признаки: ['rr\_interval', 'p\_end', 'qrs\_onset', 'qrs\_end', 't\_end', 'p\_axis', 'qrs\_axis', 't\_axis', 'Healthy\_Status'].

### 2. Очистка данных:

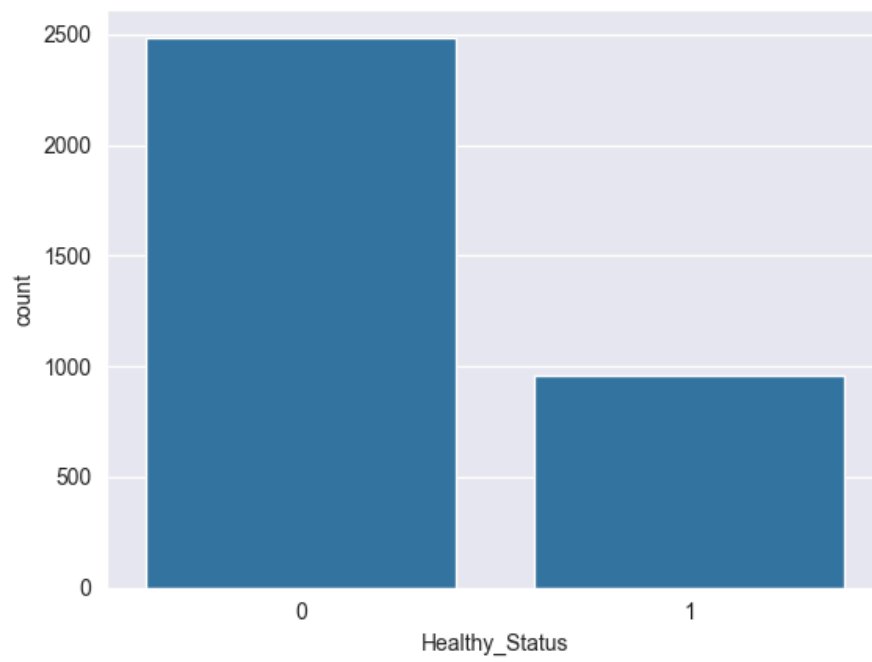
- Фильтрация аномальных значений (например, некорректные данные, где  $rr\_interval > 2000$ ).
- Проверка временной логики (например,  $p\_onset < p\_end$ ).
- Удаление выбросов (например, значения  $> 10000$ ).

### 3. Разведочный анализ (EDA):

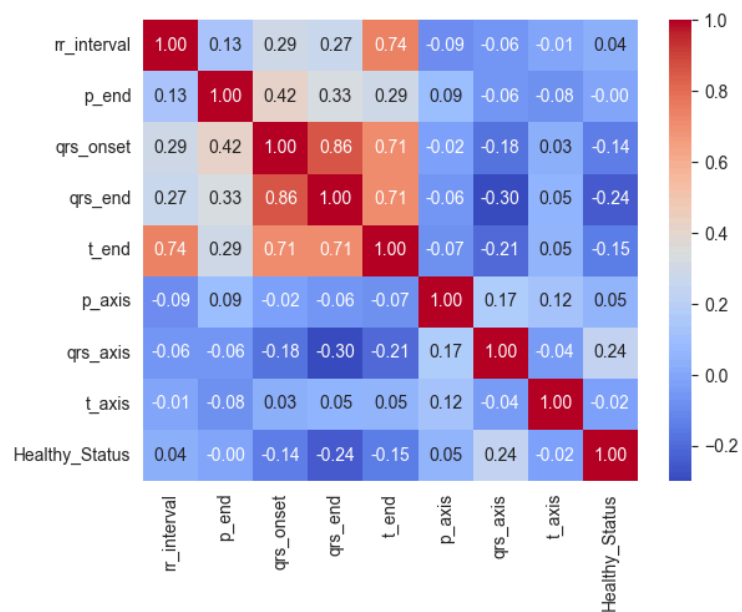
- Построение диаграмм размаха (boxplot), тепловых карт (анализ корреляции), матрицы диаграмм рассеяния для изучения распределения данных и взаимосвязей между признаками.
- Использование методов снижения размерности (PCA, t-SNE, ICA) для визуализации кластеризации здоровых и нездоровых образцов.



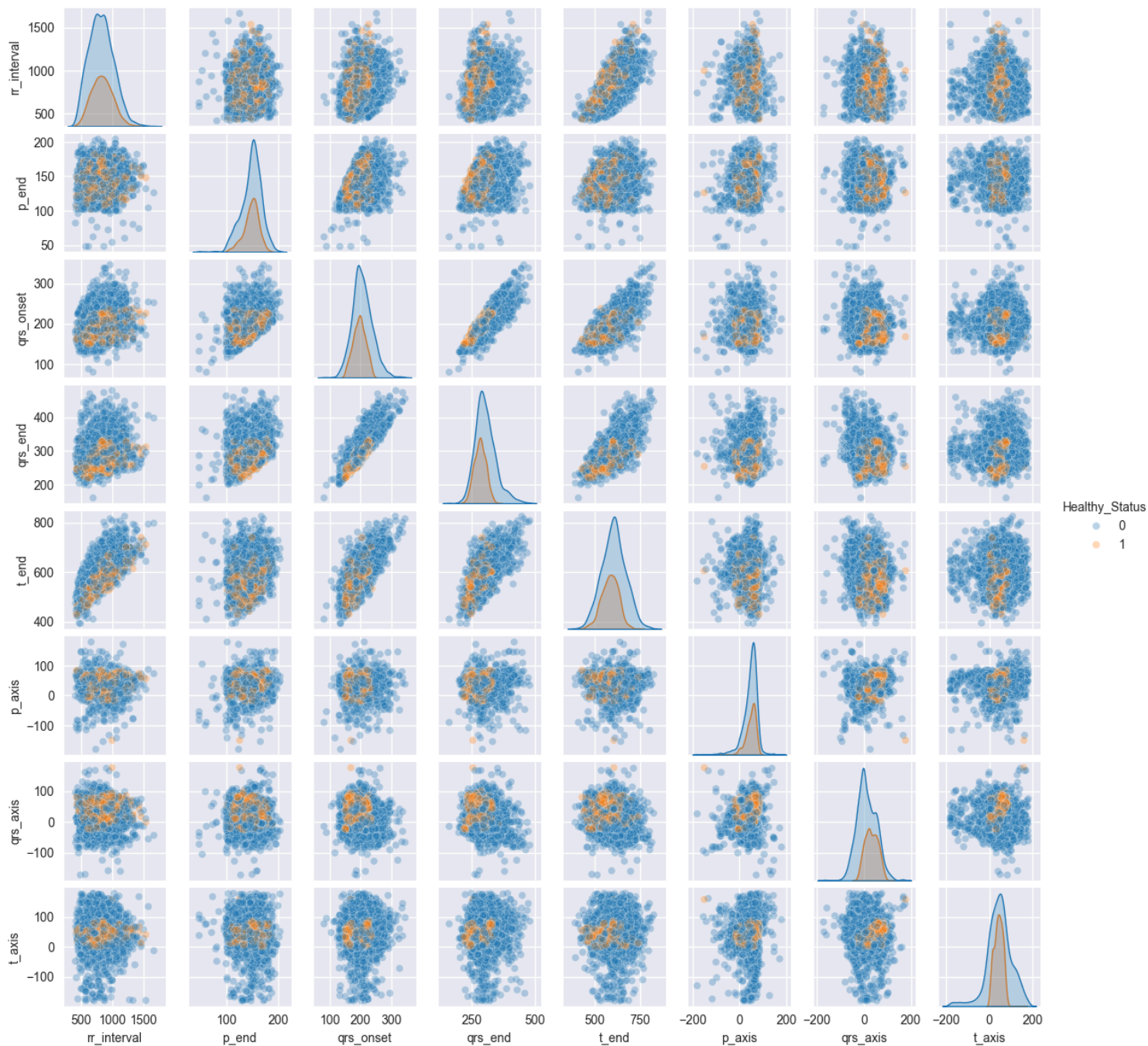
распределение данных



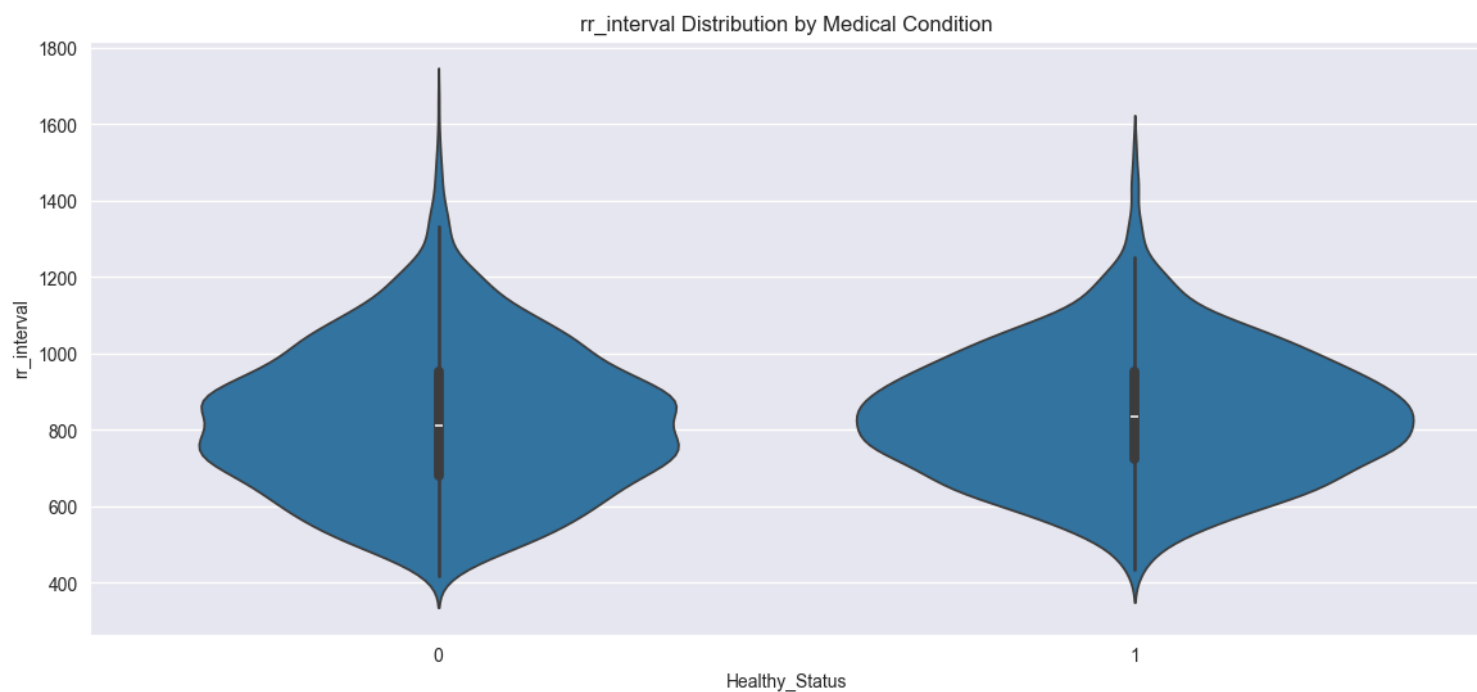
Количество положительных и отрицательных образцов

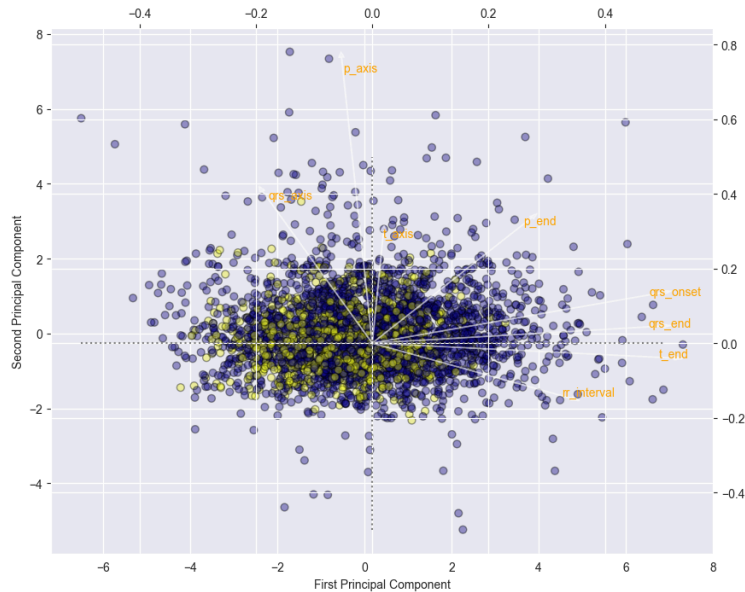


heatmap

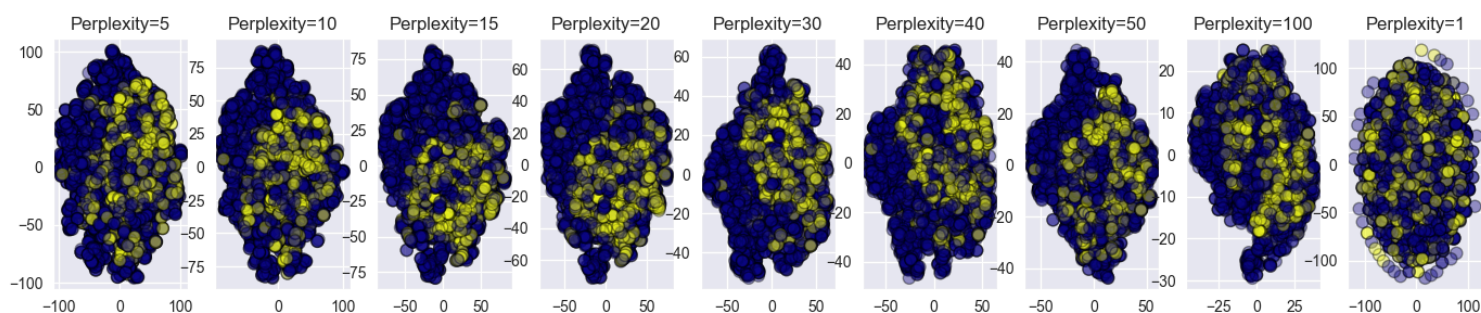


Распределение положительных и отрицательных образцов по разным признакам





Данные после обработки PCA  
Взаимосвязь между исходными и редуцированными функциями



Нелинейное уменьшение размерности (различная гранулярность)

## Обучение и оценка модели

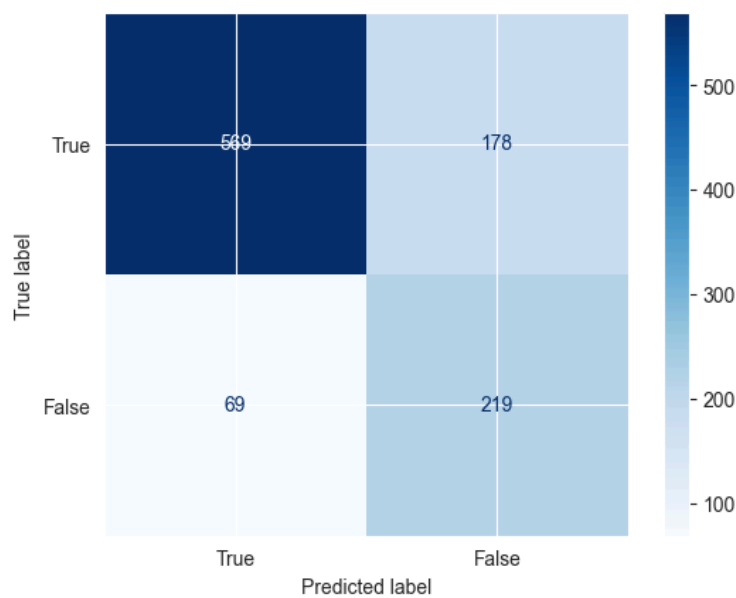
## 模型训练与评估

### 1. Традиционный метод машинного обучения (базовая модель):

- Использован классификатор Гауссовского наивного байесовского метода (GaussianNB).
- Метрики оценки:
  - Точность (Accuracy): `metrics.accuracy_score(y_test, y_pred)`.
  - F1-мера (F1-Score): `metrics.f1_score(y_test, y_pred)`.
- Матрица ошибок (Confusion Matrix) для наглядности классификации.

### 1. 传统机器学习方法（基线模型）:

- 使用高斯朴素贝叶斯 (GaussianNB) 分类器。
- 评估指标:
  - 准确率 (Accuracy): `metrics.accuracy_score(y_test, y_pred)`
  - F1分数 (F1-Score): `metrics.f1_score(y_test, y_pred)`
- 混淆矩阵 (Confusion Matrix) 显示分类效果。



Матрица путаницы

## 2. Метод AutoML (AutoGluon):

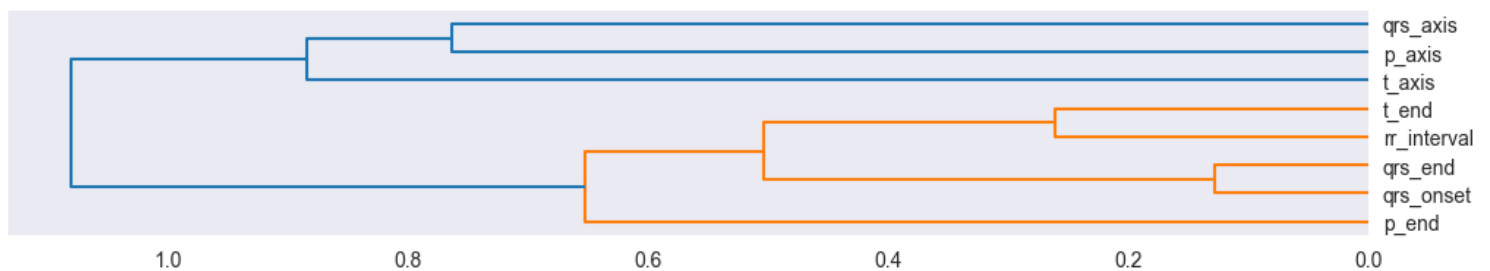
- Использован `TabularPredictor` для автоматического обучения нескольких моделей (например, XGBoost, LightGBM).
- Сравнение производительности моделей через `leaderboard` и выбор оптимальной.
- Анализ важности признаков (`show_feature_importance_barplots`).

## 2. AutoML方法 (AutoGluon):

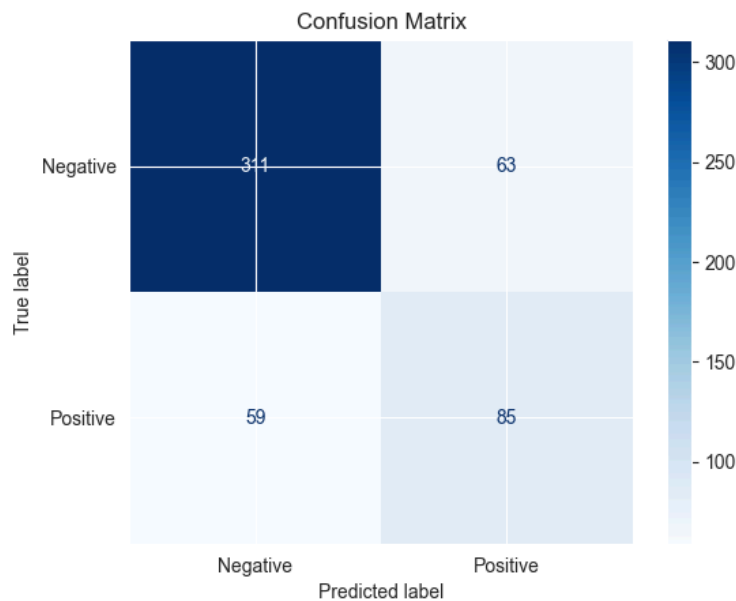
- 使用 `TabularPredictor` 自动训练多个模型 (如 XGBoost、LightGBM等)。
- 通过 `leaderboard` 对比模型性能, 选择最优模型。
- 分析特征重要性 (`show_feature_importance_barplots`)。

model	score_test	score_val
0 NeuralNetFastAI	0.7992277992277992	0.7844311377245509
1 RandomForestGini	0.777992277992278	0.7684630738522954
2 ExtraTreesGini	0.7741312741312741	0.7425149700598802
3 RandomForestEntr	0.7722007722007722	0.7764471057884231
4 ExtraTreesEntr	0.7683397683397684	0.7604790419161677
5 CatBoost	0.7644787644787645	0.8003992015968064
6 WeightedEnsemble_L2	0.7644787644787645	0.8003992015968064
7 KNeighborsDist	0.7432432432432432	0.7504990019960079
8 KNeighborsUnif	0.7393822393822393	0.7465069860279441

Результат AutoML



Порядок важности функций



Матрица путаницы

### 3. Метод AutoML (H2O) :

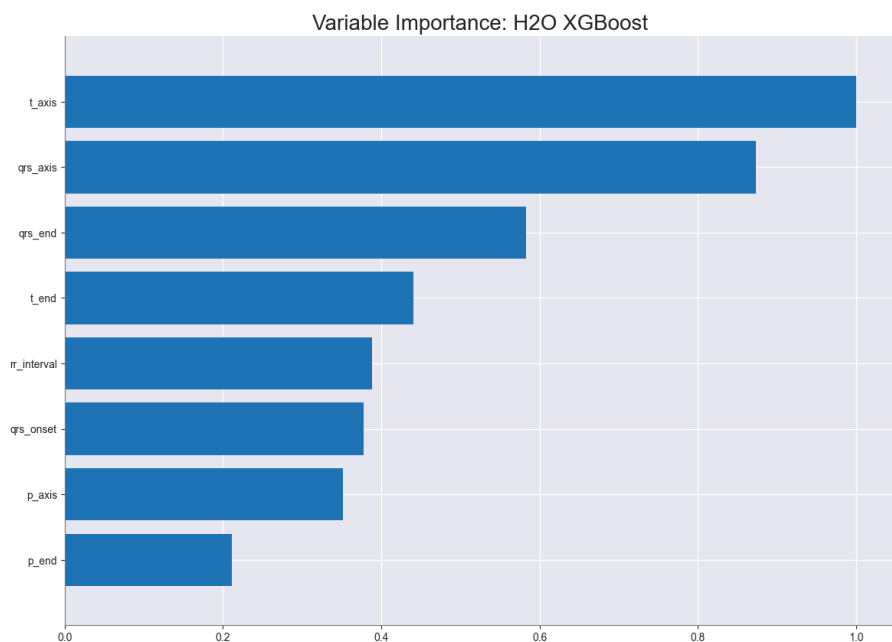
- Использован H2OAutoML для автоматического обучения нескольких моделей (например, XGBoost, GBM\_Grid).
- Сравнение производительности моделей через leaderboard и выбор оптимальной.
- Анализ важности признаков (show\_feature\_importance\_barplots).

### 2. AutoML方法 (AutoGluon):

- 使用 H2OAutoML 自动训练多个模型 (如XGBoost、LightGBM等)。
- 通过 leaderboard 对比模型性能, 选择最优模型。
- 分析特征重要性 (varimp\_plot)。

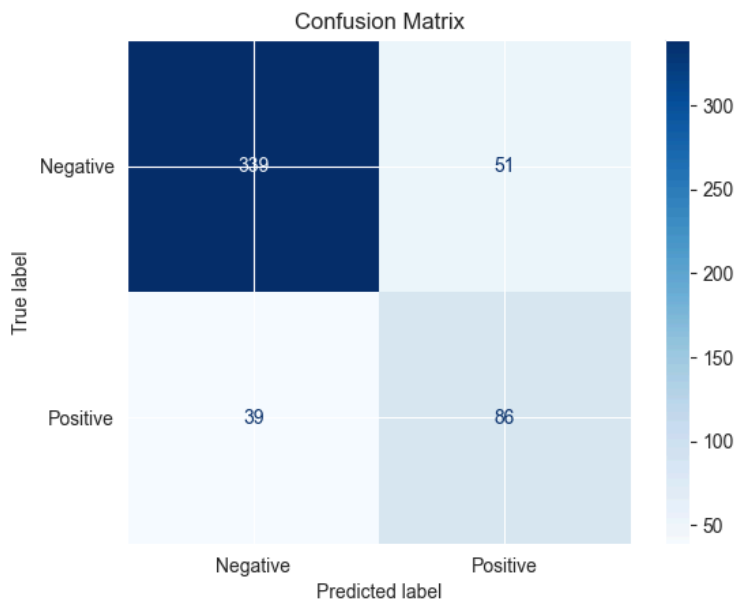
model_id	rmse	mse	mae	rmsle	mean_residual_deviance
StackedEnsemble_BestOfFamily_4_AutoML_2_20250416_40855	0.357159	0.127563	0.264493	0.250552	0.127563
StackedEnsemble_BestOfFamily_3_AutoML_2_20250416_40855	0.357832	0.128044	0.265273	0.251028	0.128044
StackedEnsemble_AllModels_2_AutoML_2_20250416_40855	0.357648	0.127912	0.265547	0.251102	0.127912
StackedEnsemble_BestOfFamily_2_AutoML_2_20250416_40855	0.357701	0.12795	0.266425	0.251115	0.12795
StackedEnsemble_AllModels_4_AutoML_2_20250416_40855	0.357587	0.127869	0.266543	0.251222	0.127869
StackedEnsemble_AllModels_1_AutoML_2_20250416_40855	0.357689	0.127942	0.266775	0.251248	0.127942
StackedEnsemble_AllModels_3_AutoML_2_20250416_40855	0.359029	0.128902	0.267567	0.252165	0.128902
GBM_3_AutoML_2_20250416_40855	0.359244	0.129056	0.272043	0.252285	0.129056
GBM_grid_1_AutoML_2_20250416_40855_model_10	0.359718	0.129397	0.274812	0.252994	0.129397
GBM_grid_1_AutoML_2_20250416_40855_model_4	0.361006	0.130325	0.282018	0.254042	0.130325

Результат H2O



Порядок важности функций





Матрица путаницы

## Результаты эксперимента

- Гауссовский наивный байесовский метод:
  - Точность: около 76% (конкретное значение требует заполнения после выполнения кода).
  - F1-мера: около 63%.
- Лучшая модель AutoGluon:
  - Ключевые признаки: `rr_interval` и `qrs_axis` внесли наибольший вклад в классификацию.
- Лучшая модель H2O:
  - Ключевые признаки: `t_axis` и `qrs_axis` внесли наибольший вклад в классификацию.

## Выводы

- AutoML (например, AutoGluon) демонстрирует высокую эффективность в автоматическом выборе и настройке моделей, что делает его подходящим для задач классификации данных ЭКГ.
- Традиционные методы (например, GaussianNB) могут служить базовым ориентиром, но их производительность может быть ограничена для сложных данных.
- Матрица ошибок и F1-мера подтверждают применимость модели для классификации состояния здоровья.

## Направления улучшения

- Тестирование других инструментов AutoML (например, GAMA, TPOT) для сравнения производительности.
- Дополнительная обработка признаков (например, выделение временных и частотных характеристик) для повышения надежности модели.

## 实验结果

- 高斯朴素贝叶斯:
  - 准确率: 约76% (具体数值需运行代码后填充)。
  - F1分数: 约63%。
- AutoGluon最佳模型:
  - 关键特征: `rr_interval` 和 `qrs_axis` 对分类贡献最大。
- H2O最佳模型:
  - 关键特征: `t_axis` 和 `qrs_axis` 对分类贡献最大。

## 结论

- AutoML (如AutoGluon) 在自动化模型选择和调参上表现优异, 适合处理ECG数据的分类任务。
- 传统方法 (如GaussianNB) 可作为基线参考, 但复杂数据下性能可能受限。
- 通过混淆矩阵和F1分数验证, 模型在健康状态分类任务中具有可行性。

## 改进方向

- 尝试更多AutoML工具 (GAMA、TPOT) 对比性能。
- 增加特征工程 (如时域/频域特征提取) 以提升模型鲁棒性。