制定解决泰伦加纳癌症意识问题的方案

Priyanka Avhad
Department of Computer Engg.
Veermata Jijabai Technological Institute
Mumbai, India 400019
pravhad_b19@ce.vjti.ac.in

Vedanti Kshirsagar
Department of Computer Engg.
Veermata Jijabai Technological Institute
Mumbai, India 400019
vakshirsagar_b19@ce.vjti.ac.in

Mahek Nakhua
Department of Computer Engg.
Veermata Jijabai Technological Institute
Mumbai, India 400019
msnakhua_b19@ce.vjti.ac.in

Urvi Ranjan
Department of Electronics Engg.
Veermata Jijabai Technological Institute
Mumbai, India 400019
urranjan_b19@el.vjti.ac.in

摘要—根据数据 [1]，2019-2020 年在泰伦加纳进行宫颈癌、乳腺癌和口腔癌筛查的女性比例分别为 3.3%，0.3%和 2.3%。尽管早期检测是减少发病率和死亡率的唯一方法，但人们对宫颈癌和乳腺癌的症状以及筛查实践的意识非常低。我们开发了一个机器学习分类模型，根据人口统计因素预测一个人是否容易患乳腺癌或宫颈癌。我们设计了一个系统，基于用户的位置或地址提供最近医院或癌症治疗中心的建议。除此之外，我们可以整合健康卡来维护所有人的医疗记录，并进行意识提升活动和运动。对于机器学习分类模型，我们分别使用了决策树分类和支持向量机分类算法来预测宫颈癌易感性和乳腺癌易感性。因此，通过设计这个解决方案，我们离我们的目标更近了一步，即传播癌症意识，从而降低泰伦加纳人民的癌症死亡率并提高他们的癌症知识水平。

I. 介绍

癌症素养对于减少全球癌症死亡率至关重要，并且对早期发现癌症也非常重要。乳腺癌是全球女性中最常见的癌症类型，包括在印度，诊断出晚期病例以及发病率和死亡率的上升使得提高女性的癌症素养变得尤为重要。

NFHS 数据集 [1] 被用于评估女性（年龄在 30 至 49 岁之间）的癌症筛查指标，包括宫颈癌、乳腺癌和口腔癌。这些数据根据居住地分为两类——农村和城市地区，并按区分类。

农村和城市地区（年龄 30-49 岁）女性癌症筛查的百分比如图 1 所示

NFHS		Urban	Rural
Women of Telangana in the age group of 30-49	Who have ever undergone a screening test for cervical cancer	2.3 %	3.9 %
	Who have ever undergone a breast examination for breast cancer	0.3 %	0.4 %
	Who have ever undergone an oral cavity examination for oral cancer	3.2 %	2.1 %

图 1. NFHS 数据集摘要，突出显示接受这些测试的女性比例较小

为了基于特伦甘纳邦的地区进行比较，设计了一个包含所有纬度和经度的新组合数据集。对三个指标（是

否曾接受过宫颈癌筛查测试（%），是否曾接受过乳腺检查以检测乳腺癌（%），是否曾接受过口腔检查以检测口腔癌（%））进行了数据可视化处理。图 2、图 3 和图 4 显示了结果，并按百分比降序排列所有地区的表格表示。

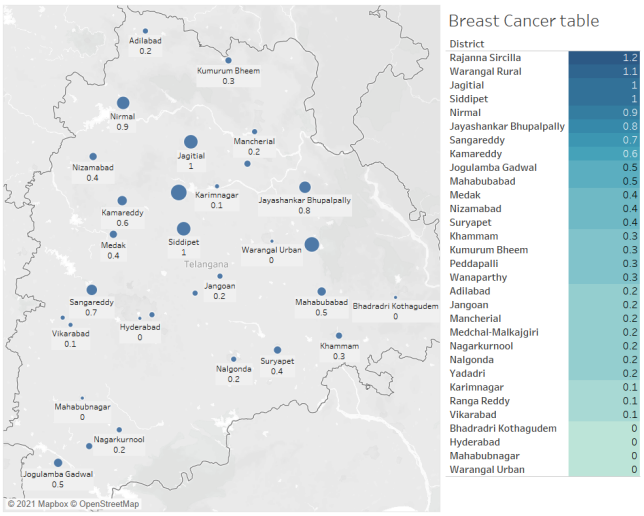

图 2. 泰伦甘纳各地区女性乳腺癌检查测试的百分比分布

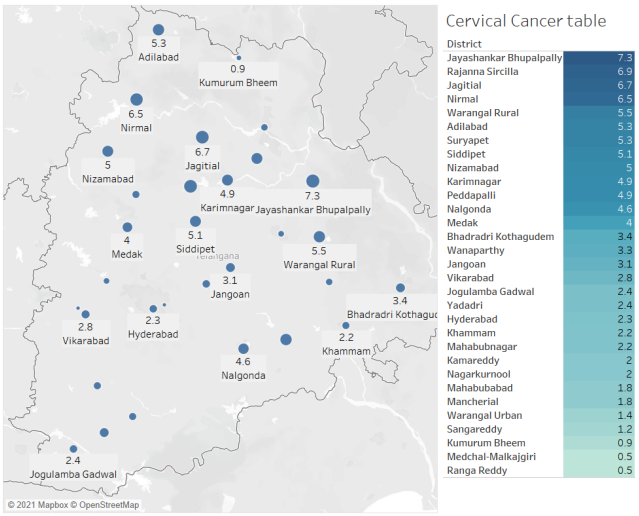

图 3. 筛查宫颈癌的女性在 Telangana 各地区的百分比分布

根据报告“癌症及相关因素概况 - 特伦甘纳邦，2021”[2]，女性癌症的五大发病部位是乳腺（35.5%）、子宫颈（8.7%）、卵巢（6.9%）、子宫体（5.5%）和肺部（4.1%）。在 0 至 74 岁的年龄段中，女性患癌的累积风险为每 7 人中有 1 人。（累积风险是在不存在任何竞争性死因且假设当前趋势持续的情况下，个体被诊断出癌症的概率）。

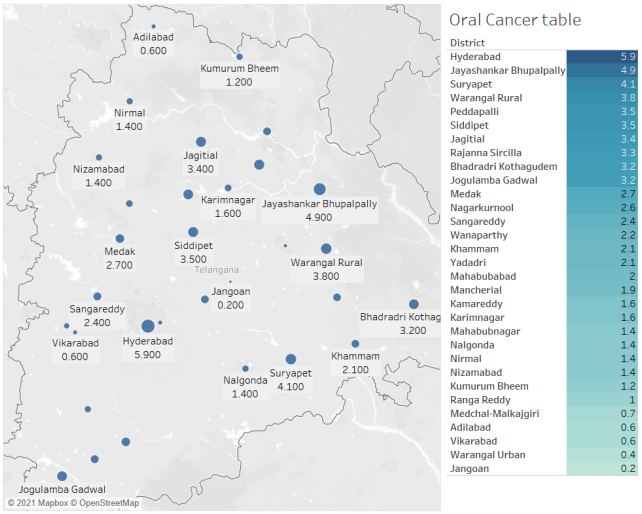

图 4. 按区划分的泰伦甘纳州女性口腔癌检查的百分比分布

在进一步研究女性选定解剖部位癌症的临床疾病范围时，考虑到所有病例，发现了图 5 中所示的比例。

	Localized Only (%)	Loco Regional (%)	Distant Metastasis (%)	Unknown Extent (%)
Breast	32	60	7	1
Cervix Uteri	71	29	-	-
Lung	30	47	21	2
Stomach	30	50	19	1
Head & Neck	44	51	4	1

图 5. 女性在特兰甘纳的解剖部位病例总结

临床疾病范围在首次出现时的百分比（%）对于选定解剖部位的癌症的一个限制是，它可能仅基于州内的 HBCR 计算得出，并不能代表整个州的情况。

预计该州在 2020 年和 2025 年的癌症病例发生率是根据性别使用 2012-2016 年的发病率数据作为参考进行计算的。对于女性而言，发现 2020 年为 25434 例，2025 年为 28708 例。

从这两项研究/报告中，我们可以得出结论，尽管存在高风险因素，但仍有一小部分女性接受了筛查测试。造成这种情况的可能原因包括对现有风险严重性的不知情、医疗设施的缺乏特别是农村地区以及经济限制。迫切需要开展国家级和全国范围内的意识提升项目，涉及社会和卫生系统的多方利益相关者，以提高印度的癌症素养。如果人们更加了解癌症的症状，并且在发现症状时尽快寻求帮助，每年可以挽救成千上万的生命。早期治疗通常更有效。

特伦甘纳已经开展的一些意识提升项目如下：

一次免费的癌症筛查和培训营在阿波罗医院举办，作为正在进行中的国际癌症会议的一部分。作为该培训营的一部分，超过 60 名 DWCRA 女性接受了如自我乳房检查等程序的培训，以便早期发现癌症。据一位肿瘤科医生 Nalini 博士介绍，她是其中一名培训师，该项目旨在向全州传播一个信息：如果能够早期检测到，癌症是一种可治愈的疾病。

Omega Bannu Hospitals 的董事总经理 Charanjith Reddy Veeramalla 博士表示，乳腺癌和宫颈癌在印度女性中最为常见，而且近年来病例数量正在增加。然而，如果通过筛查测试在早期阶段发现这些疾病是可以预防的，并且需要为妇女进行更多的筛查测试以防止乳腺癌和宫颈癌的发生。2021 年 8 月，在 Warangal 地区法律服务局（DLSA）与 Omega Bannu Hospitals 的合作下，在 DLS 大楼组织了一次免费癌症筛查测试营，旨在惠及女性司法官员、工作人员及律师。共有 80 名妇女在此次营地接受了测试。营地进行了诸如乳腺癌的乳房 X 光摄影检查（针对 40 岁以上的女性）、宫颈癌的巴氏涂片检查（针对 18 岁以上的女性）、通过血液检测进行的贫血筛查和心电图等测试。

II. 模型函数

我们模型的主要目标是根据患者对乳腺癌和宫颈癌的易感性对其进行分类。癌症在某种程度上取决于环境因素、人口统计学因素以及遗传倾向。因此，早期筛查那些有患癌风险的人可以大大减少所需治疗、恢复时间，并提供更好的生存机会。本项目开发的模型学习了这些因素与患癌风险之间的关系。它将个人分为两类：易患乳腺癌/宫颈癌和不易患这种癌症。易感个体随后可以通过临床乳房检查、乳腺 X 线摄影和超声波来早期发现癌症。

III. 数据

A. 宫颈癌

数据集来源于 Kaggle[3]，在此表示感谢。数据是在委内瑞拉加拉加斯的'Hospital Universitario de Caracas' 收集的。该数据集包括 858 名患者的个人信息、生活习惯和过往医疗记录。由于隐私问题，一些患者选择不回答某些问题（缺失值）。此数据集描述了可能导致宫颈活检的宫颈癌风险因素。该数据集由 36 列和 858 行组成。它是一个多变量数据集，包含实数和整数值。

B. 乳腺癌

数据集来源于 BCSC Research[4]，对此表示感谢。该数据集最初由乳腺癌监测联盟在 1996-2002 年间收集，指定了患乳腺癌的风险因素。此数据记录包含 16 列和 4,62,563 行。这是一个含有实数和整数值的多元数据集。

IV. 预处理和特征

A. 宫颈癌

数据集中提供的所有数值数据在使用前都进行了处理。数据集中有许多缺失值和 22 个重复值，这些都被完全移除以避免歧义。一些包含超出的空值且无相关性的列被从数据集中完全排除。所有的数值都通过 python 中的 astype() 函数转换为浮点类型。清理完数据集后，保留了数据集中没有空值和重复值的患者。经过处理后，获得了一个包含 688 名患者的数据库，并随后用于模型训练。数据集按 3:1 的比例（分别为 75%和 25%）划分为训练和验证两部分。年龄、性伴侣数量、首次性行为时间、怀孕次数、吸烟年限、激素避孕使用年限、宫内节育器使用年限、HPV 诊断等是建模的主要特征。独立参数包括处理后的数据集中除 “Dx: Cancer” 列外的所有列。依赖参数是 “Dx: Cancer” 这一列。为了避免方差，对独立特征进行了拟合。

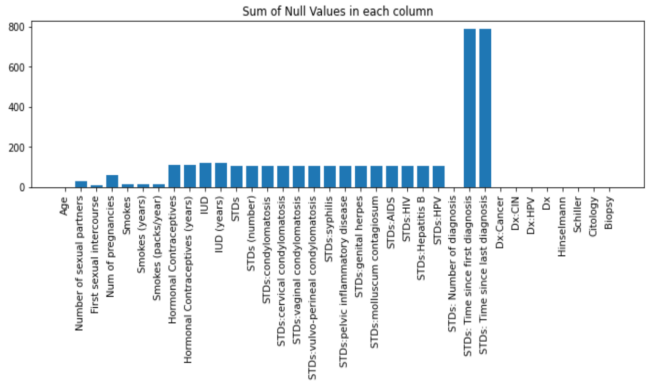

图 6. 数据集中每个特征的空值数量

B. 乳腺癌

如上所述收集的数据在用于模型拟合之前进行了清理和处理。数据集包含一个 “训练” 列，该列将数据

```
Index(['Age', 'Number of sexual partners', 'First sexual intercourse',
       'Num of pregnancies', 'Smokes', 'Smokes (years)', 'Smokes (packs/year)',
       'Hormonal Contraceptives', 'Hormonal Contraceptives (years)', 'IUD',
       'IUD (years)', 'STDs', 'STDs (number)', 'STDs:condylomatosis',
       'STDs:vaginal condylomatosis', 'STDs:vulvo-perineal condylomatosis',
       'STDs:syphilis', 'STDs:pelvic inflammatory disease',
       'STDs:genital herpes', 'STDs:molluscum contagiosum', 'STDs:HIV',
       'STDs:Hepatitis B', 'STDs:HPV', 'STDs: Number of diagnosis',
       'Dx:Cancer', 'Dx:CIN', 'Dx:HPV', 'Dx', 'Hinselmann', 'Schiller',
       'Citology', 'Biopsy'],
      dtype='object')
```

图 7. 用于训练宫颈癌 ML 分类模型的特征

分为验证和训练类别，已被移除。此外，识别了数据集中缺失的值，并删除了含有缺失值的行。还移除了重复样本（14655）。

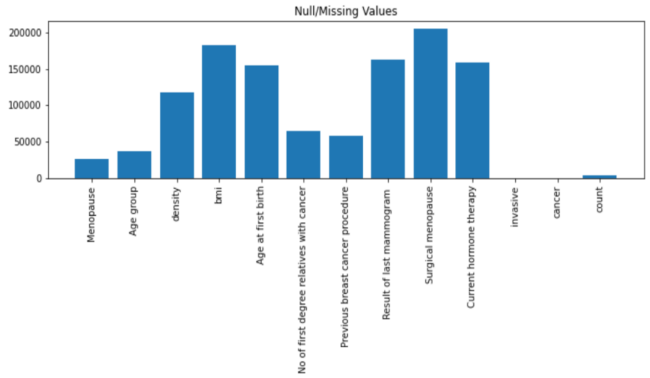

图 8. 数据集中每个特征中缺失值的数量

清洗后，剩下 15203 行数据。在训练前，数据通过 StandardScaler() 进行了缩放以达到单位方差的标准化处理。数据集按照 1:1 的比例被划分为训练集和测试集（分别为 50%和 50%）。

用于建模的主要特征是年龄组、绝经状态、侵袭性肿瘤、基础代谢指数、初次生育年龄、一级亲属中有癌症的数目、当前激素治疗、BI-RADS 密度、之前的乳腺癌手术、乳房 X 线照片的结果、外科绝经和当前激素治疗。需要预测的目标特征是癌症。

V. 探索性数据分析

A. 宫颈癌

加载数据集后，我们首先查看其维度，为 (650, 32)。查看数据集的信息以了解其特征、特征的数据类型等。之后，我们预处理和清洗数据。对特征的统计摘要有助于检查特征分布和异常值（如果有）。

- 年龄列的最大值为 84，但其他列的最大值要低得多，这可能会导致模型表现不佳，因为年龄列的影响比其他列更大。为了避免训练模型时影响的不同，标准化所有列的值。年龄列的最大值为 84，但其他列的最大值要低得多，这可能会导致模型表现不佳，因为年龄列的影响比其他列更大。为了避免训练模型时影响的不同，标准化所有列的值。
- ’怀孕次数’列中的最大值为 11，这是一个非常高的怀孕次数，有可能是影响该列中所有其他值的异常值。解决方案可能是删除这些行，但只有在我们的表现不佳时才会这样做。
- 列 ’STDs: 宫颈乳头状瘤’ 和 ’STDs:AIDS’ 只包含零，因此是无用的。去除它们就是解决方案。
- 平均值为 0.0255，因此’Dx:Cancer’列（这将是我们的因变量或预测变量）非常不平衡。如果类别是平衡的，平均值会接近 0.5。为了更好地理解这一点，我们将用一个图来展示。解决这个问题极其困难；最好的解决方案是获取更多正面数据来训练我们的模型，但这在我们的情况下是不可能的；另一个解决方案可能是删除一些负面案例以与正面案例达到平衡，但这将导致大量信息丢失。

在我们能够标准化数据之前，我们需要知道是否存在提供相同（或非常相似）信息的列，这可能会导致我们的模型表现不佳。这些信息可以通过创建相关矩阵来获得。然后我们将几个列与 Dx:Cancer 列分别进行绘制，以检查每个特征对癌症的影响。

B. 乳腺癌

加载数据集后，我们查看其维度。查看数据集的信息以了解其特征和数据类型等信息。

此后，我们对数据进行预处理和清洗。删除无用的列以及包含缺失值的行。特征的统计摘要有助于检查特征分布及是否存在异常。

下面的总结给了我们大量关于数据的宝贵信息。

- ’count’列中的最大值为 1128，远高于其他列的最大值，这可能导致模型性能不佳，因为这一列的影响比其他列更大。为了避免在训练模型时影响的不同，我们对所有列的值进行标准化。
- 平均值为 0.043，说明’癌症’列（即我们的因变量）非常不平衡。如果类别是平衡的，平均值应该是 0.5。为了更好地理解这一点，我们将用一个图表来展示。解决这个问题非常困难——最好的解决方案是获取更多正例数据来训练我们的模型，但在我们的情况下这是不可能的。另一个解决方案可能是移

```
Data columns (total 36 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Age                                 858 non-null    int64
 1   Number of sexual partners           832 non-null    object
 2   First sexual intercourse            851 non-null    object
 3   Num of pregnancies                  802 non-null    object
 4   Smokes                              845 non-null    object
 5   Smokes (years)                      845 non-null    object
 6   Smokes (packs/year)                 845 non-null    object
 7   Hormonal Contraceptives             750 non-null    object
 8   Hormonal Contraceptives (years)     750 non-null    object
 9   IUD                                 741 non-null    object
 10  IUD (years)                         741 non-null    object
 11  STDs                                753 non-null    object
 12  STDs (number)                       753 non-null    object
 13  STDs:condylomatosis                 753 non-null    object
 14  STDs:cervical condylomatosis        753 non-null    object
 15  STDs:vaginal condylomatosis         753 non-null    object
 16  STDs:vulvo-perineal condylomatosis  753 non-null    object
 17  STDs:syphilis                       753 non-null    object
 18  STDs:pelvic inflammatory disease    753 non-null    object
 19  STDs:genital herpes                 753 non-null    object
 20  STDs:molluscum contagiosum          753 non-null    object
 21  STDs:AIDS                           753 non-null    object
 22  STDs:HIV                            753 non-null    object
 23  STDs:Hepatitis B                    753 non-null    object
 24  STDs:HPV                            753 non-null    object
 25  STDs: Number of diagnosis           858 non-null    int64
 26  STDs: Time since first diagnosis    71 non-null     object
 27  STDs: Time since last diagnosis     71 non-null     object
 28  Dx:Cancer                           858 non-null    int64
 29  Dx:CIN                              858 non-null    int64
 30  Dx:HPV                              858 non-null    int64
 31  Dx                                  858 non-null    int64
 32  Hinselmann                          858 non-null    int64
 33  Schiller                            858 non-null    int64
 34  Citology                            858 non-null    int64
 35  Biopsy                              858 non-null    int64
dtypes: int64(10), object(26)
memory usage: 241.4+ KB
```

图 9. 数据集信息

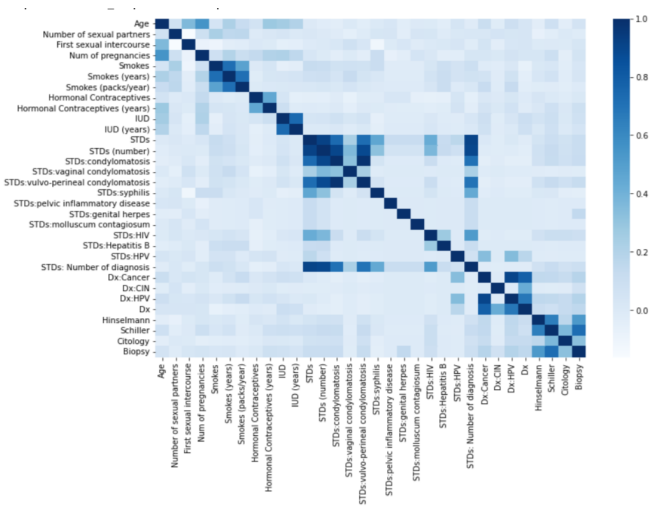

图 10. 显示所有数据集特征之间相关性的热图

除一些负例以与正例达到平衡，但这会导致大量信息的丢失。

在我们能够标准化数据之前，我们需要知道是否存在提供相同（或非常相似）信息的列，这可能会导致我们的模型表现不佳。这些信息可以通过创建相关矩阵来获得。然后我们将一些列单独与癌症列进行比较绘制图表，以检查每个特征对癌症的影响。

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 15203 entries, 29 to 181193
Data columns (total 13 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Menopause                             15203 non-null  float64
 1   Age group                             15203 non-null  float64
 2   density                               15203 non-null  float64
 3   bmi                                   15203 non-null  float64
 4   Age at first birth                    15203 non-null  float64
 5   No of first degree relatives with cancer  15203 non-null  float64
 6   Previous breast cancer procedure      15203 non-null  float64
 7   Result of last mammogram              15203 non-null  float64
 8   Surgical menopause                    15203 non-null  float64
 9   Current hormone therapy               15203 non-null  float64
 10  invasive                              15203 non-null  int64
 11  cancer                                15203 non-null  int64
 12  count                                 15203 non-null  float64
dtypes: float64(11), int64(2)
memory usage: 1.6 MB
```

图 11. 用于训练的特征的基本信息

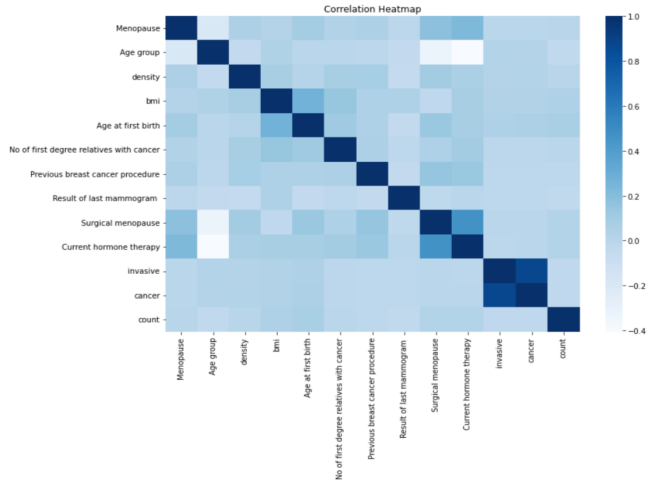

图 12. 显示所有数据集特征之间相关性的热图

VI. 分类模型

A. 支持向量分类

支持向量机（SVM）是一种监督机器学习模型，它使用分类算法解决两组分类问题。与诸如神经网络等较新算法相比，在样本数量有限（数千个样本）的情况下，它们在速度和性能方面具有两大优势。由于在高维空间中表现良好，SVM 特别适合此任务。从适用的分类问题类型来看，SVM 非常灵活，因为用户可以指定将在模型中充当决策函数的核函数。尽管有效且灵活，但当特征数量超过样本数量时，SVM 的效果会变差。由于我们的两个分类问题（宫颈癌数据集中有 36 个标签

和 688 个训练样本，在乳腺癌数据集中有 13 个标签和 15203 个训练样本）并非如此，该小组开始使用支持向量分类进行归类。

B. 随机梯度下降

随机梯度下降（SGD）是一种简单而有效的拟合线性分类器和回归器的方法，适用于凸损失函数，如（线性）支持向量机和逻辑回归。SGD 已成功应用于大规模和稀疏的机器学习问题。由于数据是稀疏的，因此该模块中的分类器可以轻松扩展到包含超过 105 个训练样本和超过 105 个特征的问题。严格来说，SGD 是一种优化技术，并不对应于特定类型的机器学习模型。它仅仅是一种训练模型的方法。对于这个问题而言，SGDs 是有用的，因为它们非常高效且易于实现。

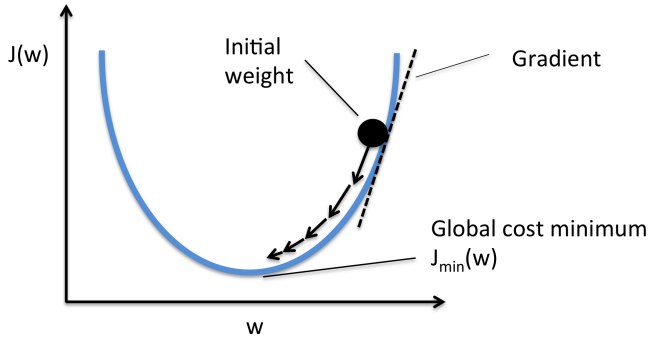

图 13. 梯度下降沿成本函数的示例

SGDClassifier 通过在 OVA（“一对一”）方案中结合多个二元分类器来实现多类分类。为每个 k 类学习一个能够区分该类与其他所有类的二元分类器。我们在测试时计算每个分类器的信心得分（即到超平面的符号距离），并选择信心最高的类别。

C. 决策树分类器

分类技术是从一组输入数据构建分类模型的系统方法。决策树分类器、基于规则的分类器、神经网络、支持向量机和朴素贝叶斯分类器等是解决分类问题的不同技术。每种技术采用一种学习算法来识别最适合描述输入数据属性集与类别标签之间关系的模型。因此，学习算法的主要目标是创建一个能够准确预测先前未知记录类别的预测模型。

决策树分类器是一种简单且广泛使用的分类技术。它采用一个直接的想法来解决分类问题。决策树分类器对测试记录的属性提出一系列精心设计的问题。每次收到答案后，都会继续提问直到得出关于记录类别标签的结论。

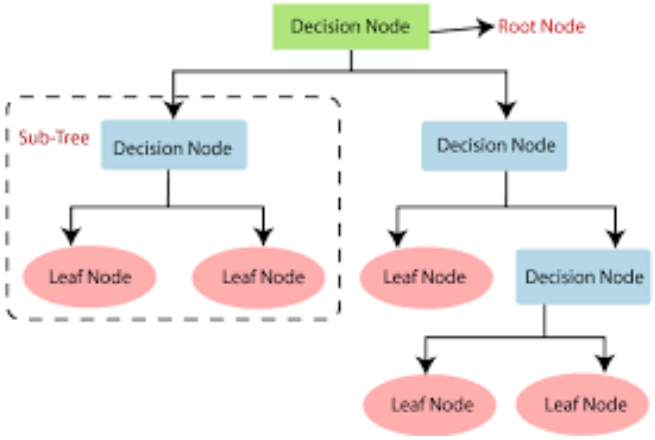

图 14. 决策树示例

D. 随机森林分类器

随机森林是一种监督学习算法。构建的森林由多个决策树组成，这些决策树通常使用“装袋”方法进行训练。装袋方法的基本思想是，学习模型的组合会提高总体结果。随机森林中的每棵树都会生成一个类别预测，得票最多的类别成为模型的预测。

随机森林在树生长过程中为模型增加了额外的随机性。在分裂一个节点时，并不是寻找最重要的特征，而是从特征的随机子集中寻找最佳特征。这导致了极大的多样性，通常会得到更好的模型。

VII. 结果

我们对分类模型的最终结果非常满意。提到的所有四个分类模型都产生了最大的训练和测试准确率。用于训练宫颈癌数据集的模型是随机梯度下降、支持向量机分类和支持决策树分类。同样，用于训练乳腺癌数据集的模型是支持向量机分类、决策树分类和随机森林分类。最终，我们为每个数据集选择了最佳分类器，其详细情况将在下文进一步解释。

A. 宫颈癌

在上述三个分类模型中，决策树分类算法产生了最好的结果。测试数据（25%）的准确率为 99.39%，而训练数据（75%）的准确率为 100%。由决策树算法生成的结果的混淆矩阵和分类报告如图 16 所示。

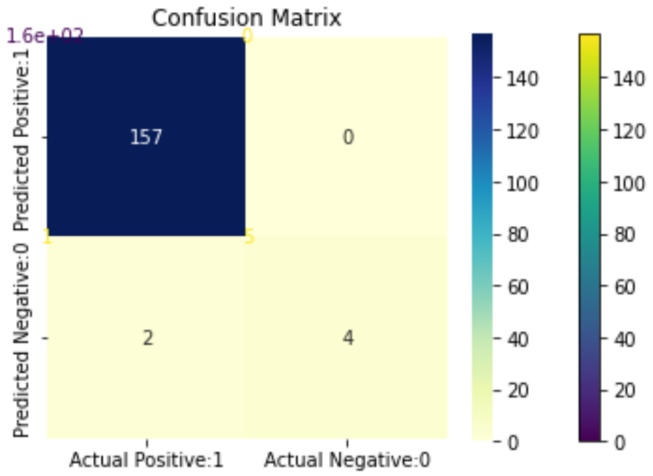

图 15. 使用决策树分类算法训练模型后获得的混淆矩阵

	precision	recall	f1-score	support
0.0	0.99	1.00	1.00	157
1.0	1.00	0.83	0.91	6
accuracy			0.99	163
macro avg	1.00	0.92	0.95	163
weighted avg	0.99	0.99	0.99	163

图 16. 使用决策树分类算法训练模型后获得的分类报告

B. 乳腺癌

三个分类模型中，支持向量分类产生了最佳结果。其测试数据（50%）的准确率为 98.89%，而训练数据（50%）的准确率为 99.04%。所获得的混淆矩阵、分类报告和特征重要性如图 18 所示。

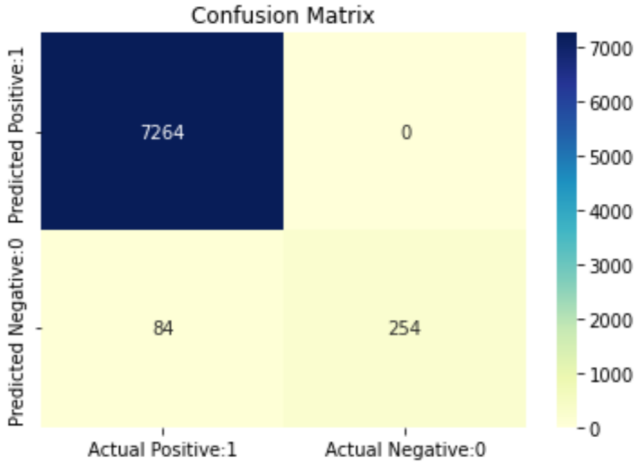

图 17. 训练模型后获得的混淆矩阵

	precision	recall	f1-score	support
0	0.99	1.00	0.99	7264
1	1.00	0.75	0.86	338
accuracy			0.99	7602
macro avg	0.99	0.88	0.93	7602
weighted avg	0.99	0.99	0.99	7602

图 18. 模型训练后获得的分类报告

VIII. 继续工作

本文提出了一种分类模型，可用于检查一个人患乳腺癌和宫颈癌的易感性。该模型可以作为一种开源应用程序的有效工具，供所有人使用。此外，该应用还将包括其他方法，以支持我们提高意识并强调其重要性的努力。

A. 最近的医院/中心建议

该应用程序将由我们的系统提供支持，该系统会根据用户的位置或地址为用户提供最近的医院或癌症治疗中心的建议。该系统集成了 2 个 API，有助于找到最佳建议。第一个来自 Position Stack 的 API 帮助查找用户的纬度和经度，第二个来自 MapMyIndia 的 API 帮助查找最近的医院或癌症治疗中心。这些建议将确保每个人在需要时也知道正确的联系来源。

B. 意识驱动和活动

癌症意识宣传活动在癌症预防计划中至关重要。这些活动的目的是提高特伦甘纳邦居民对癌症的认识。重要的是要破除人们的错误观念，告知他们关于癌症的征兆和症状以及早期检测的重要性。此外，了解癌症风险因素是这一过程中的决定性要素。

C. 与健康卡集成

2021 年 9 月 27 日，总理纳伦德拉 · 莫迪推出了数字健康身份证，该证件将提供给所有人。这将创建一个无缝的在线平台，使所有与健康相关的信息便携且易于医生访问。这可以用于与我们的系统集成以实现便捷访问。每个人健康记录将被维护，并使用 k 均值聚类算法根据位置和人口统计数据来识别活动和驱动的需求，如图 19 所示。

IX. 结论

通过对当前癌症筛查统计数据的分析，发现女性急需提高癌症知识水平。为此需要经常举办意识提升活动

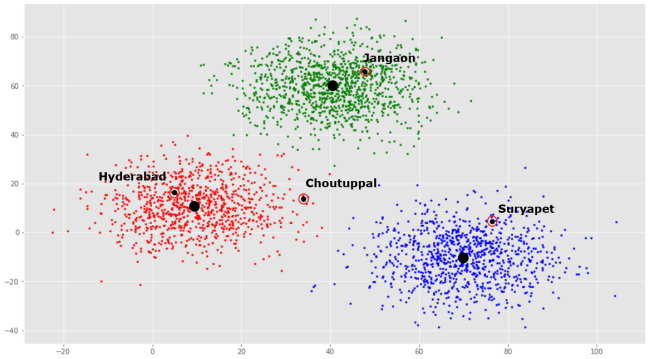

图 19. k 均值聚类算法确定活动位置的示例

和培训项目。我们设计了一个机器学习分类模型，根据人口统计因素预测一个人是否容易患乳腺癌或宫颈癌。然后触发一个建议系统，指引用户前往最近的医院。由于这个集成了易感性计算和医院推荐功能的系统是开源的，人们可以轻松检查自己的风险，并在必要时进行癌症筛查测试。为进一步推进这一目标，我们可以将其与健康卡集成，这将使患者数据更加可访问（基于患者的隐私设置），有助于根据实时统计数据制定新政策，从而提高女性的癌症知识水平。此外，通过分析各区的癌症病例和死亡率情况，可以组织新的本地化方案、意识提升活动、培训营以及财政支持政策。这些措施可以根据该地区在 Telangana 州内的癌症知识水平和城市化的程度进行定制。

参考文献

[1] NHFS dataset

[2] https://ncdirindia.org/All_Reports/State_Factsheet __21/Fact-sheet/FS_Telangana.pdf

[3] https://www.kaggle.com/franxi/predicting-cervical-cancer/data

[4] https://www.bcsc-research.org/data/rfdataset/dataset