课程设计报告

教师：金福生

课程名：大数据技术导论

姓名：于丁一

学生类型：延河班学生

学生学校与班级：中国农业大学 农学 181 班

中国大学 MOOC 账号：CAU 于丁一 2018301010111

目录

基于机器学习的森林火灾预测分析

摘要：森林火灾随着地球气候变化愈发频繁，给社会自然经济造成极大损失，如何及时准确地预测成为难题。本文本首先对森林火险相关指标做出解释，简要地描述了随机森林和决策树算法的原理，利用 python 程序利用机器学习对森林火灾的是否会发生做出预测，并分析的模型的错判率。使用的数据采集自从葡萄牙东北部的 Montesinho 国家公园、阿尔及利亚东北部 Bejaia 区域和阿尔及利亚西北部 Sidi Belabbes 区域。

关键字： 森林火灾 随机森林 决策树 机器学习 Python 分类算法

1.问题背景

森林火灾不仅影响森林保护，还会造成巨大经济损失和严重的生态破坏，给人类的生活带来灾难性影响。森林火灾的发生源于多种原因(如人为疏忽和闪电)，尽管越来越多的国家斥巨资来控制，全世界每年仍有数百万公顷的森林灭失火中。

近几年，由于传统的监视费用昂贵且受主观因素的影响较大，人们逐渐重视并发展自动化监测、预测火灾的解决方案。这些方案大致可分为三类：卫星，红外扫描仪和局部传感器[1]。由于高昂的设备成本、维护成本以及延迟问题，卫星定位不适用于所有情况。研究表明，天气条件，如气候和相对湿度，是影响火灾发生的关键因素，而自动气象站[2]通常可以提供有效数据，这些数据可以实时采集且成本低廉。

在过去，气象数据已纳入量化指标体系，用以预防火灾危险、警告公众和支持消防管理决策。特别是，加拿大森林火险天气指数(FWI)系统[3]的设计，在上世纪 70 年代计算机还十分稀缺的情况下它只需要利用手动收集的四个气象观测读数(温度，相对湿度，风速和降水量)进行简单的计算。目前该指数系统在加拿大和其他一些国家广泛使用[4]。

现今计算机技术的快速发展，使得对数据的采集越发的实效和便捷。机器学习使用自动化的数据挖掘工具分析原始数据可以为高层决策者提取有效信息。事实上，已经应用到火灾探测领域。例如采用神经网络(NN)预测人类引起的森林火灾；红外扫描仪和神经网络结合在减少森林火灾误报率方面达到 90%的成功率；使用卫星和气象数据应用逻辑回归、随机森林和决策树来探测斯洛文尼亚森林火灾[5]。

学习上述方法，本 python 程序利用机器学习对森林火灾的是否会发生做出预测，并分析的模型的错判率。 使用从葡萄牙东北部的 Montesinho 国家公园（517 条）、阿尔及利亚东北部 Bejaia 区域（122 条）和阿尔及利亚西北部 Sidi Belabbes 区域（122 条）采集的最新数据。应用决策树和随机森林对三类指标进行分析(即时间，气象指标和部分 FWI 系统指标)。将对三类不同性质的指标分别

进行基于机器学习的数据分析，如气象指标(即温度，相对湿度，风速和降雨量)与随机森林相结合，能够预测森林火灾是否会发生，构建火灾燃烧等级对未来的火灾防治和消防管理决策是非常有用的。

2.数据分析

2.1 数据介绍

本文涉及的森林火灾数据来自葡萄牙东北部的 Montesinho 国家公园、阿尔及利亚东北部 Bejaia 区域和阿尔及利亚西北部 Sidi Belabbes 区域的数据库，信息包含 10 个变量：

1）信息采集的月份（month）；

2）FWI 系统的部分指数变量：FFMC (细小可燃物湿度码)、DMC (粗腐殖质湿度码)、DC (干旱码)和 ISI (初始蔓延指数)；

3）四种可直接测量的气象数据：温度 temp（℃）、相对湿度 RH（%）、风速 WS（km/h）和降水量 RF（mm）的气象数据；

4）是否发生森林火灾：fire，只有两个值，1 表示发生，0 表示未发生。

2.2 变量解释

文中给出了四个 FWI 系统指数变量：3 个代表可燃物湿度的基本子指数，分别为细小可燃物湿度码(FFMC, fine fuel moisture code)，粗腐殖质湿度码(DMC, duff moisture code)和干旱码(DC, drought code)；1 个代表可燃物扩散速率的中间子指数，为初始蔓延速度(ISI, initial spread)。火险气候指数系统中所涉及的元素由每天测量的气温、相对湿度、风速和降水量的气象数据中计算得到。

2.2.1. 细小可燃物湿度码 FFMC

FFMC 代表的是森林中地被物干质量为 0.25 $kg \cdot m^{-2}$，厚度为 1.2 cm 的枯枝落叶和其他的已经固化的细小燃料的含水率。FFMC 是代表细小可燃物的可燃性和易燃性的指标，它受温度、降水、相对湿度和风速的影响，值随着燃料含水率的变化而改变，其核心是一个简单的水分交换的指数模型：

$$m_0 = 147.2 \times \left(101.0 - c_{FFMC}\right) / \left(59.5 + c_{FFMC}\right)$$

其中 m_0 为前一天的燃料含水率。

2.2.2. 粗腐殖质湿度码 DMC

DMC 代表的是森林地被物最上层厚度约为 7 cm，干质量为 5.00 $kg \cdot m^{-2}$ 的有机物质的含水率。DMC 用来表明中等下层落叶层和中型木质物质的燃料消耗，DMC 模型是一个简单的水分交换的指数模型：

$$M_0 = 20.00 + \ln\left[\frac{c_{DMC} - 244.73}{-43.43}\right]$$

其中 M_0 表示前一天的地表可燃物含水率。

2.2.3. 干旱码 DC

DC 代表的是森林地被物中干质量为 25.00 kg·m^{-2}，厚度为 18 cm 的深层可燃物和粗死木残体的含水率。干旱码用于衡量季节性干旱对森林燃料以及深层下层落叶层和大型段木的影响指标。

DC 模型的核心是一个简单的指数模型：

$$Q_0 = 400 \times e^{-CD/400}$$

其中 Q_0 表示前一天干旱码的湿度指标。

2.2.4. 初始蔓延指数 ISI

ISI 代表的是火灾蔓延的潜在等级，由 FFMC 和风速两个指标决定。ISI 一直是表示火灾蔓延等级的很好指标。

3.模型描述

3.1 决策树模型（DT）

决策树(decision tree)是一种基本的分类与回归方法。决策树模型呈树形结构，在分类问题中，表示基于特征对实例进行分类的过程。它可以认为是 if-then 规则的集合，也可以认为是定义在特征空间与类空间上的条件概率分布[6]。

其主要优点是模型具有可读性，分类速度快。学习时，利用训练数据，根据损失函数最小化的原则建立决策树模型。预测时，对新的数据，利用决策树模型进行分类。决策树学习通常包括 3 个步骤：特征选择、决策树的生成和决策树的修剪[6]。

决策树的典型算法有 ID3，C4.5，CART 等。ID3 算法是目前最有影响的决策树算法，是由 Quinlan 于 1986 年首次提出的。ID3 决策树算法筛选“信息增益”最大的属性划分训练数据集，基本原则是: 数据集被分裂为若干子集后，要使每个子集中的数据尽可能地“纯”，即进行分枝时系统的熵值最小，从而很大地提高算法的运算速度和精确度[7]。

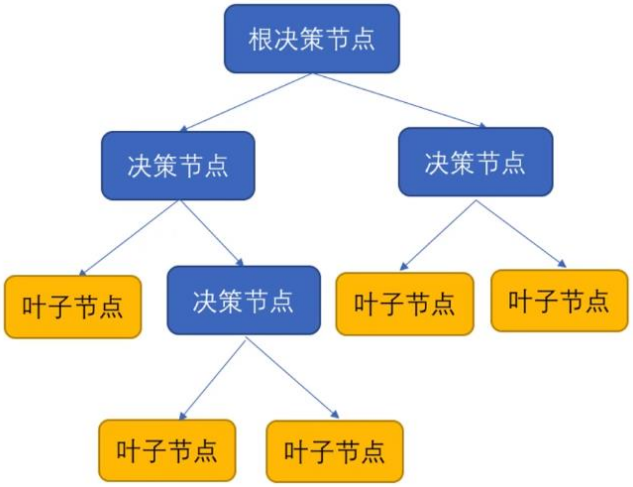

图 1 决策树基本模型

信息熵：是用来衡量信息的不确定性或混乱程度的指标，信息的不确定性（即一件事情出现不同结果的可能性）越大，则信息熵越大。熵的计算公式如下：

$$H(X) = -\sum_{i=1}^{n} P(X=i)\log_2 P(X=i)$$

其中：P(X=i)为随机变量 X 取值为 i 的概率

条件熵：是通过获得更多的信息来减小不确定性。条件熵的计算公式如下（Y 条件下，X 事件的条件熵）：

$$H(X|Y) = \sum_{v \in values(Y)} P(Y=v)H(X|Y=v)$$

$$H(X|Y=v) = -\sum_{i=1}^{n} P(X=i|Y=v)\log_2 P(X=i|Y=v)$$

信息增益：等于父节点的熵减去子节点加权熵，计算公式如下：

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

注：信息增益不是构建决策树的唯一指标，可选择其他指标如基尼不纯度等。

防止过拟合：决策树容易过拟合，表现为在训练数据集上性能好，在测试数据集上性能差。**一般可以选择，**

- Pre-pruning（预剪枝）:当树分支到一定条件下，还没有完美分支的时候，就预先停止分支。
- Post-pruning（后剪枝）:一直分支到完美的情况，然后再剪枝。

取得良好剪枝效果关键是定义一个剪枝的标准

- 使用 validation dataset（测试数据集）来评估后剪枝（ post-pruning）的效果。
- 直接使用 training dataset（训练数据集），利用统计分析来衡量剪枝或者分支带来的好处：Error estimation 误差估计、Significance testing 重要性检测（例如 Chi- square test）。
- 最小描述长度原则（ Minimum Description Length Principle）:使用一个评价

树和训练数据集编码的复杂度的度量值，当树的编码长度（size（ree）+size（ misclassification（tree）最小的时候停止。

此处不做过多赘述。

3.2 随机森林模型(RF)

随机森林(RF)是一种统计学习理论，它是利用 bootsrap 重抽样方法从原始样本中抽取多个样本,对每个 bootsrap 样本进行决策树建模，然后组合多棵决策树的预测，通过投票得出最终预测结果。大量的理论和实证研究都证明了 RF 具有很高的预测准确率，对异常值和噪声具有很好的容忍度，且不容易出现过拟合[8]。

随机森林分类(RFC)是由很多决策树分类模型$\{h(X, \Theta_k), k=1, \cdots\}$组成的组合分类模型，且参数集$\{\Theta_k\}$是独立同分布的随机向量，在给定自变量 X 下，每个决策树分类模型都由一票投票权来选择最优的分类结果。RFC 的基本思想:首先，利用 bootstrap 抽样从原始训练集抽取 k 个样本，且每个样本的样本容量都与原始训练集一样;其次，对 k 个样本分别建立 k 个决策树模型，得到 k 种分类结果;最后，根据 k 种分类结果对每个记录进行投票表决决定其最终分类，详见图 2[8]。

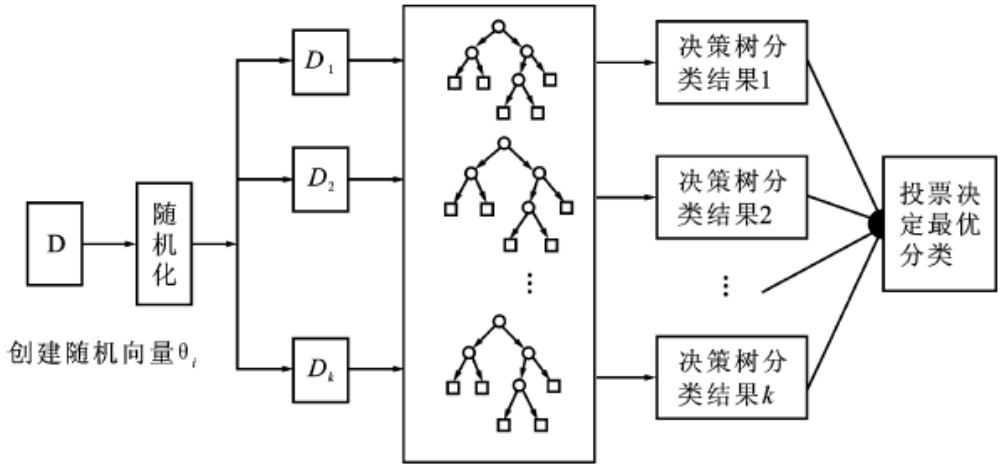

图 2 随机森林基本模型

RF 通过构造不同的训练集增加分类模型间的差异，从而提高组合分类模型的外推预测能力。通过 k 轮训练，得到一个分类模型序列$\{h_1(X), h_2(X), \cdots, h_k(X)\}$，再用它们构成一个多分类模型系统，该系统的最终分类结果采用简单多数投票法。最终的分类决策:

$$H(x) = \arg\max_{Y} \sum_{i=1}^{k} I(h_i(x) = Y)$$

其中，$H(x)$表示组合分类模型，h_i 是单个决策树分类模型,Y 表示输出变量(或称目标变量),$I(\cdot)$为示性函数。该式说明了使用多数投票决策的方式来确定最终的分类[8]。

4.Python 代码编写运行及实验过程、结果讨论

注：本部分内容详细代码请参考名为 Randomforest.py/pdf/html 文件。

4.1 数据处理与分析

首先为了方便后续分析工作的进行，在程序中引用 pandas、numpy、matplotlib、seaborn 等常用于数据分析的 Python 语言扩展程序库。读入数据生成 Dataframe 后，使用 isnull()检查缺失值，确认无缺失值后，使用 head()观察数据情况。

使用 shape、dtypes 等读取数据体量和类型，发现共 761 条数据、每条 10 个特征值；数据类型主要有 float 64 与 int 64 两类，可以直接进行处理。

接着进行描述性统计：数据整体有 54%为有火灾发生，其余 46%无火灾发生，火灾发生是最终要预测的值，在数据集中两者占比接近有利于保证结果的一般性；整体统计数据给出的信息有限，使用 groupby 以是否发生火灾为分类标准进行分组平均值统计，观察得 RF 均值在两组间差距较大。

进行相关性分析，使用 corr()、heatmap()等得出相关系数矩阵和热力图，发现与‘fire’相关性较大（相关系数绝对值大）的数据列：‘FFMC’(0.29)、‘ISI’(0.19)、‘temp’(0.14)、‘RF’(-0.11)。这四类数据可作为森林火灾预测的主要影响因素。

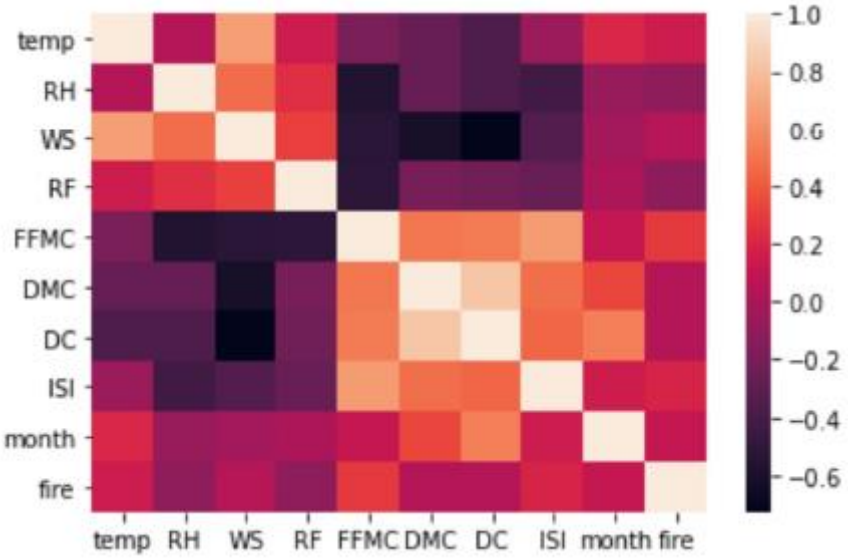

图 3 热力图

为了验证结论，使用 T-Test 检验在有无火灾的两组数据中，上述四类主要影响因素数据是否存在显著差异（以 FFMC、ISI 为例）。由于两组数据量不相同，采用的检验方法为计算出其中一组的平均值，再通过 scipy.stats 的 ttest_lsamp()将另一组数据与该平均值搭配做检验。得到的结果为 $P_{ffmc}=6.47\times10^{-76}$、$P_{isi}=1.91\times10^{-10}$，均明显小于 0.05，因此有极大把握认为在有无火灾的两组数据中对应的 FFMC、ISI 数据是有显著差异的。为了更直观的观察不同组（有火灾 fire、无火

灾 no fire）数据的分布，使用 kdeplot 绘制了概率密度分布图。

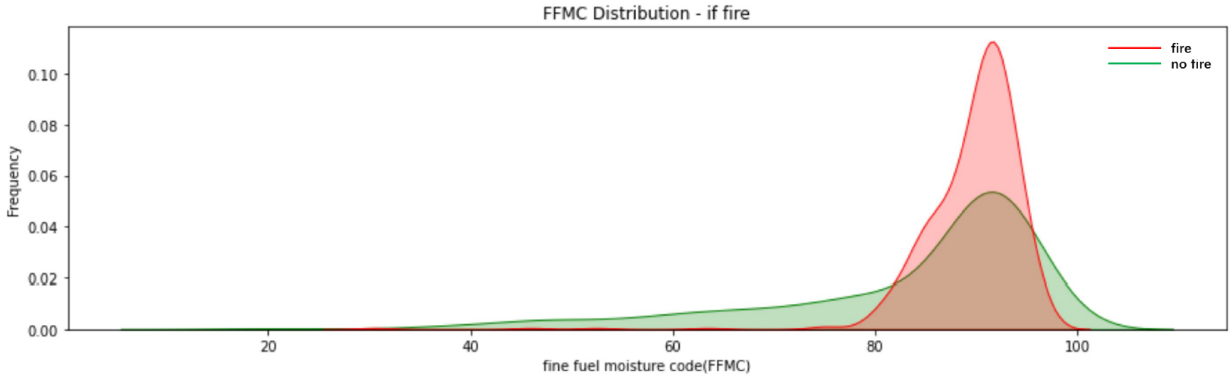

图 4-1 FFMC 概率密度分布图

可以看出在有火灾发生的案例中，FFMC 明显呈现出更加极端且集中分布的态势，其数值在 90 左右；而在无火灾案例中，FFMC 呈现出在 70~100 之间的较均匀分布。

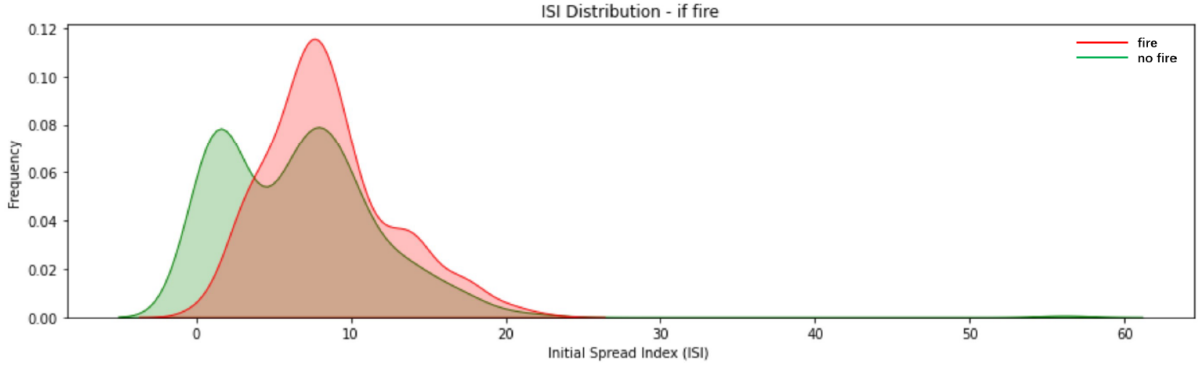

图 4-2 ISI 概率密度分布图

观察可以看出，有无火灾情况下 ISI 数据分布的峰形大不相同，有火灾时的 ISI 较多地集中在 8~10 之间。

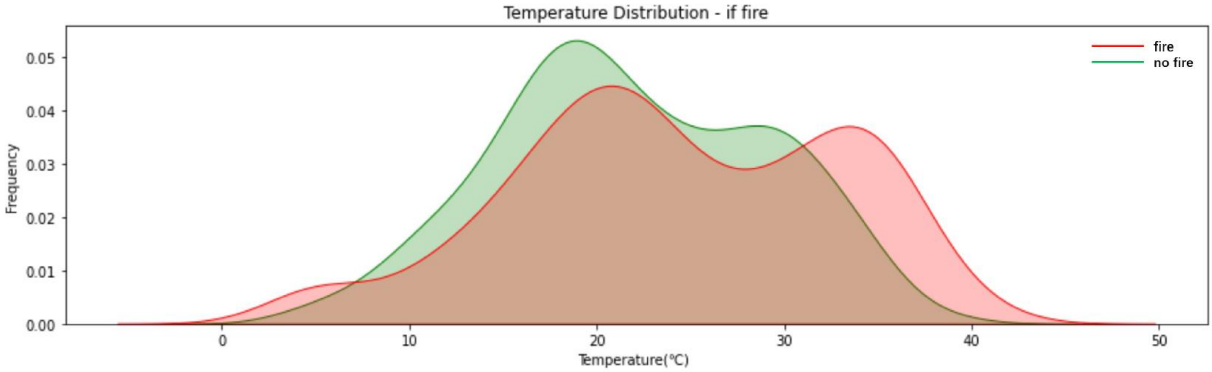

图 4-3 temp 概率密度分布图

观察可以看出，温度数据在两种情况下的概率密度曲线峰形极为接近，这符合一般情况，因为温度是最基础的外界气象数据，其集中度不应因有无火灾而骤变。但我们仍然可以得出在有火灾的情况下，温度要普遍比无火灾情况下

高出一个身位，大约 2~4℃。

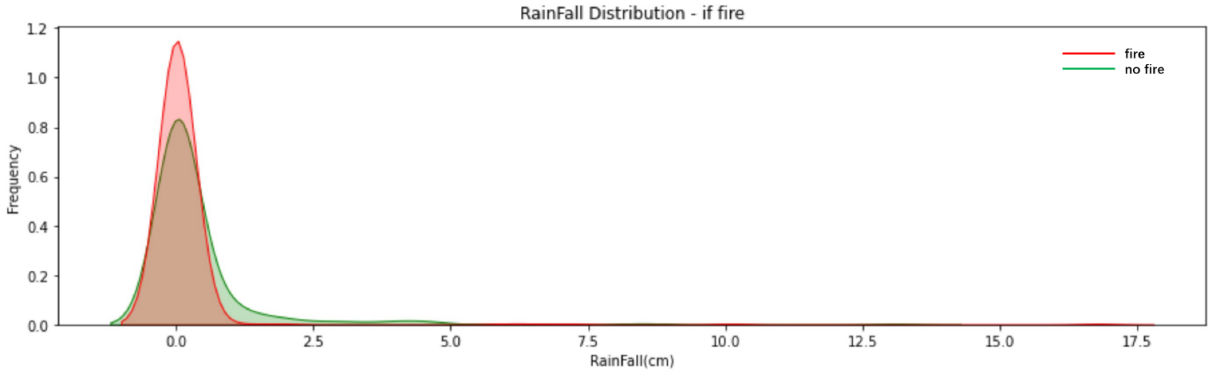

图 4-4 RF 概率密度分布图

降雨量的数据在进行分组统计均值时，有无火灾的情况呈现出了较大的差异（0.13mm 与 0.40mm），而在进行相关性分析时，RF 与 fire 的相关系数并不算高(0.11)。造成这种矛盾的原因是数据本来就收集自干旱地区，有许多时间降雨量为 0mm，大量的极端值导致数据的可用性降低。

4.2 模型的构建与评价

对数据有了一定的认识之后开始构建模型。首先将数据集分为训练集与测试集，选择测试集占比 15%。并规定测试与训练集中有无火灾案例的比例要和整体数据集相等。接着引用 sklearn 扩展程序库中的不同子库进行模拟。

首先构建通过 fit 函数构建了决策树 dtree 模型，选择的预剪枝方式为叶子节点至少要包含 1%的样本量，分节指标为信息熵。同理构建了随机森林 rf 模型，选择的分叉界限为 10 个样本量，由此防止过拟合，分节指标同样为信息熵。

使用 classification_report 检验模拟结果可以看出随机森林模型的预测准确度（accuracy）达到了 72%，决策树模型的预测准确度（accuracy）达到了 68%。为了进一步对比两个模型，通过 roc_curve 绘制两者的 ROC 图，发现随机森林模型的表现确实犹豫决策树模型，相比之下假阳性率低。

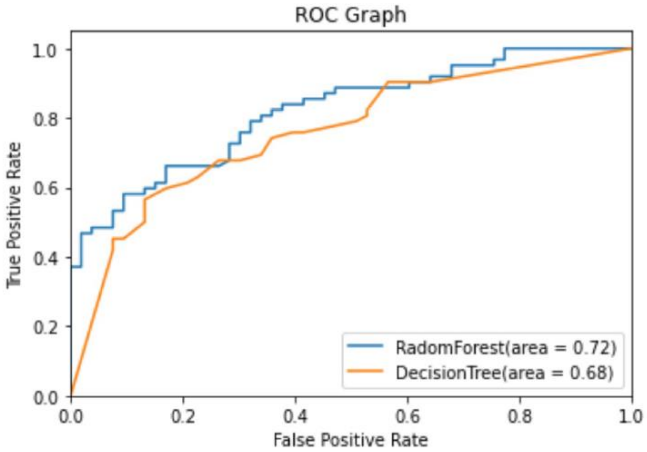

图 5 ROC 图

最后以随机森林模型为基础，通过 feature_importances 对象、argsort()等对除去 fire 之外的 9 个特征重要性进行了排序，绘制帕累托图，得出结论：进行预测时，可以首先考虑得指标是 FFMC>temp>ISI>WS>DMC，而 month 与 RF 两类数据参考价值不大。

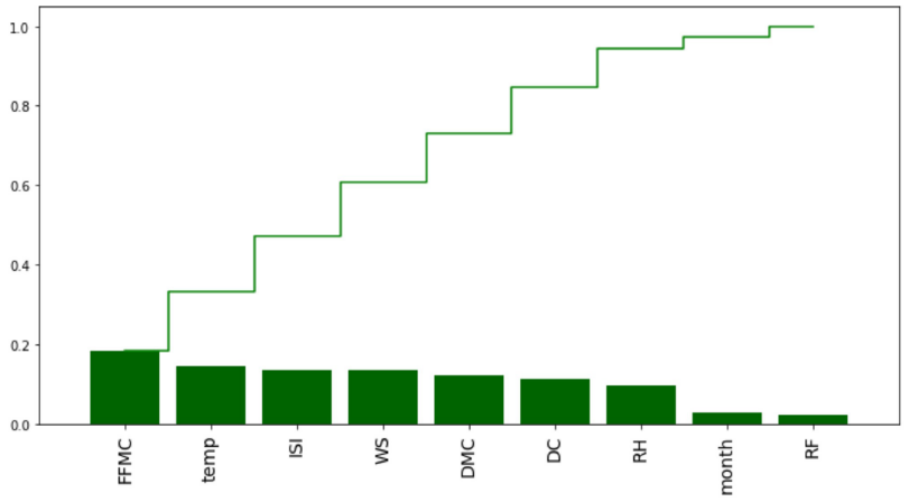

图 6 重要性占比帕累托图（柱高为单项占比，阶梯图为累计值）

注：重要性的计算方法为，假设数据有 M 个特征，使用信息熵来决定分叉情况。首先初始化一个数组 A[]，长度为 M，所有值为 0。接着遍历每一个节点，假设每一个节点是基于第 m（1≤m≤M）个特征来分叉的，这一个节点分叉造成的信息增益为 e，训练数据中通过这个分叉的数据个数为 d，令 A[m] += d*e。假设数组 A[]中所有的值求和结果为 S，将 A[]中每个元素除以 S，最终 A[]中第 m 个元素的值就是特征 m 的重要性。

5.参考文献

[1] Cortez, P., Morais, A. A Data Mining Approach to Predict Forest Fires using Meteorological Data. *associação portuguesa para a inteligência artificial*, 2007

[2] 曲智林, 胡海清. 基于气象因子的森林火灾面积预测模型. 应用生态学报, 2007(12): 2705-2709

[3] 田晓瑞, McRae Douglas J., 等. 大兴安岭地区森林火险变化及Fwi适用性评估. 林业科学, 2010(05): 127-132

[4] 袁建, 江洪, 信晓颖. 基于Fwi的浙江省森林火险等级划分. 福建农林大学学报(自然科学版), 2013(03): 283-288

[5] 丹刘. Prediction and Analysis of Forest Fire Based On Machine Learning. *Statistics and Application*, 2016(02)

[6] 李航著. 统计学习方法.北京：清华大学出版社, 2012

[7] 马红丽, 徐长英, 杨新鸣. 决策树模型在中医药领域的应用现状. 世界中医药, 2021(17): 2648-2651

[8] 方匡南, 吴见彬, 等. 随机森林方法研究综述. 统计与信息论坛, 2011(03): 32-38