西安交通大學

XI'AN JIAOTONG UNIVERSITY

大数据课程大作业报告

水质监测数据挖掘

课程名称：	大数据科学与应用技术
姓名：	李艺涵
学院：	电信学部
专业：	自动化钱 001
学号：	2206123627
指导老师：	焦在滨

2023 年 6 月 24 日

西安交通大學实验报告

专业：自动化钱 001
姓名：李艺涵
学号：2206123627
日期：2023 年 6 月 24 日
地点：

课程名称：大数据科学与应用技术　指导老师：焦在滨　成绩：
实验名称：水质监测数据挖掘　实验类型：　同组学生姓名：

一、 背景

在人类生活的基本需求中，安全的饮用水是至关重要的。我们的身体需要足够的水分来维持正常的生理功能和代谢过程，水对于体温调节、细胞功能、器官运作以及营养物质吸收都起着至关重要的作用，因此，饮用水的质量直接关系到我们的健康和生活质量。然而，许多国家和地区目前还面临着饮用水不足、水污染、供水基础设施不完善等问题。这导致了生活在这些地区的人们的健康受到严重影响，引发了许多水源传播的疾病。

水质监测是保障饮用水安全的必要手段，而本次大作业旨在通过 Python 数据挖掘和分析来揭示水体的各项物理和化学指标与可饮用性之间的关系，从而更好地辅助水质判定。

二、 数据集的获取

本次作业数据集来自开源数据集网站 Kaggle: https://www.kaggle.com/datasets/adityakadiwal/water-potability 该数据集包含了多组样本水体的如下信息：

(1) pH 值 (pH)：pH 是评估水的酸碱平衡的重要参数, 它也是水状况酸性或碱性的指标。世界卫生组织建议 pH 的最大容许限制为 6.5 至 8.5。当前的调查范围为 6.52 至 6.83，处于世界卫生组织标准范围内。

(2) 硬度 (Hardness)：硬度主要由钙和镁盐引起, 这些盐是从水经过的地质沉积物中溶解出来的。水与产生硬度的物质接触的时间长度有助于确定原始水中的硬度程度。硬度最初被定义为水由于钙和镁而沉淀肥皂的能力。

(3) 总溶解固体 (Solids)：水具有溶解多种无机和一些有机矿物质或盐类的能力，例如钾、钙、钠、碳酸氢盐、氯化物、镁、硫酸盐等。这些矿物质会给水的味道和外观带来不必要的味道和色彩。这是水使用的重要参数。总溶解固体（TDS）值的水表示水的矿化程度很高。TDS 的理想限制为 500 毫克/升，最大限制为 1000 毫克/升，这是为饮用目的规定的。

(4) 氯和氯胺含量 (Chloramines)：氯和氯胺是公共供水系统中常用的主要消毒剂, 氯胺通常是通过向氯中加入氨来处理饮用水时形成的。在饮用水中，每升氯的含量高达 4 毫克（mg/L）或 4 百万分之一（ppm）被认为是安全的。

(5) 硫酸盐含量 (Sulfate)：硫酸盐是自然界中存在的物质，存在于矿物、土壤和岩石中。它们存在于大气中、地下水、植物和食物中。硫酸盐的主要商业用途是在化工行业中。海水中的硫酸盐浓度约为每升 2700 毫克（mg/L）, 在大多数淡水供应中，硫酸盐的浓度范围为 3 至 30 毫克/升，尽管在某些地理位置上也可发现更高浓度的硫酸盐（1000 mg/L）。

(6) 导电性 (Conductivity)：纯水并不是良好的电流导体，而是良好的绝缘体; 离子浓度的增加会增强水的电导性。通常，水中溶解固体的量决定了电导率。电导率（EC）实际上是衡量溶液中离子传递电流的过程。根据世界卫生组织的标准，电导率值不应超过 400 微西门子/厘米（ S/cm）。

(7) 有机碳含量 (Organic Carbon)：水源中的总有机碳（TOC）来自于腐烂的天然有机物质（NOM）以及人造碳源。TOC 是纯水中有机化合物中碳的总量的衡量指标。根据美国环境保护署（EPA）的规定，处理/饮用水中的 TOC 浓度应低于 2 毫克/升，在用于处理的源水中应低于 4 毫克/升。

(8) 三卤甲烷含量 (Trihalomethanes)：THM（三氯甲烷类化合物）是可能在用氯处理的水中发现的化学物质。饮用水中的 THM 浓度因水中的有机物含量、处理水所需的氯量以及处理水的温度而异。在饮用水中，THM 浓度高达 80 毫克/升被认为是安全的。

(9) 浑浊度 (Turbidity)：水的浊度取决于悬浮状态下存在的固体物质的数量。它是水的光发射特性的衡量指标，该测试用于指示废水排放的胶体物质的质量。

(10) 可饮用性 (Potability)：数据集中，1 表示可饮用水，0 表示非饮用水。

三、 数据集预处理

数据集预处理部分包含了获取数据集基本信息、空缺数据填补、数据集分布检验、数据集相关性检验、离群点去除以及数据集划分等步骤。

1. 数据集基本信息

首先获取数据集大小和结构信息, 可知数据集共有 3276 组样本数据，每组 10 个指标，各项指标计数如表1所示。可以看到，除了可饮用性，其余各项指标都是连续的；同时可以观察到其中的某些指标存在缺失值。

指标	非空值计数	数据类型
pH	2785 non-null	float64
Hardness	3276 non-null	float64
Solids	3276 non-null	float64
Chloramines	3276 non-null	float64
Sulfate	2495 non-null	float64
Conductivity	3276 non-null	float64
Organic Carbon	3276 non-null	float64
Trihalomethanes	3114 non-null	float64
Turbidity	3276 non-null	float64
Potability	3276 non-null	int64

表 1: 数据集各项指标计数

其次获取数据集统计信息，如图1所示。通过各个指标的方差以及 25%、50%、75% 处的数值分布，可以观察到有离群点的存在：总溶解固体值一项具有很大的方差，硫酸盐含量和导电性两项也具有较大的方差。

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	2785.000000	3276.000000	3276.000000	3276.000000	2495.000000	3276.000000	3276.000000	3114.000000	3276.000000	3276.000000
mean	7.080795	196.369496	22014.092526	7.122277	333.775777	426.205111	14.284970	66.396293	3.966786	0.390110
std	1.594320	32.879761	8768.570828	1.583085	41.416840	80.824064	3.308162	16.175008	0.780382	0.487849
min	0.000000	47.432000	320.942611	0.352000	129.000000	181.483754	2.200000	0.738000	1.450000	0.000000
25%	6.093092	176.850538	15666.690297	6.127421	307.699498	365.734414	12.065801	55.844536	3.439711	0.000000
50%	7.036752	196.967627	20927.833607	7.130299	333.073546	421.884968	14.218338	66.622485	3.955028	0.000000
75%	8.062066	216.667456	27332.762127	8.114887	359.950170	481.792304	16.557652	77.337473	4.500320	1.000000
max	14.000000	323.124000	61227.196008	13.127000	481.030642	753.342620	28.300000	124.000000	6.739000	1.000000

图 1: 数据集基本统计信息

2. 缺失值填补

作出缺失值的可视化统计图如图2所示。可以看到，pH 值、硫酸盐含量以及三卤甲烷含量三项指标数据中存在缺失值。下面利用 K 最近邻（K-Nearest Neighbors Impulation）方法对缺失值进行填补。

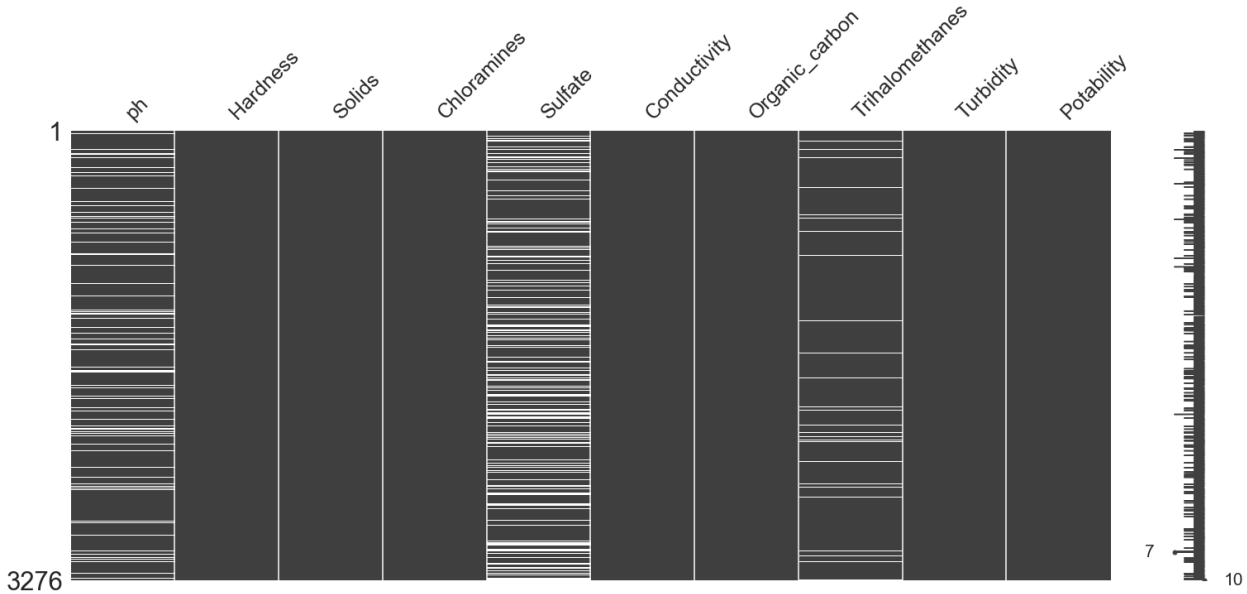

图 2: 数据集缺失信息可视化统计图

该方法使用数据集中 K 个最近邻的平均值来替换数据集中的缺失值。默认情况下，它使用欧氏距离度量来填充缺失值，如果两个样本在不缺失的特征上距离近，则认为这两个样本是相似的。设置 K 为 20，利用 sklearn 工具包中的 KNNImputer 来实现这一填补过程，填补完成后，查看各项指标缺失值计数，确认全部为 0.

3. 数据分布以及相关性检测

(1) 分布检测：首先作出 9 个连续性指标的数据分布图，如图3所示。可以看到各个指标总体上满足正态分布率。则运用夏皮罗-威尔克检验对各个数据是否满足正态分布进行进一步验证。该检验的统计量为:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{1}$$

取显著性水平 $\alpha = 0.05$，得到如表2所示结果：从表中可以看出，大部分指标实际上不符合正态分布。则进一步使用曼-惠特尼 U 检验：该检验假设两个样本分别来自除了总体均值以外完全相

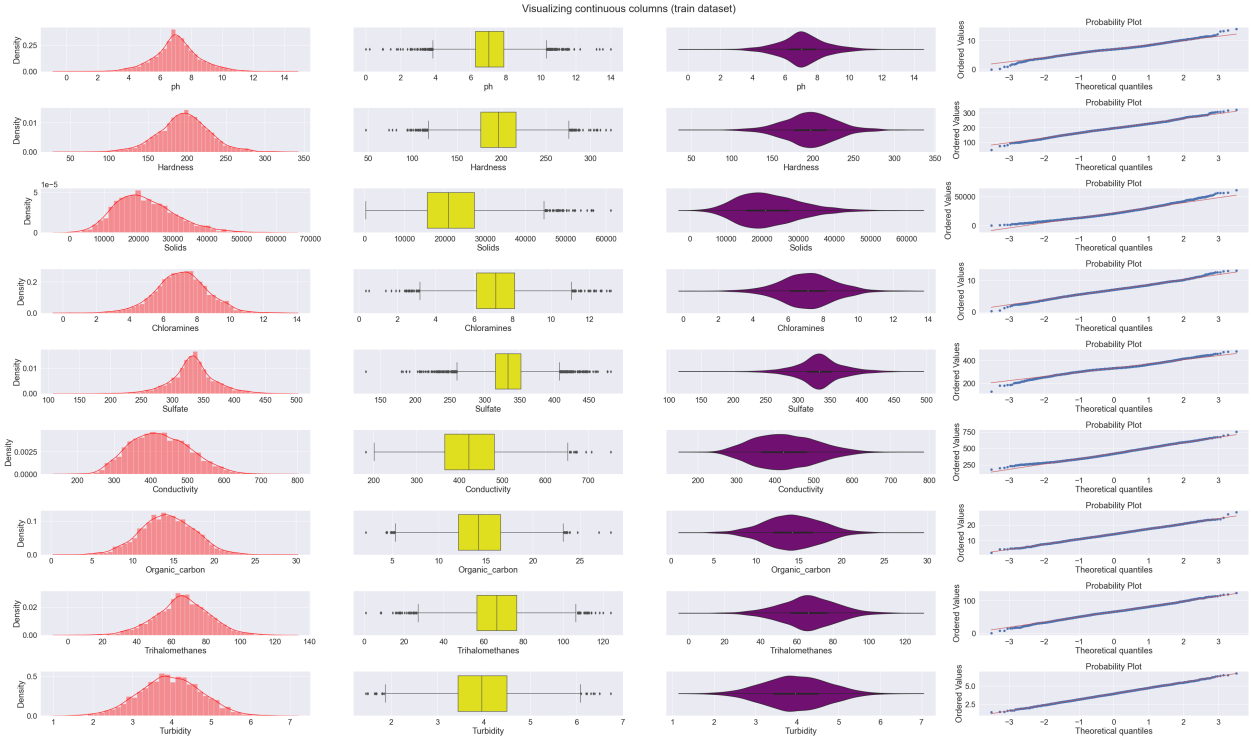

图 3: 连续性指标数据分布图。从左至右，前三列为数据分布图，第一列根据原始数据作出分布；第二列为箱式图，体现了数据分布的分位数和离群点；第三列为经过平滑的分布图。第四列为样本数据分布与标准高斯分布的对比图，蓝色的样本点越接近红色参考线，表明样本数据分布越接近标准的高斯分布。

pH	HD	SD	CL	SF	CD	OC	THM	TBD
reject	reject	reject	reject	reject	reject	pass	reject	pass
p=0.000	p=0.000	p=0.000	p=0.000	p=0.000	p=0.000	p=0.620	p=0.000	p=0.931

表 2: 夏皮罗-威尔克检验结果，HD(Hardness)，SD(Solids)，CL(Chloramines)，SF(Sulfate)，CD(Conductivity)，OC(Organic Carbon)，THM(Trihalomethanes)，TBD(Turbidity)

同的两个总体，目的是检验这两个总体的均值是否有显著的差别。则运用该检验方法验证可饮用性分别与上述 9 个指标是否抽取自同一分布，取显著性水平 $\alpha = 0.05$, 结果如表所示：由表可知，

pH	HD	SD	CL	SF	CD	OC	THM	TBD
reject	reject	reject	reject	reject	reject	reject	reject	reject
st=6111.000	st=0.000	st=0.000	st=2556.000	st=0.000	st=0.000	st=0.000	st=1278.000	st=0.000
p=0.000	p=0.000	p=0.000	p=0.000	p=0.000	p=0.000	p=0.000	p=0.000	p=0.000

表 3: 曼-惠特尼 U 检验, 各个指标缩写含义同表2, st(Statistics)

各个指标与可饮用性之间的分布存在显著差异，在接下来的机器学习中，可以将上述所有指标纳入数据集进行训练。

统计离散化指标可饮用性的分布如图4所示。观察到正样本与负样本数量上较为平衡，故判断不需要采取采样等手段进行数据平衡。

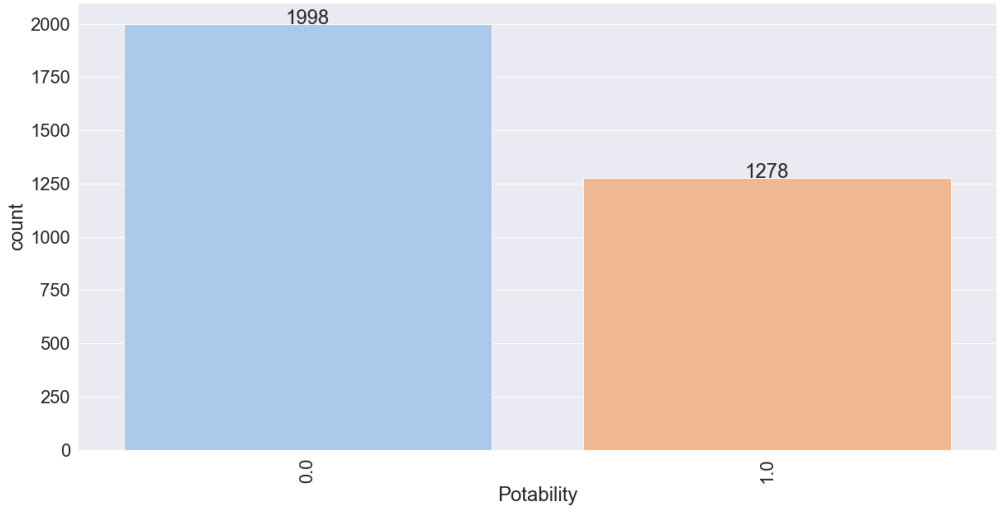

图 4: 可饮用性数据分布统计

(2) 相关性检测：首先，计算各个指标与可饮用性之间的相关性系数，结果如表4. 由表可知，没有单个指标与可饮用性存在显著关联。

pH	HD	SD	CL	SF	CD	OC	THM	TBD
-0.003382	-0.013837	0.033743	0.023779	-0.020383	-0.008128	-0.030001	0.008309	0.001581

表 4: 各指标与可饮用性之间的相关性系数, 各个指标缩写含义同表2

其次，以可饮用性为判断标准，作出各个指标之间的相关性匹配图，如图5. 计算各个指标之间的相关性系数，作出相关系数热力图如图6. 由图5和图6可知，各个指标之间没有显著的相互关联，在这个基础上，可以去除离群点并且消除各数据之间的不平衡性而不损失原数据集的信息。

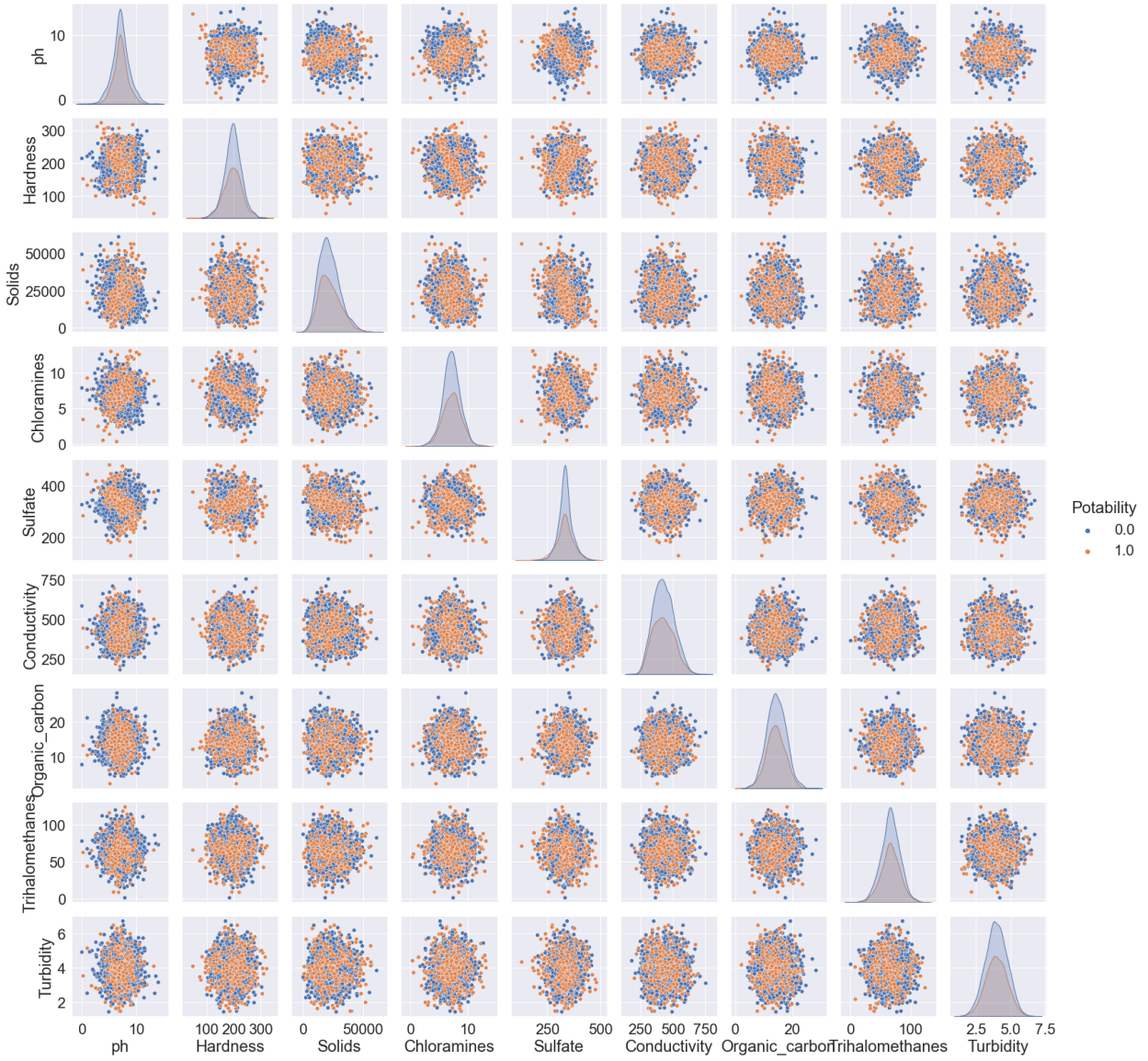

图 5: 各指标相关性匹配图

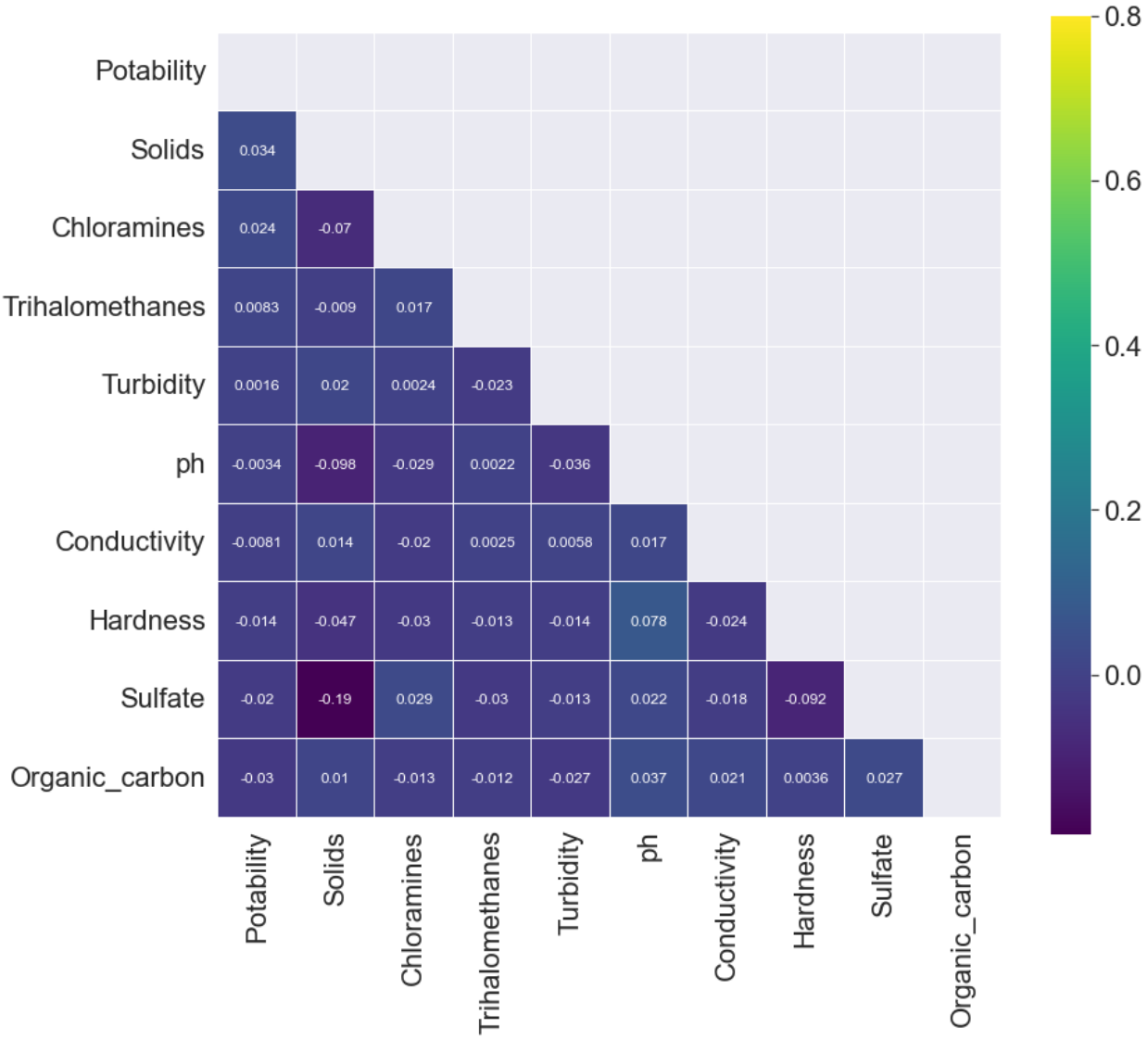

图 6: 各指标相关性系数热力图

4. 离群点检测及去除

在该部分中，使用 sklearn 工具包中的局部离群因子（Local Outlier Factor，LOF）进行无监督异常检测。每个样本的异常得分称为局部离群因子，它测量给定样本的密度相对于其邻居的局部偏差，其异常得分取决于对象相对于周围邻域的孤立程度，即局部性是由 K 个最近邻确定的，其距离用于估计局部密度。通过将样本的局部密度与其邻居的局部密度进行比较，可以识别出局部密度明显低于其邻居的样本，这些样本被判定为离群值。一般情况下，$LOF(k) < 1$ 则认为该点比周围的 K 个点局部密度高，$LOF(k)$ 1 则认为该点与周围的 K 个点局部密度临近，$LOF(k) > 1$ 则认为该点比周围的 K 个点局部密度低，若远大于 1，则可认为该点是离群点。取 $K = 1$，则可筛选出 54 组离群点，在样本总量的 1% 至 2% 之间，可以去除。去除离群点后，数据集大小为 3222 组样本。至此，数据清洗完成。

5. 数据集划分

利用清洗完毕的数据集划分训练集和测试集。划分 70% 的数据作为训练集，30% 的数据作为测试集。至此，数据预处理全部结束，进入机器学习部分。

四、 机器学习过程

在机器学习部分中，使用逻辑回归（Logistic Regression）、支持向量机（SVM）、决策树（Decision Tree）、随机森林（Random Forest）以及 K 近邻（K-nearest neighbors）五种方法进行训练，并分别比较上述五种模型的性能。机器学习部分各方法统一使用 Python sklearn 工具包实现。

1. Logistic Regression 方法

逻辑回归（Logistic Regression）是一种用于建模和预测分类问题的统计方法。它是一种广义线性模型，可用于预测二元或多元分类问题。逻辑回归的目标是根据输入特征的线性组合来估计一个样本属于某个特定类别的概率，它通过使用 Sigmoid 函数将线性组合映射到一个介于 0 和 1 之间的概率值，这个概率值表示样本属于某个类别的可能性。对于二元分类问题，如果概率大于某个阈值（通常为 0.5），则将样本分类为正类别，否则分类为负类别。逻辑回归的模型参数通过最大似然估计或梯度下降等优化算法来训练；模型的参数估计可以用于预测新样本的类别，并提供每个类别的概率。

在训练中,设置回归的最大迭代次数为 120 次;训练完成后,在测试集上测得模型准确率 $Accuracy = 0.614$，作出逻辑回归模型测试的混淆矩阵如图7所示。

2. SVM 方法

支持向量机（SVM）是一种用于分类和回归问题的监督学习算法，它通过将样本映射到高维特征空间，并找到能够最大化样本间距离的超平面来进行分类。SVM 的基本思想是寻找一个最优的超平面，能够将不同类别的样本分开，并使得两个类别的间隔尽可能大。这个最优的超平面由支持向量组成，它们是距离超平面最近的训练样本点。SVM 可以处理线性可分和线性不可分的问题：对于线性可分的问题，可以使用线性核函数，如线性、多项式和高斯核函数；对于线性不可分的问题，可以通过使用非线性核函数，如径向基函数（RBF）核，将样本映射到高维特征空间中，从而在高维空间中找到一个线性可分的超平面。

模型训练完成后，在测试集上测得模型准确率 $Accuracy = 0.615$，作出 SVM 模型测试的混淆矩阵如图8所示。

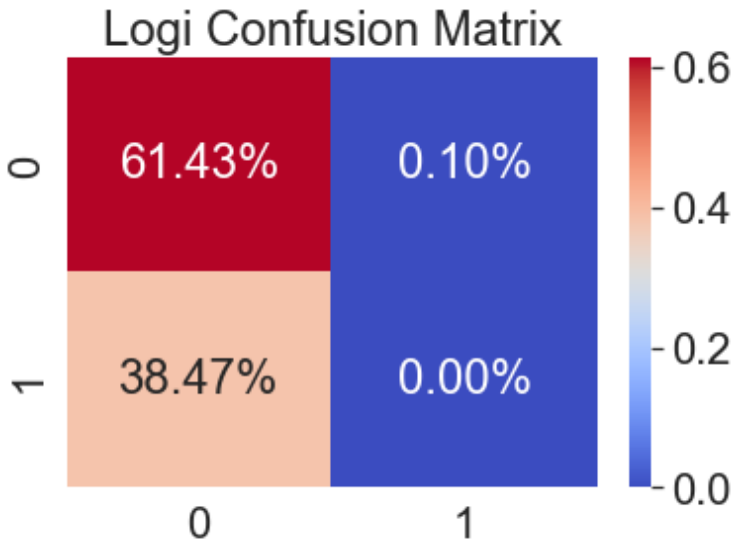

图 7: 逻辑回归模型测试结果混淆矩阵

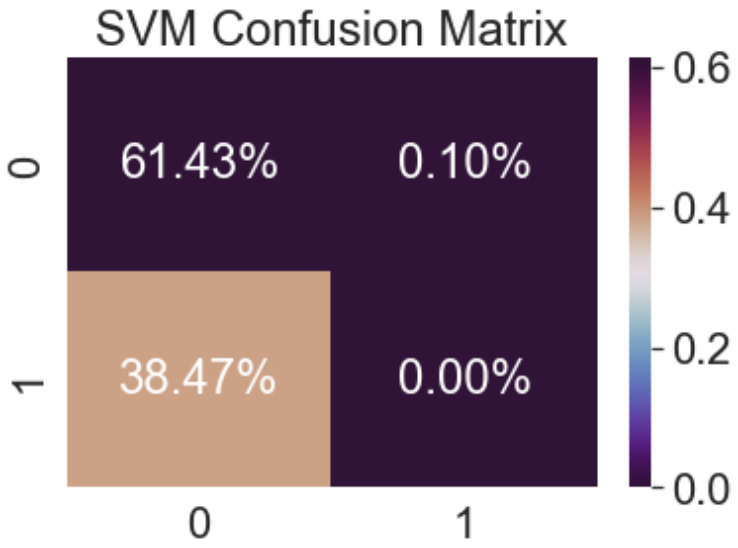

图 8: SVM 模型测试结果混淆矩阵

3. Decision Tree 方法

决策树（Decision Tree）是一种用于分类和回归问题的监督学习算法，它通过构建一个树状结构来对数据进行分类或预测。决策树的构建过程是基于特征的条件划分，根据不同的特征和其取值，将数据集分割成不同的子集。每个内部节点代表一个特征的判断条件，每个叶节点代表一个类别或预测值。决策树的构建过程基于一些指标，如信息增益、基尼系数等，来选择最优的划分特征和划分点。在实现过程中，为了防止过拟合，设置决策树生成的最大高度为 5，并将其结构可视化如图9所示。

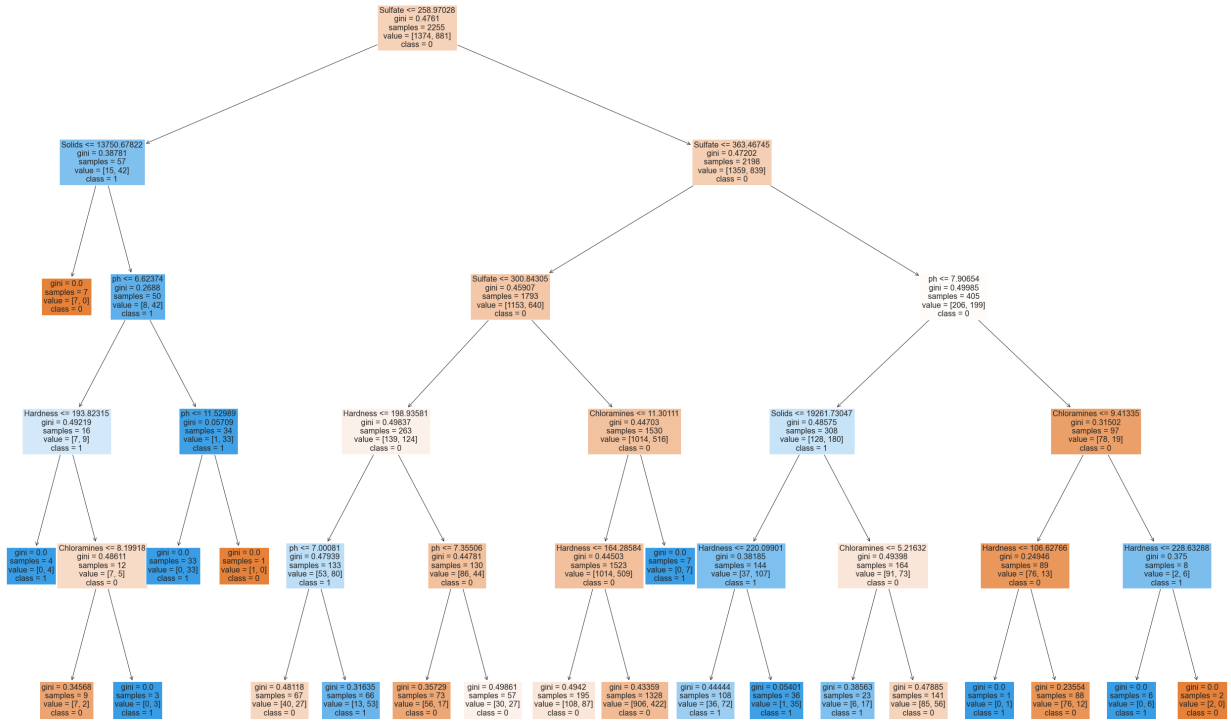

图 9: 决策树模型可视化

模型训练完成后，在测试集上测得模型准确率 $Accuracy = 0.633$，作出决策树模型测试的混淆矩阵如图10所示。

4. Random Forest 方法

随机森林（Random Forest）是一种并行的集成学习方法，通过组合多个决策树来进行分类和回归任务。它利用随机采样和随机特征选择的方法构建一组决策树，并通过集体投票或平均来获得最终的预测结果。随机森林的构建步骤包括：

(1) 随机选择训练样本：从原始训练集中随机选择一定数量的样本（有放回抽样），构成一个子集用于构建每棵决策树。

(2) 随机选择特征：对于每棵决策树的节点，在一个随机选择的特征子集中选择最佳的划分特征。

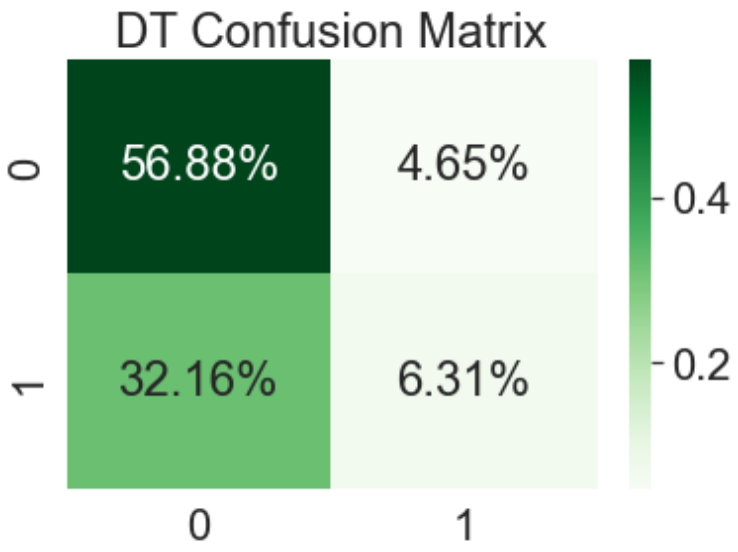

图 10: 决策树模型测试结果混淆矩阵

(3) 构建决策树：使用选定的特征进行决策树的构建，直到达到预定义的停止条件，如树的深度达到最大值或节点中的样本数量小于某个阈值。

(4) 重复步骤 2 和步骤 3，直到生成指定数量的决策树。

(5) 预测：对于分类问题，通过投票来确定最终的预测结果；对于回归问题，通过平均预测值来得到最终结果。

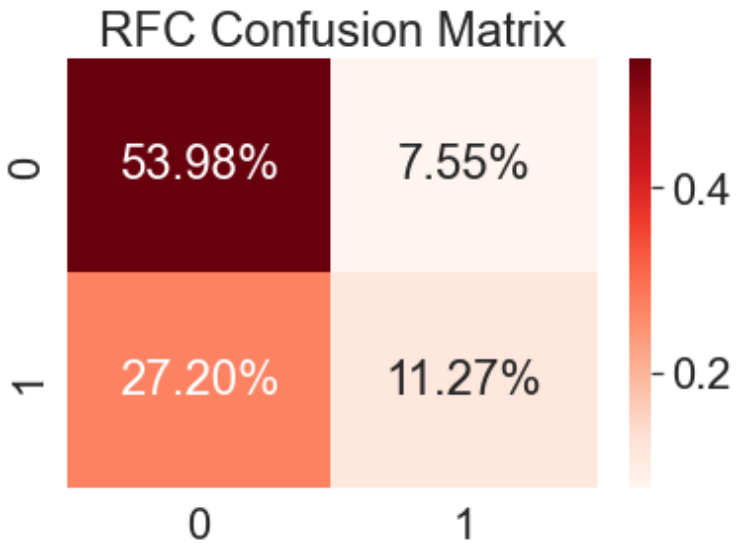

图 11: 随机森林模型测试结果混淆矩阵

模型训练完成后，在测试集上测得模型准确率 $Accuracy = 0.647$，作出随机森林模型测试的混淆矩阵如图11所示。可以观察到随机森林模型在这一任务中表现较为出色。

5. K-Neighbors 方法

K 近邻（K-nearest neighbors）方法是一种常用的非参数化监督学习算法，常用于分类和回归问题。它的基本思想是根据样本之间的相似性，将未知样本分配给其 K 个最近邻居中占据最多数量的类别或者基于平均值预测其数值。其关键点包括：

(1) 距离度量：K 近邻方法通常使用欧氏距离、曼哈顿距离、闵可夫斯基距离等度量来计算样本之间的相似性。

(2) K 值选择：K 值代表在分类或回归时考虑的最近邻居的数量。选择适当的 K 值对模型的性能具有重要影响，一般通过交叉验证或其他评估指标来确定最佳的 K 值。

(3) 投票或平均预测：对于分类问题，K 近邻方法根据 K 个最近邻居的类别进行投票，并将样本分配给占据多数的类别；对于回归问题，K 近邻方法计算 K 个最近邻居的平均预测值作为未知样本的预测结果。

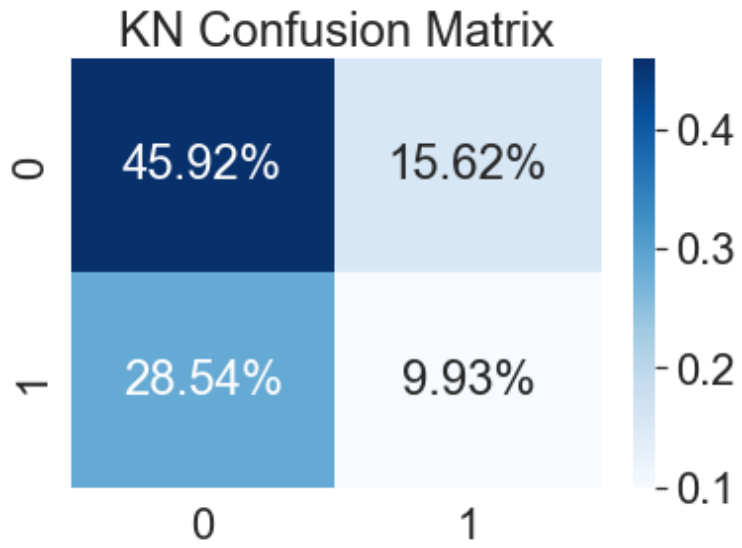

图 12: K 近邻模型测试结果混淆矩阵

模型训练完成后，在测试集上测得模型准确率 $Accuracy = 0.558$，作出 K 近邻模型测试的混淆矩阵如图12所示。可以观察到 K 近邻模型在这一任务中的性能并不可观。

五、 模型评估

1. 准确率

统计五种方法的准确率，作出柱状图方便对比，如图13所示。可以看出，在单个学习器中，决策树的准确率最高，K 近邻方法的准确率最低；随机森林作为集成学习模型，其准确率高于其他的任何一种模型。

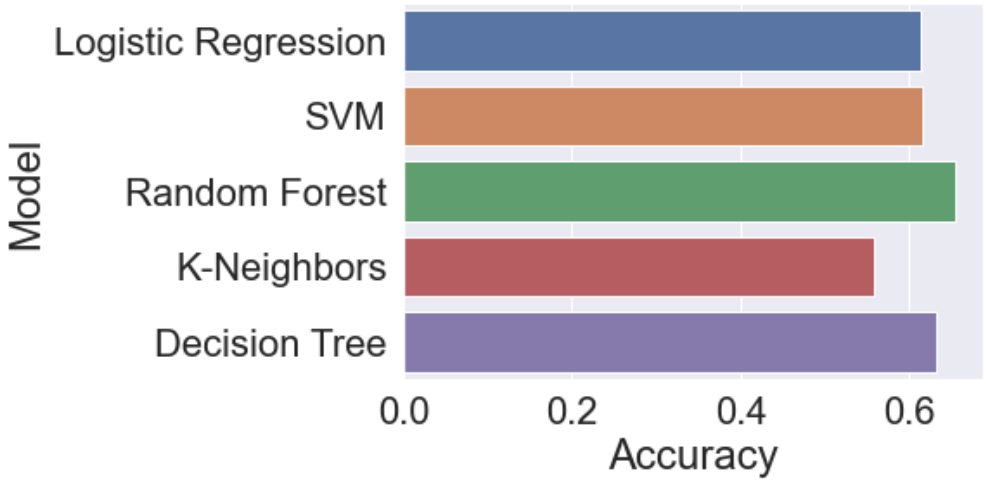

图 13: 各模型准确率比对图

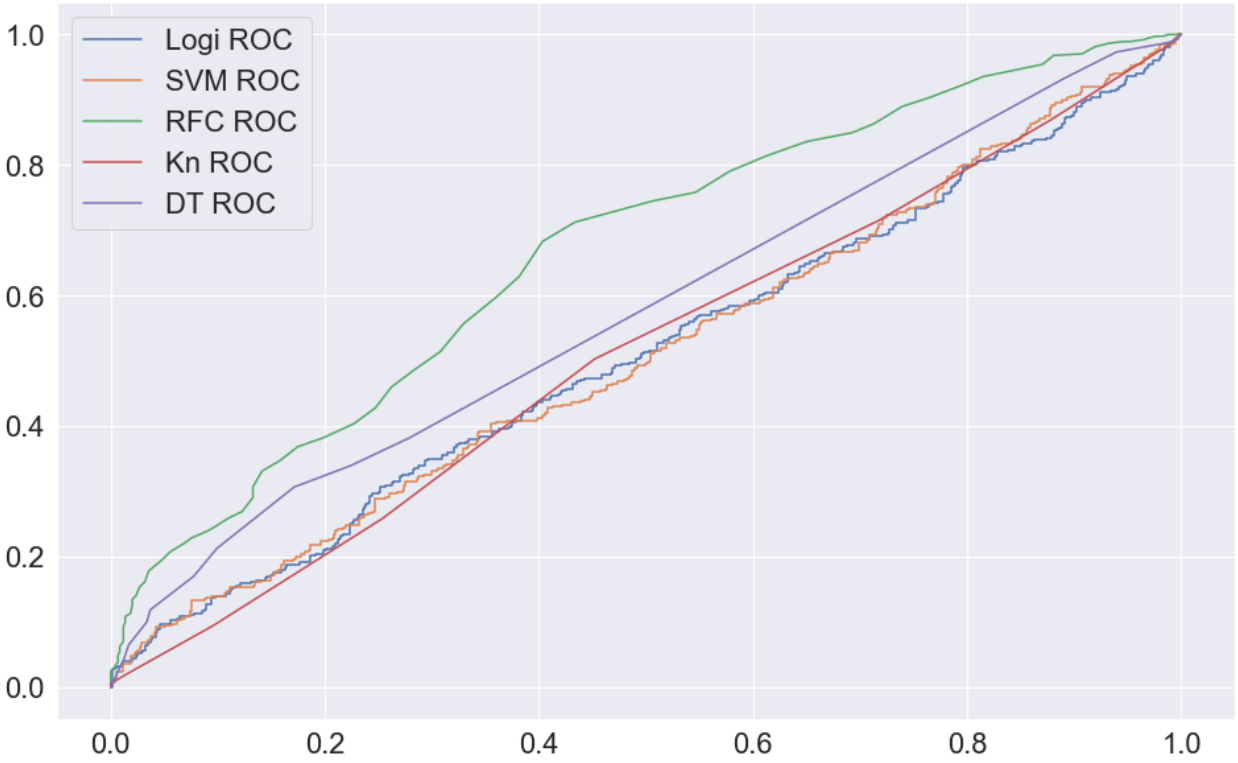

图 14: 各模型 ROC 曲线比对图

2. ROC 曲线

可以看出，随机森林和决策树模型的 ROC 曲线下面积较大，可以包络其他几种方法的 ROC 曲线；逻辑回归、SVM 模型以及 K 近邻方法的 ROC 曲线下面积接近。

六、 总结

在本次大作业中，我完成了从原始数据到构建识别模型的全过程。这其中涉及了原始数据处理（包括获取数据集基本信息、空缺数据填补、数据集分布检验、数据集相关性检验、离群点去除以及数据集划分等步骤）、机器学习（包括了逻辑回归、支持向量机、决策树、随机森林以及 K 近邻五种模型）以及模型性能评估几个步骤。从这个过程中，我对于数据挖掘的基本步骤、数据的可视化辅助以及机器学习的基本方法有了更加清晰深刻的认知。

在整个过程中，我发现数据预处理是数据科学中不可或缺的一步。现实中，数据的获取往往并不简单，得到的原始数据往往也有残缺值、不平衡等问题，因此如何充分得利用每一条数据就变得非常重要。根据研究需要处理残缺值、检查数据集平衡性并判定是否需要重采样以及检查数据集的分布和关联性都是必不可少的环节。

同时，在训练模型的过程中，我也明显地观察到了随机森林这一集成学习方法的效果好于其他的单个学习器，这也进一步加深了我对于集成学习“好而不同”这一要求的认知。

数据科学是当下蓬勃发展的学科，它为我们的生活也带来了许多便利。在这样的学科发展背景下，我也希望它能帮助到更多这个世界上处于困难中的人，希望如鉴定安全健康的饮用水这样微小却又对于许多贫困地区来说困难的事情也能得到数据科学的辅助。

七、 代码

Jupiter Notebook 纯代码见后方附录页。完整的 Jupiter Notebook 已上传至个人 Github 账号，可在如下链接中查看：https://github.com/YihanLi126/Water-Quality-Classify