东北大学秦皇岛分校
Northeastern University at Qinhuangdao

《医疗数据处理实践》课程设计报告

泰坦尼克号幸存者数据分析及基于神经网络填充缺失值的预测模型

学　　院	数学与统计学院
专　　业	数据科学与大数据技术
班级序号	**200221**
学　　号	**202015140**
姓　　名	周华
指导教师	王子健、张建波
开始日期	2022 年 5 月 28 日
结束日期	2022 年 7 月 6 日

泰坦尼克号幸存者数据分析及基于神经网络填充缺失值的预测模型

摘　　要

泰坦尼克号是英国白星航运公司 20 世纪 10 年代建设的一搜豪华游艇。其排水量达到了惊人的 4.6 万吨，是当时世界上体积最大，最豪华的客运轮船。但正是这艘号称“永不沉没”的泰坦尼克号，在 1912 年从英国驶往美国时，在大西洋与冰山相撞并沉没！

根据英国贸易委员会公布的数据显示，在灾难发生时，泰坦尼克号共搭载 2224 人，其中 710 人生还，1514 人不幸罹难，其中乘客约有 1317 人，共 498 人幸存；男性船员约有 885 人，共 192 人幸存；女性船员 23 人，共 20 人幸存。

本文通过数据分析与机器学习算法，对泰坦尼克号数据进行数据预处理，数据可视化，特征工程，模型调参，模型优化，有效提取了泰坦尼克号数据中的信息，并对幸存者建立了预测模型，模型准确率高达 85.86%。

第一步：首先进行导包，读取数据，然后查看数据基本信息，并对每个属性的数据进行剖析。**第二步**：对多个属性之间的关系进行分析，探索与可视化，得出幸存率主要与性别，年龄，船舱有关的结论。**第三步**：首先对无用信息、非数值型数据、缺失数据进行处理，然后训练神经网络模型并对 Age 数据进行预测填充。**第四步**：进行模型的训练与验证，模型调参与模型融合。

关键词：**泰坦尼克号**；**特征工程**；**参数优化**；**数据分析**；**神经网络**

目　　录

1 绪论

1.1 研究背景

泰坦尼克号是英国白星航运公司 20 世纪 10 年代建设的一搜豪华游艇。其排水量达到了惊人的 4.6 万吨，是当时世界上体积最大，最豪华的客运轮船。但正是这艘号称“永不沉没”的泰坦尼克号，在 1912 年从英国驶往美国时，在大西洋与冰山相撞并沉没!

根据英国贸易委员会公布的数据显示，在灾难发生时，泰坦尼克号共搭载 2224 人，其中 710 人生还，1514 人不幸罹难, 其中乘客约有 1317 人，共 498 人幸存；男性船员约有 885 人，共 192 人幸存；女性船员 23 人，共 20 人幸存。

大数据比赛是一项综合 Python 编程，数据分析，机器学习的比赛。常见的大数据比赛包括 Kaggle 比赛，“钉钉杯”大数据挑战赛，微信大数据挑战赛。大数据比赛一般流程包括数据初步探索，数据可视化探索，特征工程，模型建立与验证，模型调参，模型融合等。

1.2 研究内容

本课程设计的**研究内容**就是使用数据分析比赛的流程对泰坦尼克号幸存者的数据进行分析，提取有效信息，得出有效结论；并在数据分析的基础上，建立预测模型，对测试集的数据进行预测。

1.3 研究目的

本课程设计的**研究目的**就是通过对泰坦尼克号的数据分析和模型建立，从而熟练掌握 jupyterlab，anaconda，Python 常用包的使用和方法；深刻理解数据分析，机器学习的步骤和算法。

1.4 研究思路

本课程设计第一步在导入数据后进行数据的初步探索，包括查看数据的基本属性，缺失值数据，统计数据，每个属性的基本信息；第二步通过数据可视化，进一步探索属性与属性之间关系并提取有效信息，得出有效结论，为特征工程做准备；第三步进行特征工程，包括数据预处理，缺失值处理，模型训练和模型验证；第四步使用网格搜索思想对模型参数进行优化；第五步使用 Stacking 对模型进行融合。

具体研究思路如下：

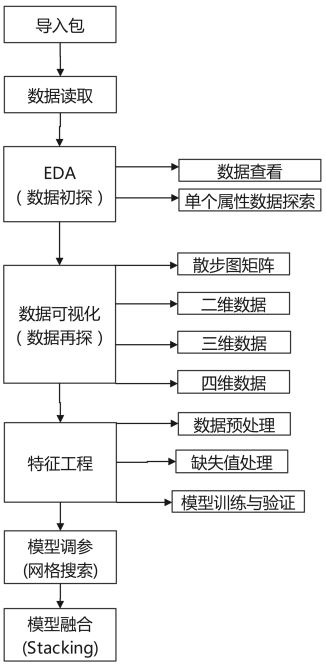

图 1.1　论文研究思路

2 EDA(数据初探)

2.1 导入包

```
#导入常用包
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

#导入深度学习框架pytorch
import torch
import torch.nn as nn
import torch.nn.functional as F

#导入机器学习包
from sklearn.linear_model import LogisticRegression   #逻辑回归
from sklearn.svm import SVC, LinearSVC                #SVC
from sklearn.ensemble import RandomForestClassifier   #随机森林
from sklearn.neighbors import KNeighborsClassifier    #KNN
from sklearn.naive_bayes import GaussianNB            #贝叶斯
from sklearn.linear_model import Perceptron           #感知机
from sklearn.linear_model import SGDClassifier        #SGD
from sklearn.tree import DecisionTreeClassifier       #决策树
```

图 2.1　导入常用的包

首先导入本课程设计需要的包，其中常用包包括了 numpy，pandas，matplotlib，seaborn；pytorch 是一个深度学习框架，本文中用于对缺失数据 Age 的预测填充；scikit-learn 的包

包括了逻辑回归，SVC，随机森林，KNN，朴素贝叶斯，感知机，SGD 和决策树，本文中用于训练并预测泰坦尼克号乘客的幸存与否。

2.2 数据读取

```
train=pd.read_csv('./input/06.titanic-train.csv')
test=pd.read_csv('./input/06.titanic-test.csv')
combine = [train, test]
```

图 2.2　读取训练集和测试集

分别读取数据集和测试集，其中 conbine 将 train 和 test 合并，便于查看数据和处理数据。

2.3 查看数据

首先查看数据，包括数据的行数列数，训练集前 5 行，基本信息，缺失值和统计数据。其中基本信息如图 2.3, 其他数据信息见附录 3.3.1。

```
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
```

图 2.3　训练集基本信息

训练集共有 891 行；训练集包含 12 个属性，分别为乘客 Id，是否幸存，客舱等级，姓名，性别，年龄，同代亲属数，不同代亲属数，船票编号，床票价格，客舱号，登船港口；其中 Age 缺失 177 个数据，Cabin 缺失 687 个数据，Embarked 缺失两个数据；通过查看统计信息，没有发现异常值。

2.4 单个属性数据探索

对训练集 Survived，Pclass，Sex，Age，SibSp，Parch，Fare，Embarked 共 8 个属性进行计数并绘制直方图，可以提取以下信息：

- 幸存者 342 人，遇难 549 人，幸存者比例为 38.38%；
- 三等仓人数最多，为 55.1%，一等舱为 24.2%，二等舱 20.7%；
- 男性 577 人，女性 314 人，男性更多，占 64.76%；
- 20 岁-40 岁的人较多；
- 68% 的人没有同级亲属，23% 的人有一个同级亲属，同级亲属有两个以上的很少；
- 76.1% 的人没有非同级亲属，13% 的人有一个非同级亲属；
- 大部分的船票在 100 美元以下；
- 72% 的人从 S 上船，18% 从 C 上船，只有 8% 的人从 Q 上船。

其中，对 Survived 的探索结果如图 2.4，Pclass，Sex，Age，SibSp，Parch，Fare，Embarked 的探索结果见附录 3.3.2。

```
0    549
1    342
Name: Survived, dtype: int64
```

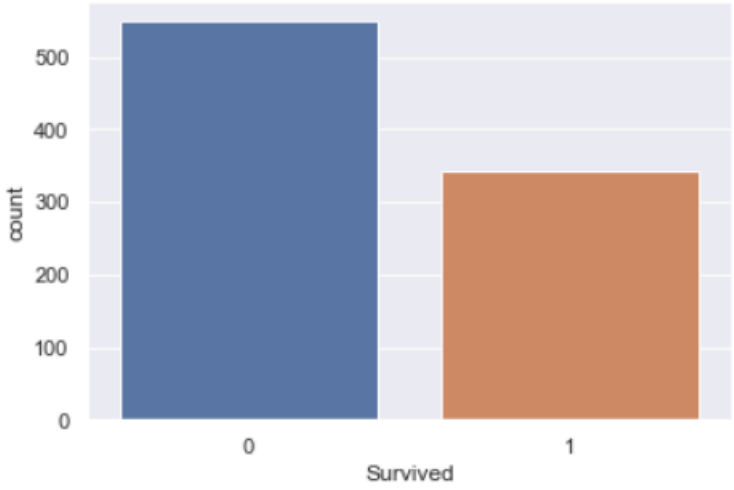

图 2.4　对单属性 Survived 的探索结果

3 数据可视化（数据再探）

3.1 数据总体概览

通过绘制散布图矩阵，查看所有数据的整体情况。如下图所示：

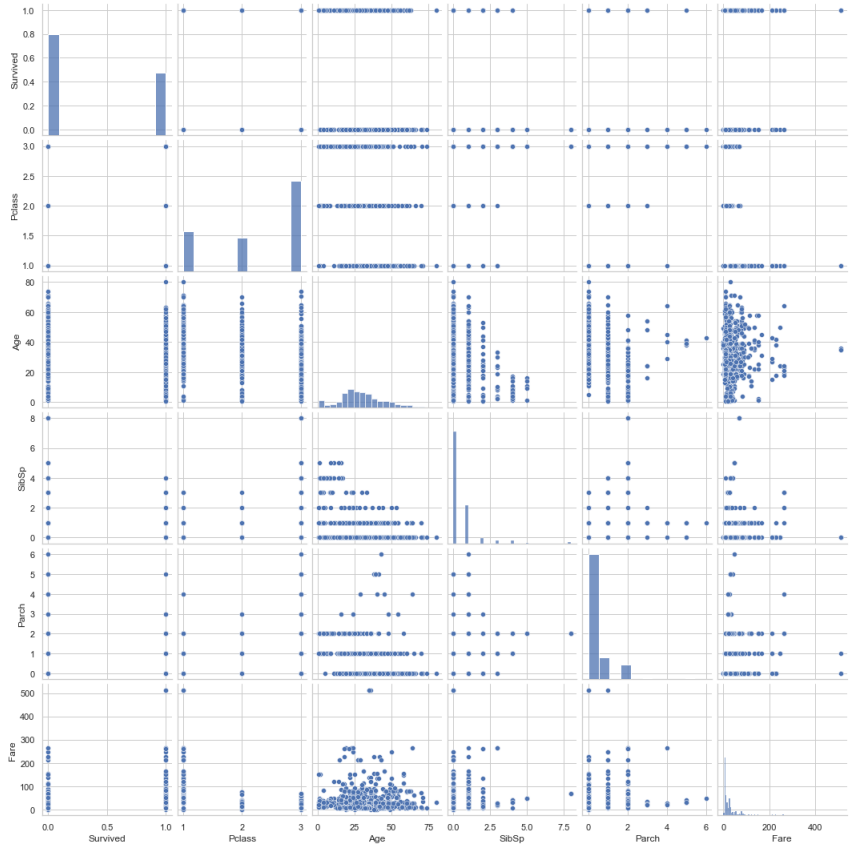

图 3.1　训练集散布图矩阵

其中散步图的矩阵的对角线位置在 2.4 中已经进行了阐述，下面进行非对角线数据的探索。

3.2 二维数据探索

通过绘制 Fare 与 Pclass，Sex 与 Pclass，Pclass 与 Survived，Sex 与 Survived，Age 与 Survived，SibSp 与 Survived，Parch 与 Survived，Fare 与 Survived，Embarked 与 Survived 的散点图或者计数图，可得以下结论：

- 费用与船舱等级具有较高的相关性，费用越多，船舱等级可能越高，比尔森系数为-0.55；
- 各个船舱，男性乘客均多余女性乘客；

- 一等舱的幸存率最高，三等舱幸存率最低，幸存率与船舱有重要关系；
- 女性的生存率为 74.2%, 远高于男性的 18.9%，性别与生存与否具有重要关系；
- 孩子和老人的生存率明显高于死亡率，而成年的死亡率明显高于生存率，年龄与生存与否具有重要关系；
- 船费更高的倾向于更高的的生存率。

其中 Sex 与 Survived 计数图，Age 与 Survived 的条形图如 3.2所示，其他的二维数据图见附录 3.4.2。

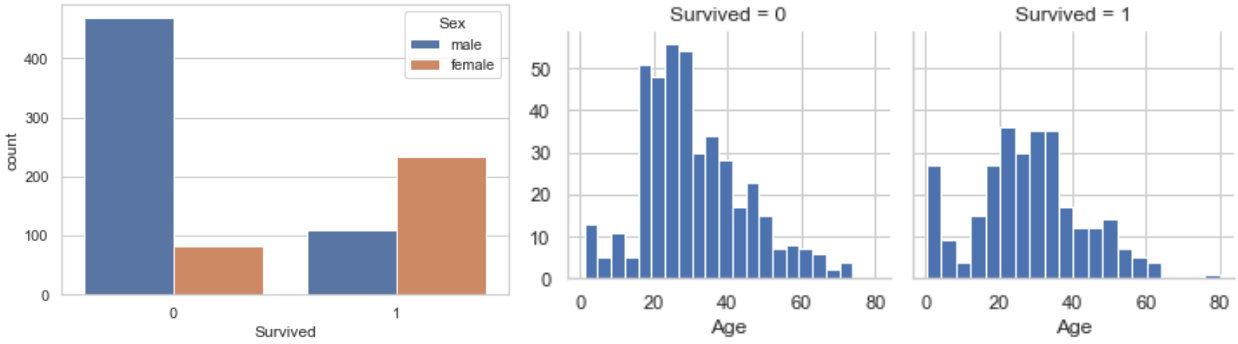

图 3.2　Sex 与 Survived 计数图，Age 与 Survived 的条形图

3.3 三维数据探索

通过绘制 Sex 与 Survived,Pclass；Age 与 Survived,Pclass；SibSp 与 Survived,Pclass；Parch 与 Survived,Pclass；Embarked 与 Survived,Pclass 的计数分布条形图，可以得到以下结论：

- 各个等级船舱，女性生存率高于男性; 高等级船舱生存率更高；
- 船舱一，死亡数较少；船舱二，死亡数中等；船舱三，孩子死亡少，成年死亡多

其中 Sex 与 Survived,Pclass、Age 与 Survived,Pclass 如图 3.3：其他三维数据见附录 3.4.3

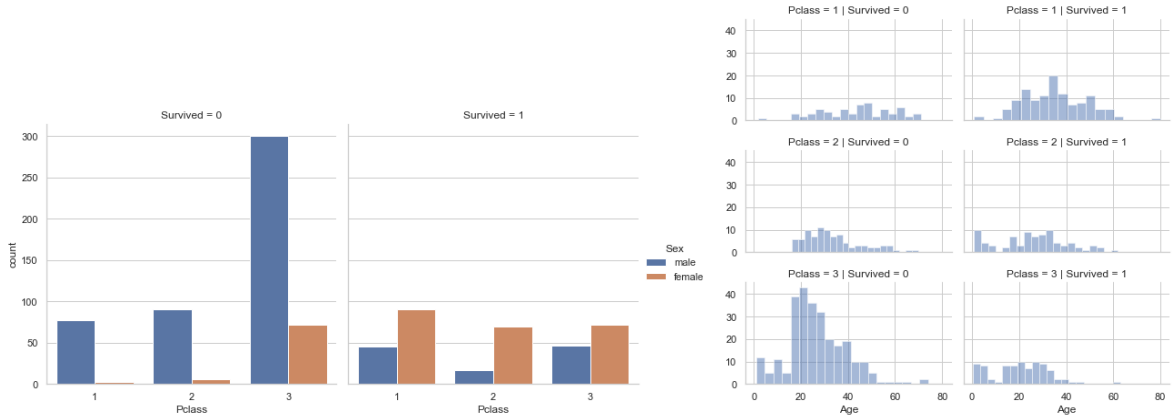

图 3.3　Sex 与 Survived,Pclass 及 Age 与 Survived,Pclass 计数分布条形图

3.4 四维数据探索

通过绘制 Embarked,Sex,Pclass 与 Survived、Fare,Sex,Embarked 与 Survived 计数分布条形图如下：

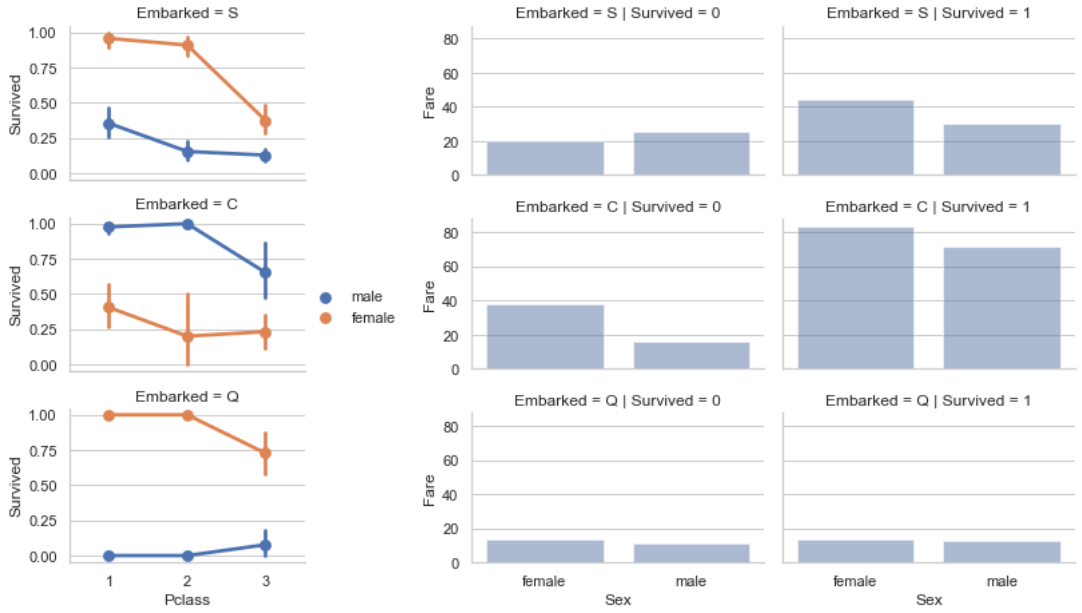

图 3.4　Embarked,Sex,Pclass 与 Survived 及 Fare,Sex,Embarked 与 Survived 计数分布条形图

从 Embarked,Sex,Pclass 与 Survived 计数分布条形图中可以看出 Q 上岸的人，女性生存率高于男性；一等舱，二等舱的生存率比三等舱高相关性较弱；从 Fare,Sex,Embarked 与 Survived 计数分布条形图中可以看出幸存的人倾向于更高的船费；Q 上岸的人倾向于更高的船费；男性倾向于更高的船费。

4 特征工程

4.1 数据预处理

(1) 剔除无用数据 ticket,PassengerId 信息无用，cabin 缺失值过多，将二者剔除

(2)Sex 装换为值 Sex 是一个字符串型的数据，对其进行编码，男性赋值为 0.5，女性赋值为-0.5。

(3)Embarked 缺失值填充并装换为值 考虑到 Embarked 只缺失了两个数据，通过查找相似数据的众数，确定 C 作为缺失数据的填充值，C 赋值为-0.4，S 赋值为 0.1，Q 赋值为 0.6

(4) 合并相似属性 SibSp 与 Parch 表示的都是亲属，将其合并，建立新属性 $family$，表示亲戚的数量。三者的关系为：

$$family = SibSp + Parch + 1$$

(5) 提取 Name 的有效信息 Name 中包含了特定的称呼，称呼反映了乘客的性别，年龄，职业等信息。首先提取 Name 中的称呼并进行计数，发现 99% 的称呼为 Miss，Mr，Mrs，Master。因此将称呼分为以下五类并进行赋值：

$$"Mr" = 1, "Miss" = 2, "Mrs" = 3, "Master" = 4, "other" = 5$$

4.2 缺失值处理

4.3 Fare 的填充及其归一化

(1) 填充缺失数据 测试集中的 Fare 缺失一个数据，使用中位数对其填充。

(2) 归一化 对数据进行归一化，公式如下：

$$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

4.4 age 的预测填充及其归一化

age 中缺失的数据较多，因此考虑建立神经网络对其进行训练和预测。提取训练集和测试集中 age 非缺失的数据作为训练集，缺失数据作为预测集。神经网络输入包括 Pclass，Sex，Age，family，Fare，Embarked，输出为 Age，搭建的神经网络及数据集划分见附录 3.5.2，对数据进行训练和预测。

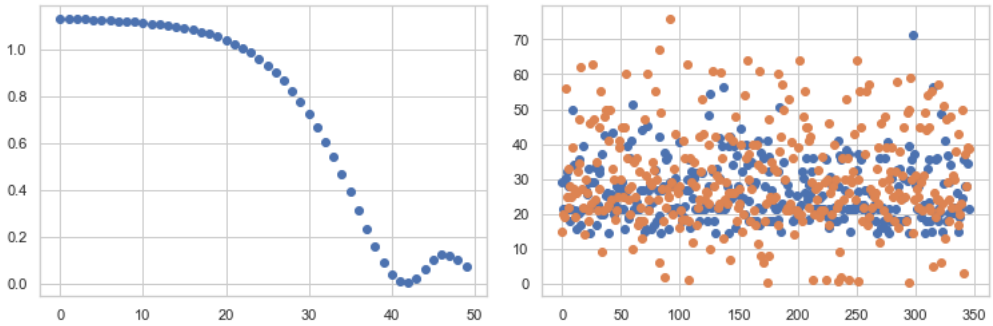

图 4.1　损失值变化及在测试集上的效果

从图中可以看出，损失值随着训练次数的增多而逐渐下降。神经网络在测试集上的预测效果相比真实数据相对集中，可以认为神经网络预测效果良好。

4.5 模型训练与模型验证

选用逻辑回归，SVC，随机森林，KNN，朴素贝叶斯，感知机，SGD 和决策树 9 种算法，使用训练集的前 700 个数据作为训练样本，剩余数据作为测试样本，9 个模型在测试集上的得分如图 4.2, 模型代码见附录 3.5.3。

	Model	Score
0	Support Vector Machines	85.34
7	Linear SVC	82.72
2	Logistic Regression	82.20
4	Naive Bayes	81.68
3	Random Forest	81.15
1	KNN	80.10
5	Perceptron	79.58
6	Stochastic Gradient Decent	78.01
8	Decision Tree	75.92

图 4.2　各个模型得分情况

从结果可以看出，支持向量机的得分最高，准确率为 85.34%

5 模型调参 (网格搜索)

选用随机森林，KNN，决策树，支持向量机等受参数影响较大的模型，通过网格搜索的思想对其进行优化。最终优化结果如图 5.1, 代码见附录 3.6。

模型	优化前	优化后	提升效果	最佳参数
随机森林	81.15%	83.77%	2.62%	n_estimators=10, max_features=3
KNN	80.10%	83.77%	3.67%	n_neighbors = 7
决策树	75.92%	85.34%	9.42%	max_depth=3
支持向量机	85.34%	85.86%	0.52%	C=0.5

图 5.1　参数优化情况情况

6 模型融合 (Stacking)

选用随机森林，KNN，决策树，支持向量机, 逻辑回归并使用最优参数搭建 Stacking 模型，Stacking 模型的示意图如下：

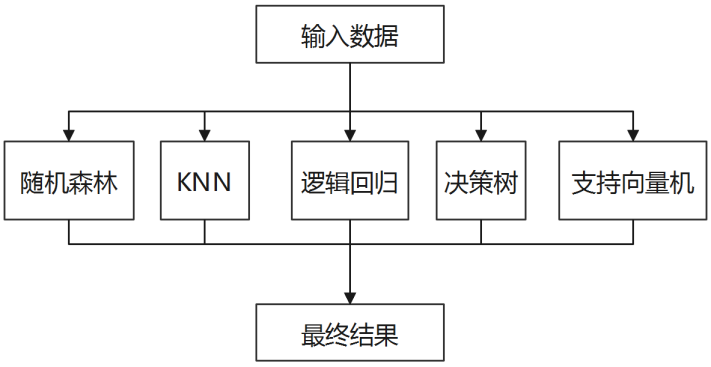

图 6.1　Stacking 模型

使用 Stacking 对模型进行融合的准确率为 85.86%。最终使用 Stacking 模型对测试集进行预测，得出测试集上乘客的幸存与否，提交结果。

7 结论

根据数据分析以及模型的预测效果，可以发现在泰坦尼克号乘船事故中，女性、孩子和老人的生还率远远大于男性和成年。一等舱的生还率大于二等舱和三等舱。根据查阅资料显示，在泰坦尼克号发生事故后，船上首先进行了妇孺的疏散，因此妇孺具有更大的机会幸存，本文的分析结果与实际情况相同。

本文通过数据分析与机器学习算法，对泰坦尼克号数据进行数据预处理，数据可视化，特征工程，模型调参，模型优化，有效提取了泰坦尼克号数据中的信息，并对幸存者建立了预测模型，模型准确率高达 85.86%。

参考文献

[1] 周志华. 机器学习 [J]. 清华大学出版社, 2016, 8(28): 1–415.