# SpotHitPy：使用 Spotify 进行基于机器学习的歌曲热门度预测研究

Ioannis Dimolitsas*、Spyridon Kantarelis†和 Afroditi Fouka‡雅典国立技术大学电气与计算机工程学院，希腊雅典 邮箱： *jdimol@netmode.ntua.gr， †spyroskanta@ails.ece.ntua.gr， ‡afroditifouka@mail.ntua.gr

arXiv:2301.07978v1 [cs.SD] 19 Jan 2023

摘要——**本研究探讨了热门歌曲预测问题，旨在预测哪些歌曲会成为 Billboard 热门歌曲。我们收集了一个包含近 18500 首热门歌曲和非热门歌曲的数据集，并使用 Spotify Web API 提取了它们的音频特征。我们在该数据集上测试了四种机器学习模型。我们能够以约 86% 的准确率预测歌曲在 Billboard 上的表现。效果最好的算法是随机森林和支持向量机。**

索引词——**机器学习、音乐信息检索、热门歌曲科学、二元分类、数据挖掘、数据收集、特征选择**

## 一、引言

热门歌曲科学（HSS）是音乐信息检索（MIR）领域的一个活跃研究课题。该课题主要关注预测一首歌曲能否成为热门歌曲。因此，热门歌曲预测对音乐人、唱片公司和音乐供应商都很有用，因为热门歌曲能够吸引大量观众。热门歌曲有助于唱片公司和音乐供应商提高利润，也能帮助艺人向广大观众传递自己的理念。我们的方法基于这样一个假设：热门歌曲在音频特征方面具有相似性。我们使用 Spotify Web API[1]收集了热门歌曲和非热门歌曲及其特征，构建了一个数据集。随后，我们使用机器学习方法和算法来预测一首歌曲是否热门。

本文其余部分安排如下：第二节介绍热门歌曲科学领域的相关研究。第三节介绍我们收集数据集的方法。第四节介绍我们使用的机器学习方法和算法。第五节对所提出的框架进行全面评估，第六节总结全文。

## II. 相关工作

正如我们所提到的，热门歌曲预测是音乐信息检索领域的一个活跃研究课题。Raza 和 Nanath [1] 得出结论，目前还没有能够在歌曲发行前预测其能否走红的万能公式。研究人员已经提出了各种方法。Li- Chia Yang 等人 [2]将最先进的深度学习技术引入基于音频的热门歌曲预测问题。Zangerle 等人 [3] 提出在深度神经网络中结合歌曲的低层和高层音频特征，以区分低层和高层特征，并考虑它们的特殊性。Elena Georgieva 等人 [4]使用了五种机器学习算法，成功预测歌曲在 Billboard 上的成绩，准确率约为 75%。Middlebrook 和 Sheik [5] 测试了四种机器学习模型，在预测歌曲的 Billboard 成绩方面达到了 88% 的准确率。

表I：数据集的特征规格

| 特征 | 规格 |
|---|---|
| id | 歌曲唯一的 Spotify 曲目 ID |
| 艺术家 | 歌曲的艺人名称 |
| 流行度 | 一个介于 0 到 100 之间的值，其中 100 表示最受欢迎 |
| 显式 | 一个布尔值，表示某个音轨是否含有露骨歌词 |
| 专辑类型 | 专辑的类型：可以是“专辑”、“单曲”，或“合辑” |
| 可舞性 | 一个介于 0.0 到 1.0 之间的数值，用于描述一首音轨适合跳舞的程度 |
| 能量 | 一个介于 0.0 到 1.0 之间的数值，表示一种感知上的强度和活跃度的度量 |
| key | 音轨所处的音乐调性 |
| 响度 | 音轨的整体响度，以分贝（dB）为单位 |
| mode | 表示一首音轨的调式类型（大调=1 或小调=0）音轨 |
| 口语性 | 一个从 0.0 到 1.0 的数值，用于描述其中包含的音轨中出现的口语词汇 |
| 原声度 | 一个介于 0.0 到 1.0 之间的值，用于预测音轨是否为原声 |
| 器乐性 | 一个介于 0.0 到 1.0 之间的值，用于预测音轨是否是纯音乐还是包含人声 |
| 现场感 | 一个介于 0.0 到 1.0 之间、用于描述……存在情况的数值音轨中观众的存在情况 |
| 情感效价 | 一个介于 0.0 到 1.0 之间、用于描述音乐……的数值音轨传达的积极程度 |
| 速度 | 音轨的整体估算速度，单位为每分钟节拍数（BPM） |
| 时长 毫秒 | 音轨的时长，单位为毫秒 |
| 拍号_ | 音轨的估算整体拍号 |

## III. 数据集与特征

本节将讨论与数据收集和处理相关的见解。具体而言，本节介绍用于数据收集的工具和方法，同时说明样本特征的筛选与提取过程。此外，下文还将详细介绍并采用数据集准备技术，例如归一化和数据增强。

[1]https://developer.spotify.com/documentation/web-api/

*A.* 获取数据

首先，我们获取了数据集准备所必需的数据，其中包括从2011年到2021年每一年的公告牌百强热门歌曲。这些数据通过 Billboard API，利用 Python 库 *bill-board.py*[2] 获取。具体而言，我们构建了一个包含 1000 首热门歌曲的集合（这些歌曲曾进入某一年的公告牌百强榜），每首歌曲都由其标题及对应的艺术家（某些情况下为多位艺人）确定。随后，我们根据该集合使用 Python 库 *spotipy*[3] 向 Spotify API 发送 HTTP请求，以获取与这 1000 首公告牌热门歌曲相关的信息。对于每个请求，Spotify API 都会以 JSON 格式返回一组包含 10 个相关对象的数据。每个对象都包含“曲目 ID”、“艺术家ID”和歌曲“流行度”等基本歌曲特征。因此，我们选取流行度最高的对象用于构建数据集。之后，我们使用 Spotify 上的随机歌曲填充数据集。为此，我们生成由随机字符组成的查询，并将其作为曲目标题。这些查询随后作为 HTTP请求发送至 Spotify API。每条响应消息中包含 50 首歌曲，我们从中随机选择 10 首。总共发送了 2000 条此类请求，从 Spotify 收集了 20,000 首音轨。与热门歌曲的情况一样，对于非热门歌曲，响应消息作为 JSON 对象返回，其中同样包含这些基本特征。

基于上述数据，我们进一步丰富了各音轨对应的特征。为此，我们使用了音频特征*spotipy* 功能。对于每个获取的音轨对象，我们依据*id* 特征（该特征对每个对象而言都是唯一的）调用相应的 API调用，以获取更多特征。表 I总结了从 Spotify 提取的每条音轨的特征及其对应说明。收集到的数据构成了分类器将使用的数据集的基础。不过，要为此目的创建合适的数据集，还需要进行进一步处理。

*B.* 数据集准备

首先，我们进行了数据清理，以便仅保留 Spotify 中不重复的音轨。此外，我们剔除了缺少特定特征数据的音轨对象。经过这一初步处理，得到约 18,000 首 Spotify歌曲，其中有 861 首曾在 2011年至2021年间进入公告牌百强榜。显然，这个样本实例并不平衡，因为“热门”类别和“非热门”类别的分布比例并不均衡。若未妥善处理，数据不平衡会影响分类预测。因此，在继续后续步骤之前，必须平衡数据集。解决这一问题的一种方法是为代表不足的类别生成新样本。

[2]https://github.com/guoguo12/billboard-charts [3] https://spotipy.readthedocs.io/en/2.19.0/

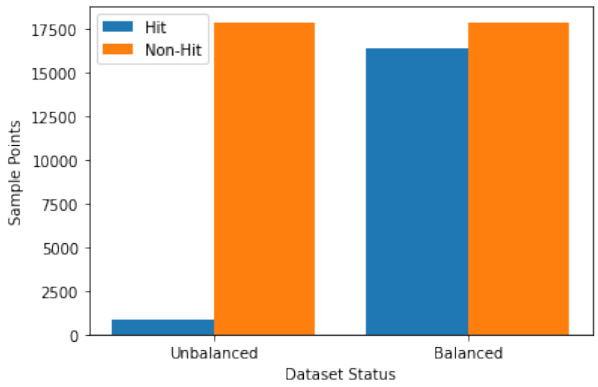



图1：不平衡数据集和平衡数据集

最简单直接的策略是通过对当前可用样本进行有放回的随机抽样来生成新样本。*RandomOverSampler* 提供了这种方案，我们使用了 sklearn Python 库中的这一技术[4]。

另一个需要解决的问题是，Spotify 提供的音频特征不在同一尺度上。如表I所示，有些特征是布尔值，有些是分贝值，有些取值范围为 0 到 100，还有些范围为 0 到 1.0。我们使用 sklearn Python 库中的缩放方法，成功将所有数值和布尔值缩放到 0 至 1.0 之间。

表II：主成分分析与特征系数

| 主成分 | 方差最大的特征 | 系数 |
|---|---|---|
| PC1 | 显式 | 0.728131 |
| PC2 | mode | -0.729190 |
| PC3 | key | -0.900311 |
| PC4 | 原声度 | 0.651091 |
| PC5 | 情感效价 | -0.810128 |
| PC6 | 可舞性 | -0.546361 |
| PC7 | 流行度 | -0.713593 |
| PC8 | 速度 | -0.834668 |
| PC9 | 器乐性 | 0.711965 |
| PC10 | 现场感 | 0.604598 |
| PC11 | 时长 毫秒 | 0.044111 |
| PC12 | 能量 | -0.454230 |
| PC13 | 响度 | -0.139564 |
| PC14 | 口语性 | 0.239800 |
| PC15 | 拍号 | -0.050520 |

标准化后，我们采用主成分分析，以识别能够反映更多分类问题相关信息的特征，同时降低数据集的维度。在进行主成分分析之前，必须先进行标准化，因为主成分分析对初始变量的方差非常敏感。也就是说，如果初始变量的取值范围差异较大，取值范围较大的变量就会占据主导地位，而取值范围较小的变量则会受到影响，从而导致有偏结果。主成分是由初始变量线性组合构成的新变量。在我们的 15 维

[4]https://scikit-learn.org/stable/

数据集中，得到了 15 个主成分，如表 II 所示，其中还列出了每个主成分中方差最大的特征。

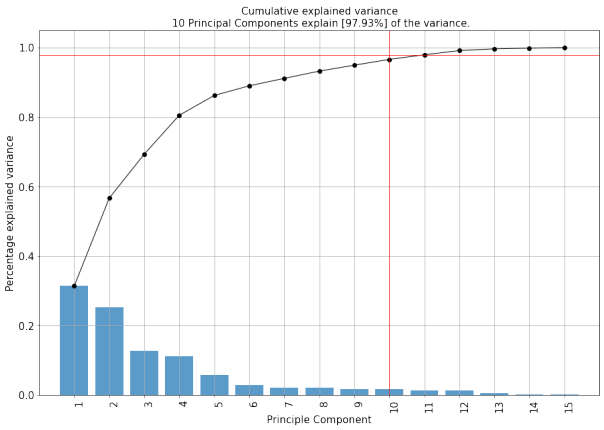



图 2：主成分分析累计解释方差

特征组合的确定方式是：使新特征（主成分）彼此不相关，并将初始特征中的大部分信息压缩到前几个主成分中。图 2 展示了方差随主成分变化的情况。如图所示，前 10 个主成分贡献了显著的方差，约为 98%，因此其余主成分（11–15）可以忽略。

## IV. 方法

本节将介绍本文所采用的不同机器学习方法：（i）随机森林 *(RF)*，（ii）支持向量机 *(SVM)*，（iii）逻辑回归 *(LR)*，（iv）*k*近邻 *(kNN)*

### *A.* 随机森林

随机森林 (RF) 模型 [6]是分类中最受欢迎的方法之一，旨在克服决策树的过拟合问题。具体而言，这类模型由一组通用的分类树构成，这些树使用数据集的不同方面以及随机选取的数据进行训练。其目的是降低剩余训练数据的总体方差。换言之，该方法试图通过依次进行二元划分，从初始数据集中生成一个同质子集。RF 模型的优点包括准确率高、训练简单且对离群值具有鲁棒性。其缺点之一是，得到的函数通常是离散的，而非平滑的。模型训练完成后，可通过对所有单棵分类树 $g_b$ 对 y 的预测结果取平均，得出对未见样本 y 的预测结果，如下所示。

$$f = \frac{1}{B} sum_{b=1}^{B} g_b(y) \tag{1}$$

### *B.* 支持向量机

支持向量机（SVM） [7]最初是为解决逻辑回归或分类问题而创建的。该方法是随机森林方法的推广，其目的是找出最优超平面，将数据划分为两个可识别且定义明确的类别。当训练数据稀疏时，SVM 能有效提升拟合质量，但同时也会增加计算开销。高斯径向基函数（RBF）通常用作算法的核：

$$e^{-g\|(x-y)^2\|}, \tag{2}$$

其中，g 是核参数。

### *C.* 逻辑回归

当目标变量为类别型时，逻辑回归（LR） [8] 是一种常用的分类算法。该方法旨在建立特征与期望输出之间的关联。一般而言，逻辑回归问题有两种类型：二元问题和多分类问题。逻辑回归模型的基本形式使用逻辑函数来确定二元因变量。该函数将对数优势重新表示为有意义的概率，最终的标签值为“0”和“1”。逻辑函数由以下方程描述

$$P(x) = \frac{1}{1 + e^{\frac{-(x-m)}{s}}} \tag{3}$$

其中，m 分别是位置参数，s 是尺度参数。

### *D. k*近邻

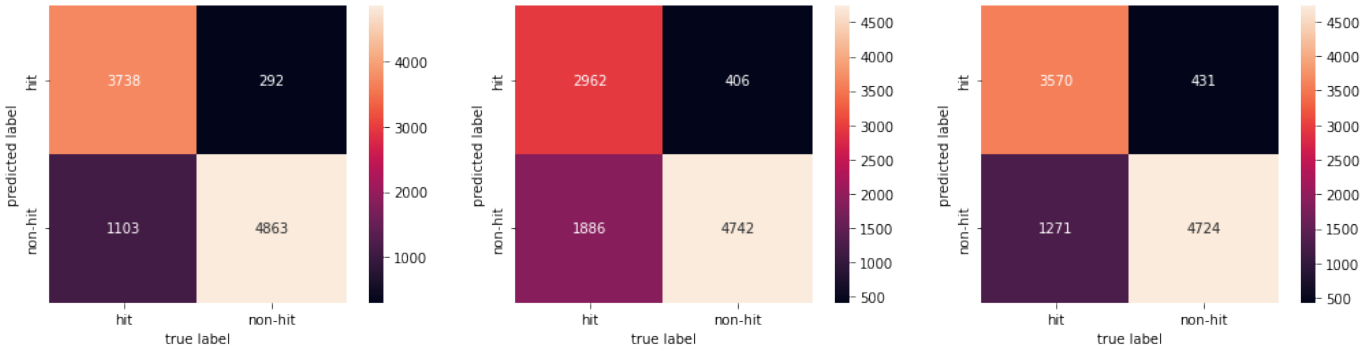
k近邻算法 [9] 是一种非参数分类算法，属于监督学习问题。随后，该算法应用于标记数据，并将数据划分为多个类别。它通常用作分类器，以简单有效而广为人知。不过，它也可用于回归问题。在前一种情况下，算法会使用汉明距离等度量来确定对象所属的类别；在后一种情况下，通常使用欧几里得距离。最后，参数 k 的取值对算法的性能十分重要，因为它会影响类别之间的边界。一般而言，参数 k 没有严格的取值规则，而是取决于数据。在二分类问题中，可以将与某个样本最相似的 K 个样本中出现频率最高的类别作为输出。例如，类别 0 的概率可按如下方式估算。

$$P(class = 0) = \frac{count(class = 0)}{count(class = 0) + count(class = 1)} \tag{4}$$

## V. 实验/结果/讨论

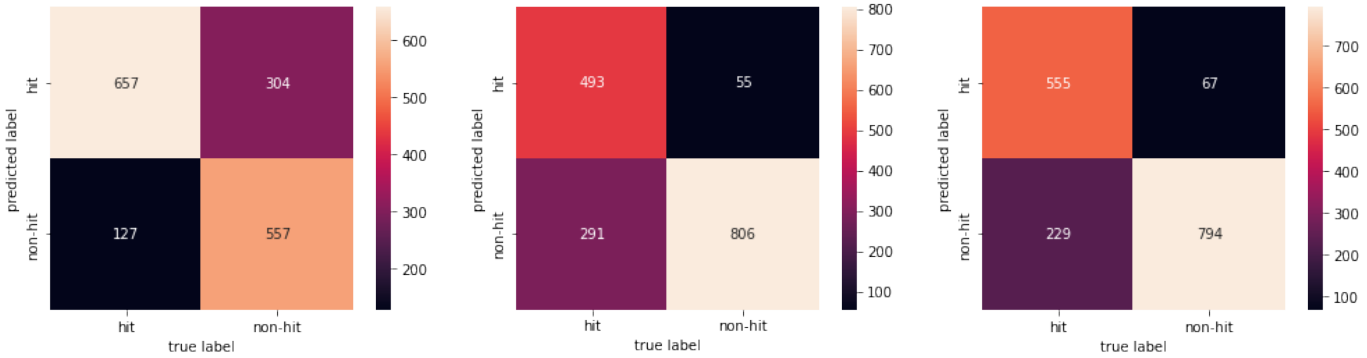
本节介绍训练过程，并对第四节所述的机器学习模型进行全面评估。

首先，我们根据数据集的最终版本创建了训练集、验证集和测试集。该最终版本本质上是经过尺度变换的数据，并确保每个类别都有平衡的样本点。




(a) 随机森林分类器　(b) 优化的 k近邻分类器　(c) 优化的支持向量机分类器

图 3：验证集的混淆矩阵




(a) 随机森林分类器　(b) 优化的 k近邻分类器　(c) 优化的 SVM 分类器

图4：测试集的混淆矩阵

表 III：验证集上的模型指标

| 分类器 | 指标 | | |
|---|---|---|---|
| | 准确率 | 精确率 | 召回率 |
| 支持向量机 | 0.73 | 0.70 | 0.83 |
| 逻辑回归 | 0.66 | 0.64 | 0.76 |
| **随机森林** | **0.86** | **0.82** | **0.94** |
| k近邻（n=25） | 0.76 | 0.71 | 0.90 |
| 支持向量机 - opt | 0.83 | 0.79 | 0.92 |
| k近邻 (n=25) - opt | 0.80 | 0.71 | 0.90 |

我们按照 7:3 的比例将平衡数据集拆分为训练集和验证集。因此，我们使用了约 25,000 首音乐曲目来训练所讨论的模型。测试集由 861 首热门歌曲和 861 首随机选取、各不相同的非热门歌曲组成。

在评估中，我们使用三种不同的指标：准确率、精确率和召回率[5]。

准确率是一个通常用于描述模型在所有类别上的表现的度量。它的计算方式是正确预测的数量 与预测结果总数之比：

$$Accuracy = \frac{True_{positive} + True_{negative}}{Total\ number\ of\ predictions}$$

精确率的计算方式是 被正确分类的正样本数量与被分类为正样本的样本总数（无论分类正确与否）之比：

$$Precision = \frac{True_{positive}}{True_{positive} + False_{positive}}$$

召回率的计算方法是将正确分类为正类的正样本数量除以正样本总数：

$$Recall = \frac{True_{positive}}{True_{positive} + False_{negative}}$$

我们在验证测试中的结果如表 III 所示。我们观察到，表现最佳的分类器是随机森林、优化后的支持向量机和优化后的 k近邻。为了提升 k近邻算法的性能，我们使用了交叉验证，并发现邻居数的最佳值为 25。对于支持向量机模型，正则化参数设为 10，使用径向基函数核。这些分类器在验证测试中的性能混淆矩阵如图 3 所示。为了获得更好的结果，我们使用了

[5]https://blog.paperspace.com/deep-learning-metrics-precision-recall-准确率/

在支持向量机和 k近邻分类器上采用了流水线方法。如表 III 所示，使用流水线使这两种算法的所有指标平均提升了 9%。

为了评估我们的方法的适应性，我们在测试集上测试了表现最佳的三个分类器。结果指标如图5所示，混淆矩阵如图4所示。可以看出，随机森林的精确率高于支持向量机和 k近邻，而另外两个模型的召回率较高，准确率也更高。

考虑到音乐趋势会随着年份不断变化， [10],热门歌曲预测问题具有很高的复杂性。在本研究中，我们使用了过去十年的热门歌曲，以聚焦于当代音乐趋势。

模型的召回率较高，意味着假阴性的数量较少。这是因为从 Spotify 收集的音频特征可以反映歌曲的声音，却不受歌曲发行年份的音乐趋势影响。因此，使用这些特征训练模型，能够更准确地预测一首歌是否不是热门歌曲，而不是预测它是否会成为热门歌曲。要提高模型的精确率，也就是让模型更准确地预测一首歌是否会成为热门歌曲，就需要加入能够考虑当前音乐趋势的特征。

因此，结果的影响可以归纳为以下问题：

*(i) 2011*年的头号热门歌曲如果在*2021*年发行，还会成为热门歌曲吗？

*(ii)* 预测*2011*年的头号热门歌曲在*2021*年仍会成为热门歌曲，是否意味着过拟合？

*(iii)* 如果一首*2011*年的非热门歌曲的特征与*2021*年热门歌曲的主导特征相匹配，它能否在*2021*年成为热门歌曲？

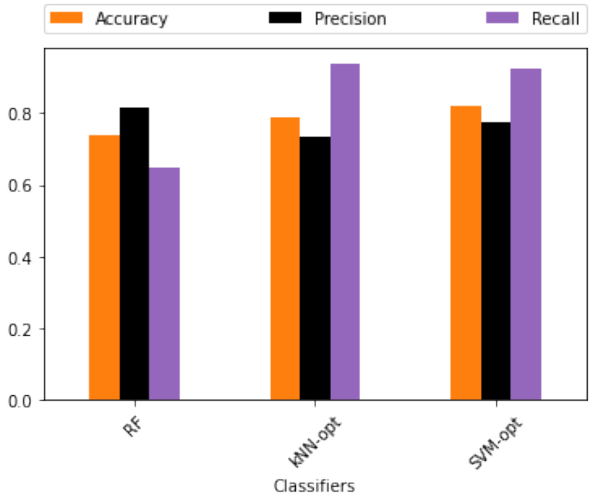



图 5：测试集的模型指标

## VI. 结论与未来工作

我们的研究表明，在热门歌曲预测中表现最佳的机器学习算法是随机森林，其准确率达到86%。随机森林在验证集和测试集上都取得了较高的精确率，因此适用于热门歌曲预测问题。支持向量机和k近邻算法在我们的测试集上显示出更高的准确率，这表明它们在规模更大的实验中可能更有效，因此值得进一步探索和优化。

此外，进一步探索并收集更多歌曲元数据，将有助于选择更好的特征。获取歌曲旋律和和声方面的信息，应能更好地理解歌曲结构，从而得出更好的结果。此外，训练时也可以考虑情绪或情感等特征。

受第五节所讨论问题的启发，探索参数并定义能够体现音乐趋势的特征，是解决热门歌曲预测问题时另一项有用的工作。

## 致谢

本研究是在雅典国立技术大学电气与计算机工程学院数据科学与机器学习硕士项目的机器学习课程背景下开展的。

我们感谢埃因霍温理工大学博士后研究员 Giannis Delimpaltadakis 提供有益的讨论，并祝愿他在学术生涯的下一阶段一切顺利。

## 参考文献

[1] A. H. Raza 和 K. Nanath，“利用机器学习预测热门歌曲：是否存在先验秘密公式？”，载于*2020*年数据科学、人工智能与商业分析国际会议 *(DATABIA)*，2020年，第111–116页。[2] L.-C. Yang、S.-Y. Chou、J.-Y. Liu、Y.-H. Yang 和 Y.-A. Chen，“重新审视利用卷积神经网络进行基于音频的热门歌曲预测问题”，载于*2017*年*IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，2017年，第621–625页。[3] E. Zangerle、M. V¨otter、R. Huber 和 Y.-H. Yang，“热门歌曲预测：利用低层和高层音频特征”，载于*ISMIR*，2019年。[4] E. Georgieva、M. S.uta 和 N. S. Burton，“Hitpredict：利用 Spotify 数据预测热门歌曲，斯坦福大学计算机科学229课程：机器学习”，2018年。[5] K. Middlebrook 和 K. Sheik，“歌曲热门度预测：利用 Spotify 数据预测 Billboard 热门歌曲”，*arXiv*，卷 abs/1908.08609，2019年。[6] T. K. Ho，“随机决策森林”，载于第三届文档分析与识别国际会议论文集，第1卷。IEEE，1995年，第278–282页。[7] C. Cortes 和 V. Vapnik，“支持向量网络”，机器学习，第20卷，第3期，第273–297页，1995年。[8] D. R. Cox，“二进制序列的回归分析”，皇家统计学会学报：*B*辑（方法论），第20卷，第2期，第215–232页，1958年。[9] K. Taunk、S. De、S. Verma 和 A. Swetapadma，“用于学习与分类的最近邻算法简要综述”，载于*2019*年智能计算与控制系统国际会议 *(ICCS)*，2019年，第1255–1260页。[10] M. Interiano、K. Kazemi、L. Wang、J. Yang、Z. Yu 和 N. Komarova，“ 排行榜内外当代歌曲的音乐趋势与成功可预测性 ”，英国皇家学会开放科学，第5卷，171274，2018年5月，第。