# **基于数据挖掘的森林火灾预测分析**

摘要：

森林火灾，是指失去人为控制，在林地内自由蔓延和扩展，对森林、森林[生态系统](https://baike.baidu.com/item/%E7%94%9F%E6%80%81%E7%B3%BB%E7%BB%9F?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/%E6%A3%AE%E6%9E%97%E7%81%AB%E7%81%BE/_blank)和人类带来一定危害和损失的[林火行为](https://baike.baidu.com/item/%E6%9E%97%E7%81%AB%E8%A1%8C%E4%B8%BA/58670985?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/%E6%A3%AE%E6%9E%97%E7%81%AB%E7%81%BE/_blank)。森林火灾是一种突发性强、破坏性大、处置救助较为困难的[自然灾害](https://baike.baidu.com/item/%E8%87%AA%E7%84%B6%E7%81%BE%E5%AE%B3/81488?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/%E6%A3%AE%E6%9E%97%E7%81%AB%E7%81%BE/_blank)。随着全球范围性的温室效应，全球气温正逐步升高，为对抗温室效应，森林已经成为我们急需保护的资源，但是火灾时刻威胁着森林资源。为了帮助对抗温室效应、保护森林，本研究通过选取葡萄牙蒙特西尼奥自然公园的 517 起火灾的记录，采用数据挖掘技术对数据进行线性回归分析、聚类分析等操作，得到森林火灾发生的主要因素之间怎样的搭配更容易发生森林火灾，最终为林业管理者预测森林火灾发生、森林火灾管控、降低人员财产损失等方面提供相关建议，具有较高的经济价值和学术价值。

关键词：

葡萄牙蒙特西尼奥自然公园 温度 湿度 聚类分析 泛化能力

1. 引言

1.1 研究背景

随着全球范围性的温室效应，全球气温正逐步升高，森林火灾的发生次数也在逐渐增加。森林火灾是森林生态系统天敌，它会给森林生态带来灾难性的后果。森林火灾在毁灭大片的森林的同时，还会烧伤、烧死大量的野生动物，降低森林更新能力，导致森林涵养水源的能力减弱，引发土壤的酸化和贫瘠，致使生态遭受破坏。

1.森林火灾发生的过程中，大量的植物被烧毁，降低了森林中植被分布密度，破坏原有的森林植被结构；还引发植物类型的更替，一些低价值的灌木、杂草取代了以前经济生态价值较高的高大乔木，导致森林经济价值降低。

2.由于植被遭到破坏，失去了保持水土、涵养水源的作用，大量土壤裸露，最终造成洪涝、干旱、山体滑坡、风沙扬尘等灾害的发生。

3.因树木被林火烧伤，树木结构受到破坏，对外部的病虫害抵御能力降低，虫害肆虐，加快了树木死亡，使局域小气候受到干扰，需要漫长的时间才能恢复。

4.森林火灾在烧毁大量经济林区和林业设施的同时，威胁到分布于林区附近的居民区，危及广大人民生命财产安全；同时森林火灾能烧死大量的野生动物，毁坏野生动物栖息地，降低野生动物种群数量；火灾发生时产生的大量粉尘及烟雾污染，会对周围大气环境造成巨大的影响；与此同时，扑救森林火灾要消耗大量的社会资源，同时影响工农业的正常生产活动，有时还会威胁人身安全，造成人员伤亡，影响社会的和谐稳定。

1.2问题的提出

出于降低森林火灾对人身财产安全以及保护生态的需求，本次设计通过分析往年森林火灾发生的因素，利用数据挖掘技术进行分析，找到影响森林火灾的主要因素之间怎样的搭配更容易发生森林火灾，进而对我们日常的森林火灾防控做出指导。

1.3 国内外研究文献综述

森林火灾预测研究自２０世纪３０年代以来己有７０多年的历史，在这70多年的时间里，各种各样的预测方法相继产生。国外对森林火灾的预测早期有多因子预报林火、利用空气湿度预报林火，近几年来预测方法主要包括美国国家森林火灾风险评估系统（ＷＤＲＳ）和澳大利亚火灾风险等级（ＦＤＭ）以及加拿大森林火险等级系统（ＣＦＦＤＲＳ）等。国内对森林火灾预报研究薄弱，在国外的研究基础上，结合国内的森林情况，确定了以下几种预测方法：“综合指标法”、“风速补正综合指标法”、“双指标法”、“８０１火险尺法”和“全国森林火险天气等级”等。

1.4研究意义

对森林火灾预测分析的研究可以让我们知道当地影响火灾的主要因素以及森类火灾频发的时间段，为林业管理者预测森林火灾发生、森林火灾管控、降低人员财产损失等方面提供相关建议，具有较高的经济价值和学术价值。比如我们发现温度低和气候干燥的时候很容易发生森林火灾，那我们要做好气温与湿度监测，发现气温和温度达到易发生森林火灾的时候应及时做好灭火准备，比如增加巡山的力度和开展消防员森林火灾演练等等。这样做的目的可以有效地减少森林火灾带来的损失甚至发生的频率。具体来说，进行森林火灾预测分析对于保护生态环境、促进经济发展和维护社会安全具有重要的意义。

1.保护生态环境。森林火灾会破坏生态环境，引发土壤侵蚀、水土流失、生物多样性下降等问题。通过对森林火灾的预测分析，可以采取及时的预防措施，保护生态环境。

2.经济发展。森林资源是一项重要的自然资源，对发展林业和旅游业都有很大的帮助。发生森林火灾不仅会影响林业产值，还会影响旅游业和相关产业的发展。因此，进行森林火灾预测分析对促进经济发展具有重要意义。

3.社会安全。森林火灾会威胁人们的生命和财产安全。通过对森林火灾的预测分析和预防措施的及时实施，可以减少火灾造成的人员伤亡和财产损失。

1. 问题描述

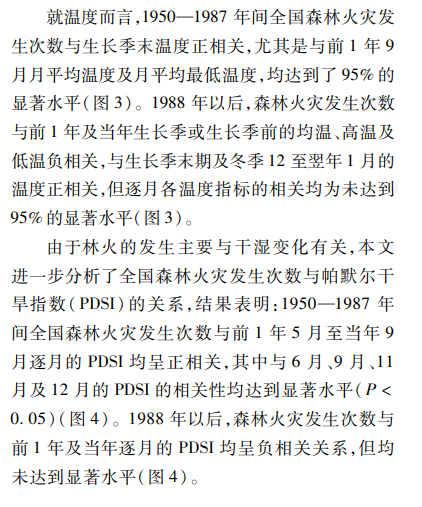
根据与森林火灾发生因素有关的文献知，森林火灾发生的主要因素是温度和湿度。为了搞明白森林火灾发生的主要因素之间怎样的搭配更容易发生森林火灾，我分析了关于葡萄牙蒙特西尼奥自然公园的 517 起火灾数据，其中数据集有十三个属性，分别为X：编号；Y：火灾发生年份；Month：火灾发生月份；Day：火灾所发生的工作日；FFMC：细小可燃物浓度；DMC：硫化物浓度；DC：空气污染物浓度；ISI：空气质量指数；Temp：火灾发生时的气温；RH：火灾发生时的湿度；Wind：风向；Rain：火灾发生时的降雨量；Area：过火面积。本次研究主要是通过对温度和湿度的聚类分析得到怎样的温度和湿度条件下森林火灾更容易发生。

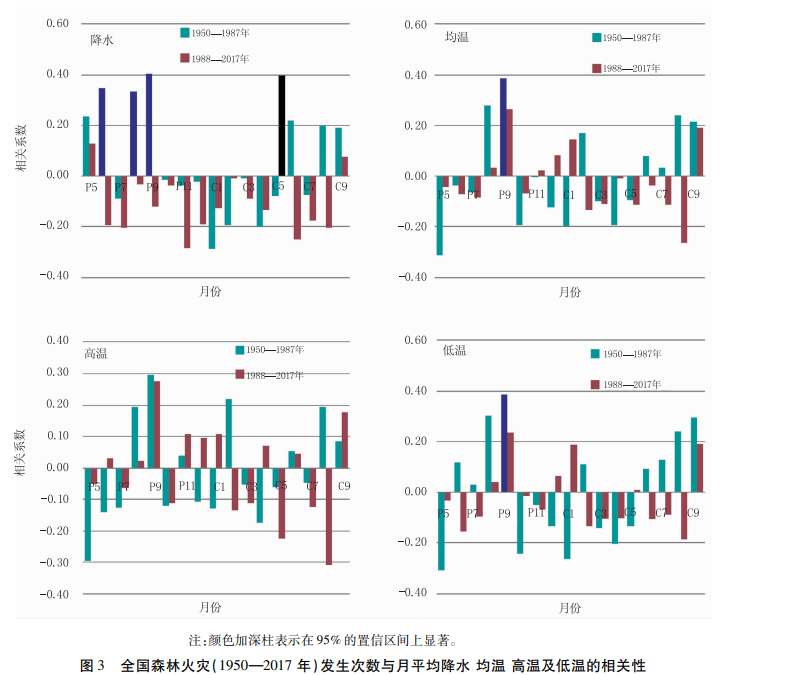
1. 模型搭建

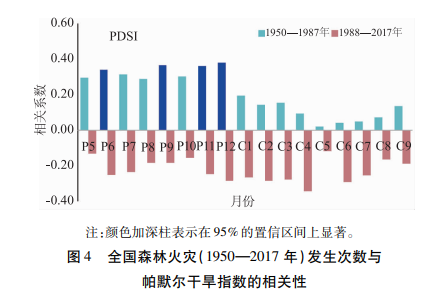
3.1 特征工程

根据以往的相关研究知火灾发生的主要影响因素是温度、干湿度和月份，比如杜建华、宫殷婷和蒋丽伟的关于中国森林火灾发生特征及其与主要气候因子的关系研究，其中就指明了全国森林火灾发生次数与生长季末温度呈相关关系和全国森林火灾发生次数与帕默尔干旱指数（ＰＤＳＩ）呈相关关系（即林火的发生与干湿变化有关），同时也指明了不同的月份，温度和湿度也有所不同，进而火灾发生的次数也有所不同。为了研究葡萄牙蒙特西尼奥自然公园火灾发生与温度、湿度和月份具体的关系，因此选择了temp、RH和month三个特征来研究它们是怎么影响森林火灾的发生。

由相关性分析得到temp与RH之间的相关性是-5.27390，相关性不是很强，因此不采取线性回归的模型来分析两者对森林火灾发生的影响因素，采取聚类分析的方法来分析两者不同的结合对森林火灾发生的影响因素。







3.2数据预处理

3.2.1 数据冲洗

数据集（数据来源：https://tianchi.aliyun.com/dataset/dataDetail?dataId=92968）中共有12个维度，518条记录。数据记录由于是官方整理的数据，所以信息较为完整，无需进行补全或删除操作，实验中只需根据需求进行切片分析即可。

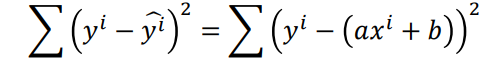
3.2.2 属性归约

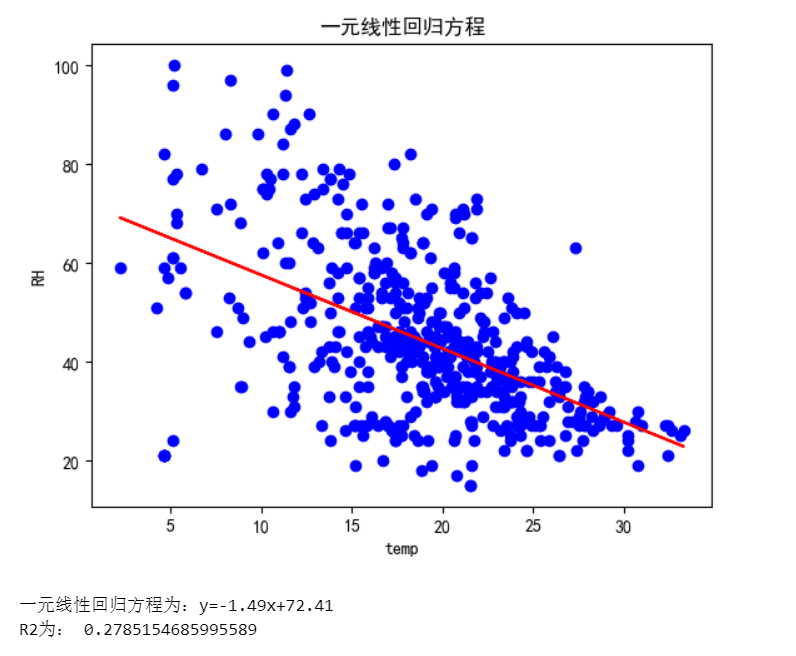
由特征工程的分析知，采取temp、RH和month三个属性来分析他们对森林火灾的具体影响，因此提取数据数据集中的三列数据，进而来分析他们对森林火灾的影响情况。

3.3 机器学习模型搭建

3.3.1 线性回归模型

一元线性回归模型也称为简单线性回归模型，其形式可以通过如下公式表达：y=a\*x+b，其中y为因变量，x为自变量，a表示回归系数，b表示截距。其中为实际值，为预测值，一元线性回归的目的就是拟合出一条线来使得预测值和 实际值尽可能的接近，如果大部分点都落在拟合出来的线上，那么该线性回 归模型则拟合较好。通常用预测值与实际值之间差值的平方和来衡量模型的好坏。





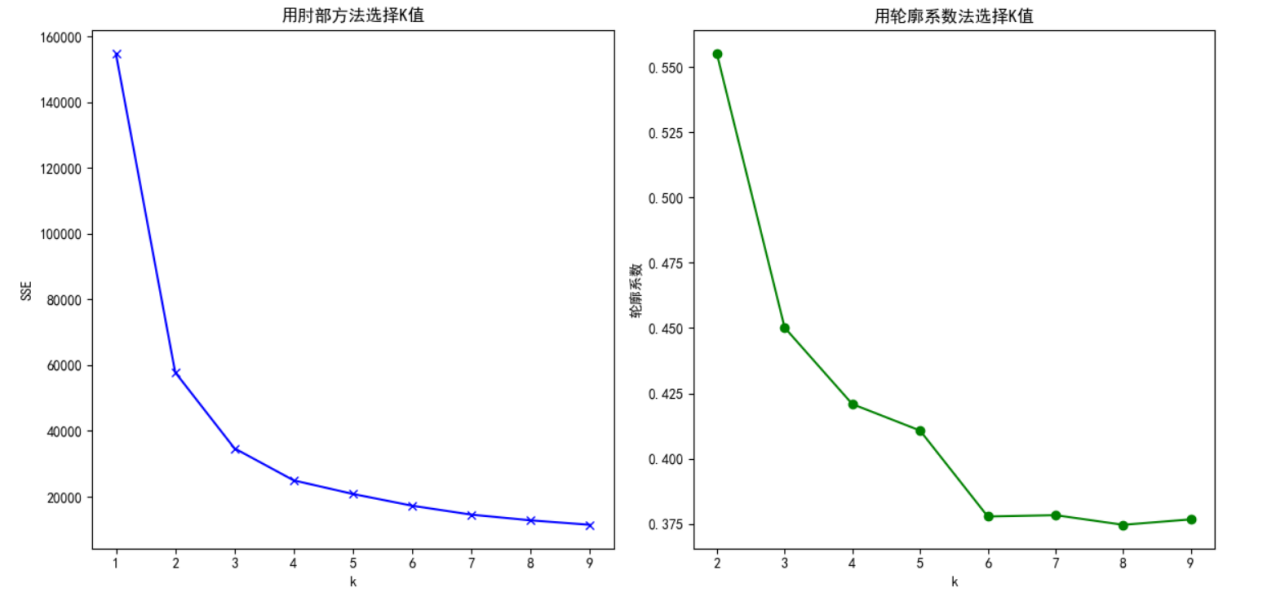
根据temp与RH的线性回归模型中R^2为0.28知temp与RH之间的线性相关不是很强，因此两者之间不能简单的说他们呈负相关的时候容易发生火灾。所以进行以下的聚类分析。

3.3.2 聚类分析

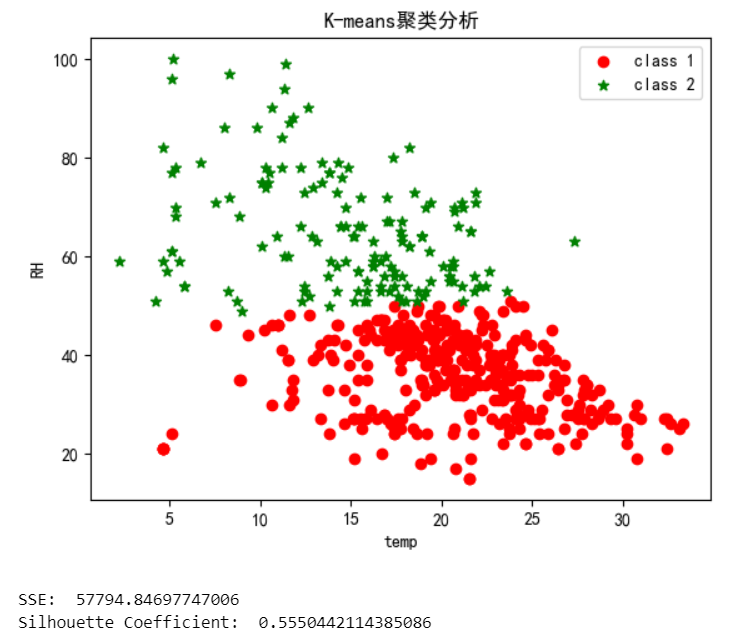
3.3.2.1 K-means聚类分析

KMeans算法的K代表类别数量，Means代表每个类别内样本的均值，所以 KMeans算法又称为K-均值算法。KMeans算法以距离作为样本间相似度的度量标准，将距离相近的样本分配至同一个类别。样本间距离的计算方式可以是欧氏距离，曼哈顿距离，余弦相似度等，KMeans算法通常采用欧氏距离来度量各样本间的距离。

KMeans算法的核心思想：对每个样本点计算到各个中心点的距离，并将该样本点分配给最近中心点代表的类别，一次迭代完成后，根据聚类结果更新每个类别的中心点，然后重复之前操作再次迭代，直到前后两次分类结果没有差别。



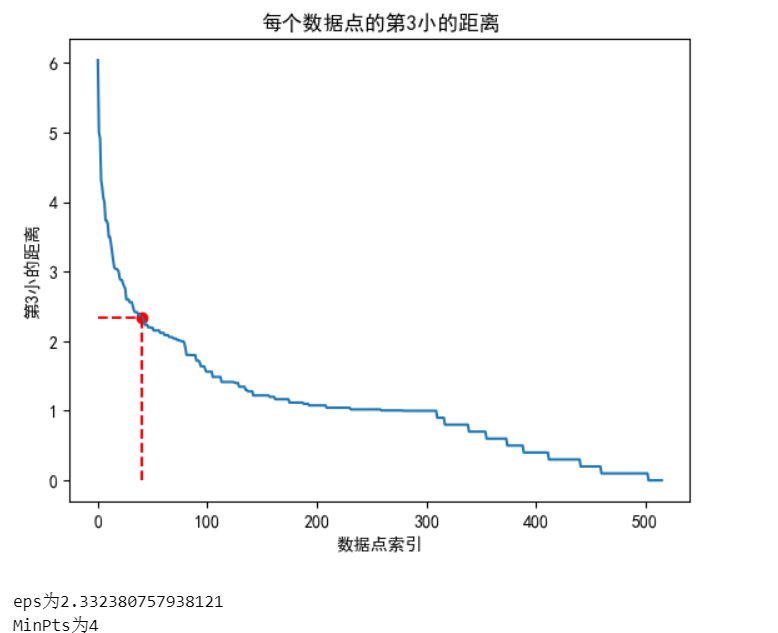
通过肘部方法和轮廓系数的结合分析来获得合适的k值，首先肘部方法的拐点k为2，在轮廓系数为最大值时k为2，因此最合适的k为2。

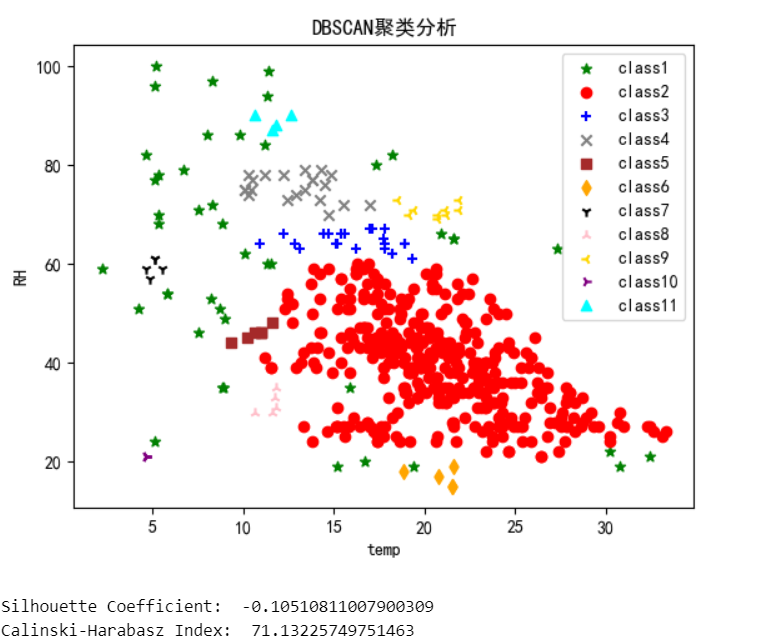


由图知温度与湿度对火灾发生的影响主要分为两种，其中温度高湿度低的时候很容易发生火灾，当温度低和湿度高也会发生火灾只不过没有温度高湿度低的时候发生的集中。根据SSE为57794.85，轮廓系数为0.56，可以粗略得到用K-means聚类分析的可信度较强。

3.3.2.2 DBSCAN聚类分析

DBSCAN 是一种以密度为基础的空间聚类算法，可以用密度的概念剔除不属于任意类别的噪声点。该算法将簇定义为密度相连的点的最大集合，将具有足够密度的区域划分为簇，并可以发现任意形状的簇。

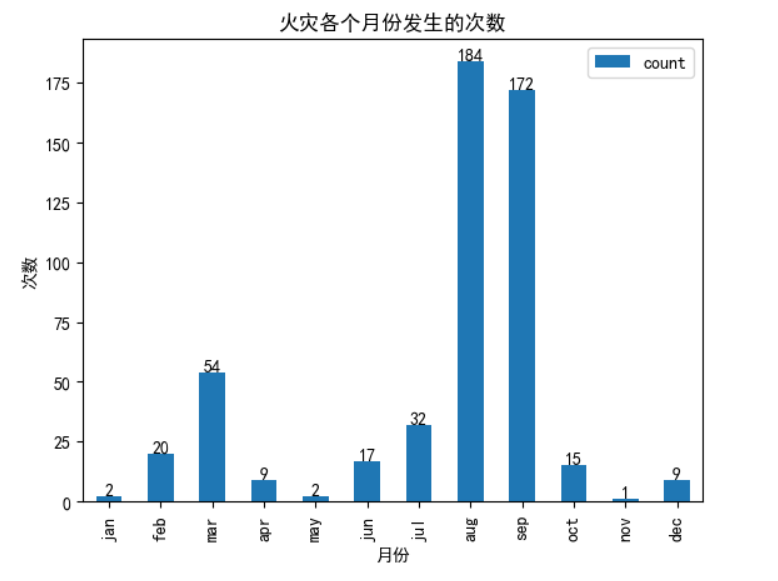




由图知温度与湿度对火灾发生的影响分为11类，但由图可以明显的看出其中class2占比居多，class3和class4占比相对居中，class1为噪声点，其余的簇占比相对较少。根据DBSCAN聚类分析可以看出大部分火灾的发生是在高温湿度低的情况下，部分发生在温度和湿度都正常状态下的情况。根据轮廓系数为-0.11和CH指数为71.13，总体看来，DBSCAN聚类分析的可信度一般。

3.3.2.3 GroupBy聚类分析

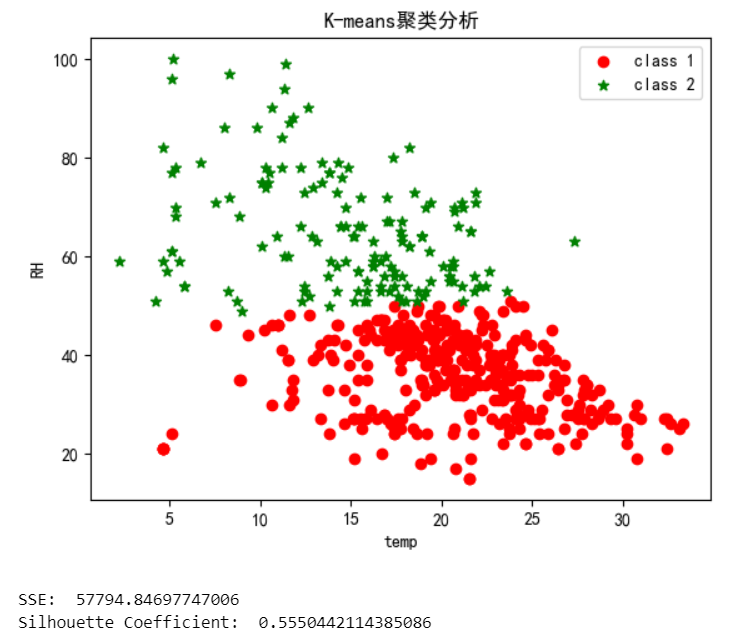
GroupBy会对所有的数据先根据指定列进行分组，然后针对分组后的数据在组内进行聚合运算(计数等)，最后再将聚合后的每组数据进行汇总就得到了我们想要的结果。



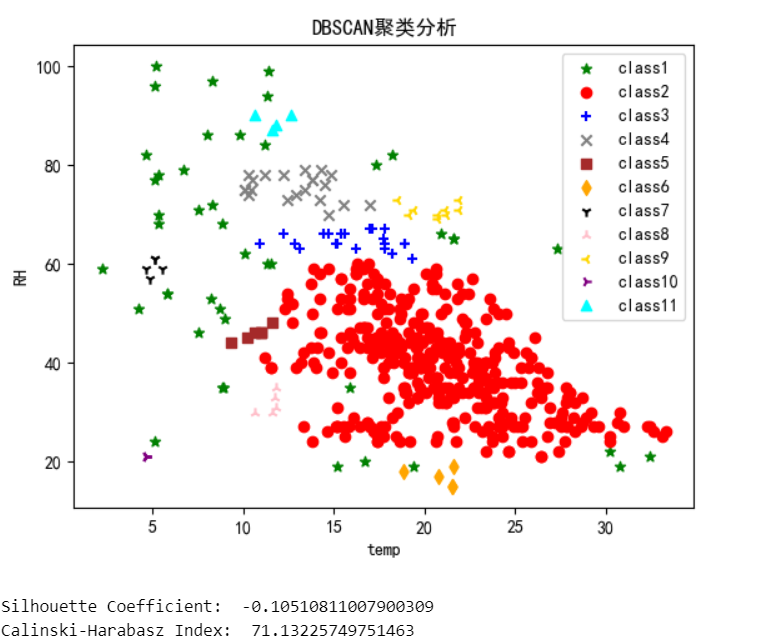
由图可以直观的看到火灾发生较为频繁的时间为8-9月，其他月份相对较少。由于葡萄牙位于北半球的欧洲地区，属海洋性气候与地中海气候的过渡地区，4-10月为旱季，8-9月气温较高，同时由于旱季还未结束，降水较少，空气中湿度降低，即温度较高湿度较低，由上图的聚类分析可知很容易发生火灾。

3.4 算法比较研究

3.4.1 K-means算法



3.4.2 DBSCAN算法



3.4.3 比较研究

K-means算法更适合常规数据集，而DBSCAN算法更适合形状类似同心圆的数据。根据葡萄牙蒙特西尼奥自然公园的 517 起火灾的记录数据知他属于常规数据集而不属于同心圆的数据集，因此采取K-means算法更合适。由上两种不同的算法结果可知K-means算法能较好地对temp和RH的数据进行分类，而DBSCAN不能较好的对temp和RH的数据进行分类。

K-means需要事先确定K值，即将数据聚为几类，DBSCAN不需要事先确定K值但无法确定最优值。由肘部方法和轮廓系数知K=2时，K-means算法可以很好的对数据进行分类，temp高RH低的时候很容发生火灾，temp低RH高的时候也会发生火灾然而没有temp高RH低的时候集中。由每个数据点第三小的距离得到eps和MinPts的值，DBSCAN算法将数据分为十一类，由图可以看出这十一类中最具意义的一类在于temp高RH低的时候更容易发生火灾。

结合K-means和DBSCAN聚类分析的结果和评价指标可以初步得到温度高和湿度低的时候更容易发生火灾，同时，需要注意火灾无时不在，需要时刻准备着，不可掉以轻心，在温度高和湿度低的时候更需要注意，加强防备。

4. 结论

4.1 研究目的

根据以往的相关研究知火灾发生的主要影响因素是温度、干湿度和月份，比如杜建华、宫殷婷和蒋丽伟的关于中国森林火灾发生特征及其与主要气候因子的关系研究，其中就指明了全国森林火灾发生次数与生长季末温度呈相关关系和全国森林火灾发生次数与帕默尔干旱指数（ＰＤＳＩ）呈相关关系（即林火的发生与干湿变化有关），同时也指明了不同的月份，温度和湿度也有所不同，进而火灾发生的次数也有所不同。为了研究葡萄牙蒙特西尼奥自然公园火灾发生的主要因素（温度和湿度）之间怎样的搭配更容易发生森林火灾，本次实验研究了温度与湿度对森林火灾发生的影响和森林火灾在不同月份的发生情况。

通过对葡萄牙蒙特西尼奥自然公园森林火灾的数据的分析可以得到，夏秋季节是火灾高发季节，干燥指数越高（即湿度越低）、温度越高，火灾发生的概率越大。查阅资料发现，森林火灾的发生与高温、连续干旱、大风有密切关系，与分析得出来的结论大体相符，而且葡萄牙为地中海气候，夏季炎热干燥，冬季低温湿润，与夏季为火灾高发季节相符。同时，需要注意火灾在任何一个时刻都可能发生，必须时刻保持警觉，并做好充分准备，以免在紧急情况下缺乏应对经验。

由上面可知夏季高温、干燥天气更容易发生火灾且火灾在任何时刻都可能发生，因此我们应做好以下几点：

一、提高认识，加强森林防火宣传教育。提高公众对森林火灾危害性的认识，普及森林防火知识和自防自救技能。通过媒体、社区、学校等渠道，广泛传播森林防火信息，形成全社会共同关注、共同参与的防火氛围。

二、强化督查，严格控制火源。在森林内严禁吸烟、野炊、燃放烟花爆竹等一切可能引发火灾的行为。加强林区巡查，及时发现并制止违规用火行为。同时，对于必要的生产用火，要严格按照规定进行审批和监管，确保用火安全。

三、强化措施,加强森林防火设施建设。在林区重要地段和关键区域，设置防火隔离带、防火林带等，有效阻隔火势蔓延。加强森林防火通讯和监控系统的建设，提高火灾预警和应急响应能力。

四、强化森林防火队伍建设。建立专业的森林防火队伍，配备必要的装备和器材，进行定期培训和演练，提高森林防火队伍的专业素质和应对能力。同时，加强林区群众参与森林的积极性，形成群防群治的良好局面。

五、加强森林防火法规制度建设。完善森林防火法规体系，明确各级政府和有关部门的职责和义务。加大对违规用火行为的处罚力度，形成有效的法律震慑。

六、科学扑救，安全有效处置森林火灾。要牢固树立“以人为本，安全第一”的思想，认真吸取森林火灾人员伤亡事故教训，把握夏秋高温气侯下扑救森林火灾的安全要领，把确保扑火人员安全放在森林火灾扑救指挥工作的首要位置。森林之灾发生后，各村场负责同志要第一时间报告镇森林防火指挥所，并亲临现场扑救，禁止群众无组织地扑救森林火灾，确保不发生人员伤亡事故。

4.2 研究局限性

本次研究整体效果很好，实现了最初设定的目标，找到主要因素对森林火灾的影响。本实验通过聚类分析得到：温度高和湿度低的时候很容易发生火灾，温度低湿度高的时候也容易发生火灾只不过没有温度高湿度低的时候集中，同时火灾在任何情况下都可能发生，需要时刻注意；除此之外，葡萄牙蒙特西尼奥自然公园发生森林火灾的时间点主要集中在八九月份。实验的目的虽然达到了，但是本实验还具有一定的局限性：特征选择说服力不够、聚类分析不够精准和泛化能力弱。

特征选择说服力不够。参考相关文献，选择温度、湿度和月份三个特征，没有针对具体的数据来进行选择特征，因此选择的特征不一定很符合此次数据的特征。特征选择要选择那些对因变量解释程度强的自变量，根据参考文献知温度和湿度对森林火灾的发生的影响很大，因此我选择这两个特征，同时温度和湿度与月份相关度很高，因此又选择了月份这个特征。

聚类分析不够精准。聚类分析要求不断的迭代，直到前后两次实验结构差距不大，本次实验关于K-means和DBSCAN的聚类分析缺乏迭代，最终的分类结果的精准度会有一点问题。虽然聚类结果不够精准，但也可以大概的看出每一类的特点，不影响大方面规律的寻找，但是虽然不影响规律的寻找，但分类不精准总是会造成一些错误的理解。

泛化能力弱。此次研究针对的是葡萄牙蒙特西尼奥自然公园的 517 起火灾的记录数据，数据面比较小，泛化能力弱，对指导葡萄牙蒙特西尼奥自然公园预防火灾具有很高的价值，其他森林与葡萄牙蒙特西尼奥自然公园的环境有一定的差别，因此此次研究的结果对于其他森林的指导意义就有限，但参考意义还是很大的。

4.3 研究展望

每个森林所处的地方不同，那么它发生火灾的影响因素将会有所不同，而本次的研究的泛化能力弱，为了让研究能对所有的森林火灾防控都具有较高的指导作用，未来的研究方向将会是增加研究的泛化能力。比如采取不同森林的火灾数据，找到不同森林火灾的主要影响因素之间的怎样的配合更容易引发森林火灾的共同点，进而根据研究的结果对我们的森林火灾防控提出更加合理的建议。森林火灾带来的危害是巨大的，在未来希望此次研究不是只适合于指导葡萄牙蒙特西尼奥自然公园的森林火灾防控，而是适合于所有的森林的火灾防控，从而减少森林火灾的发生！

参考文献：

[1]杜建华, 宫殷婷, 蒋丽伟.中国森林火灾发生特征及其与主要气候因子的关系研究[J]. 林业资源管理, 2019, 0(2): 7-14.

[2]宋志龙, 王海晖, 吴泽鹏, 杨峰峰, 靳邦鑫, 王振师.西方国家森林火灾研究架构与技术建树[J]. 林业世界, 2022, 11(2): 47-55

[3]吴有声.林火监测与预警在森林防灭火中的运用[J].农家参谋，2021（1）：181-182.

[4]高艳霞.林火监测与预警在森林防灭火中的运用探究[J].南方农业，2020，14（27）：86-87.

[5]王光耀.基于机器学习的火灾检测方法研究[D].大连理工大学,2015.

[6]舒立福, 田晓瑞.国外森林防火工作现状及展望[J]. 世界林业研究, 1997, (2): 29-37.

[7]《河南省全面加强新形势下森林防灭火工作实施方案》[N].河南日报,2024.

[8]张恒,李慧,赵鹏武.内蒙古森林火灾发生风险及驱动因素的影响[J/OL].生态学报,2024.

[9]Rahman A R ,White B ,Ma C .The effect of growth, deforestation, forest fires, and volcanoes on Indonesian regional air quality[J].Journal of Cleaner Production,2024.

[10]Ma T ,Wang G ,Guo R , et al.Forest fire susceptibility assessment under small sample scenario: A semi-supervised learning approach using transductive support vector machine.[J].Journal of environmental management,2024.