GIS 辅助下机器学习在森林覆盖类型识别中的应用

内 容 摘 要

本文探讨了基于机器学习技术的森林覆盖类型预测方法，以期为自然资源管理和生态保护提供科学依据。研究使用的数据集来源于 UCI 机器学习仓库中的“森林覆盖类型”数据集，涵盖了科罗拉多州北部罗斯福国家森林内四个受人类活动干扰较少的荒野区域，共计七种森林覆盖类型。实验首先对数据进行了预处理，包括特征选择和标准化，随后应用了对数几率回归、随机森林和多层感知机三种机器学习模型进行分类预测。实验结果显示，随机森林模型在测试集上取得了最高准确率（约 95.51%），显著优于其他模型，表明其在处理复杂生态数据集的分类任务中具有卓越效能。特征重要性分析揭示了海拔、距最近道路距离等因素在预测中的关键作用。本研究不仅为森林覆盖类型预测提供了实用模型，也为深入理解生态过程与森林分布关系、制定有效保护策略奠定了基础。

关键词：森林覆盖类型预测　随机森林　对数几率回归　多层感知机　生态管理

ABSTRACT

This study delves into the application of machine learning techniques for predicting forest cover types, aiming to inform natural resource management and ecological conservation strategies. Utilizing the " Covertype" dataset from the UCI Machine Learning Repository, which encompasses four pristine wilderness areas in northern Colorado's Roosevelt National Forest with seven distinct forest cover types, this research embarks on an analytical journey through pre-processing steps including feature selection and standardization, followed by predictive modeling using logistic regression, random forests, and multilayer perceptron (MLP). The outcomes highlight random forests as the most proficient model, achieving an accuracy of approximately 95.51% on the test set, underlining its prowess in handling complex ecological datasets for classification tasks. Feature importance analysis underscores the pivotal roles of factors such as elevation and proximity to roads in prediction accuracy. Beyond providing a practical predictive framework for forest cover types, this study deepens our understanding of ecological processes and vegetation distribution, thereby informing efficacious conservation measures.

KEY WORDS: Forest Cover Prediction　Random Forests　Logistic Regression　Multilayer Perceptron　Ecological Management

目　　录

GIS 辅助下机器学习在森林覆盖类型识别中的应用

在当前自然资源管理与生态保护的迫切需求下，准确预测森林覆盖类型对于理解生态系统动态、制定有效的保护措施及促进可持续林业管理至关重要。随着地理信息系统（GIS）与遥感技术的飞速发展，利用地图学变量预测森林覆盖类型成为可能。本研究依托的 “森林覆盖类型”数据集涵盖了科罗拉多州北部罗斯福国家森林内四个独特荒野区域的详尽信息，这些区域展现了自然状态下森林覆盖类型的多样性，为探究生态过程与植被分布提供了理想场所。

实验的主要驱动力源自先前研究中 Blackard 与 Dean 所开展的工作，他们的研究表明神经网络模型相较于传统判别分析方法，在预测森林覆盖类型方面展现出更高的精确度。这一发现不仅揭示了先进计算技术在生态预测模型中的潜力，也激发了对模型优化与新方法探索的兴趣。鉴于此，本文拟进一步检验并比较多种机器学习模型在预测森林覆盖类型上的效能，旨在挖掘最优化预测模型，同时深入分析影响森林分布的关键环境因素。

一、实验任务

（一）数据集描述

本实验采用的数据集来源于 UCI 机器学习仓库中的“森林覆盖类型”（Covertype）数据集。该数据集于 1998 年 7 月 31 日捐赠，旨在通过对地图学变量（不涉及遥感数据）的分析，对森林区域的像素进行分类，将其划分为 7 种不同的森林覆盖类型。

数据集包含来自美国林务局（USFS）和美国地质调查局（USGS）的原始数据，记录了科罗拉多州北部罗斯福国家森林内四个荒野区域的森林覆盖情况，总面积约为 581012 个 30×30米的观测单元。这些区域由于人类活动干扰较少，其森林覆盖类型主要受自然生态过程影响，使得现存的森林覆盖类型更多地反映了生态过程而非森林管理实践。

数据集包含 12 个量化特征，包括海拔、坡向、坡度、日照阴影、土壤类型等，以未缩放的形式提供，12 个定量特征的概述见表 1。除此以外，还有 54 列数据，为荒野区域（4 列）和土壤类型（40 列）两个定性变量。样本的目标变量为 7 类森林覆盖类型，分别为 Spruce/Fir、Lodgepole Pine、Ponderosa Pine、Cottonwood/Willow、Aspen、Douglas-fir 和 Krummholz。

表 1　森林覆盖类型数据集定量特征概述

名称	单位	描述
海拔	米	地面高度的垂直距离
坡向	方位角	地表某一点与真北方向之间的角度
坡度	度	垂直高度变化与水平距离的比例
距最近地表水体的水平距离	米	到最近地表水体的水平距离
距最近地表水体的垂直距离	米	到最近地表水体的垂直落差
距最近道路的水平距离	米	到最近道路的直线水平距离
上午 9 点的阴影指数	0 至 255	夏至时上午 9 点太阳光阴影强度指数
正午的阴影指数	0 至 255	夏至时正午太阳光阴影强度指数
下午 3 点的阴影指数	0 至 255	夏至时下午 3 点太阳光阴影强度指数
距最近火点的水平距离	米	到最近野火发生点的直线水平距离

在初步了解数据集中十个定量特征的基本信息后，进一步得出这些变量的平均值与标准差，如表 2 所示。平均值能反映各特征的整体集中趋势，而标准差则量化了数据的离散程度。通过对比不同特征的统计量，可以初步评估数据的均衡性与变异性，这对后续的数据预处理及模型建立至关重要，为科学决策和算法优化奠定基础。

表 2　定量特征统计摘要表

名称	平均值	标准差
海拔	2959.36	279.98
坡向	155.65	111.91
坡度	14.1	7.49
距最近地表水体的水平距离	269.43	212.55
距最近地表水体的垂直距离	46.42	58.3
距最近道路的水平距离	2350.15	1559.25
上午 9 点的阴影指数	212.15	26.77
正午的阴影指数	223.32	19.77
下午 3 点的阴影指数	142.53	38.27
距最近火点的水平距离	1980.29	1324.19

（二）实验目的

本实验旨在利用该数据集开发和评估机器学习模型，以预测特定地理区域内的森林覆盖类型。具体而言，本实验通过分析诸如海拔、坡向、距最近水源的水平距离、日照阴影指数、距道路的水平距离、距最近野火点的距离等定量特征，以及荒野区域和土壤类型的定性特征，建立一个分类模型。此模型应能够基于给定的环境特征预测出七种森

林覆盖类型之一。

实验旨在探索不同特征对森林覆盖类型预测的重要性，验证模型在未受人类活动显著影响的自然环境中分类准确性的有效性，并对比不同算法在该任务上的性能。此外，实验还意在为理解生态过程与森林覆盖分布之间的关系提供实证基础，以及评估模型在实际森林管理和生态保护策略制定中的应用潜力。

二、实验过程

（一）技术方案

1．对数几率回归（Logistic Regression）

Logistic 回归是一种广义线性模型（Generalized Linear Model, GLM）。GLM 是一类统计模型的框架，它扩展了传统线性模型，允许因变量（响应变量）遵循除正态分布之外的如二项分布、泊松分布等的其他分布。GLM 的核心思想是通过连接函数（link function）将线性预测与观测到的响应变量的分布参数联系起来。

Logistic 回归用于处理二分类问题，即因变量只有两个可能的结果（例如成功/失败，是/否）。Logistic 回归中的连接函数是 logit 函数，它将线性预测映射到$(0,1)$区间内，代表事件发生的概率。逻辑回归的模型可以表达为：

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

其中，$\text{logit}(p) = ln\left(\frac{p}{1-p}\right)$，$p$是事件发生的概率，$\beta_i$是系数，$X_i$是自变量。

多项对数几率回归是逻辑回归的直接扩展，它直接处理多分类问题，不需要拆分成多个二分类模型。在多项逻辑回归中，因变量遵循多项分布，且模型通过一个 Softmax 函数将线性预测映射到概率分布上，它确保了所有类别的概率之和为 1，且每个概率值在 0 到 1 之间，适合表示多分类的概率分布。

本实验面临着七种森林覆盖类型的分类问题，这与多项对数几率回归处理多分类的能力相契合，能够输出各类别的概率，直接对不同森林覆盖类型的概率分布建模，适应了数据的复杂多样性与分类需求。

2．随机森林（Random Forest）

随机森林是一种高度灵活且强大的机器学习模型，其核心在于将三个关键概念融合为一个统一的框架：决策树的构建、集成学习策略以及随机性引入的机制。

（1）决策树（decision tree）

决策树是随机森林算法的基学习器，通过递归地对数据集进行划分来实现预测。在每一个划分节点，算法依据特征选择准则来确定最佳分割特征及其阈值。其中，信息增

益（information gain）是一个常用指标，其数学表达式为：

$$\mathrm{Gain}(D,a)=\mathrm{Ent}(D)-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\mathrm{Ent}(D^v)$$

公式中，$\mathrm{Ent}(D)$表示信息熵，$\mathrm{Ent}(D^v)$表示使用某离散属性a对样本集D划分后第v个分支节点包含的取值为的样本a^v的信息熵。通过选择最大信息增益的特征进行分割，决策树能够有效降低数据的不确定性，逐步细化数据空间，直到满足停止条件，如达到预设的树深度或节点内样本数量的最小限制。

决策树有很多学习算法，ID3 使用信息增益构建决策树；C4.5 在 ID3 的基础上做了改进，采用增益率来处理连续属性和缺失值；Gini 指数是另一种评判标准，常见于 CART 决策数算法。剪枝是预防决策树模型过拟合的重要手段，其中预剪枝预防过拟合，限制树的大小；后剪枝则在生成完整树后删除不重要分支，提升泛化能力。

（2）集成学习（ensemble learning）

随机森林是集成学习的代表性算法。集成学习是一种机器学习方法，其核心思想是通过结合多个模型的预测来提高整体预测的准确性和稳定性。这种方法认为，虽然单一的学习器可能因为过拟合、偏差或方差问题而在某些数据子集上表现不佳，但将多个学习器的预测结果综合起来，可以有效降低错误率，提升模型的泛化能力。集成方法通常能够实现比单个模型更强大的性能，这也是“众人拾柴火焰高”理念在机器学习领域的体现。

集成学习主要分为两大类。Boosting 以一种序列化的方式逐步构建模型，它从一个弱学习器开始，每次迭代都重点学习前一轮预测错误的数据，逐步提高对困难样本的重视程度。通过这种方式，每一个后续的模型都在尝试修正前面模型的错误，最终所有模型的加权组合形成了一个强预测器。AdaBoost、Gradient Boosting 和 XGBoost 是此方法中的代表。Bagging 通过数据采样的方式创建多个不同的训练集，然后基于这些训练集分别训练出多个基学习器，最后通过投票或平均的方式来组合这些基学习器的预测结果。由于每个基学习器独立训练，因此可以减少模型间的相关性，从而降低方差，增强模型稳定性。随机森林便是 Bagging 方法的一个典型应用。

（3）随机森林（Random Forest）

随机森林作为集成学习中的一种重要算法，是基于 Bagging 原理的一种改进。它不仅在训练过程中对训练集进行有放回的抽样（bootstrap sampling）来创建多个不同的子集，而且在决策树的构建过程中对特征进行随机选择，进一步增强模型的泛化能力。

随机森林算法在本实验的森林覆盖类型分类任务中展现出了显著的效能和高度的

适应性。本实验需要在包含众多特征如海拔、坡度、土壤类型等的高维度数据集中进行精细分类。随机森林凭借其强大的数据处理能力，能够自动识别并侧重于那些对分类结果最具影响力的特征，从而在复杂的环境变量中精确筛选出关键信息。考虑到森林覆盖类型分类中可能蕴含的非线性关系，传统的单一模型难以充分捕捉这些细致变化，而随机森林通过集成多棵决策树，每棵树基于数据的不同子集学习，协同构建出广泛的决策边界，能有效揭示并利用这些非线性模式，增强了对生态系统复杂相互作用的理解和预测。此外，随机森林算法可以评估并排序特征的重要性。这为实验结果的分析提供了量化视角，帮助我们清晰认识到在预测森林覆盖类型时起决定性作用的生态因素。

3．多层感知机（Multilayer Perceptron, MLP）

（1）神经网络（neural network）

神经网络是一种受生物学启发的计算模型，旨在模仿大脑中神经元网络的结构和功能。它由大量简单的处理单元（称为神经元或节点）组成，这些单元通过加权连接相互作用，能够学习并解决复杂的非线性问题。每个神经元执行加权输入求和并通过激活函数转换，对应的公式为：

$$y = f\left(\sum_i w_i x_i + b\right)$$

其中，f是激活函数，w_i是权重，x_i是输入，b是偏置项。

神经网络通常包括输入层、一个或多个隐藏层以及输出层。信息以加权求和及非线性转换的方式从前一层向后一层传播，最后在输出层得到预测结果。

（2）前馈神经网络（Feedforward Neural Network, FNN）

前馈神经网络是最早且最基础的神经网络架构之一，其特点是通过一系列单向传播的层级结构学习数据特征。信息从输入层流向输出层，途中经过一个或多个隐藏层的非线性转换，每层使用激活函数增强网络表达复杂模式的能力。这种简单而高效的架构，使其成为处理分类和回归任务的理想选择，特别是在数据间存在复杂关系且无需考虑时间序列依赖的场景。

（3）多层感知机（MLP）

MLP 是具有至少一个隐藏层的前馈神经网络，擅长捕捉复杂输入数据的非线性模式。MLP 的核心优势在于其能够拟合任意复杂的非线性关系，这得益于其多层结构，尤其是隐藏层的存在。隐藏层的引入使得网络能够学习输入数据的抽象表示，从而提高了模型在分类和回归任务中的表现。

MLP 中的每个神经元执行以下操作：

$$\mathrm{Net}_j = \sum_{i=1}^{n} w_{ij} x_i + b_j$$

$$y_j = f\left(\mathrm{Net}_j\right)$$

其中，w_{ij}是第i个输入到第j个神经元的权重，x_i是第i个输入，b_j是第j个神经元的偏置项，Net_j是第j个神经元的净输入，f是非线性激活函数，例如 ReLU 或 sigmoid，y_j是该神经元的输出。

本研究针对森林覆盖类型分类问题，包含多维地理与生态特征。MLP 模型因能捕捉高维数据中的非线性关系、自动提炼关键特征、适应多分类任务且具有一定解释性，尤为适配本研究需求，以精准预测七类森林覆盖类型。

（二）数据预处理

为了确保模型训练的有效性和准确性，本实验对原始数据进行了一系列预处理。尽管“森林覆盖类型”数据集未声明存在缺失值，作为预处理的第一步，本实验仍然执行了缺失值检查，以确保数据的完整性和后续分析的准确性。为评估模型的泛化能力，本实验将数据分割为训练集和测试集。分割采用了一个标准比例，即 80%的数据用于模型训练，而剩下的 20%保留作为独立的测试集，用于评估模型在未见数据上的表现。考虑到不同特征之间可能存在量纲的差异，本实验对数值型特征实施了特征标准化操作。该步骤基于训练集计算每个特征X的均值μ和标准差σ，进而转换数据，使得每个特征的分布均值为 0，方差为 1。标准化的公式如下：

$$X_{\mathrm{std}} = \frac{X - \mu}{\sigma}$$

其中，X_{std}是标准化后的特征值。这样做不仅加速了算法的收敛速度，还使得不同特征在模型中的贡献度可比。对于测试集，采用了训练集得到的转换参数进行转换，避免了数据泄露的风险。

（三）数据分析

1．基于多项对数几率回归的森林覆盖类型分类模型

本实验采用 Python 3.11 环境进行多项对数几率回归模型的实证分析，主要分为通过 pandas 库加载原始数据和使用 scikit-learn 库进行建模两部分。

在正式建模前，本实验通过 Lasso 回归进行特征选择。Lasso 回归通过引入 L1 正则化，鼓励模型学习出稀疏的系数，使得许多不重要的特征权重接近零。本研究使用了较小的正则参数$C = 0.01$，以促进特征稀疏化，同时设定最大迭代次数为 1000，确保找到最优解。通过 Lasso 回归拟合数据，我们获得了每个特征的系数，基于这些系数的绝对值，筛选出了显著特征，对应的序号为 1、2、3、4、5、6、8、9、10、11、12、14、16、

17、18、21、22、23、24、26、27、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、51、52、53、54。

表 3 多类别分类性能评估表

标签	精确度	召回率	F1 分数	支持度
1	0.71	0.70	0.70	42,368
2	0.75	0.80	0.77	56,661
3	0.66	0.81	0.73	7,151
4	0.60	0.41	0.49	549
5	0.17	0.01	0.01	1,899
6	0.46	0.21	0.28	3,473
7	0.73	0.56	0.63	4,102
宏平均	0.58	0.50	0.52	116203
加权平均	0.71	0.72	0.71	116203

之后，本文利用筛选出的关键特征，构建了多项对数几率回归模型，在测试集上对森林覆盖类型进行预测。模型在测试集上的整体准确性达到了 72.18%，表明模型在大部分情况下能够准确预测森林覆盖类型，但仍有提升空间。其中，支持度（Support）表示每个类别在测试集中出现的次数，宏平均（Macro avg）不考虑各类别的样本量，对每个类别的指标（精确度、召回率、F1 分数）计算平均值；加权平均（Weighted avg）则考虑每个类别的样本量，为每个类别的指标赋予相应的权重。它们的计算公式如下：

$$\text{Macro avg} = \frac{1}{N}\sum_{i=1}^{N} \text{Metric}_i$$

$$\text{Weighted avg} = \frac{\sum_{i=1}^{N} (\text{Support}_i \times \text{Metric}_i)}{\sum_{i=1}^{N} \text{Support}_i}$$

其中，N是类别的数量，Metric_i是第i个类别的精确度、召回率或 F1 分数；Support_i是第i个类别的支持度。

观察分类性能评估表 3 可知，模型在一些类别上表现优异，如 Spruce/Fir、Lodgepole Pine 和 Ponderosa Pine，这些类别的 precision、recall 和 f1-score 均超过 0.7，表明模型在这几个类别上表现出色；在另一些类别上有较大提升空间，如 Cottonwood/Willow、Aspen、Douglas-fir 的 precision 和 recall 较低，特别是 Aspen 的 f1-score 仅为 0.01，说明模型在识别这些类别时存在问题，可能是数据稀缺或特征难以区分导致。

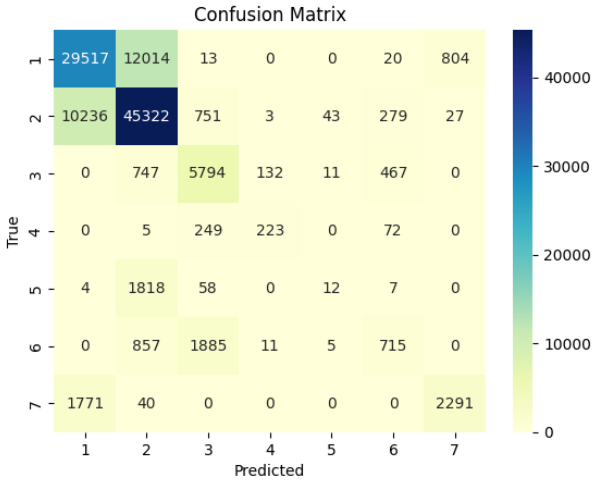

图 1　多项对数几率回归混淆矩阵

图 1 的混淆矩阵显示了各类别预测与实际类别之间的错误分配，如 Aspen（类别 5）误分类较多，意味着模型可能将 Aspen 与其他类别混淆。这强调了特征选择与模型优化的必要性，特别是在低频类别上的表现不佳类别。

2．基于随机森林的森林覆盖类型分类模型

本实验同样采用 Python 3.11 环境进行随机森林模型的实证分析，主要分为通过 pandas 库加载原始数据和使用 scikit-learn 库进行建模两部分。

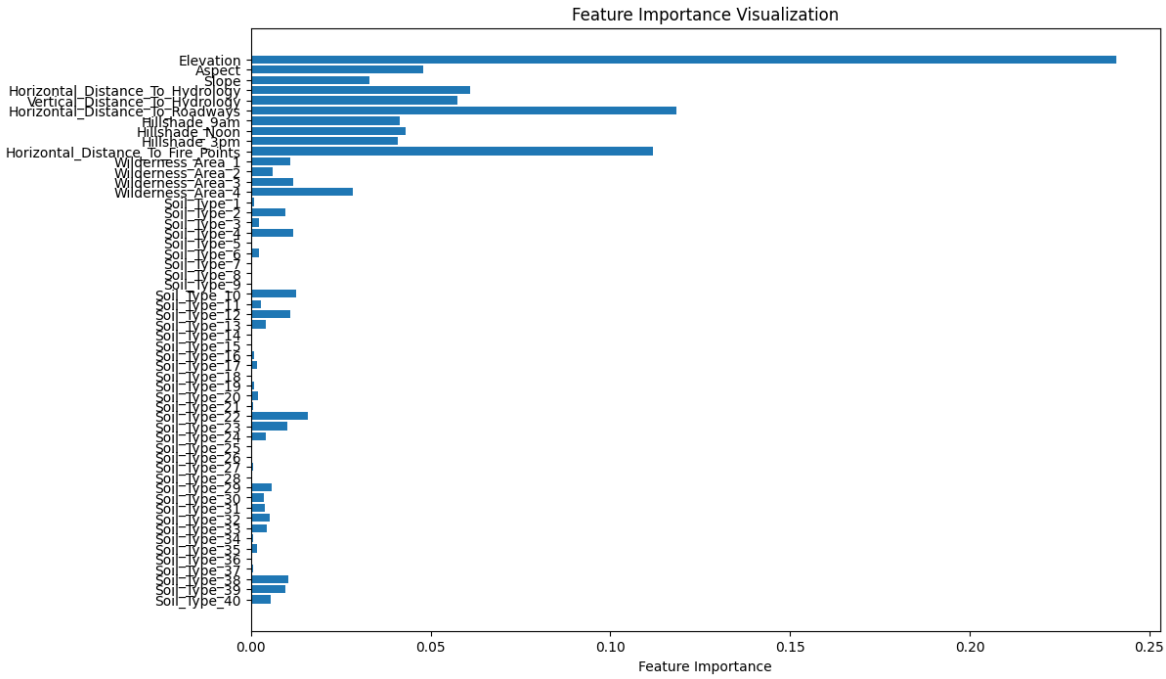

图 2　特征重要性条形图

具体来说，本实验设置 100 棵树作为基学习器以提升模型的稳定性和性能。其他参数均采用默认值，即使用 Gini 指数衡量特征划分的质量，决策树在生长过程尽可能生长直到所有叶节点都纯净或者达到其他停止条件（节点分裂时所需的最少样本数为 2 及叶子节点上最少的样本数量为 1），在构建每棵树时均使用自助抽样（bootstrap sampling）。

模型训练完成后，得出各特征的重要性如图 2。观察特征重要性条形图可以发现，不同特征对于分类决策的贡献程度存在显著差异，例如，某些环境因素如海拔高度、距最近道路、最近火点的水平距离等在决定特定树种的分布中起到较为关键的作用。另外，原数据集中定性变量被拆分成多个哑变量，原本该变量携带的信息被分散到了这些新的特征中。这意味着，尽管单独来看每个哑变量的重要性可能不高，但并不能否认它们作为一个整体可能对模型预测能力有显著贡献。随机森林中的特征重要性评估通常基于单个特征的分裂效果，可能未能完全捕捉到这种联合效应。

表 4 多类别分类性能评估表

标签	精确度	召回率	F1 分数	支持度
1	0.97	0.94	0.95	42,557
2	0.95	0.97	0.96	56,500
3	0.94	0.96	0.95	7,121
4	0.91	0.85	0.88	526
5	0.94	0.77	0.85	1,995
6	0.94	0.90	0.92	3,489
7	0.97	0.96	0.96	4,015
宏平均	0.95	0.91	0.93	116203
加权平均	0.96	0.96	0.95	116203

接着，将模型在测试集上进行测试。模型评估结果显示，测试集上的准确率达到了约 95.51%，表明模型能够有效区分不同类型的森林覆盖。表 4 进一步揭示了模型在各类别上以及总体上的精确度、召回率和 F1 分数，显示了模型在大多数类别上的表现均衡且强劲。

总体而言，模型在分类不同树种方面表现非常出色，能够有效地识别大多数树种。在所有类别中，模型在 Spruce/Fir、Lodgepole Pine、Ponderosa Pine、Douglas-fir 和 Krummholz 的分类上表现较为优秀，精确度、召回率和 F1 分数都相对较高，显示出模型对这类树种的分类具有较高的准确性和可靠性。然而，模型在 Cottonwood/Willow 和 Aspen 这两个类别的分类上表现相对较差。对于 Cottonwood/Willow，模型的精确度和召回率较低，F1 分数为 0.88；对于 Aspen，虽然精确度较高，但召回率仅为 0.77，导致 F1

分数为 0.85。这说明需要针对这两个类别进行进一步的优化和调整，以提高模型的分类性能。

最后，通过绘制如图 3 所示的混淆矩阵来直观展示模型预测结果与实际标签之间的对应关系。进一步证实了模型在分类任务上的高效性和准确性。混淆矩阵中较低的错分类率显示了模型在各种森林覆盖类型之间的良好区分能力。

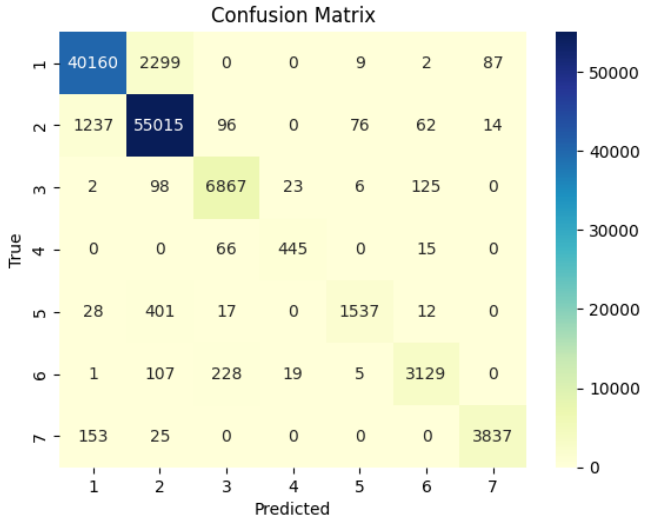

图 3　随机森林混淆矩阵

3．基于多层感知机的森林覆盖类型分类模型

本实验同样采用 Python 3.11 环境进行多层感知机模型的实证分析，主要分为通过 pandas 库加载原始数据和使用 scikit-learn 库进行建模两部分。

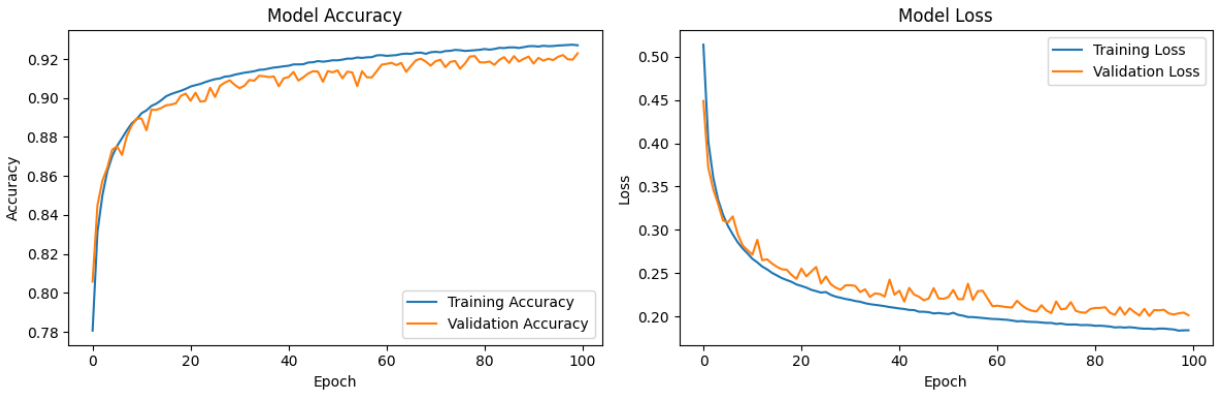

图 4 训练与验证阶段准确率及损失变化

经不断测试和调整，最终构建了一个含两个隐藏层的 MLP 模型，其中输入层有 128 个节点，第一个隐藏层有 64 个节点，均使用 ReLU 作为激活函数；输出层对应 7 个类别，每个类别由一个节点表示，并采用 Softmax 激活函数，以输出各类别的概率分布。模型使用 Adam 优化器，学习率为 0.001，损失函数为稀疏分类交叉熵（Cross Entropy），以最大化多分类任务的准确性。模型训练持续了 100 个周期，每批次处理 32 个样本。

经过训练，模型在测试集上的准确率达到 92.29%，显示了良好的泛化能力。进一步绘制了图 4 训练与验证阶段的准确率与损失曲线，结果显示模型学习稳定，没有出现过拟合现象。

表 5　多类别分类性能评估表

标签	精确度	召回率	F1 分数	支持度
1	0.93	0.91	0.92	42,368
2	0.93	0.94	0.94	56,661
3	0.92	0.93	0.92	7,151
4	0.84	0.76	0.80	549
5	0.80	0.76	0.78	1,899
6	0.84	0.87	0.85	3,473
7	0.89	0.96	0.92	4,102
宏平均	0.88	0.87	0.88	116203
加权平均	0.92	0.92	0.92	116203

通过计算如表 5 所示分类报告，我们获得了模型在各个类别上的精确度（precision）、召回率（recall）以及 F1 分数，整体的宏平均 F1 分数达到 0.88，表明模型总体上性能优异。其中，Cottonwood/Willow 和 Aspen 的召回率均为 0.76，相较于其他几个主要类别较低，意味着模型在预测为 Cottonwood/Willow 和 Aspen 的样本中，有相当一部分实际属于该类别的样本被错误地分类到了其他类别，显示出模型在识别这两类森林覆盖类型上存在一定的困难。这两个类别（Cottonwood/Willow 和 Aspen）的预测表现不佳，可能是由于它们在生态特征上与其他类别有较多重叠，或者样本量（支持度）较少导致模型学习不够充分。

此外，图 5 还展示了对应的混淆矩阵，更加直观地反映了模型预测类别与实际类别之间的匹配情况，其中多数类别预测正确，但某些类别（如类别 4 和 5，亦即 Cottonwood/Willow 和 Aspen）存在一定的预测误差，提示未来研究中可能需要进一步优化这些类别的识别策略。

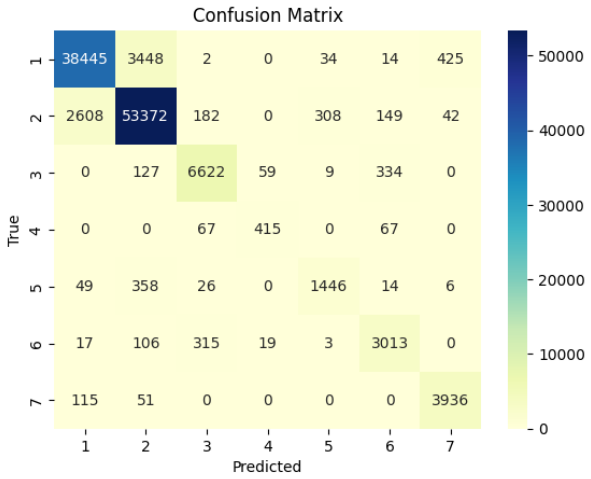

图 5　多层感知机混淆矩阵

4．模型比较与优化

通过比较三种机器学习模型在森林覆盖类型预测任务上的性能，发现对数几率回归模型在测试集上的准确率达到了 72.18%，但在如 Cottonwood/Willow、Aspen 和 Douglas-fir 类别上表现较差；随机森林模型在测试集上的准确率达到了 95.51%，在大多数类别上表现均衡且强劲，但 Cottonwood/Willow 和 Aspen 两个类别的分类效果稍差；多层感知机模型在测试集上的准确率达到 92.29%，，但在 Cottonwood/Willow 和 Aspen 上召回率较低，存在一定的识别困难。总体来说，Cottonwood/Willow 和 Aspen 两个类别是模型识别的难点，后续可以采用过采样或下采样等方法改善类别不均衡问题，提高模型对少数类别的识别效果。分别来看，对数几率回归模型在面对高度非线性的森林覆盖类型数据时显得力不从心，因为其基于线性假设的模型结构难以充分捕捉数据中复杂的非线性关系；随机森林模型则通过集成学习策略，在每棵决策树中使用随机特征选择，从而捕捉到数据的非线性模式，表现出色；多层感知机模型则通过隐藏层提取高维特征之间的非线性关系，虽然整体效果不如随机森林，但在捕捉非线性关系方面表现良好。

本实验所使用的三个模型后续均可进一步优化。在对数几率回归模型方面，可以通过特征工程和模型扩展来提高其对非线性关系的捕捉能力，例如构建交互特征来增强非线性表达能力；随机森林模型可以通过模型调优和与其他模型的集成来优化其性能，如调整树的数量、树的最大深度等参数，以及尝试与梯度提升树、XGBoost 等模型进行集成；对于多层感知机模型，可以通过网络结构优化和深度学习来提高其表示能力，如调整隐藏层的节点数和层数、选择合适的激活函数，并考虑使用更复杂的网络结构来充分挖掘数据中的空间和时间特征。

除此以外，还有一些更加现代化的模型可以考虑引入到本实验中，例如在模型中引入自注意力机制（Self-Attention Mechanism），让模型能够自动学习不同特征对分类的重要性；Graph Neural Networks（GNNs）将森林覆盖预测问题视为图结构中的节点分类问题，特别适用于具有复杂空间关系的场景。选择合适的模型，并进行针对性的优化，对于提高预测准确性具有重要意义。

三、实验结果与结论

本次实验针对 UCI 机器学习仓库中的“森林覆盖类型”数据集，采用多种机器学习模型（对数几率回归、随机森林、多层感知机）进行了森林覆盖类型预测的实证分析，旨在探索模型的效能并深入理解森林覆盖类型的关键影响因素。实验基于科罗拉多州北部罗斯福国家森林内四个荒野区域的数据，涵盖了七种不同的森林覆盖类型。

在处理多分类任务时，通过多项对数几率回归实现，测试集上的准确率为 72.18%。模型在 Spruce/Fir、Lodgepole Pine、Ponderosa Pine 等类别上表现良好，但在 Cottonwood/Willow、Aspen 等类别上召回率较低，可能与类别间生态特征的重叠或样本量不足有关。特征选择通过 Lasso 回归进行，提高了模型的解释性和效率。随机森林模型表现出极高的准确度，达到 95.51%，在多数类别上展现出均衡且强劲的分类能力。模型的特征重要性分析揭示了其在区分不同森林覆盖类型方面的高效性，尤其是海拔、距最近道路距离等特征在分类决策中起到了关键作用。多层感知机模型达到 92.29%的测试准确率，尽管略低于随机森林，但整体性能依然出色。MLP 模型通过两层隐藏层结构捕捉了特征间的非线性关系，展示了较好的预测能力。实验结果为模型优化提供了方向，特别是在处理类别不平衡和提高少数类识别上。未来研究可考虑采用过采样、下采样技术，或引入更先进的模型如自注意力机制和图神经网络（GNNs），以进一步提升预测精度和模型的鲁棒性。

在完成这项关于森林覆盖类型预测的实验研究后，我深刻体会到机器学习技术在生态学研究中的巨大潜力和价值。通过对数几率回归、随机森林、多层感知机等多种模型的比较，我认识到没有一劳永逸的解决方案，模型的选择需根据问题特性灵活调整。此外，模型的参数调优也是提升性能的关键，如学习率、树的数量、网络结构等都需要细致考量。特征的重要性分析让我明白，模型的预测能力很大程度上取决于输入特征的选取。例如，海拔、坡向、距水源距离等地理因素在森林覆盖类型预测中扮演着关键角色，在设计模型时，应深入理解生态背景知识，精选对预测目标有直接影响的特征。Cottonwood/Willow 和 Aspen 两个类别的识别难题凸显了类别不平衡问题对模型性能的

影响，在未来的研究中应采取如过采样、下采样或成本敏感学习等措施来平衡各类别样本，提高模型对少数类别的识别能力。

此次研究不仅是一次技术探索，更具有深远的生态学意义。准确预测森林覆盖类型有助于我们更好地理解生态系统动态，制定更科学的保护措施，促进可持续林业管理。这也提醒我们，技术进步应服务于生态保护，为人类与自然和谐共存提供智慧解决方案。

参考文献

[1]Blackard, J. A., & Dean, D. J. (2000). Comparison of neural network and discriminant analysis for predicting forest cover types from cartographic variables [J]. Computers & Geosciences, 26(1), 1-14.
[2]Breiman, L. (2001). Random forests [J]. Machine Learning, 45(1), 5-32.
[3]Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning [M]. Springer.
[4]Zhang, J., & Chen, S. (2016). Multi-class classification using deep fully connected neural networks [C] // 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5649-5653).