埃塞俄比亚育龄妇女破伤风类毒素免疫的数据挖掘

Kedir Hussein Abegaz[1,*], Emiru Merdassa Atomssa2
1 埃塞俄比亚 巴勒戈巴 马德达瓦拉布大学健康科学学院公共卫生系生物统计学和健康信息学
2 埃塞俄比亚 奥罗米亚州 吉姆比西沃勒加地区卫生部生物统计学和卫生信息学

摘 要：破伤风类毒素（TT）疫苗用于育龄妇女，以预防新生儿破伤风和因破伤风导致的孕产妇死亡。在全球范围内，破伤风每年造成 5%的产妇死亡和 14%的新生儿死亡。数据挖掘是从大量数据中发现有趣模式和知识的过程。因此，本研究的目的是识别最佳分类器，并使用数据挖掘算法从 TT 数据集预测模式。本研究的数据是 2011 年埃塞俄比亚人口与健康调查（EDHS）中的破伤风类毒素数据集，并使用选择、处理、转化、挖掘和解释的知识发现过程进行分析。WEKA 3.6.1 工具用于分类、聚类、关联和属性选择。分类器在训练数据上的准确率相对高于测试数据，多层感知器是我们的破伤风类毒素数据集中最好的分类器。在具有 10 倍的交叉验证中，正确分类的最好的是通过天真贝叶斯 63.30%，最不准确的是通过 k 近邻 60.52%。使用天真贝叶斯的单个数据实例测试是通过创建测试 1、测试 2、测试 3 和测试 4 数据测试实例来完成的，其中三个数据实例预测正确，但其中一个错误分类。在一般关联中获得的最大置信度为 0.98。但是，在 class 属性中，它是 0.72。母亲的识字状况具有较高的信息增益，值为 0.046。因此，基于 TT 疫苗接种数据的最佳算法是多层感知器分类器，其准确率为 67.28%，构建模型所需的总时间为 0.01 秒。与其他分类器相比，多层感知器分类器的平均误差最低，为 32.72%。这些结果表明，在测试的机器学习算法中，多层感知器分类器有可能显著改进用于破伤风类毒素 EDHS 数据的传统分类方法。

关键词：数据挖掘；韦卡；分类聚类；破伤风类毒素（TT）；电子海图

Data Mining of Access to Tetanus Toxoid Immunization Among Women of Childbearing Age in Ethiopia

Kedir Hussein Abegaz[1,*], Emiru Merdassa Atomssa[2]

1Biostatistics and Health Informatics, Public Health Department, College of Health Sciences, Madda Walabu University, Bale Goba, Ethiopia

2Biostatistics and Health Informatics, West Wollega Zonal Health Department, Gimbi, Oromia, Ethiopia

Abstract: Tetanus toxoid (TT) vaccine is given to women of childbearing age to prevent neonatal tetanus and maternal mortality attributed to tetanus. Globally, tetanus is responsible for 5% of maternal deaths and 14% of neonatal deaths annually. Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. Thus, the aim of this study was to identify the best classifier, and to predict the pattern from the TT data set using the data mining algorithms technique. The data for this study were the Tetanus Toxoid data set from the Ethiopian Demographic and Health Survey (EDHS) 2011, and analyzed using the Knowledge discovery process of Selection, Processing, Transforming, mining, and interpretation. The WEKA 3.6.1 tool was used for classification, clustering, association and attribute selection. The accuracy rate of the classifiers on training data is relatively higher than on test data and the multilayer perceptron is the best classifier in our data set on Tetanus toxoid. In the cross-validation with 10 folds, correctly classified best are by naïve Bayesian 63.30% and the least accurate were by k-nearest neighbor 60.52%. Single data instance test using Naïve Bayesian was done by creating test 1, test 2, test 3, and test 4 data test instance, three of them are correctly predicted but one of them incorrectly classified. The maximum confidence attained in the general association is 0.98. But, in the class attribute, it is 0.72. The literacy status of the mother has high information gain with the value 0.046. As a conclusion, the best algorithm based on the TT vaccination data is multilayer perceptron classifier with an accuracy of 67.28% and the total time taken to build the model is at 0.01 seconds. Multilayer perceptron classifier has the lowest average error at 32.72% compared to others. These results suggest that among the machine learning algorithm tested, multilayer perceptron classifier has the potential to significantly improve the conventional classification methods for use in EDHS data of Tetanus toxoid.

Keywords: Data mining; WEKA; Classification; Clustering; Tetanus toxoid (TT); EDHS

1.引言

为育龄妇女接种破伤风类毒素（TT）疫苗，以预防新生儿破伤风和因破伤风导致的孕产妇死亡。在许多发展中国家，婴儿早期死亡的主要原因往往是由于分娩期间没有遵守卫生程序。因此，妇女接受一定剂量的破伤风类毒素，以防止新生儿破伤风[1]。破伤风是由破伤风梭菌厌氧生长过程中产生的毒素引起的。感染是通过暴露在任何破损的皮肤或死亡组织（如伤口）的环境中，或当脐带被细菌的孢子切割时获得的。世界卫生组织估计，即使来自发达国家，也只报告了5%的新生儿破伤风病例
监控系统[2]。

在全球范围内，破伤风每年造成 5%的产妇死亡和 14%的新生儿死亡，在一些非洲国家新生儿死亡的比例高达 25%[3-5]。截至 2012 年 12 月，孕产妇和新生儿破伤风仍然是 30 个国家的公共卫生问题，主要在非洲和亚洲[2]。在撒哈拉以南非洲，估计每年有多达 70000 名新生儿死于新生儿破伤风[5]。埃塞俄比亚是世界上新生儿破伤风发病率和死亡率最高的国家之一，原因是破伤风类毒素免疫覆盖率低，加上约 90%的分娩是在不卫生的条件下进行的。1999 年，世界卫生组织估计埃塞俄比亚约有 17875 例新生儿破伤风病例和 13406 例非传染性支气管炎死亡病例，使该国占全球非传染性支原体死亡的 4.6%[3]。

埃塞俄比亚的扩大免疫计划（EPI）始于 1980 年，至今仍是卫生部支持的初级卫生保健中最重要的组成部分。到 2011 年底，作为 100 多个国家常规免疫计划的一部分，推出了预防母婴破伤风（MNT）的疫苗。2011 年，至少接种两剂破伤风类毒素疫苗的接种率估计为 70%，估计 82%的新生儿通过免疫接种预防新生儿破伤风[3]。然而，到目前为止，孕产妇和新生儿破伤风仍然是 36 个国家的公共卫生问题，主要是在非洲和亚洲。

埃塞俄比亚的育龄妇女 TT 疫苗接种计划遵循世界卫生组织为发展中国家建议的计划[6]。在分娩前为母亲接种 TT 可以保护母亲和新生儿免受破伤风的侵害，产前护理是常规 TT 免疫的主要方案切入点。孕妇在怀孕期间应至少接种两剂，除非她已经从之前的 TT 疫苗中获得免疫力。五剂 TT 可以确保在整个生育期甚至更长时间内得到保护。

数据挖掘是从大量数据中发现有趣模式和知识的过程。这是一个年轻且快速发展的领域，也称为数据知识发现（KDD），用于从各种应用程序中的数据中发现有趣的模式[7]。

医疗保健行业是世界上规模最大、增长最快的行业之一，拥有大量的医疗保健数据。该健康护理数据包括有关客户、他们的治疗和资源管理数据的相关信息。信息丰富而海量。通过数据挖掘技术的应用，可以发现医疗数据中隐藏的关系和趋势。数据挖掘技术比医疗研究中使用的更有效。在本研究中，我们使用了几种数据挖掘技术；EDHS 11 指定破伤风类毒素免疫数据集的分类、聚类、关联和异常值检测技术。

该项目的主要目标是确定最佳分类器，并使用破伤风类毒素疫苗接种的数据挖掘算法和工具从 TT 数据集预测模式，并将技术领域公共卫生和医疗领域连接起来，为社区服务。选择这项研究的理由是，在埃塞俄比亚，卫生工作者没有将累积的医疗数据用于预测目的。这一问题导致了医疗系统环境中的时间和精力的损失，并且在没有基于证据的信息进行规划和干预的情况下花费了大量的精力和成本。

2.知识发现流程

在这项研究中，我们使用了在 TT 数据集上测试过的不同数据挖掘技术。所使用的标准是所使用的每种分类技术的准确率和错误率的百分比。基于最高的分类准确率和较少的错误率来选择适合于特定数据集的技术。

应用数据挖掘，用于从 EDHS 2011 数据集中发现隐藏但有用的知识。这个过程必须有一个模型来控制它的执行步骤。根据标准流程，从数据中发现知识，预测埃塞俄比亚育龄妇女破伤风类毒素免疫接种情况，指导我们分析过程，并暴露出否则可能被忽视的方面。图 1（改编自[7]）显示了我们从数据中发现知识的基本阶段。选择阶段从 EDHS 2011 的整个数据集生成目标数据集。预处理解决了有关噪声、不完整和不一致数据的问题。下一阶段是将预处理的数据转换为适合执行所需数据挖掘任务的形式。在数据挖掘阶段，运行一个过程来执行所需的任务并生成一组模式。

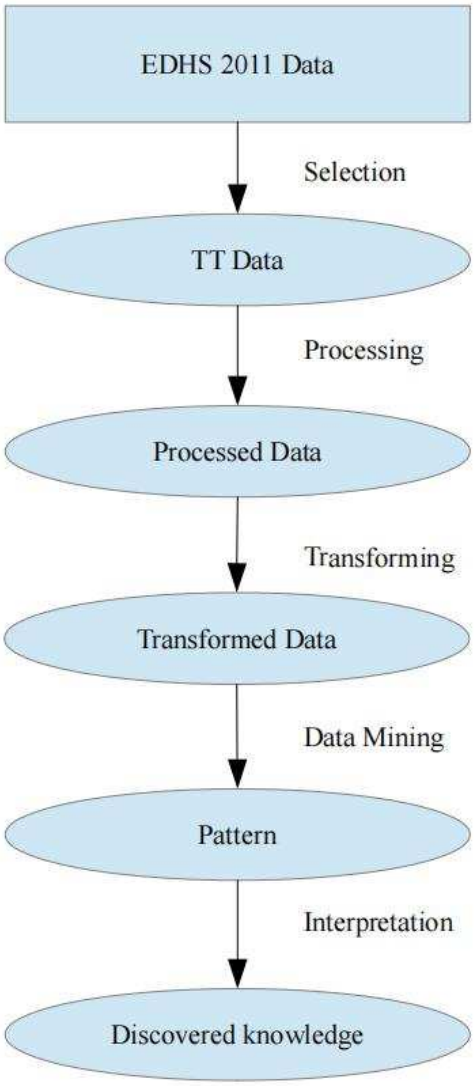

图 1.EDHS 11 的 TT 数据的 KDD 过程。

然而，并非所有模式都有用。解释和评估所有发现的模式的目的是只保留那些对用户感兴趣和有用的模式，而丢弃其余的模式。剩下的模式代表发现的知识。

3.方法

本研究的方法是应用于埃塞俄比亚 DHS 2011 年破伤风类毒素数据的实用研究方法。

3.1.数据理解

2011 年 EDHS 由中央统计局（CSA）在埃塞俄比亚卫生部（EMoH）的主持下，与 DHS 措施和 ICF 国际合作进行。对于这项特定的研究，数据集是从 DHS 网站请求和访问的 https://dhsprogram.com 在正式在线注册并提交项目名称和详细项目说明后。

3.2.数据预处理

2011 年的 EDHS 数据集被用作本研究的来源，WEKA 3.6.1 机器学习工具被使用。在这个工具中，我们应用了不同的分类算法，聚类并预测了一个有用的结果，这将对健康护理规划人员、新用户和新研究人员非常有用。本研究中使用的数据是 TT 免疫数据。它的维度为 7033 行和 12 列。这些数据是为了培训和测试而处理和安排的。只有 80%的总体数据用于训练，其余 20%用于测试所选分类方法的分类准确性。

对数据值和属性进行修改、添加和/或删除、过滤、记录、删除缺失值、转换和整合，以便机器学习技术在研究分析步骤中使用。最后，数据以“.csv”文件格式保存，并以“.arff”文件格式存储。

3.3.分类

分类是数据挖掘技术之一，用于对属于同一类的实例进行分组[8]。分类还提取描述重要数据类的模型。这类模型称为分类器，预测分类类标签（名词性、序数性）。该分类有许多应用，包括欺诈检测、目标营销、绩效预测、制造业和医疗诊断。这种分类是如何工作的？数据分类有两个步骤：；首先，由学习步骤组成，即构建分类模型。以及分类步骤，即模型用于预测给定数据的类标签[7]。本研究中使用的分类方法是根据数据的类别对数据进行分类，将数据放在属于同一类的单个组中。方法如下：

The attributes and their name in the analysis	The categories	Th
Place of Residence as "Residence"	Urban	13
	Rural	56
Access ro radio as "Radio"	Yes	27
	No	42
Access to Television as "Television"	Yes	83
	No	62
Mother's religion as "Religion"	Orthodox Christian	24
	Muslim	30
	Protestant	13
	Catholic	65
	Others	15
Mother's Ethnic group as "Ethnicity"	Oromo	22
	Amhara	14
	Tigrian	82
	Others	24
Literacy status of mothers as "Literacy_status"	Unable to read	54
	Able to read	15
Distance to health facility as "Distance_to_HF"	A big problem	50
	Not a big problem	20
Level of husband's education as "hus_education"	No education	36
	1° school	25
	2° and above	82
Women's age in category as "Women_age"	15-24	17
	25-34	34
	35-49	18
Marital status of the mothers as "Marital_status"	Single	13
	Married	64
	Widowed	14
	Divorced	29
Head of the household as "hh_head"	Male	57
	Female	13
Vaccinated with TT as "is_vaccinated" (The target attribute for this study)	yes	33
	No	36

表 1.属性列表及其在 EDHS 2011 中的命名。

（a） 决策树（J48）方法

它是一个类似于树结构的流程图。其中每个分支表示测试的结果，每个内部节点表示对属性的测试，每个叶节点持

有一个类标签。树中最顶端的节点是根节点。该方法使用除法和征服算法将根节点分割为两个分区的子集，直到出现在树中的叶节点[8, 9]。

（b） K 近邻分类器

K-Nearest Neighbor 是最简单的分类器之一，它使用先前已知的数据点发现未识别的数据点，即最近邻居[10]。当给定大的训练集时，它是劳动密集型的，并且它已被广泛应用于模式识别领域。这种分类器方法基于类比学习，通过将给定的测试元组与与其相似的训练元组进行比较。训练元组由 m 个属性描述。每个元组表示 m 维空间中的一个点。这样，所有训练元组都存储在 m 维模式空间中。当给定未知元组时，k 近邻分类器在模式空间中搜索最接近未知元组的 k 个训练元组。这些 k 个训练元组是未知元组[7, 10]的 k 个“最近邻居”。

（c） 贝叶斯分类方法

贝叶斯分类器是基于贝叶斯定理的统计分类器，是一种概率学习方法。他们可以预测类成员概率，例如给定元组属于特定类的概率[10, 11]。贝叶斯分类器在应用于大型数据库时也表现出了较高的准确性和速度。天真贝叶斯分类器假设属性值对给定类的影响与其他属性的值无关。这种假设被称为类条件独立性。它是为了简化所涉及的计算，在这个意义上，被认为是“天真的”[7]。

（d） 多层前体

多层分类器是一种简单的两层神经网络分类器，没有隐藏层。

3.4.分级器性能评估

这是为了评估分类器在预测元组的类标签时的“准确性”。我们将考虑类元组或多或少均匀分布的情况，以及类不平衡的情况。

混淆矩阵；包括准确性、敏感性、特异性和精密度。我们需要知道另外四个术语，它们是计算许多评估指标时使用的“构建块”。理解它们将使人们容易理解各种措施的含义。

真阳性（TP）：这些是指分类器正确标记的阳性元组。TP 是真阳性数。真否定（TN）：这些是由分类器正确标记的否定元组。TN 是真阴性的数量。假阳性（FP）：这些是错误标记为阳性的阴性元组。FP 是误报的数量。假阴性（FN）：这些是被错误标记为阴性的阳性元组。FN 是假阴性的数量。

	Predicted Class of TT		Totul
Actual class of TT	Yes	No	P
Yes	TP	FN	N
No	FP	TN	P
Total	P'	N'	P+N

表 2.EDHS 2011 TT 数据的混淆矩阵。

3.5.交叉验证

在 K 折叠交叉验证中，初始数据被随机划分为 K 个互斥折叠 D1、D2…DK，每个折叠大小大致相等。培训和测试进行了 k 次。通常，分层 10 倍交叉验证用于估计精度，即使由于其相对较低的偏差和方差，计算能力允许使用更多倍。

3.6.聚类

大多数聚类算法要求用户输入他们想要的聚类数量[12]。因此，在本研究中，我们使用了五个集群。基于最大化同一类中对象之间的相似度（即类内相似度）和最小化不同类对象之间的相似性（即类间相似度）的原则，使用简单 K-Means 发现可接受的类[7]。“k-means（）算法是如何工作的？”k-means 算法将簇的质心定义为簇内点的平均值。首先，它随机选择 D 中的 k 个对象，每个对象最初表示一个簇平均值或简单的中心。对于每个剩余对象，基于对象和簇平均值之间的欧几里得距离，将一个对象分配给其最相似的簇。然后，k 均值算法迭代地提高簇内变化。对于每个集群，它使用上一次迭代中分配给集群的对象来计算新的中心。然后使用更新的中心作为新的簇中心重新分配所有对象。迭代一直持续到任务稳定为止，也就是说，当前一轮中形成的集群与上一轮中的集群相同。

3.7.协会

包含单个谓词的关联规则称为一维关联规则。这是为了识别所选属性发生的频率以及破伤风类毒素疫苗接种的机会，基于称为支持的阈值，识别频繁属性集。另一个阈值是置信度，它是使用 Apriori 算法在事务中出现属性的条件概率。

3.8.属性选择

在属性选择中，针对类属性和可视化使用了具有 Ranker T 的 InfoGainAttributeEval。

4.结果和讨论

在选定的 7037 名母亲中，3351 名母亲接受了 TT 免疫。5680 名母亲来自埃塞俄比亚农村，其中更多的母亲（3484 名）年龄在 25-34 岁之间。（表 1）

如图 2 所示，分类器对训练数据的准确率相对较高。这表明，学习机对破伤风类毒素疫苗接种数据集的准确性和性

能的结果因此是可靠的，可以用作分类器检测能力的良好指标。多层感知器是我们数据集中最好的分类器。

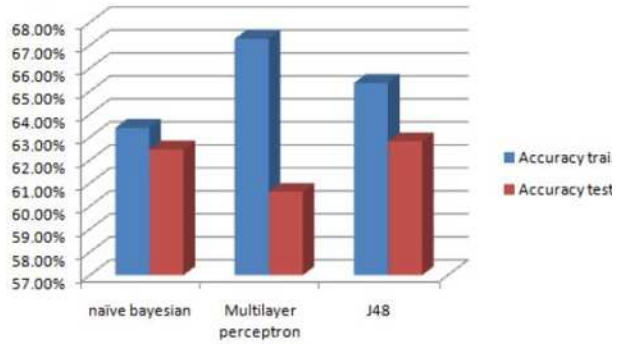

图 2.所选学习算法在训练和测试数据。

使用交叉验证的评估（10 倍），通过天真贝叶斯法正确分类的最佳值为 63.30%，最不准确的值为 K 近邻法 60.52%。（表 3）

Simple K-Means 首选此项目的聚类方法，我们通过单击 Simple K-Means 调整了聚类算法的属性。我们在这里感兴趣的算法的唯一属性是“簇数字段”，它根据给定的指令告诉我们五个簇中有多少个簇。简述如下：

第 0 组——这一群体有 1534（27%）例，其中包括居住在农村的母亲，没有收音机，没有电视，宗教信仰正统，奥罗莫族，识字状况无法阅读母亲，母亲认为离卫生设施的距离是个大问题，丈夫的教育没有教育，15-24 岁的妇女，母亲婚姻状况已婚，hh_head 男性，tt 接种了疫苗。

第 1 组——这一组由 1004 例（18%）组成，其中母亲生活在农村，没有收音机，没有电视，是的，宗教正统，奥罗莫族，识字状况无法阅读母亲，母亲认为离卫生设施的距离是一个大问题，丈夫的教育没有教育，15-24 岁的妇女，母亲已婚，户主为男性，tt 接种了是的。

第 2 组——这一组包括 2063 例（37%），其中母亲生活在农村，无法收听广播，无法收看电视，宗教信仰信奉新教，其他种族，母亲的识字状况无法阅读，母亲认为与卫生设施的距离是一个大问题，丈夫的小学教育，25-34 岁的女性，母亲的婚姻状况已婚，hh-head 男性，tt 接种了疫苗。

第 3 组——这一组由 633 名（11%）母亲组成，这些母亲生活在农村，能收听广播，不能收看电视，宗教信仰信奉新教，其他种族的母亲，母亲的识字状况无法阅读，母亲认为离卫生设施的距离是一个大问题，丈夫的小学教育，25-34 岁的妇女，母亲的婚姻状况已婚，hh-head 男性，tt 接种疫苗。

第 4 组——这一组由 633 名（11%）母亲组成，这些母亲住在农村，可以收听广播，不能收看电视，宗教信仰正统，阿姆哈拉族，母亲的识字状况无法阅读，母亲认为离卫生设施的距离是一个大问题，丈夫的教育没有教育，35-49 岁的妇女，母亲的婚姻状况已婚，hh_head 女性，tt 接种了疫苗。

如表 3 中所解释的，决策树 J48 正确地预测分类，因为实际分类是 1363。1320 名母亲被分类为已接种疫苗[是]，预测为未接种疫苗[否]，630 名被分类为实际未接种疫苗，但通过 J48 算法预测为已接种。对于幼稚和多层感知器，请参见（表 3 和 4）。通过创建测试 1、测试 2、测试 3 和测试 4 数据测试实例，使用天真贝叶斯进行了单个数据实例测试，其中三个数据实例被正确预测，一个数据实例分类错误。

The Classifiers	Correctly classified	Incorrectly classified	Time Taken
Decision Tree (J48)	62.59%	37.41%	0.87Sec
K-nearest neighbors	60.52%	39.48%	0.00Sec
Naïve Bayesian	63.30%	36.70%	0.01Sec
Multilayer perceptron	60.94%	39.06%	46.12Sec

表 3.正确和错误分类，以及使用十倍交叉验证的分类器算法加载时间，EDHS 2011。

在一般协会中获得的最大置信度为 0.98，居住协会=农村 marital_status=已婚 4132==>电视=第 4064 号 conf:(0.98)”

在类属性中获得的最大置信度为 0.72，居住关联=城市 marital_status=已婚 959==>tt_vaccated=yes 689 conf:（0.72）”

置信度为 70.5%时发现的五条最佳规则：

1.居住地=城市 marital_status=已婚 959==>tt_vaccated=是 689；配置文件：（0.72）

2.住宅=城市 1099==>tt_vaccated=yes 787 conf:（0.72）

3.literalcy_status= 能够读取 marital_status= 已婚 1107==>tt_vaccated=yes 787 conf:（0.71）

4.literalcy_status=能够读取 1244==>tt_vaccated=yes 884 conf:（0.71）

5.radio=yes distance to_HF= 问题不大 804==>tt_vaccated=yes 569 conf:（0.71）

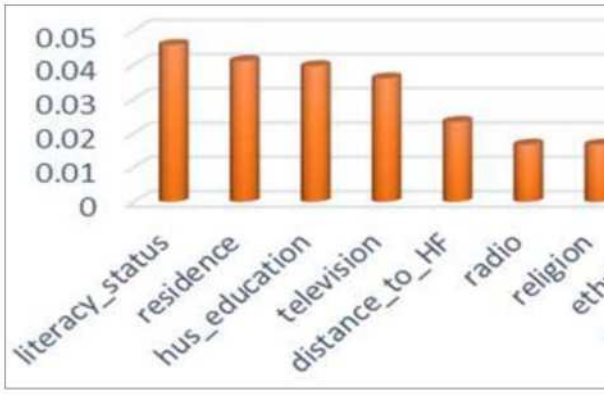

图 3.通过“Rank+InfoGainAttributeEval”算法获得的信息，EDHS 2011。

Algorithm Types	CCI	ICI	TT vaccina
Decision tree (J48)			
Training	65.36%	34.64%	Yes
Test	62.83%	37.17%	No
Bayesian naive			
Training	63.41%	36.59%	Yes
Test	62.47%	37.53%	No
Multilayer perceptron			
Training	67.28%	32.72%	Yes
Test	60.61%	39.37%	No

表 4.所选算法的详细精度，EDHS 2011。

母亲的识字状况信息获取率很高（0.046），其次是母亲的识字情况（0.041），获得的信息最少的是户主（0.0000147）。（图 3）

5.结论和建议

在这项研究中，数据挖掘工具和算法（J48、k-neast 和 Bayes）用于选择训练和测试数据，用于分类，k-means 方法用于聚类，一维关联规则用于识别最佳关联。挖掘工具、学习的具体方法具有特点，我们开发了 WEKA 方法，该方法基于选择文件和选择属性来转换“.csv”文件，并使用 WEKA 性能描述了特征。我们的工作扩展到利用数据集在所有部分中的数据挖掘工具的实现，以实现更好的准确率，并提高分析大型数据集时的效率。

因此，基于 TT 疫苗接种数据的最佳算法是多层感知器分类器，其准确率为 67.28%，构建模型所需的总时间为 0.01 秒。与其他分类器相比，多层感知器分类器的平均误差最低，为 32.72%。这些结果表明，在测试的机器学习算法中，多层感知器分类器有可能显著改进用于医学数据的传统分类方法。

参考文献

[1] Central Statistical Agency (CSA) [Ethiopia] and ICF, *Ethiopia Demographic and Health Survey 2016: Key Indicators Report*.2016: Addis Ababa, Ethiopia, and Rockville, Maryland, USA, CSA, and ICF.

[2] WHO, *Maternal immunization against tetanus: Standards for Maternal and Neonatal Care*. 2006, Department of making pregnancy safer.

[3] Central Statistical Agency (CSA) [Ethiopia] and ICF, *Ethiopia Demographic and Health Survey 2011: Key Indicators Report.* 2012: Addis Ababa, Ethiopia, and Rockville, Maryland, USA, CSA, and ICF.

[4] *Validation of neonatal tetanus elimination in Andhra Pradesh Weekly Epidemiological Record*, 2004. 79: p. 292-297.

[5] Fauveau V et al., *Maternal tetanus: magnitude, epidemiology, and potential control measures.* International Journal of Gynecology and Obstetrics, 1993. 40: p. 3-12.

[6] WHO, Standards for maternal and Neonatal care: *Integrated management of pregnancy and chield birth*. 2007, Department of making pregnancy safer.

[7] Han, J., M. Kamber, and J. Pei, eds. *Data mining concepts and techniques.* Third ed. 2013, Morgan Kaufmann Publishers: Waltham, Mass.

[8] G. Rasitha Banu, *A Role of decision Tree classification data Mining Technique in Diagnosing Thyroid disease.* International Journal of Computer Sciences and Engineering, 2016. 4 (11).

[9] Ian H. Witten and Eibe Frank, eds. *Data Mining: Practical Machine Learning Tools and Techniques.* Second edition. 2005, Morgan Kaufmann publications.

[10] Parvez Ahmad, Saqib Qamar, and Syed Qasim Afser Rizvi, *Techniques of Data Mining In Healthcare: A Review.* International Journal of Computer Applications, 2015. 120 (15).

[11] P. L. Geenen, et al., *Constructing naive Bayesian classifiers for veterinary medicine: A case study in the clinical diagnosis of classical swine fever.* Research in Veterinary Science, 2010. 91: p. 64-70.

[12] Yi Peng, et al., *Application of Clustering Methods to Health Insurance Fraud Detection*. 2006.