中国石油大学（北京）

2025— 2026 学年秋季学期

《数据挖掘技术与应用》

结课报告

题　　目：帕尔默企鹅数据集的探索学习

小组成员：　胡林森

郑　智

周　程

成　　绩：

中国石油大学（北京）克拉玛依校区

2025 年　11 月　25 日

撰写说明：

1. 为了提高学生的大数据存储和分析的应用能力和实践水平，客观考察学生的大数据分析能力，以小组任务完成一个大数据存储和分析应用案例。小组成员自行组合，每组人数 3 人。不够 3 人或者无法自行组合，由老师安排。
2. 任务选题自定（可自主选择校园、生活服务、公共交通、消费娱乐、环境监测等领域），但需覆盖数据挖掘全流程核心内容：数据获取、numpy/pandas 数据预处理、数据可视化、关联规则挖掘/分类模型构建/聚类分析。
3. 数据收集可以是公开的数据集或者模拟数据，体现一定的数据量，数据预处理包括确保数据的一致性，数据挖掘方法不限，结果展示包括报表或者可视化仪表盘等。
4. 任务体现一定工作量，体现团队合作精神，明确每个成员的分工和贡献。
5. 撰写体例参照目录。
6. 报告需要提交电子版和打印版，提交时间为结课时间一周内。
7. 报告成绩评定结合小组成绩和个人成绩。个人成绩根据小组分数，按照贡献度进行成员内部分配。例如，某小组成绩为 80，该小组的成员总分为 80*3=240，成员 ABC 根据贡献分配 240 分，个人分不超 100 分。
8. 严禁抄袭，一经发现，报告为 0 分。

注：依据参与度、合作度和贡献度，报告个人贡献。各位在提交报告时，只需填写成员和任务贡献%这一列，不但填自己的贡献，还要填小组其他成员的贡献，且小组内各个成员取得一致。

注

组号	成员	任务贡献%	个人分数	备注
6	A	30%	72	<=100
	B	30%	72	<=100
	C	40%	96	<=100
总计	80*3	100%	240	小组得分 80

目　录

一、选题与组内任务分配

1.1 选题分析

题目：

本文使用 Gorman 等人 2014 年论文中[1]的由 Dr. Kristen Gorman 和南极洲巴勒莫站长期生态研究计划（LTER）收集并提供的数据，获取自 https://github.com/allisonhorst/palmerpenguins,是近年来在数据科学和机器学习领域受到关注的一个数据集，数据集目标是提供一个优秀的数据集用于数据探索与可视化，作为鸢尾花（Iris）数据集的替代方案。

数据集包含了对南极洲不同地区生活的企鹅种群的研究数据，主要用于数据探索和可视化，以及预测、分类任务。

本次实践依托 github 进行协作、代码托管。项目地址：https://github.com/GALA-Lin/2025-Fall-Term-Data-Mining-Course-Final-Report。

表 1.1 数据集结构

英文字段名	中文字段名	描述
species	种类	Gentoo：巴布亚企鹅；Adelie：阿德利企鹅；Chinstrap：帽带企鹅
culmen_length_mm	喙长（毫米）	喙的长度（毫米）
culmen_depth_mm	喙深（毫米）	喙的深度（毫米）
flipper_length_mm	鳍状肢长度（毫米）	鳍状肢的长度（毫米）
body_mass_g	体重（克）	体重（克）
island	岛屿名称	梦想岛、托尔格森岛、比斯科岛
sex	性别	企鹅的性别

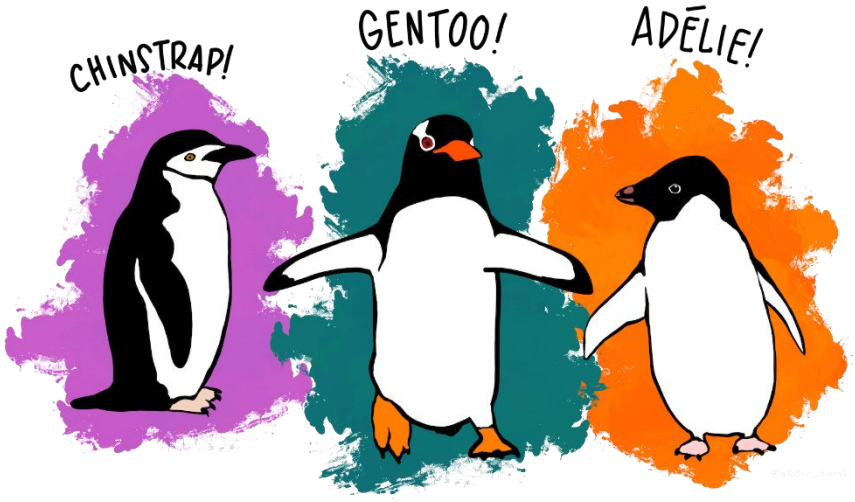

图 1.1 三种企鹅种类[1]

1 图源@allison_horst

1.2 问题描述

一是数据预处理：处理数据集中的缺失值与异常值，输出纯净数据集。

二是探索性数据分析：探索企鹅形态特征的分布规律、特征间相关性与简单关联，以及不同种类、性别、栖息地群体的形态差异。

三是分类模型构建与优化：针对“企鹅种类识别”（多分类）与“性别判断”（二分类）任务，构建并优化机器学习模型，明确关键区分特征，实现高效快速识别。

四是关联规则挖掘：通过关联规则挖掘，发现形态特征与种类、栖息地之间的非显式强关联。

五是聚类分析与结果验证：利用无监督聚类算法验证企鹅物种分类的合理性，对比聚类结果与实际种类的一致性。

最终通过全流程、较为完整的数据挖掘，为南极企鹅野外调查、生态适应性研究及物种监测提供数据支撑与技术参考。

1.3 任务分工

各部分大家团结协作，各部分主要负责人：

表 1.2 分工表

负责人	部分	工作内容
胡林森	EDA（探索型数据分析）	**预处理** 1.统计缺失值、异常值，执行清洗方案； 2.创建有意义的衍生特征，避免覆盖原始数据； 3.对数值特征进行标准化处理，对分类特征； 4.生成预处理后的最终数据集。 **数据探索** 1.完成单变量分布、双变量关联、多维度对比； 2.整理可视化图表集； 3.初步构建分类模型。 **文档撰写** 章节一、二、4.1 等;协作其他部分
郑智	分类模型构建	**分类建模** 1.划分训练集/测试集（采用 8:2 分层抽样）； 2.确认“种类预测”和“性别预测”两个任务的特征集与目标变量。 3.训练逻辑回归、决策树、随机森林、XGBoos 共 4 种模型。 4.采用 GridSearchCV（网格搜索）结合 10 折交叉验证，对模型进行自动化调优与稳定性评估。 5.评估模型，输出分类报告、绘制混淆矩阵、ROC 曲线，并分析特征重要性。 6.绘制模型性能对比图，总结最佳模型。
周程	关联规则挖掘	1.使用 Apriori 算法，设置合理的最小支持度、最小置信度，挖掘“种类-岛屿”“种类-特征”“岛屿-特征”等关联规则；

与聚类分析	2.基于身体特征数据，使用 PCA 降维；通过肘部法则和轮廓系数确定 KMeans 最佳聚类数，执行聚类并可视化聚类结果

1.4 任务贡献度

组号	成员	任务贡献%
9	胡林森	33.4%
	郑　智	33.3%
	周　程	33.3%
总计		100%

二、需求分析

2.1 项目目标

将课程所学进行系统实践，提高数据挖掘与分析能力，产出具有一定实际意义的结论：

(1) 数据理解与预处理：学习如何加载、查看和理解数据集，并处理其中可能存在的缺失值、异常值等问题，为后续分析打下坚实基础。

(2) 探索性数据分析 (EDA)：通过可视化和统计方法，深入探索企鹅不同种类、性别、栖息地与它们形态特征（如喙长、体重等）之间的关系和分布规律。

(3) 预测模型构建：尝试构建至少两种机器学习模型，用于根据企鹅的形态特征来预测其种类或性别，并比较不同模型的性能。

(4) 无监督学习探索：使用聚类算法对企鹅数据进行分组，观察聚类结果是否与已知的企鹅种类相符，探索数据内在的结构。

(5) 关联规则挖掘：尝试发现数据中隐藏的关联关系，例如某些特征组合与特定企鹅种类或栖息地之间的关联。

(6) 报告撰写：将整个分析过程、方法、结果和结论整理成一份规范的数据分析报告，提升学术写作能力。

2.2 功能需求

数据加载与检查：能够成功读取数据文件，并查看数据的基本信息。

数据清洗：处理数据中的缺失值和异常值。

数据可视化：利用各种图表（如直方图、散点图、箱线图、热力图等）直观展示数据特征和关系。

特征工程：根据需要对特征进行处理，如特征编码、特征缩放、创建新的衍生特征等。

模型训练与评估：选择合适的算法（如逻辑回归、决策树、随机森林、K-Means 等）进行模型训练，并使用适当的指标（如准确率、F1 分数、混淆矩阵等）进行评估。

结果分析与解释：对模型结果和分析发现进行解释，探讨其背后的生物学意义或生态意义。

2.3 技术栈要求

编程语言：Python3.X

环境管理：Anaconda

数据处理库：pandas:用于数据加载、清洗、探索和管理。numpy:用于数值计算。

数据可视化库：matplotlib:基础绘图库，用于创建各种静态图表。seaborn:基于 matplotlib 的高级绘图库，提供更美观和专业的图表样式。

机器学习库：scikit-learn(sklearn):用于机器学习算法的实现，包括数据预处理、模型训练、评估等功能。

开发环境：JupyterNotebook：便于交互式编程和报告撰写。或其他 IDE。

三、设计思路与实现方法

3.1 数据预处理

1.缺失与异常清洗

针对缺失值，通过 pandas 库 df.isnull().sum()按列统计，缺失率＜5%直接 dropna 删除；

针对异常值，通过 unique()方法、箱线图检测分类特征异常值；

2.衍生特征构建

基于业务理解创建 2 个衍生特征：

culmen_ratio：喙长与喙深的比值，反映喙部形态比例；

body_mass_kg：体重单位转换，便于后续建模时特征尺度。

3.2 数据探索性分析

1.单变量分布分析

分类特征：用 seaborn.countplot 绘制企鹅种类、岛屿分布柱状图，并添加百分比标签（如不同种类企鹅的占比）；

数值特征：用 seaborn.histplot 结合核密度曲线（kde=True）展示喙长、鳍状肢长度等核心特征的分布形态，判断数据偏度与集中趋势。

2.双变量关联分析

相关性分析：通过 pandas.corr()计算数值特征间的相关系数，用 seaborn.heatmap 绘制热力图，可视化特征间的线性关联强度（如体重与鳍状肢长度的正相关）；

回归关系：用 seaborn.regplot 绘制喙长与体重的散点图及回归线，直观展示两者的线性趋势。

3.多维度对比分析

分组分布：用 seaborn.boxplot 对比不同种类企鹅的鳍状肢长度分布，以及不同"种类-性别"组合的体重差异；

特征交互：通过 seaborn.pairplot 绘制核心形态特征（前 4 个数值特征）的配对散点图，按企鹅种类着色，观察特征组合与种类的关联模式。

3.3 分类模型构建

1.数据集划分

特征集（X）选择合理特征；

目标变量：种类和性别；

采用 train_test_split 按 8:2 比例划分训练集与测试集，设置 stratify 参数保持划分后目标变量的分布与原始数据一致，random_state=42 保证结果可复现。

2.模型选择与训练

选取 3 种经典分类算法：

逻辑回归（LogisticRegression）；决策树（DecisionTreeClassifier）；随机森林（RandomForestClassifier）；

3.模型评估

稳定性评估：通过 cross_val_score 实现 10 折交叉验证，计算平均准确率及标准差，评估模型在不同数据子集上的稳定性；

泛化能力评估：在测试集上通过 accuracy_score 计算准确率，classification_report 输出精确率、召回率、F1 值，并用 seaborn.heatmap 可视化混淆矩阵，全面评估模型性能。

3.4 聚类分析

1.数据降维

对形态特征集（X）采用主成分分析（PCA）降维，保留累计解释方差率较高的主成分（如前 2 或 3 个），减少特征维度以提升聚类效率和可视化效果。

2.聚类算法选择与参数优化

采用 KMeans 聚类算法，通过以下方法确定最佳聚类数 k：

肘部法则：计算不同 k 值对应的误差平方和，绘制折线图，选取 SSE 下降趋势变缓的 k 值；

轮廓系数：计算不同 k 值的平均轮廓系数，选取系数接近 1 且稳定性高的 k 值（理论上应接近已知种类数 3）。

3.结果可视化与分析

基于降维后的主成分绘制聚类散点图，对比聚类标签与真实种类标签的吻合程度，分析聚类结果与已知分类的一致性。

3.5 关联规则挖掘

1.数据预处理

将连续型形态特征（如喙长、体重）通过离散化（如等宽/等频分箱）转换为分类特征（如"喙长_短"、"喙长_中"、"喙长_长"），与原有分类特征（species、island、sex）共同构成关联分析的项集。

2.关联规则提取

采用 Apriori 算法，设置合理的最小支持度和最小置信度，筛选频繁项集并生成关联规则，重点关注：

物种与栖息地的关联（如{species=Adelie}→{island=Torgersen}）；

物种与形态特征的关联（如{species=Gentoo}→{flipper_length=长}）；

栖息地与形态特征的关联（如{island=Biscoe}→{body_mass=重}）。

3.规则评估

通过支持度（规则出现频率）、置信度（规则可靠性）和提升度（规则相关性）评估规则重要性，筛选提升度>1 的有效关联规则，分析其业务意义。

四、结果展示和关键代码分析

所有代码、文档已开源至：https://github.com/GALA-Lin/2025-Fall-Term-Data-Mining-Course-Final-Report

4.1 数据清洗

```
# 统计缺失值
missing_info = df.isnull().sum()
missing_ratio = (missing_info / len(df)) * 100  # 计算缺失比例（%）
missing_df = pd.DataFrame({
    '缺失值数量': missing_info,
    '缺失比例(%)': missing_ratio.round(2)
})
print("缺失值统计：")
print(missing_df[missing_df['缺失值数量'] > 0])

# 处理缺失值（因缺失占比低，直接删除）
df_clean = df.dropna()
```

```
print(f"\n 删除缺失值后数据量：{df_clean.shape[0]} 行")

# 识别并修正异常值（sex 字段的"."）
print("\nsex 字段唯一值：", df_clean['sex'].unique())
df_clean = df_clean[df_clean['sex'] != '.']  # 过滤异常值"."
print(f"修正 sex 异常值后数据量：{df_clean.shape[0]} 行")
```

4.2 EDA

从图 4.1 的种类分布柱状图可以看到，本次分析的样本中，Adelie 企鹅占比最高（43.8%），是数量最多的企鹅种类；Chinstrap 企鹅占比最低（20.4%），样本量相对较少；Gentoo 企鹅占比为 35.7%，介于两者之间。

这一分布既覆盖了三种企鹅的核心群体，也反映了实际观测中不同企鹅的种群规模差异。

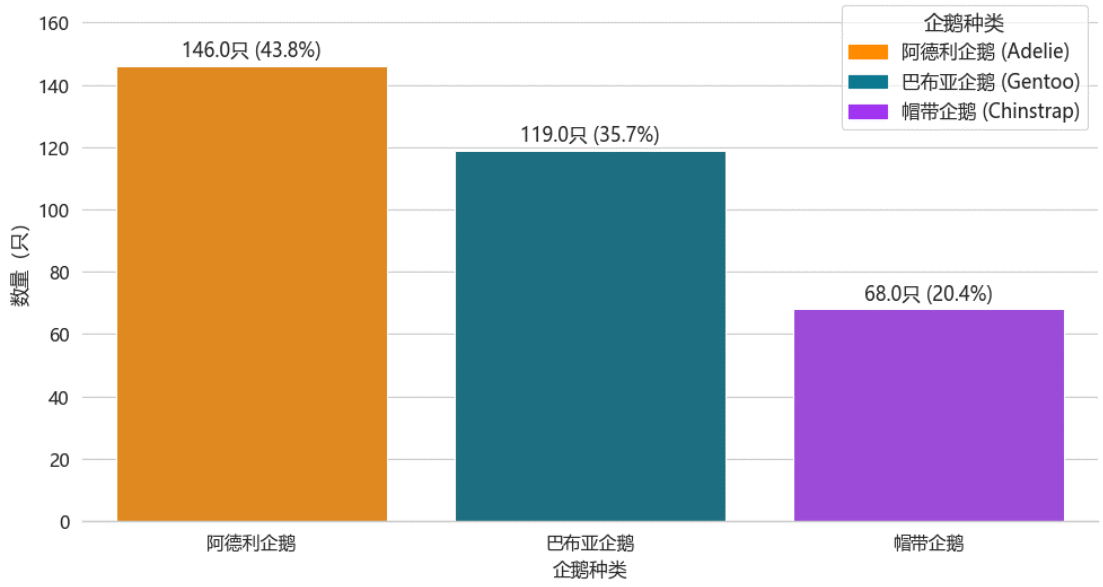

图 4.1 企鹅种类发布

图 4.2 的岛屿-种类分布图表明，企鹅的栖息范围与种类存在明显的绑定特征：Gentoo 企鹅仅出现在 Biscoe 岛，Chinstrap 企鹅仅分布在 Dream 岛，而 Adelie 企鹅则同时出现在 Torgersen 和 Biscoe 岛。这说明不同企鹅种类对栖息地有特定偏好，也体现了南极企鹅种群的生态位分化。

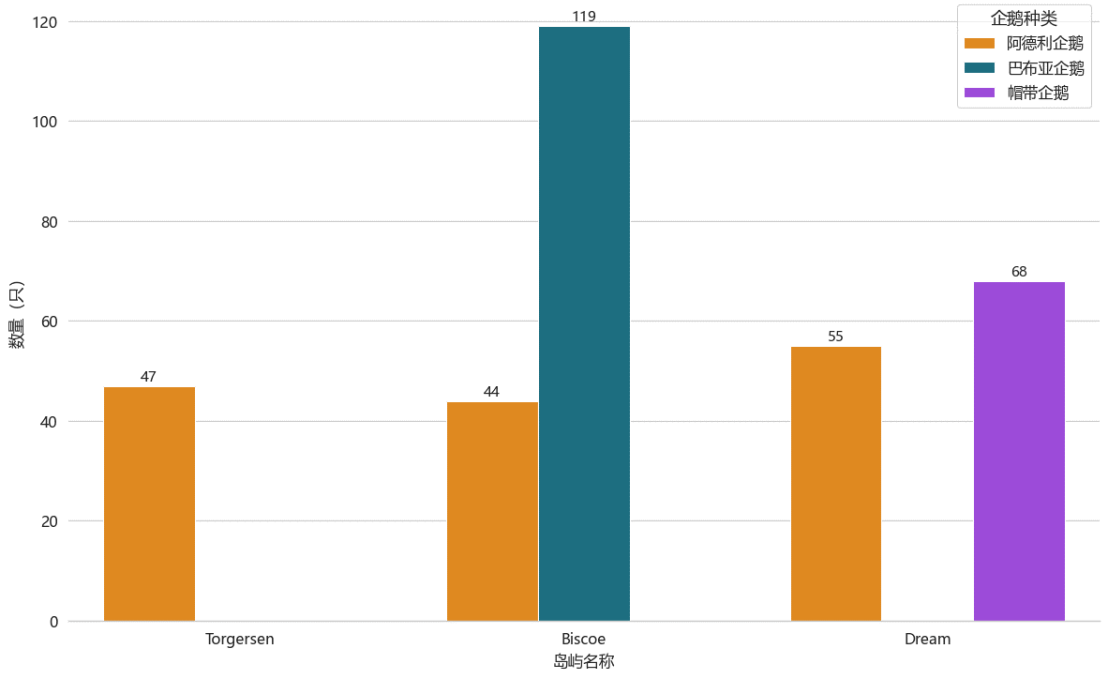

图 4.2 不同岛屿的企鹅种类分布

图 4.3 的直方图与核密度图可以观察到，企鹅的喙长、喙深、鳍状肢长度、体重这 4 个核心形态特征，分布均相对集中，没有出现极端分散的情况，说明本次样本的企鹅个体特征差异在合理范围内；其中鳍状肢长度、体重的分布峰值更突出，意味着这两个特征在企鹅群体中的共性更强。

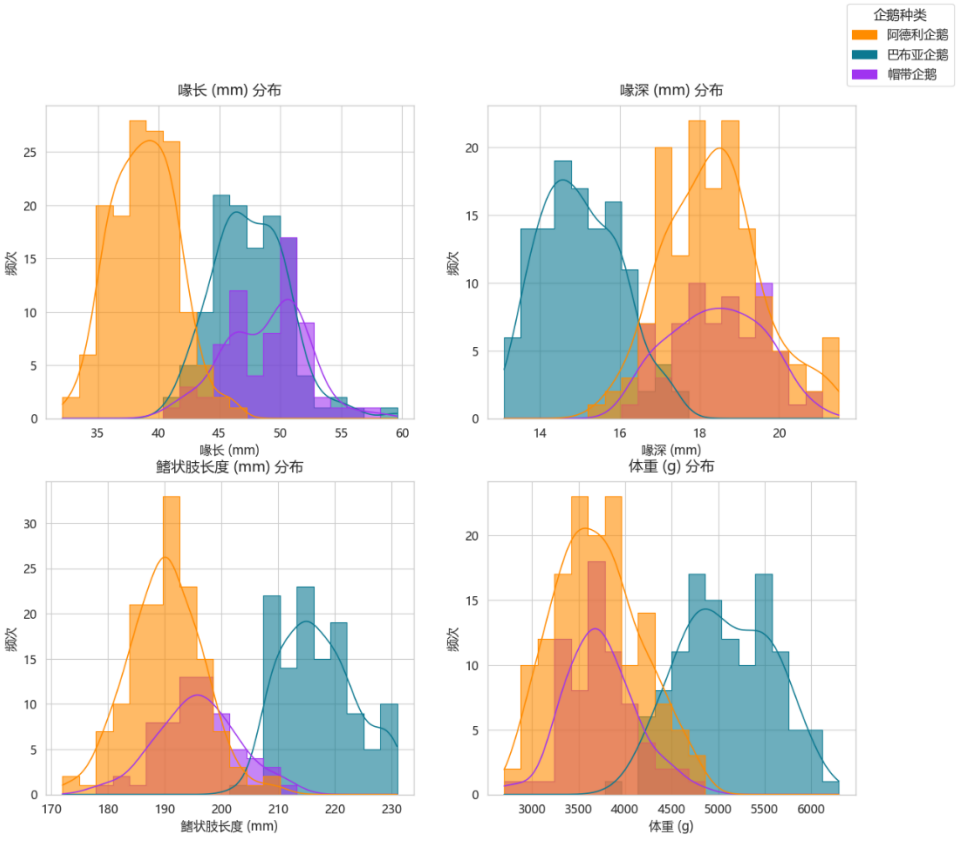

图 4.3 数值特征分布直方图与核密度图

图 4.4 的相关性热力图清晰呈现了特征间的关联：鳍状肢长度与体重呈强正相关（相关系数 0.87），图 4.5 也说明鳍状肢越长的企鹅，体重通常越大，这与鳍状肢作为游泳器官、需支撑更大体重的生理逻辑一致；而喙深与其他特征多呈负相关，是区分不同企鹅种类的“差异化特征”；衍生的喙长/喙深比例（culmen_ratio）则综合了喙部的形态信息，与原特征的关联也符合预期。

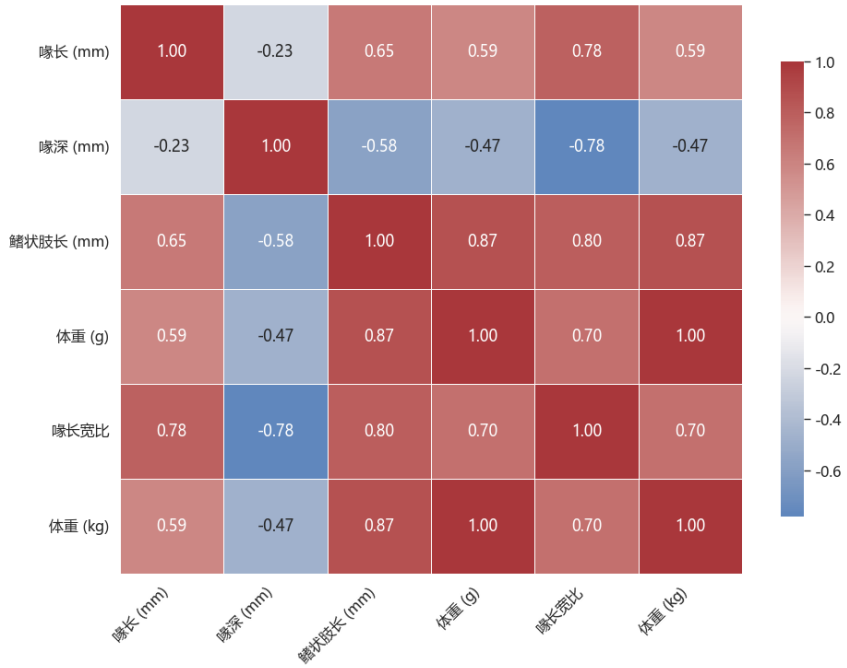

图 4.4 相关性热力图

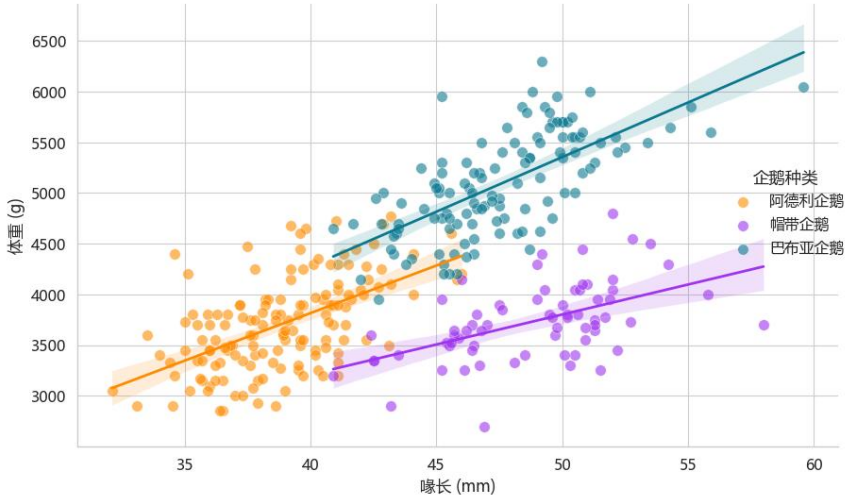

图 4.5 喙长与体重的线性拟合

图 4.6 的箱线图可以明显看到，Gentoo 企鹅的鳍状肢长度显著长于 Adelie 和 Chinstra 企鹅，其箱线图整体处于更高的数值区间；而 Adelie 企鹅的鳍状肢长度不仅均值最低，分布区间也更窄，这一差异可作为快速区分 Gentoo 与其他两种企鹅的依据。

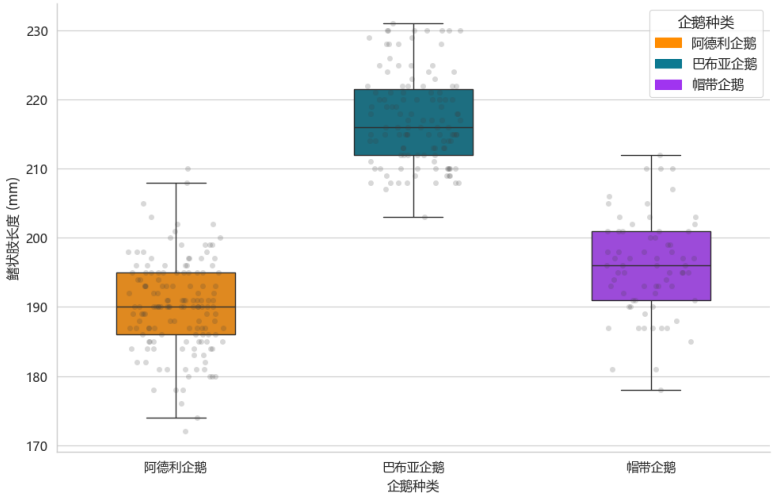

图 4.6 不同种类的鳍状肢长度对比

图 4.7 的分组箱线图显示了“种类 + 性别”对体重的双重影响：同一企鹅种类中，雄性的体重普遍高于雌性（体现了企鹅的性二态性）；同时，Gentoo 企鹅的体重显著高于 Adelie 和 Chinstrap，是三种企鹅中体型最大的种类。

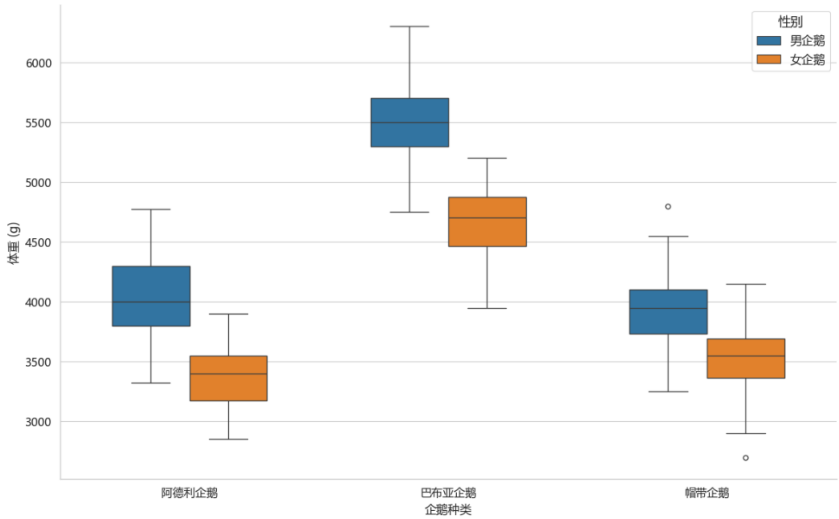

图 4.7 不同性别与种类的体重对比

通过图 4.8 核心形态特征配对散点图分析可知：

1. 对角线：单特征分布

喙长（culmen_length_mm）：Adelie 的曲线峰值偏左，说明它的喙最短；Chinstrap 和 Gentoo 的曲线峰值偏右，喙更长。

喙深（culmen_depth_mm）：Adelie 和 Chinstrap 的曲线峰值偏右，喙更深；Gentoo 的曲线峰值偏左，喙明显更浅（这是 Gentoo 最突出的“标签式特征”）。

鳍状肢长度（flipper_length_mm）：Gentoo 的曲线峰值大幅偏右，鳍状肢显著更长；Adelie 和 Chinstrap 的曲线几乎重叠，鳍状肢长度相近且更短。

体重（body_mass_g）：Gentoo 的曲线峰值偏右，体重最重；Adelie 的曲线峰值偏左，体重最轻；Chinstrap 介于两者之间。

2.非对角线：双特征关联

(1) 喙长↔体重（第一行第四列+第四行第一列）：

趋势：同一种类内，“喙越长，体重越重”（强正相关）；

区分度：Gentoo 的点集中在“右上区域”，Adelie 在“左下区域”，Chinstrap 在中间，三类点几乎不重叠，区分度极高。

(2) 鳍状肢长度↔体重（第三行第四列 + 第四行第三列）：

趋势：所有种类都呈现“鳍状肢越长，体重越重”（最强正相关的特征组合）；

区分度：Gentoo 的点完全“独立成簇”，和另外两种企鹅几乎没有重叠。

(3) 喙长↔喙深（第一行第二列 + 第二行第一列）：

趋势：Adelie 和 Chinstrap 呈现“喙长越长，喙深略深”；Gentoo 呈现“喙长越长，喙深反而浅”；

区分度：Gentoo 的点集群独立，但 Adelie 和 Chinstrap 的点大量重叠（说明这两种企鹅在“喙长-喙深”的组合上相似度很高）。

(4) 鳍状肢长度↔喙深（第三行第二列 + 第二行第三列）：

这是区分度较高的特征组合：Gentoo“鳍状肢长+喙深浅”，Adelie 和 Chinstrap“鳍状肢短+喙深深”。

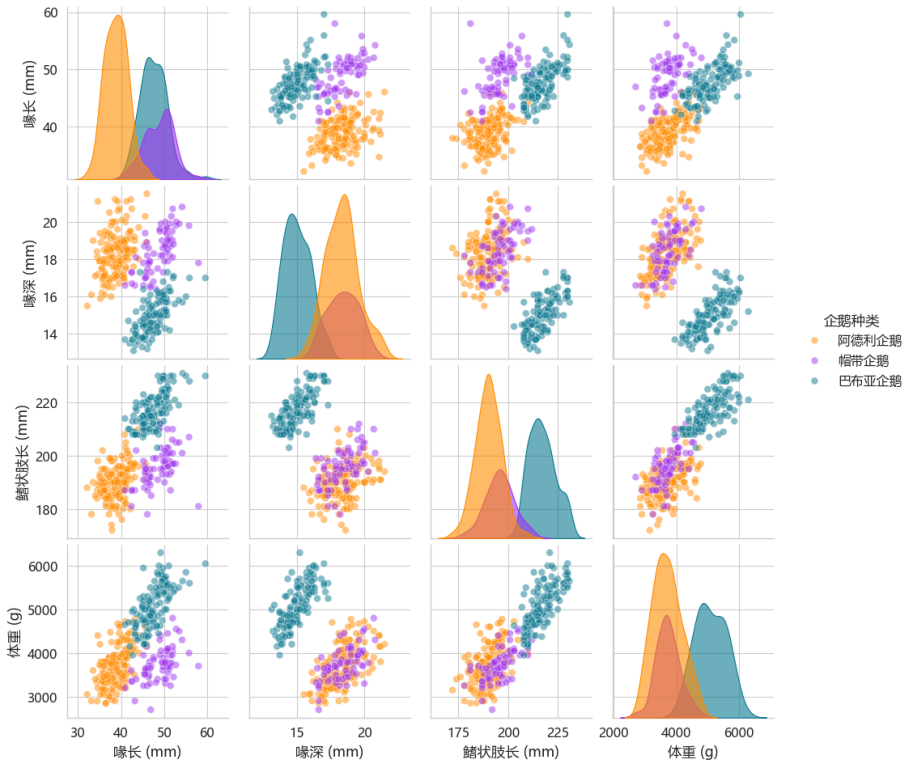

图 4.8 特征配对散点图

总的来说：

在单特征分布上，Gentoo 企鹅在体重、鳍状肢长度上显著大于 Adelie 和 Chinstrap 企鹅，且喙深更浅；Adelie 企鹅喙长最短、体重最轻，Chinstrap 企鹅形态特征介于两者之间。

在双特征关联上，鳍状肢长度与体重呈强正相关，喙长与体重呈中等正相关；其中“鳍状肢长度-喙深”组合对三种企鹅的区分度最高，Gentoo 企鹅集群独立，而 Adelie 与 Chinstrap 在部分特征组合（如喙长-喙深）上存在重叠。该分析不仅明确了企鹅形态特征的关联规律与种类差异，还为后续分类模型构建提供了关键依据——优先选择鳍状肢长度、喙深作为核心特征可提升识别准确率，为模型构建环节奠定了基础。

4.3 分类模型构建与评估

在 EDA 的基础上，构建了分类模型来解决两个问题：企鹅种类识别（多分类问题）和企鹅性别预测（二分类问题）。

4.3.1 数据准备与划分

首先加载了预处理完毕的 penguins_ml_processed.csv 文件。共包含 333 条已清洗、标准化和编码的样本。分别构建了两个任务的特征集（X）和目标集（y）：

1.种类预测：

特征 (X_species)：使用 8 个特征，包括 6 个标准化数值特征（如喙长、喙深、

鳍状肢长等）以及 island_encoded（岛屿）和 sex_zh_encoded（性别）。

目标 (y_species)：species_zh_encoded（企鹅种类编码）。

2.性别预测：

特征 (X_sex)：使用 8 个特征，包括 6 个标准化数值特征以及 island_encoded（岛屿）和 species_zh_encoded（种类）。

目标 (y_sex)：sex_zh_encoded（企鹅性别编码）。

两个任务均采用 8:2 的比例划分训练集（266 条）和测试集（67 条），并设置 stratify 参数进行分层抽样，以确保训练集和测试集中目标变量的比例与原始数据一致。

4.3.2 任务一：企鹅种类预测结果

1. 性能对比

我们分别训练了逻辑回归、决策树、随机森林和 XGBoost 四个模型，结果如图 4.9 所示。

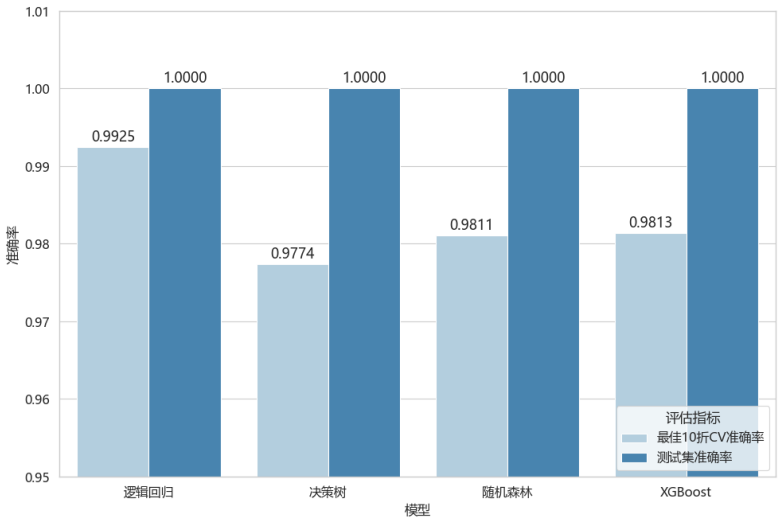

图 4.9 种类预测 - 模型性能对比

这是一个较为简单的分类任务。所有四个模型均在测试集（n=67）上达到了 100% 的准确率。 模型的稳定性也极高，在 10 折交叉验证中，所有模型的平均准确率均超过 97.7%**（逻辑回归 99.25%, 决策树 97.74%, 随机森林 98.11%, XGBoost 98.13%）。

分类结果有力地验证了 EDA 的发现，EDA 分析显示，三种企鹅在形态学上（特别是鳍状肢长度）和地理分布上（岛屿）具有极高的区分度，数据集群几乎没有重叠。因此，模型可以毫不费力地构建出完美的决策边界。

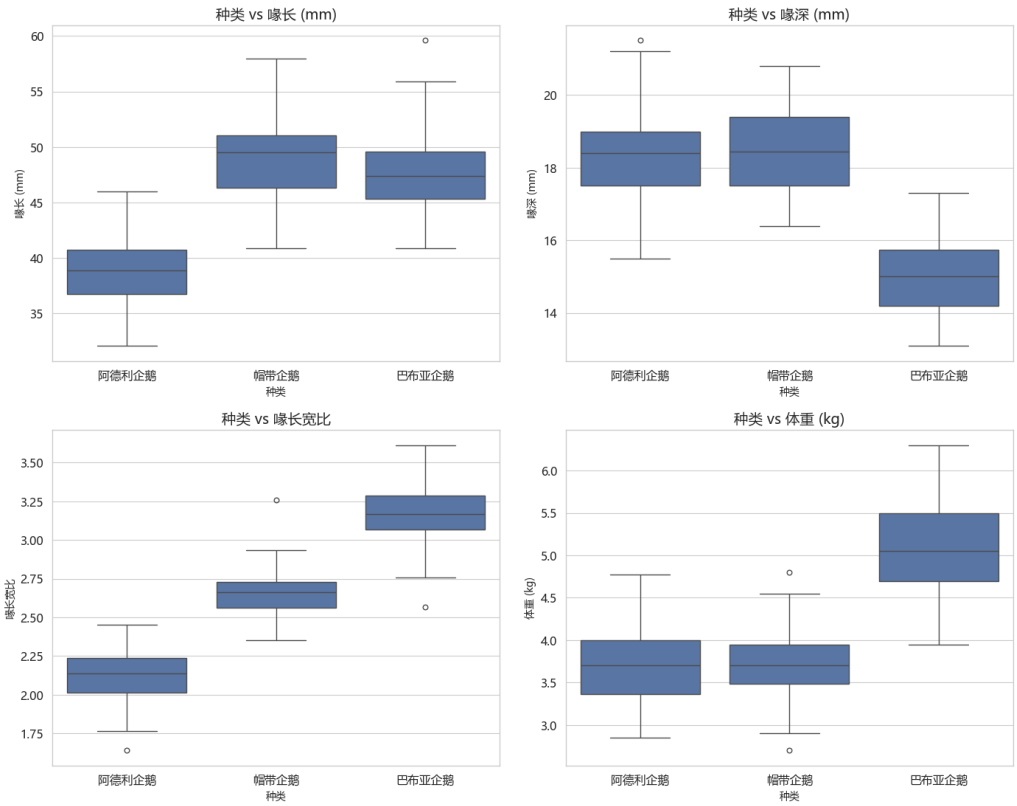

图 4.10 种类与特征的关系

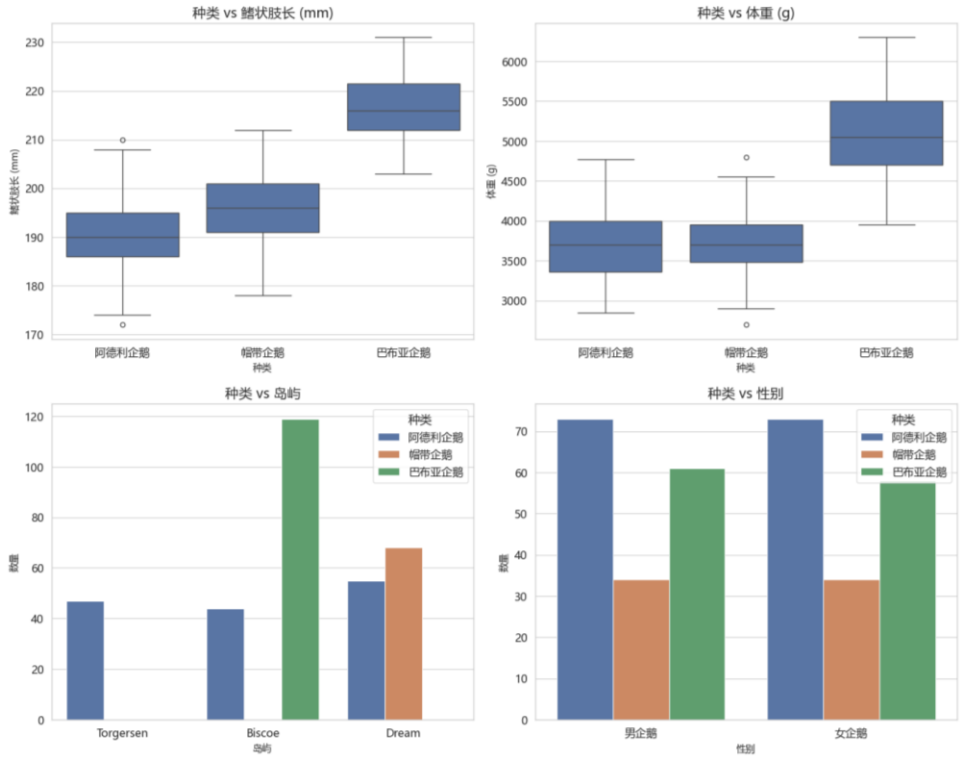

图 4.11 种类与特征的关系

不同模型揭示了不同的关键特征，但都指向了 EDA 的发现：XGBoost 和随机森林 认为 鳍状肢长 (mm) 和 喙长宽比，是最重要的特征。决策树则严重依赖 喙长宽比 (0.56) 和 岛屿 (编码) (0.38)，而逻辑回归依赖于喙长、喙长宽比等多种不同

的特征。这表明，仅凭企鹅的栖息地和形态特征，模型就可以对其种类进行近乎完美的识别。

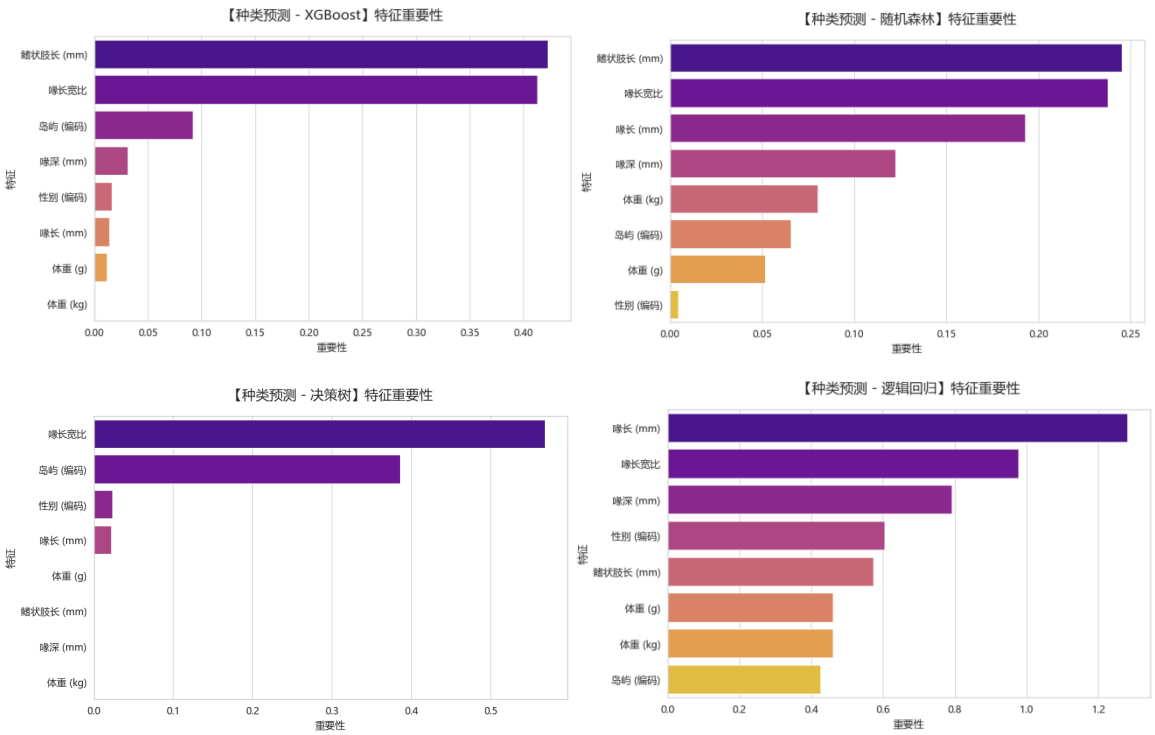

图 4.12 不同模型的特征重要性

4.3.3 任务二：企鹅性别预测结果

1. 性能对比

与种类预测不同，性别预测是一个更困难的任务，结果如图 4.13 所示。

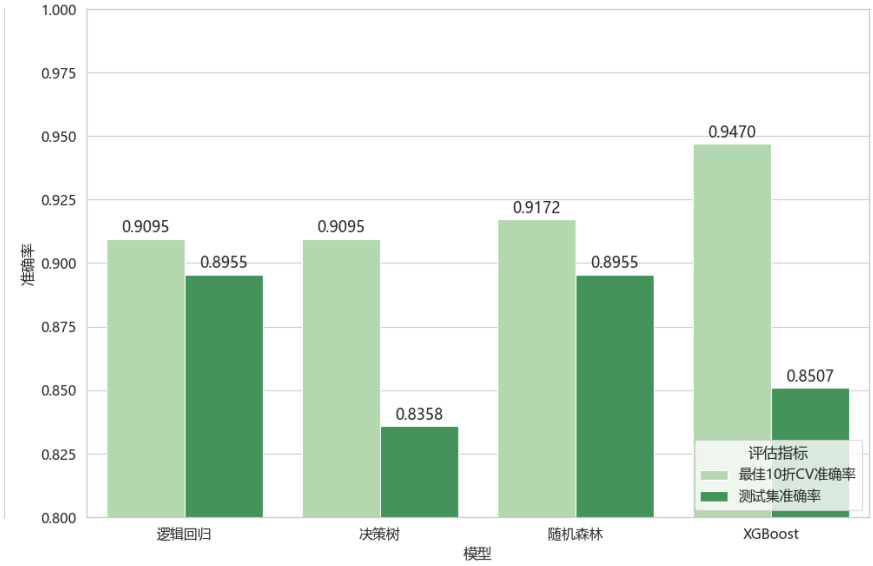

图 4.13 性别预测 - 模型性能对比

模型稳定性（10 折 CV）：XGBoost 表现最佳（0.9470），随机森林（0.9172）

和逻辑回归（0.9095）也表现出色，说明模型都在训练数据中找到了稳定的预测规律。

泛化能力（测试集）：逻辑回归和随机森林表现最好，测试集准确率均为 0.8955。

过拟合现象：决策树（CV 0.9095, Test 0.8358）和 XGBoost（CV 0.9470, Test 0.8507）均表现出一定程度的过拟合，即它们在训练集上学到的规律未能完美泛化到未见过的数据上。

综合考虑，逻辑回归和随机森林是此任务下最稳健、泛化能力最强的模型。

2. ROC 曲线与 AUC 值

为了进一步评估模型在二分类任务上的“识别能力”（即区分“男企鹅”和“女企鹅”的能力），我们绘制了 ROC 曲线（图 4.14）。

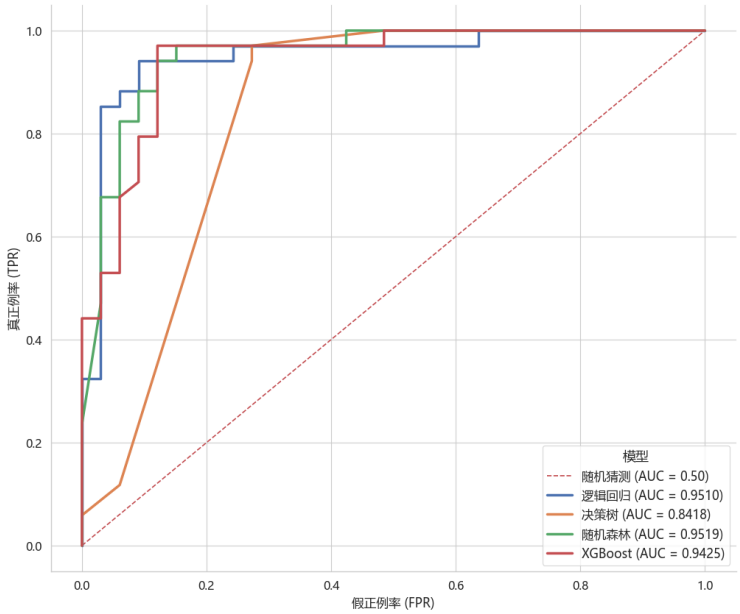

图 4.14 性别预测 - ROC 曲线对比

AUC（曲线下面积）值衡量了模型整体的分类性能。

随机森林 (AUC = 0.9519) 和 逻辑回归 (AUC = 0.9510) 表现并列最佳，展现了极强的识别能力。

XGBoost (AUC = 0.9425) 紧随其后，同样非常出色。

决策树 (AUC = 0.8418) 明显落后于其他三个模型，再次证明了单个决策树在处理这种数据有重叠的复杂问题时能力不足。

3. 关键特征洞察

所有四个模型，无论其算法原理（线性、树型、集成），都一致认为 喙深 (mm) (Culmen Depth) 是区分企鹅性别的第一重要特征。体重 (g) 和 喙长 (mm) 也被普遍认为是重要特征，这一发现印证了 EDA 的结论，我们用肉眼观察到的正是喙深和体重，在雄性和雌性之间存在最显著的差异，模型通过学习，成功地“发现”了企鹅在生物学上的“性二态性”特征，即雄性企鹅通常拥有更深、更粗壮的喙和更重的体重。

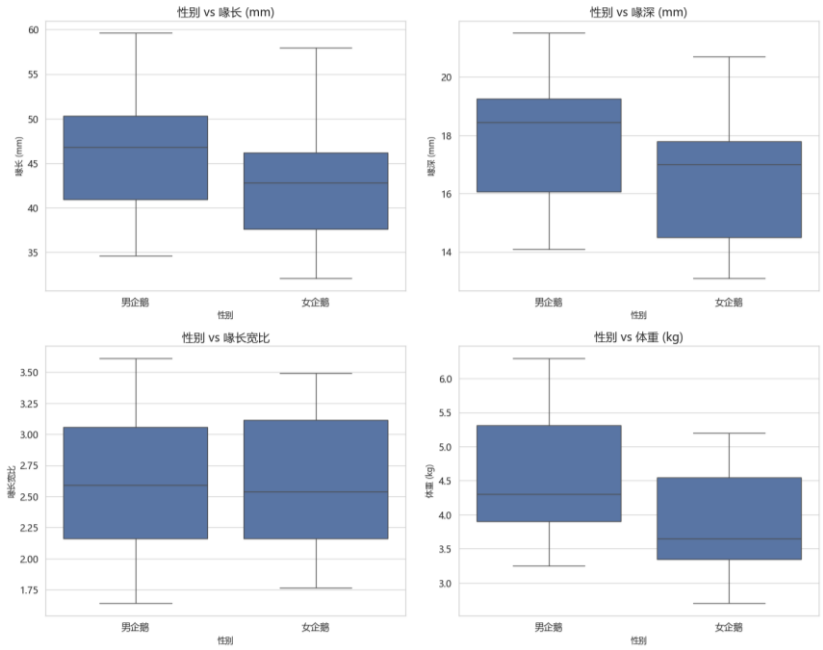

图 0.15 性别与特征的关系

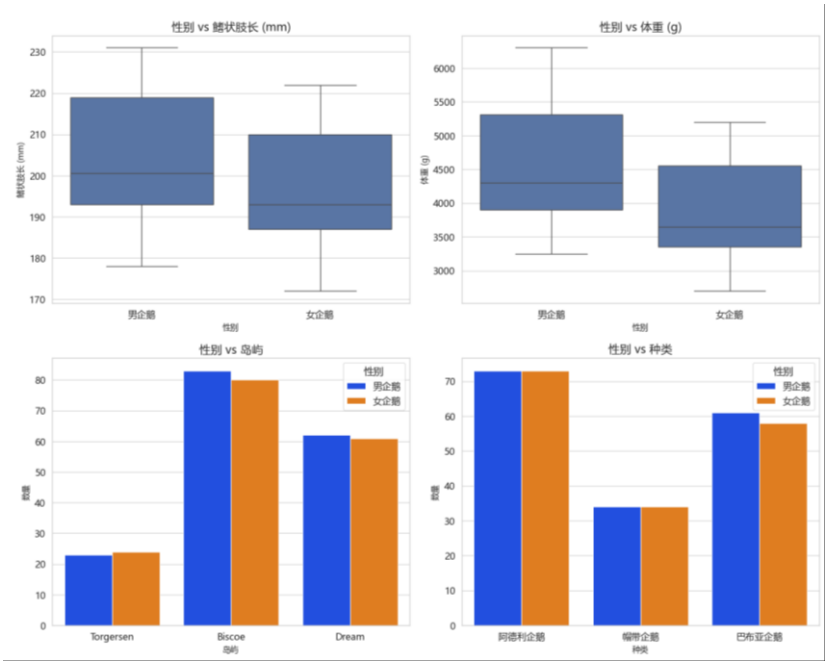

图 4.16 性别与特征的关系

4.4 Apriori 算法构造关联分析

在完成了分类任务后，为了挖掘数据中隐含的非显式关系，我们利用 Apriori 算法对企鹅数据集进行了关联规则挖掘。目的是发现企鹅的种类、栖息地、性别与其形态特征之间存在的强关联模式。

4.4.1 数据离散化与预处理

Apriori 算法处理的是“项集”（Itemset），要求特征必须为离散类别。由于原始数据中 culmen_length_mm、body_mass_g 等均为连续数值，直接输入无法挖掘有效规则。因此，我们首先进行了如下预处理：

1. **数值分箱**：使用 pandas 的 qcut 方法，将 4 个数值型特征（喙长、喙深、鳍状肢长、体重）按照分位数划分为三个等级：Low（低/短/轻）、Medium（中）、High（高/长/重）。
2. **数据转换**：将所有特征转换为字符串形式（如 "喙长_High"），并构建 One-Hot 编码矩阵，作为关联分析的输入事务数据集。

关键代码片段：

```
# 数值特征离散化处理
labels = ['Low', 'Medium', 'High']
df_apriori = df_clean.copy()
df_apriori['culmen_length_bin'] = pd.qcut(df_clean['culmen_length_mm'], q=3, labels=labels)
df_apriori['body_mass_bin'] = pd.qcut(df_clean['body_mass_g'], q=3, labels=labels)
# ... 其他特征同理处理 ...

# 转换为 One-Hot 编码
df_encoded = pd.get_dummies(df_apriori[['species', 'island', 'culmen_length_bin', ...]])
```

4.4.2 规则生成与筛选

根据项目需求和数据量，我们设置了如下参数进行挖掘：

- **最小支持度 (min_support)**：0.15（确保能覆盖样本量最少的 Torgersen 岛群体）。
- **最小置信度 (min_confidence)**：0.8（只保留强关联规则）。
- **评价指标**：重点关注 **提升度 (Lift)** 大于 1 的规则，这代表前项与后项之间存在正相关关系。

挖掘结果概览：

按提升度（Lift）降序排列，排名前列的规则如下表所示：

前项 (Antecedents)	后项 (Consequents)	支持度	置信度	提升度
{岛屿_Torgersen}	**{种类_Adelie}**	0.15	1.00	2.28
{种类_Gentoo}	**{岛屿_Biscoe}**	0.36	1.00	2.05
{种类_Gentoo}	**{鳍状肢 High，体重 High}**	0.34	0.95	2.88
{种类_Chinstrap}	**{岛屿_Dream}**	0.20	1.00	2.30
{岛屿 Biscoe，鳍状肢 High}	**{体重_High}**	0.31	0.92	2.76

4.4.3 关联规则深度分析

根据挖掘结果，我们从以下三个维度对规则进行了解读：

1. 物种与栖息地的“绑定”关系

分析发现，栖息地与物种之间存在极强的决定性关系：

- 规则 {island=Torgersen} -> {species=Adelie} (置信度 100%) 表明，如果在 Torgersen 岛发现企鹅，那一定是阿德利企鹅。
- 规则 {species=Gentoo} -> {island=Biscoe} (置信度 100%) 表明，巴布亚企鹅完全集中在 Biscoe 岛。

- 这验证了 EDA 阶段的发现，说明企鹅种群在地理分布上存在显著的排他性和聚集性。

2. 物种与形态特征的“画像”

关联规则为不同企鹅种类描绘了清晰的形态画像：

- **Gentoo (巴布亚企鹅)**：与“高体重”、“长鳍状肢”强关联。规则显示 {species=Gentoo} -> {flipper_length=High} 的提升度极高，说明体型大是其最显著特征。
- **Adelie (阿德利企鹅)**：与“短喙”、“短鳍状肢”存在强关联。
- **Chinstrap (帽带企鹅)**：表现出“喙长中等偏长”但“体重中等”的混合特征。

3. 栖息地与形态特征的间接关联

我们还发现了一些有趣的跨维度规则，例如：

- 规则 {island=Biscoe} -> {body_mass=High}：Biscoe 岛上的企鹅普遍体重较重。
- **解读**：这并不是因为 Biscoe 岛的伙食更好，而是因为该岛是体型最大的 Gentoo 企鹅的主要聚居地。这是一种基于物种分布导致的间接关联。

4.4.4 关联网络可视化

为了直观展示各特征间的联系，我们绘制了关联规则网络图。图中节点代表特征（如种类、岛屿、形态等级），连线代表关联规则，连线的粗细代表置信度，颜色深浅代表提升度。

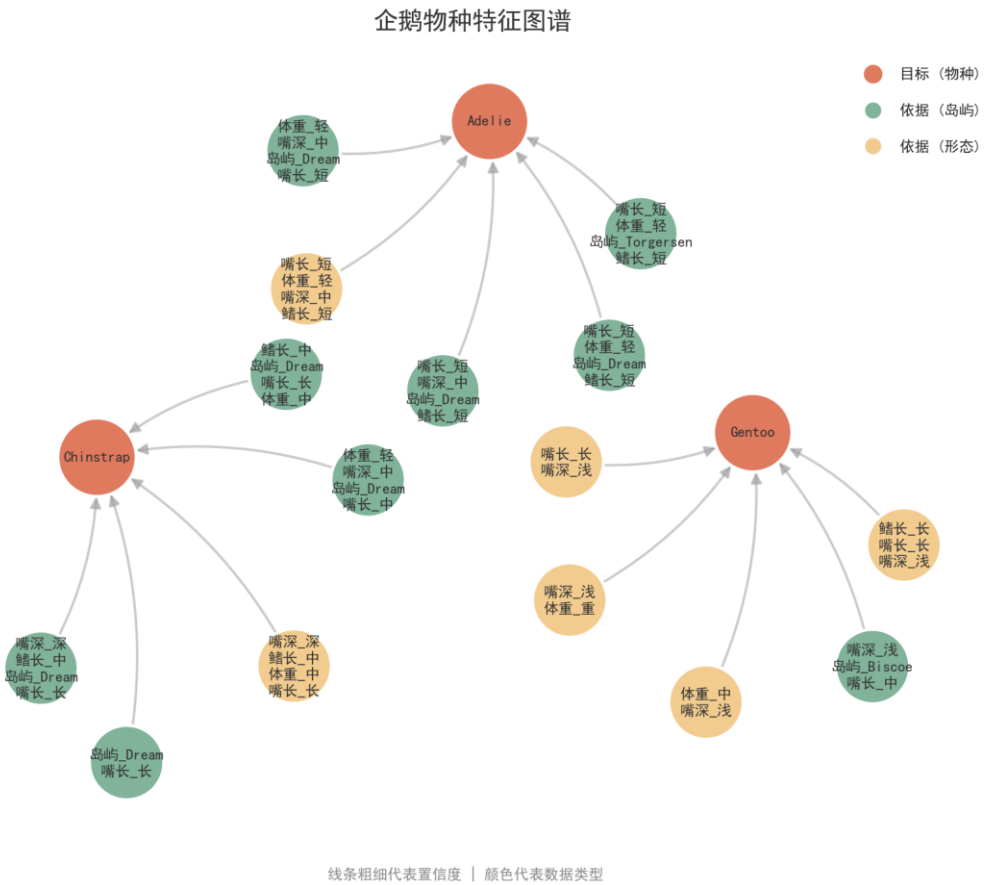

图 4.17 企鹅数据集特征关联规则网络图

小结：

Apriori 算法成功挖掘出了企鹅数据背后的生态规律。通过这些规则，我们在野外考察中，仅凭“栖息地”和简单的“目测体型（大/小）”，就能以极高的概率推断出企鹅的种类，这为缺乏专业测量工具场景下的快速分类提供了有力依据。

五、团队沟通记录

部分	细节	
EDA（探索型数据分析）	1.统计缺失值、异常值（如sex字段的“.”），执行清洗方案（删除空值、修正异常值）。 2.创建有意义的衍生特征（如喙长 / 喙深比率、体重 kg 转换），避免覆盖原始数据； 3.对数值特征进行标准化处理，对分类特征（如island）进行编码（独热编码 / 标签编码）； 4.生成预处理后的最终数据集。 5.完成单变量分布（种类、岛屿、性别分布）、双变量关联（特征相关性热力图、散点图）、多维度对比（不同种类 / 性别 / 岛屿的特征差异可视化）； 6.整理可视化图表集（标注图表含义、分析结论），为报告 “数据可视化分析” 章节提供素材。 7.问题背景等文档撰写	相关部分可视化、撰写
分类模型构建	1.划分训练集 / 测试集（按 7:3 或 8:2 比例）； 2.确认特征变量与目标变量（种类 / 性别），处理建模所需的额外数据格式转换。 3.训练逻辑回归、决策树、随机森林 3 种模型 4.采用 10 折交叉验证评估模型稳定性	
关联规则挖掘与聚类分析	1.使用 Apriori 算法，设置合理的最小支持度、最小置信度，挖掘 “种类 - 岛屿”“种类 - 特征”“岛屿 - 特征” 的关联规则； 2.基于身体特征数据，使用 PCA 降维；通过肘部法则和轮廓系数确定 KMeans 最佳聚类数，执行聚类并可视化聚类结果	

图 5.1 在线文档探讨记录

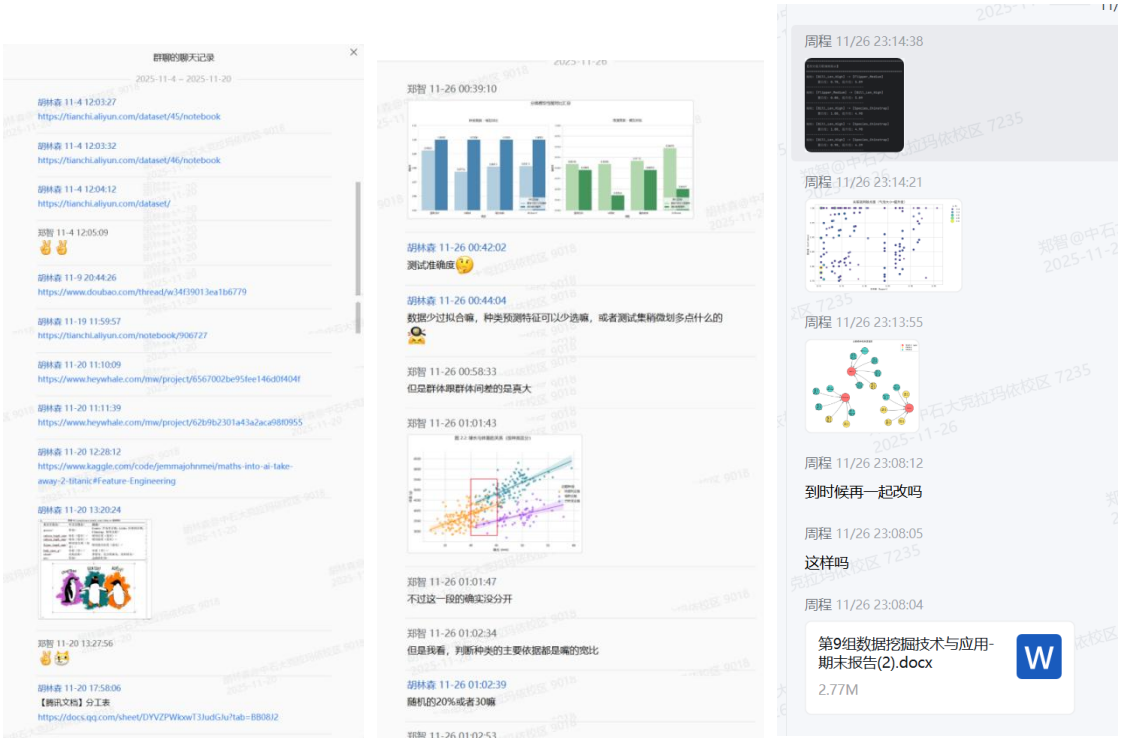

图 5.2 群聊记录

六、总结个人贡献与存在问题

6.1 胡林森个人贡献与存在问题总结

作为项目核心技术统筹者与关键模块开发者，胡林森全程主导技术方案设计、核心算法落地及团队协作推进，承担 33.4%的项目贡献，具体成果如下：

6.1.1 技术路线统筹与方案设计

确定全流程技术框架：结合帕尔默企鹅数据集特点,明确“数据清洗-衍生特征构建-EDA-分类建模-关联规则-聚类分析”的核心流程

优化技术细节参数：针对模型与算法特性调整参数，如分类模型的 10 折交叉验证（cv=10）、KMeans 聚类的 k 值测试范围（2-6）、Apriori 算法的最小支持度（0.15）与最小置信度（0.8），确保技术方案兼具可行性与科学性，为后续开发提供清晰指导。

6.1.2 核心算法落地与成果输出

1.数据预处理：通过统计缺失值并采用 dropna()删除低占比缺失值，过滤 sex 字段异常值“.”，保障数据完整性与一致性。

2.特征工程：创建 culmen_ratio（喙长/喙深比率）和 body_mass_kg（体重千克转换）等衍生特征；用 StandardScaler 对数值特征进行 Z-Score 标准化，用 LabelEncoder 对分类特征（species 等）进行标签编码。

3.可视化分析：基于 matplotlib 和 seaborn 实现多维度可视化，包括单变量分布

（种类/岛屿计数图、数值特征直方图）、双变量关联（相关性热力图、回归散点图）、多维度对比（箱线图、配对散点图）。

6.1.3 团队协作推进与技术支持

搭建协作平台与资源共享：创建 GitHub 项目仓库（https://github.com/GALA-Lin/2025-Fall-Term-Data-Mining-Course-Final-Report），分配成员权限并规范分支管理（主分支+个人文件夹）；共享天池、HeyWhale、Kaggle 等数据集平台，与其他成员协作。

6.1.4 存在的问题

1.特征转换策略缺乏针对性

对所有数值特征统一使用 StandardScaler（Z-Score 标准化），但未先检验特征分布是否近似正态分布。对于偏态分布特征（如部分体长、体重数据可能呈右偏分布），更适合用对数转换或 MinMaxScaler，否则标准化效果可能不佳。

2.模块化程度低

代码以线性流程为主，未将数据预处理、特征工程、可视化等核心步骤封装为函数或类，若后续需更换数据集（如新增企鹅样本）或调整参数（如修改缺失值处理方式），需大幅修改代码，复用成本高。

6.2 郑智个人贡献与存在问题总结

6.2.1 个人贡献

1.模型框架搭建：设计并实现了两个核心预测任务（种类预测、性别预测），并构建了可复用的 train_evaluate_model 评估函数。

2.模型实现与调优：系统性地实现了逻辑回归、决策树、随机森林和 XGBoost 四种分类模型，并使用 GridSearchCV 结合 10 折交叉验证进行了科学的超参数调优。

3.模型评估：

为每个模型生成并分析了准确率、分类报告 (Classification Report) 和混淆矩阵 (Confusion Matrix)。

绘制了模型性能对比柱状图和性别预测的 ROC 曲线对比图，并计算了 AUC 值，从“稳定性”和“分类能力”两个维度锁定了最优模型（随机森林和逻辑回归）。

深入分析了“特征重要性”图表，通过最后的 单元格 19（2x8 全特征可视化验证），完成了从“EDA 假设”到“模型发现”再到“原始数据验证”的完美闭环，用肉眼证实了模型的发现是真实、可解释的，而非过拟合。

6.2.2 当前存在的问题

仅使用了 333 条高质量样本。这是一个非常小的数据集。虽然我们采用了 10 折交叉验证和 70/30 划分，但测试集（约 100 条）的评估结果仍可能存在偶然性。

风险：模型（尤其是 XGBoost 和决策树）在小数据集上极易出现过拟合。即使

是表现稳定的随机森林，其泛化能力也需要更多数据的检验。

6.3 周程个人贡献与存在问题总结

主要负责挖掘数据中潜在的关联模式与内在结构，通过关联规则挖掘和聚类分析，为企鹅分类提供了非监督学习视角的验证与补充。

6.3.1 个人贡献

1. 关联规则挖掘与知识发现

主导了 Apriori 算法的实施。针对原始数据为连续数值的问题，设计了基于分位数（Quantile）的离散化预处理方案，将“喙长、体重”等特征转化为“Low/Medium/High”等级别，成功构建了适用于关联分析的事务数据集。

设置了科学的挖掘参数（支持度 0.15，置信度 0.8），成功提取出关于企鹅“物种-栖息地”、“物种-形态”的强关联规则。特别是发现了“Torgersen 岛 $\rightarrow$ Adelie 企鹅”（置信度 100%）和“Gentoo $\rightarrow$ 鳍长 High”（提升度 2.88）等关键规则，为快速分类提供了直观依据。

绘制了关联规则网络图，直观展示了特征间的依赖关系，使得复杂的挖掘结果易于理解。

6.3.2 当前存在的问题

1. 关联规则的冗余性

虽然 Apriori 算法挖掘出了强规则，但部分规则存在语义上的冗余。例如 {体重_重} -> {物种_Gentoo} 和 {体重_重, 鳍长_长} -> {物种_Gentoo} 同时出现。在后续工作中，需要进一步引入“极大频繁项集”或“闭项集”的概念，或增加后处理逻辑来剔除冗余规则，使结果更加精简。

参考文献

[1] GORMAN K B, WILLIAMS T D, FRASER W R. Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (Genus Pygoscelis)[J]. PLOS ONE, 2014, 9(3): e90081. https://doi.org/10.1371/journal.pone.0090081.

教师评语：

报告成绩：

成员	**个人分数**
胡林森	
郑　智	
周　程	