به نام خداوند بخشنده مهربان

عنوان پروژه

پیش‌بینی نجات مسافران تایتانیک با استفاده از رگرسیون لجستیک

عنوان درس

یادگیری ماشین

استاد

دکتر الهام قصرالدشتی

دستیاران آموزشی

مهرداد قصابی

مریم صفوی

گردآورنده

سید حسین حسینی

بهار ۱۴۰۴

دانشکده مهندسی کامپیوتر

دانشگاه اصفهان

فهرست مطالب:

1. مقدمه
2. بارگذاری داده‌ها و کتابخانه‌ها
    - 2.1. کتابخانه‌های مورد استفاده
    - 2.2. بارگذاری مجموعه داده‌های آموزشی و آزمون
3. بررسی و توصیف داده‌ها (EDA بخش اول)
    - 3.1. ابعاد و اطلاعات کلی داده‌ها
    - 3.2. آمار توصیفی ویژگی‌های عددی
    - 3.3. آمار توصیفی ویژگی‌های دسته‌ای
    - 3.4. بررسی مقادیر گمشده
4. مصورسازی داده‌ها (EDA بخش دوم)
    - 4.1. تحلیل ویژگی‌های عددی
        - 4.1.1. توزیع داده‌ها (هیستوگرام و نمودار چگالی)
        - 4.1.2. شناسایی داده‌های پرت (نمودار جعبه‌ای)
        - 4.1.3. بررسی روابط دو به دو (نمودار جفتی)
    - 4.2. تحلیل ویژگی‌های دسته‌ای
        - 4.2.1. فراوانی دسته‌ها (نمودار میله‌ای شمارشی)
5. تحلیل همبستگی و انتخاب ویژگی‌های اولیه
    - 5.1. ماتریس همبستگی ویژگی‌های عددی

- 5.2. شناسایی ویژگی‌های با همبستگی پایین
- 5.3. حذف دستی ستون‌های غیرمفید

6. رسیدگی به داده‌های گمشده
   - 6.1. پر کردن مقادیر گمشده در ویژگی‌های عددی
   - 6.2. پر کردن مقادیر گمشده در ویژگی‌های دسته‌ای
7. آماده‌سازی داده‌ها برای مدل
   - 7.1. جداسازی ویژگی‌ها و متغیر هدف
   - 7.2. استانداردسازی ویژگی‌های عددی
   - 7.3. رمزگذاری یک-هات ویژگی‌های دسته‌ای
8. پیاده‌سازی و آموزش مدل رگرسیون لجستیک
   - 8.1. ساختار کلاس LogisticRegressionScratch
   - 8.2. آموزش مدل و بررسی همگرایی
9. پیش‌بینی و ایجاد فایل خروجی

1. مقدمه

فاجعه تایتانیک یکی از مشهورترین حوادث دریایی تاریخ است. مجموعه داده مسافران تایتانیک به طور گسترده‌ای در زمینه علم داده و یادگیری ماشین برای وظایف طبقه‌بندی (Classification) مورد استفاده قرار می‌گیرد. هدف این پروژه، استفاده از تکنیک‌های تحلیل داده و یادگیری ماشین برای ساخت مدلی است که بتواند احتمال زنده ماندن یک مسافر را بر اساس ویژگی‌های او پیش‌بینی کند. در این پروژه، تمرکز بر پیاده‌سازی مدل رگرسیون لجستیک از ابتدا (بدون استفاده مستقیم از کتابخانه‌های آماده مانند Scikit-learn برای خود مدل) است.

2. بارگذاری داده‌ها و کتابخانه‌ها

2.1. کتابخانه‌های مورد استفاده

برای انجام این پروژه، از کتابخانه‌های استاندارد پایتون در حوزه علم داده استفاده شده است:

- pandas: برای کار با DataFrame ها و خواندن و نوشتن فایل‌های CSV.
- numpy: برای عملیات عددی و کار با آرایه‌ها.
- seaborn و matplotlib.pyplot: برای مصورسازی داده‌ها و ترسیم نمودارها.
- sklearn.preprocessing (شامل StandardScaler و OneHotEncoder): برای پیش‌پردازش ویژگی‌ها.
- sklearn.compose (شامل ColumnTransformer): برای اعمال تبدیل‌های مختلف به ستون‌های مختلف.
- sklearn.pipeline (شامل Pipeline): برای ساخت زنجیره‌ای از مراحل پیش‌پردازش.
- Io: برای عملیات ورودی/خروجی (مانند گرفتن خروجی info() در یک رشته).
- Warnings: برای مدیریت نمایش هشدارها.

2.2. بارگذاری مجموعه داده‌های آموزشی و آزمون

داده‌ها از دو فایل CSV مجزا خوانده می‌شوند:

- train.csv: شامل ویژگی‌های مسافران و ستون هدف (Survived) که برای آموزش مدل استفاده می‌شود.
- test.csv: شامل ویژگی‌های مسافران که برای پیش‌بینی و ارزیابی مدل (در مسابقات Kaggle) استفاده می‌شود. ستون هدف در این مجموعه داده وجود ندارد.

```
Training DataFrame successfully read from file '/content/Titanic/train.csv'.
First few rows of the training DataFrame:
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S
```

3. بررسی و توصیف داده‌ها (EDA - بخش اول)

3.1. ابعاد و اطلاعات کلی داده‌ها

- مجموعه داده آموزشی دارای 891 سطر و 12 ستون است.
- مجموعه داده آزمون دارای 418 سطر و 11 ستون است (فاقد ستون Survived).

```
============================================================
          Training Data Description
============================================================

1. Shape (Rows, Columns) of Training data:
(891, 12)

2. Basic Information for Training data (Columns, Non-Null Counts, Dtypes):
<class 'pandas.core.frame.DataFrame'>
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
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

3.2.آمار توصیفی ویژگی‌های عددی

آمار توصیفی (میانگین، انحراف معیار، چارک‌ها و غیره) برای ویژگی‌های عددی مانند PassengerId, Survived, Pclass, Age, SibSp, Parch, Fareمحاسبه شد.

3. Numerical Features Summary Statistics for Test data:

	PassengerId	Pclass	Age	SibSp	Parch	Fare
count	418.00	418.00	332.00	418.00	418.00	417.00
mean	1100.50	2.27	30.27	0.45	0.39	35.63
std	120.81	0.84	14.18	0.90	0.98	55.91
min	892.00	1.00	0.17	0.00	0.00	0.00
25%	996.25	1.00	21.00	0.00	0.00	7.90
50%	1100.50	3.00	27.00	0.00	0.00	14.45
75%	1204.75	3.00	39.00	1.00	0.00	31.50
max	1309.00	3.00	76.00	8.00	9.00	512.33

- **مثال تحلیل :**میانگین سن مسافران در داده‌های آموزشی حدود 29.7 سال است. توزیع کرایه دارای چولگی به راست است و مقادیر بسیار بالایی نیز مشاهده می‌شود.

3.3.آمار توصیفی ویژگی‌های دسته‌ای

برای ویژگی‌های دسته‌ای مانند Name, Sex, Ticket, Cabin, Embarked، تعداد مقادیر یکتا، رایج‌ترین مقدار و فراوانی آن بررسی شد.

4. Categorical/Object Features Summary Statistics for Test data:

	Name	Sex	Ticket	Cabin	Embarked
count	418	418	418	91	418
unique	418	2	363	76	3
top	Peter, Master. Michael J	male	PC 17608	B57 B59 B63 B66	S
freq	1	266	5	3	270

- **مثال تحلیل :**جنسیت مسافران (Sex) دارای دو دسته (male, female) است و بیشتر مسافران مرد بوده‌اند. ستون Embarkedدارای سه بندر اصلی است که بیشترین تعداد مسافران از بندر Sسوار شده‌اند.

3.4.بررسی مقادیر گمشده

تعداد مقادیر گمشده برای هر ستون محاسبه و نمایش داده شد.

```
5. Missing Values per Column in Test data:
Age       86
Fare       1
Cabin    327
dtype: int64
```

- **مجموعه داده آموزشی:**
  - Age: 177مقدار گمشده.
  - Cabin: 687 مقدار گمشده (تعداد بسیار زیاد).
  - Embarked: 2 مقدار گمشده.
- **مجموعه داده آزمون:**
  - Age: 86 مقدار گمشده.
  - Fare: 1 مقدار گمشده.
  - Cabin: 327مقدار گمشده.

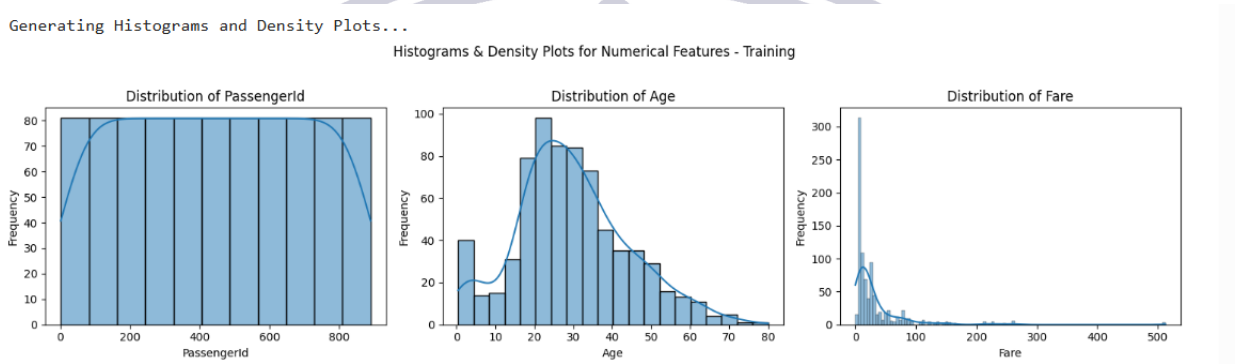
4.مصورسازی داده‌ها (EDAبخش دوم)

برای درک بهتر توزیع داده‌ها و روابط بین آن‌ها، از نمودارهای مختلف استفاده شد.

4.1.تحلیل ویژگی‌های عددی

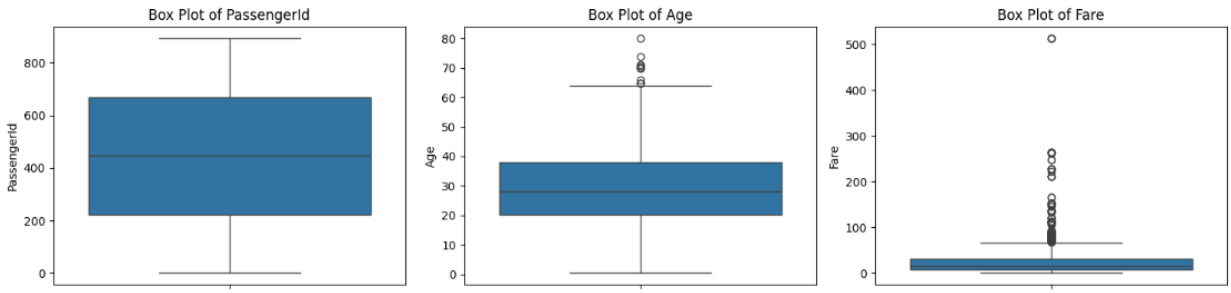
4.1.1.توزیع داده‌ها (هیستوگرام و نمودار چگالی)

نمودار هیستوگرام و KDE برای ویژگی‌های عددی مانند (PassengerIdکه بعداً حذف می‌شود(، Ageو Fareترسیم شد.



4.1.2.شناسایی داده‌های پرت (نمودار جعبه‌ای)

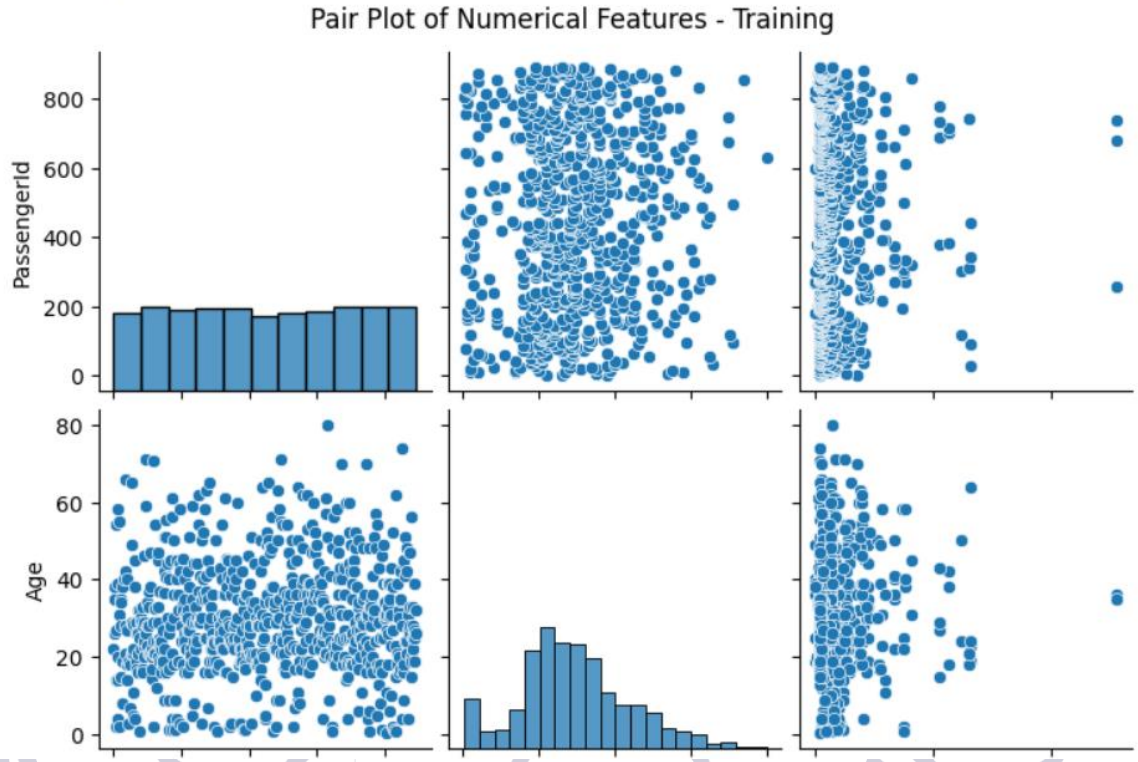
نمودار جعبه‌ای برای PassengerId, Ageو Fareبه منظور شناسایی داده‌های پرت ترسیم شد.



4.1.3.بررسی روابط دو به دو (نمودار جفتی)

نمودار جفتی برای ویژگی‌های عددی ترسیم شد تا روابط بین آن‌ها و توزیع هر کدام به صورت همزمان نمایش داده شود.

```
Generating Pair Plot...
```

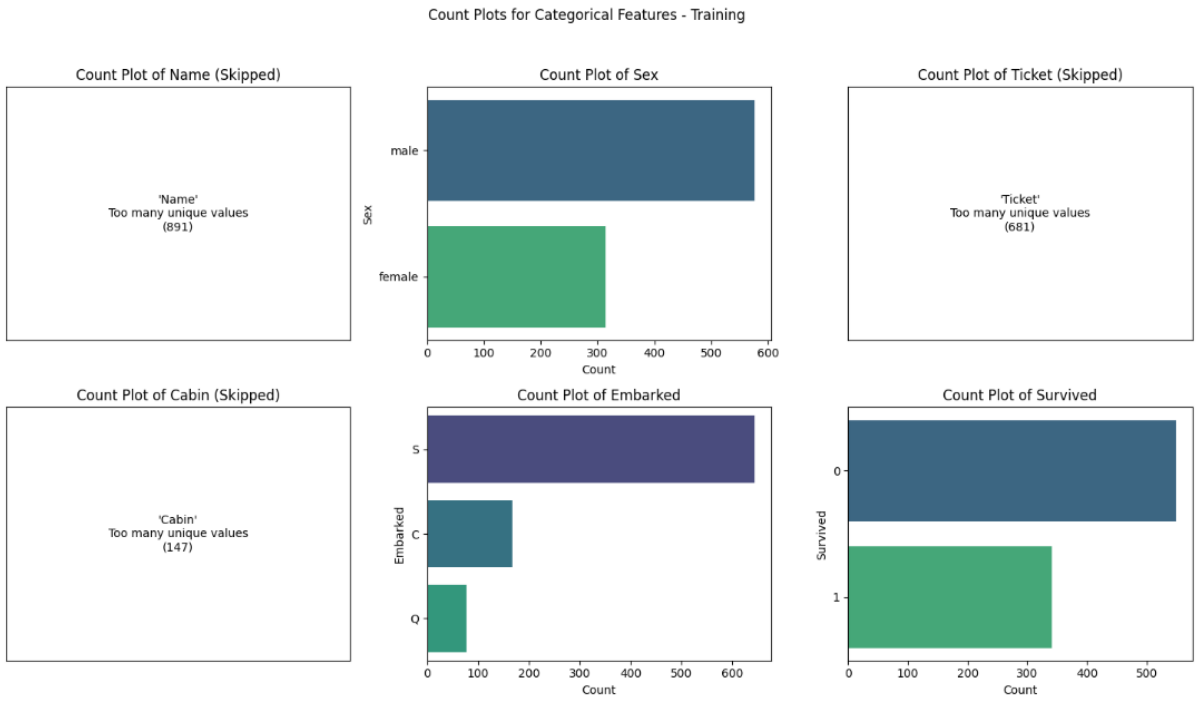


4.2.تحلیل ویژگی‌های دسته‌ای

4.2.1.فراوانی دسته‌ها (نمودار میله‌ای شمارشی)

نمودار میله‌ای شمارشی برای ویژگی‌های دسته‌ای مانند Sex, Embarked, Survived, Pclass, SibSp, Parchترسیم شد.

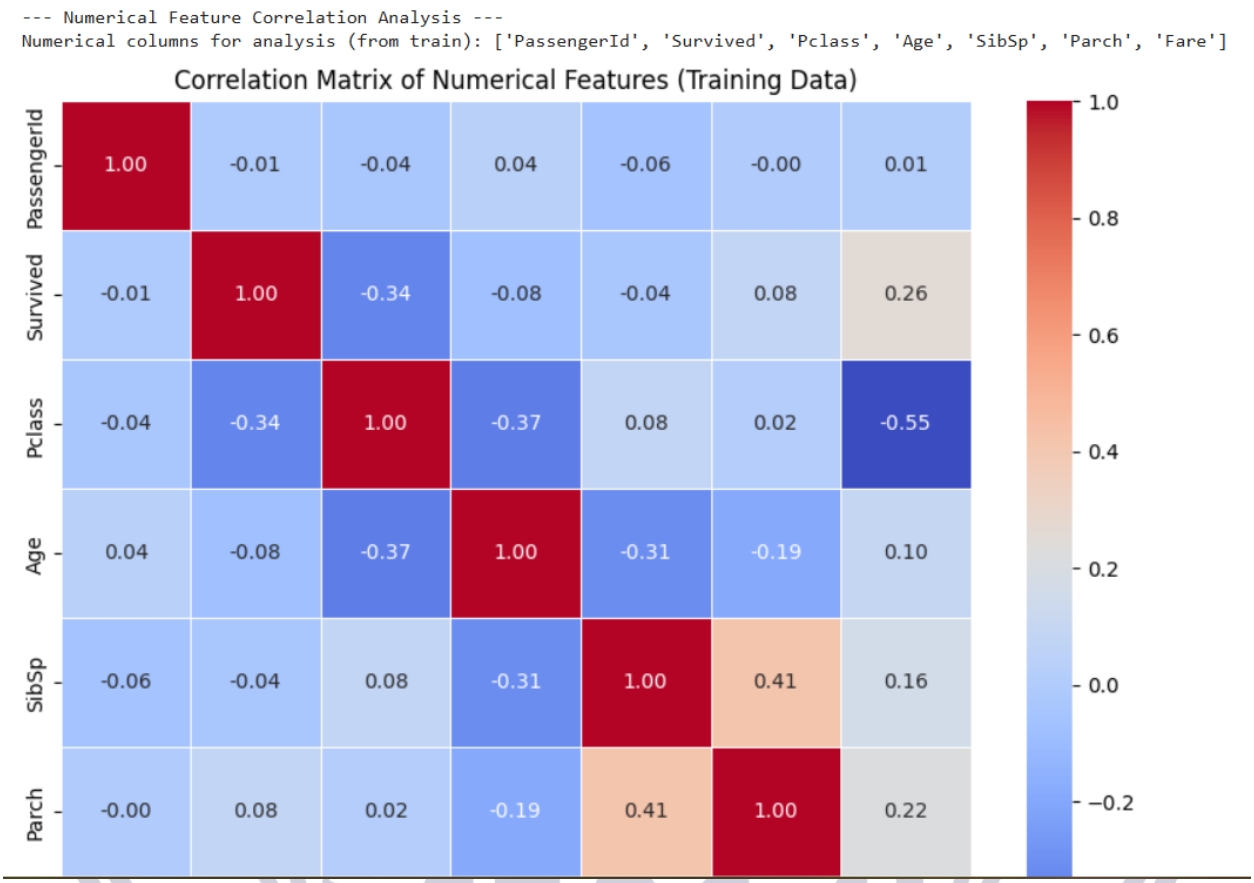
```
Generating Count Plots...
  Skipping count plot for 'Name' (too many unique values: 891).
  Skipping count plot for 'Ticket' (too many unique values: 681).
  Skipping count plot for 'Cabin' (too many unique values: 147).
```



5.تحلیل همبستگی و انتخاب ویژگی‌های اولیه

5.1.ماتریس همبستگی ویژگی‌های عددی

ماتریس همبستگی بین ویژگی‌های عددی در داده‌های آموزشی محاسبه و با استفاده از نقشه حرارتی نمایش داده شد.



تحلیل همبستگی با ستون هدف(Survived) :

- Pclass:همبستگی منفی قابل توجهی با Survivedدارد (حدود −0.34)، به این معنی که مسافران کلاس‌های پایین‌تر شانس کمتری برای نجات داشته‌اند.
- Fare:همبستگی مثبت متوسطی با Survivedدارد (حدود 0.26)، نشان‌دهنده شانس بیشتر نجات برای مسافرانی که کرایه بیشتری پرداخت کرده‌اند.
- Parchو Ageهمبستگی‌های ضعیف‌تری با Survivedنشان می‌دهند.

- SibSpو PassengerIdهمبستگی بسیار پایینی با Survivedدارند.

5.2. شناسایی ویژگی‌های با همبستگی پایین

ویژگی‌های عددی که قدر مطلق همبستگی آن‌ها با ستون هدف کمتر از آستانه 0.01 بود، شناسایی شدند. در این مورد، PassengerIdاین شرط را داشت.

5.3. حذف دستی ستون‌های غیرمفید

بر اساس تحلیل‌های اولیه و دانش دامنه، ستون‌های زیر به دلیل عدم ارائه اطلاعات مفید برای پیش‌بینی یا داشتن تعداد زیادی مقادیر یکتا (که مدیریت آن‌ها در مدل ساده دشوار است) برای حذف انتخاب شدند:

- Ticket: دارای تعداد زیادی مقدار یکتا و فرمت‌های مختلف.
- Cabin: دارای تعداد بسیار زیادی مقدار گمشده و همچنین تعداد زیادی مقدار یکتا.
- PassengerId: یک شناسه یکتا است و برای مدل‌سازی مفید نیست.
- Name: اگرچه ممکن است اطلاعاتی مانند عنوان (Title) از آن استخراج شود، اما در این پیاده‌سازی ساده، حذف شده است.

ستون‌های شناسایی شده (Cabin, Name, PassengerId, Ticket) از هر دو مجموعه داده آموزشی و آزمون حذف شدند.

6. رسیدگی به داده‌های گمشده

پس از حذف ستون‌های اولیه، به مقادیر گمشده در ستون‌های باقیمانده رسیدگی شد.

6.1. پر کردن مقادیر گمشده در ویژگی‌های عددی

- **Age**:مقادیر گمشده در ستون Age(177 مورد در آموزش، 86 مورد در آزمون) با *میانگین* سن مسافران در مجموعه داده آموزشی (29.70 سال) پر شدند.
- **Fare**:مقدار گمشده در ستون Fareدر مجموعه داده آزمون (1 مورد) نیز با *میانگین* کرایه در مجموعه داده آموزشی (32.20) پر شد.

6.2.پر کردن مقادیر گمشده در ویژگی‌های دسته‌ای

- **Embarked**:مقادیر گمشده در ستون Embarkedدر مجموعه داده آموزشی (2 مورد) با *مد* (رایج‌ترین بندر) که Sبود، پر شدند.

پس از این مرحله، هیچ مقدار گمشده‌ای در مجموعه داده‌های آموزشی و آزمون باقی نماند.

7. آماده‌سازی داده‌ها برای مدل

برای اینکه داده‌ها برای مدل رگرسیون لجستیک قابل استفاده باشند، مراحل زیر انجام شد:

7.1. جداسازی ویژگی‌ها و متغیر هدف

در مجموعه داده آموزشی، ستون Survivedبه عنوان متغیر هدف (y_train_orig) و سایر ستون‌ها به عنوان ویژگی‌ها (X_train_orig) جدا شدند.

7.2.استانداردسازی ویژگی‌های عددی

ویژگی‌های عددی باقیمانده (Pclass, Age, SibSp, Parch, Fare) با استفاده از StandardScaler استانداردسازی شدند. این کار باعث می‌شود که ویژگی‌ها دارای میانگین 0 و انحراف معیار 1 شوند و از تسلط ویژگی‌هایی با مقیاس بزرگتر بر مدل جلوگیری شود.

7.3.رمزگذاری وان-هات ویژگی‌های دسته‌ای

ویژگی‌های دسته‌ای باقیمانده (Sex, Embarked) با استفاده از OneHotEncoderبه فرمت عددی تبدیل شدند. پارامتر handle_unknown='ignore'برای مدیریت دسته‌های جدید احتمالی در داده‌های آزمون استفاده شد.

از ColumnTransformerبرای اعمال این تبدیل‌ها به صورت سازمان‌یافته استفاده گردید. در نهایت، داده‌های آموزشی و آزمون پیش‌پردازش شده به آرایه‌های NumPy تبدیل شدند (X_train_prepared_np, y_train_np, X_test_prepared_np).

8.پیاده‌سازی و آموزش مدل رگرسیون لجستیک

یک مدل رگرسیون لجستیک به صورت سفارشی از ابتدا پیاده‌سازی شد.

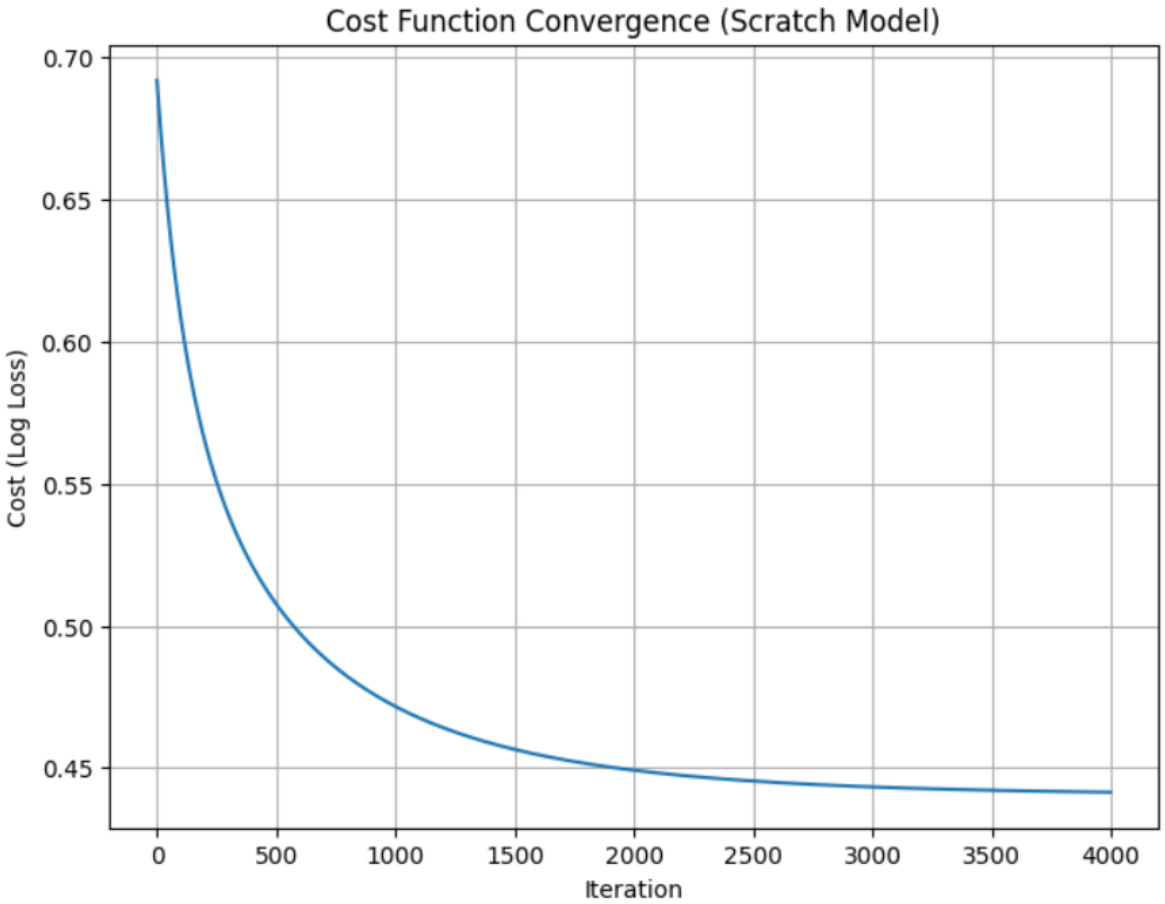
8.1.ساختار کلاس LogisticRegressionScratch

کلاس شامل متدهای زیر است:

- __init__:برای مقداردهی اولیه هایپرپارامترها مانند نرخ یادگیری، تعداد تکرارها، و اینکه آیا عرض از مبدا اضافه شود یا خیر.
- add_intercept_:برای افزودن یک ستون از یک‌ها به ماتریس ویژگی‌ها برای محاسبه عرض از مبدا.
- sigmoid_:برای محاسبه خروجی تابع سیگموئید.
- cost_function_:برای محاسبه تابع هزینه (Log Loss) که معیاری از خطای مدل است.
- Fit: قلب مدل که با استفاده از الگوریتم گرادیان کاهشی، وزن‌های بهینه را برای ویژگی‌ها پیدا می‌کند.
- predict_proba:برای پیش‌بینی احتمال تعلق هر نمونه به کلاس مثبت (نجات یافته).
- Predict:برای پیش‌بینی برچسب کلاس (0 یا 1) بر اساس یک آستانه (پیش‌فرض 0.5).

8.2.آموزش مدل و بررسی همگرایی

مدل LogisticRegressionScratchبا نرخ یادگیری 0.01 و 4000 تکرار بر روی داده‌های آموزشی پیش‌پردازش شده، آموزش داده شد.



همانطور که در نمودار همگرایی تابع هزینه مشاهده می‌شود، با افزایش تعداد تکرارها، مقدار تابع هزینه به تدریج کاهش یافته و به یک مقدار تقریباً ثابت همگرا شده است. این نشان می‌دهد که مدل به خوبی آموزش دیده و وزن‌های مناسبی برای ویژگی‌ها پیدا کرده است. هزینه نهایی حدود 0.4414 بوده است.

9.پیش‌بینی و ایجاد فایل خروجی

پس از آموزش مدل، از آن برای پیش‌بینی احتمال نجات مسافران در مجموعه داده آزمون استفاده شد. برچسب‌های کلاس (0 یا 1) با استفاده از آستانه 0.5 تعیین شدند. سپس، یک DataFrame شامل ستون‌های PassengerId(با شروع از 892) و Survived(پیش‌بینی‌های مدل) ایجاد و در فایلی با نام Saving_Titanic_Passengers_From_Disaster.csvذخیره شد. این فرمت برای ارسال نتایج به مسابقات Kaggle مناسب است.

```
--- Making predictions on test data with the scratch model ---
Sample predictions from the scratch model: [0 0 0 0 1 0 1 0 1 0]

Output file (with PassengerId starting from 892 and column 'Survived') saved to 'Saving_Titanic_Passengers_From_Disaster.cs
v'.
First few rows of the output file:
   PassengerId  Survived
0          892         0
1          893         0
2          894         0
3          895         0
4          896         1
```

10.نتیجه‌گیری و پیشنهادات

در این پروژه، با موفقیت یک مدل رگرسیون لجستیک از ابتدا برای پیش‌بینی نجات مسافران تایتانیک پیاده‌سازی و آموزش داده شد. مراحل پیش‌پردازش داده‌ها، از جمله رسیدگی به مقادیر گمشده و تبدیل ویژگی‌های دسته‌ای، نقش مهمی در آماده‌سازی داده‌ها برای مدل داشتند. تحلیل همبستگی و حذف ویژگی‌های غیرمفید نیز به ساده‌سازی مدل کمک کرد.

11.امتیاز سایت Kaggle

12.LeaderBoard Kaggle

موفق باشید 😊😁