1. **Цель**

Целью данной лабораторной работы является обработка кардиологического датасета и построение бинарного классификатора по признаку Healthy\_Status с применением AutoML-фреймворков. Основные задачи включают:

- загрузку и фильтрацию 5000 записей ЭКГ данных;

- формирование признакового пространства по выбранным параметрам;

- обучение моделей с помощью AutoML;

- оценка качества классификации по матрице ошибок и F1-мере.

1. **目的**

本实验旨在处理心电图（ECG）数据集，并基于Healthy\_Status特征构建一个二元分类器。主要目标包括：

- 加载并筛选5000条心电图数据；

- 构建由特定特征组成的特征空间；

- 使用AutoML框架训练分类模型；

- 通过混淆矩阵和F1-分数评估模型性能。

1. **Методы**
2. Подготовка данных

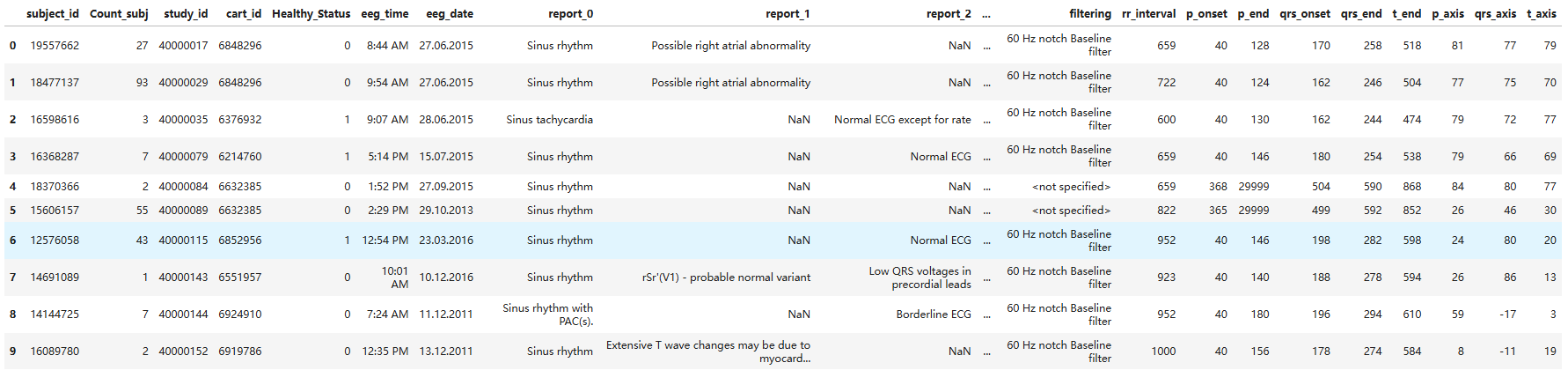
Из датасета «модуль 3...» были выбраны 5000 строк и следующие столбцы:

2. **方法**

1) 数据准备

从提供的“模块3...”数据集中加载5000条记录，选取以下字段：





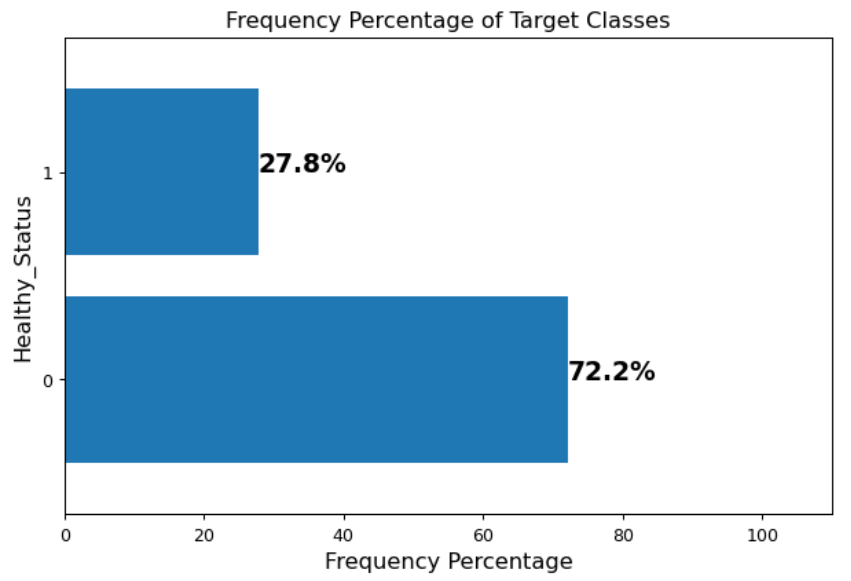
Очищение данных от выбросов осуществлялось с помощью логических условий:

应用逻辑筛选条件以清洗异常值（通过去除异常值、合并和清理文本报告数据，使用 Word2Vec模型将文本转化为数值表示，并将其作为新的特征用于后续分析）：



Предварительно обработанные характеристики ЭКГ были классифицированы с использованием модели Гаусса Naive Bayes, после чего была проанализирована эффективность модели.

用Gaussian Naive Bayes模型对预处理后的ECG特征进行分类，并对模型进行性能分析。



1. Обучение модели AutoML

Изучены следующие AutoML-фреймворки:

- H2O AutoML: прост в применении, есть графический интерфейс, хорошо масштабируется.

- TPOT: генетический подход к построению моделей.

- AutoSklearn: использует байесовскую оптимизацию, высокое качество, требует больше времени.

Выбор: H2O AutoML как оптимальный по соотношению качества и простоты настройки.

2) AutoML模型训练

对比以下AutoML框架：

- H2O AutoML：易用，支持图形界面，适合大数据；

- TPOT：基于遗传算法构建模型；

- AutoSklearn：基于贝叶斯优化，准确率高但耗时较长。

选择结果：H2O AutoML平衡了效果与使用便捷性。

Здесь используется фреймворк AutoML от H2O для автоматического перебора нескольких комбинаций моделей (GLM, GBM, DRF, DeepLearning, StackedEnsemble и т. д.)

По умолчанию в качестве основы для оценки модели используются AUC и показатель F1.

使用H2O的AutoML框架自动尝试多种模型组合（GLM、GBM、DRF、DeepLearning、StackedEnsemble等）。

默认采用AUC、F1-score为模型评估依据。

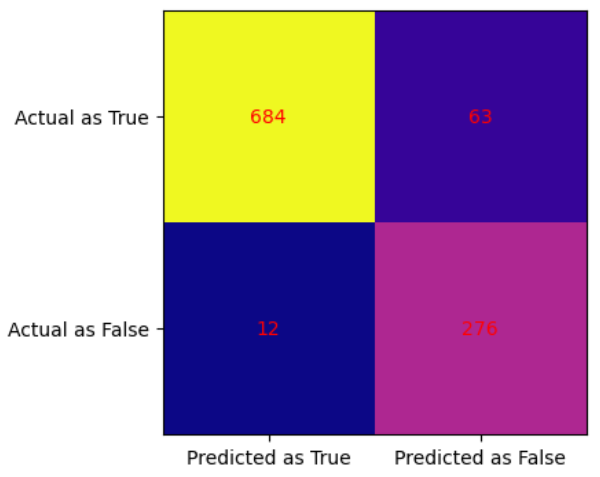
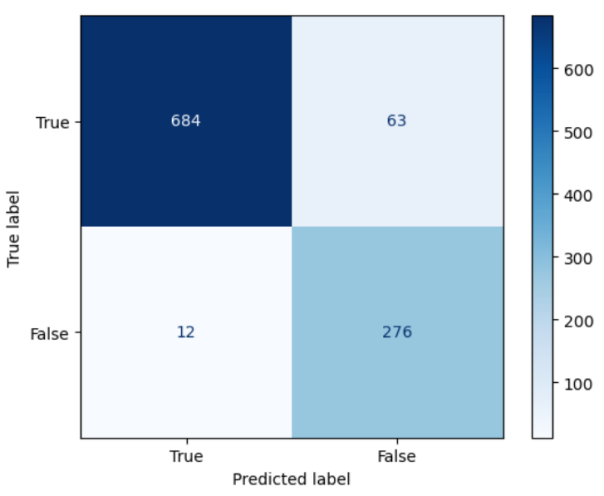


1. Матрица путаницы и максимальные результаты F1

Матрица путаницы: оценивает эффективность модели классификации, показывая истинно положительные, ложно положительные, истинно отрицательные и ложно отрицательные результаты.

3)混淆矩阵与最大F1分数结果

混淆矩阵：评估分类模型的性能，显示真阳性、假阳性、真阴性和假阴性。



F1-мера: это гармоническое среднее значение точности и полноты, подходящее для оценки несбалансированных наборов данных.

F1分数：综合考虑了精确率和召回率的调和平均数，适用于不平衡数据集的评估。

1. **Сравнение моделей**

**3.模型对比**

Следующие экспериментальные данные показывают, что модель GBM\_5 имеет наилучшие характеристики на тестовом наборе:

- В тестовом наборе только 3 отрицательных класса были классифицированы неправильно (почти идеальная классификация)

- Оценка F1 достигла 0,9881, что является одним из важнейших показателей в оценке

- AUC достигает 0,9998, и модель обладает сильной дискриминационной способностью.

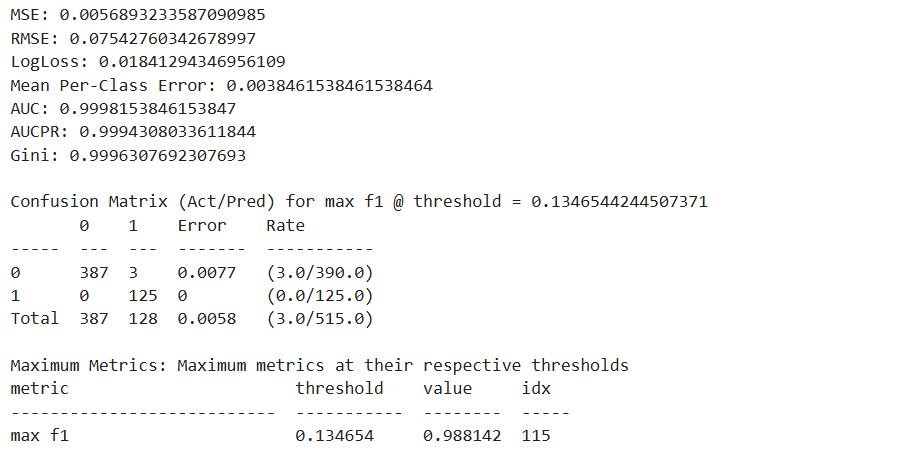
以下实验数据说明， GBM\_5模型在测试集上评估性能最优：

- 在测试集上，仅有3个负类被误分（几乎完美分类）

- F1分数达到0.9881，这是评估中最重要的指标之一

- AUC也高达0.9998，模型具有极强区分能力

1747075129553



Лучшая производительность модели (на тестовом наборе):

最优模型性能表现（在测试集上）：

| **Значение /指标** | **Показателя/数值** | **Описание/说明** |
| --- | --- | --- |
| **AUC** | 0.9998 | Модель обладает сильными возможностями классификации.  模型分类能力极强 |
| **F1-score** | 0.9881 | Учитываются как точность, так и скорость отклика, почти идеальная.  准确率与召回率兼顾，几乎完美 |
| **Accuracy** | ≈ 0.994 | Общая точность классификации высокая.  整体分类正确率高 |
| **LogLoss** | 0.01841 | Вероятность выходных данных модели стабильна и достоверна.  模型输出概率稳定且可信 |
| **Ошибка матрицы путаницы**  **混淆矩阵误差** | Ложноположительный результат/假阳性FP = 3  ложноотрицательный результат/假阴性FN = 0 | Почти все нездоровые люди были успешно идентифицированы (Полнота = 1,0)  非健康者几乎全部被成功识别（Recall = 1.0） |

1. **Заключение**

В этом проекте используется фреймворк H2O AutoML для автоматического выбора и обучения нескольких моделей классификации для данных цифровых характеристик ЭКГ с целью решения задачи бинарной классификации (определения того, является ли человек «здоровым») и оценки эффективности на основе таких показателей, как матрица неточностей и показатель F1. Были получены следующие результаты:

- H2O AutoML может эффективно и автоматически искать и определять оптимальную модель классификации

- Характеристики ЭКГ имеют значительную дискриминационную силу для оценки состояния здоровья

- Окончательная модель GBM\_5 имеет точность ≈ 99,4% и оценку F1 ≈ 0,988, что делает ее готовым к развертыванию высококачественным классификатором.

**5.结论**

本项目针对ECG数字特征数据，利用H2O AutoML框架自动选择和训练多个分类模型，解决二分类问题（判断个体是否为“健康”），并基于混淆矩阵、F1分数等指标进行性能评估。得出以下结果：

- H2O AutoML能够有效自动搜索并识别最优分类模型

- ECG特征对于健康状态判断具有显著判别力

- 最终模型GBM\_5准确率≈ 99.4%，F1-score ≈ 0.988，是一个可部署的高质量分类器

1. Ссылки на литературу

[1] L. van der Maaten, G. Hinton. Visualizing data using t-SNE, JMLR, 2008.

[2] I. T. Jolliffe, Principal Component Analysis, Springer, 2002.