گزارش پروژه نهایی : پیش‌بینی بقای مسافران تایتانیک

محمد عظیم بصیری – علی دنکوب – عرفان طباطبایی

بهمن 1404

گزارش پروژه نهایی : پیش‌بینی بقای مسافران تایتانیک

معرفی مسئله و انتخاب مسیر

مسیر انتخابی : مسیر ۲ – یادگیری ماشین صنعتی (Classic Machine Learning)

تعریف مسئله :

هدف این پروژه، طراحی و پیاده‌سازی یک مدل یادگیری ماشین برای پیش‌بینی زنده‌ماندن یا فوت مسافران کشتی تایتانیک است. ورودی مدل شامل ویژگی‌های دموگرافیک و اطلاعات سفر مسافر مانند کلاس سفر(Pclass) ، جنسیت (Sex) ، تعداد خواهر/برادر و همسر (SibSp) ، تعداد والدین/فرزندان (Parch) و هزینه بلیط (Fare) است . خروجی مدل یک برچسب دودویی شامل ۰ (فوت) و ۱ (زنده‌ماندن) خواهد بود .

معیار اصلی ارزیابی، دقت (Accuracy) مدل بر روی داده‌های آزمون است. با این حال، برای تحلیل دقیق‌تر از معیارهای Precision، Recallو F1-Score نیز استفاده شده است. هدف پروژه دستیابی به دقتی بالاتر از ۸۰ درصد تعیین شد.

---

2. داده‌ها و تحلیل اکتشافی (EDA)

مجموعه داده :
در این پروژه از دیتاست کلاسیک Titanic استفاده شده است که شامل اطلاعات ۸۹۱ مسافر (از مجموع ۲۲۲۴ نفر) می‌باشد. منبع این داده‌ها پلتفرم Kaggle است.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

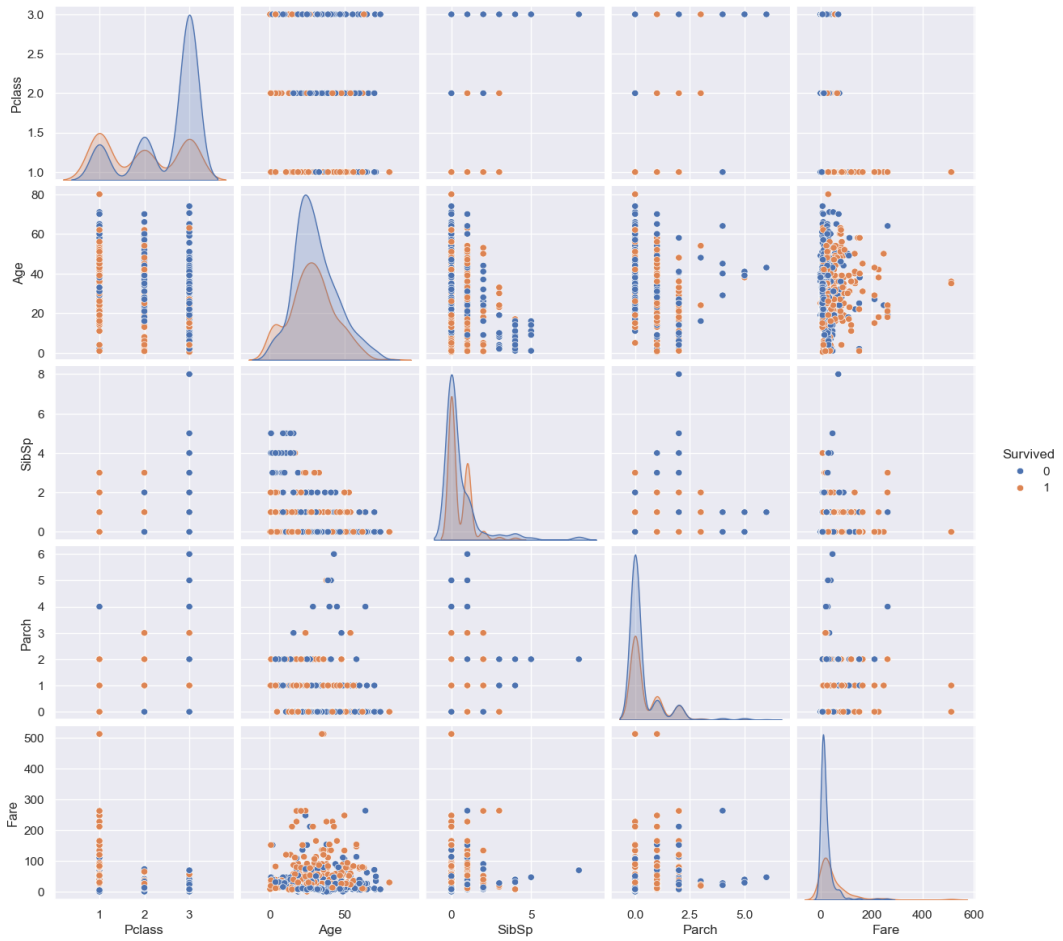
تحلیل داده‌ها :

در فایل eda.ipynb تحلیل جامعی بر روی داده‌ها انجام شد. مهم‌ترین یافته‌ها به شرح زیر است :

مقادیر گمشده : ستون‌های Age و Cabin دارای مقادیر گمشده بودند. با توجه به حجم بالای داده‌های گمشده در ستون Cabin و اهمیت کمتر آن در مدل‌های کلاسیک، این ستون به همراه Name و Ticket حذف گردید.

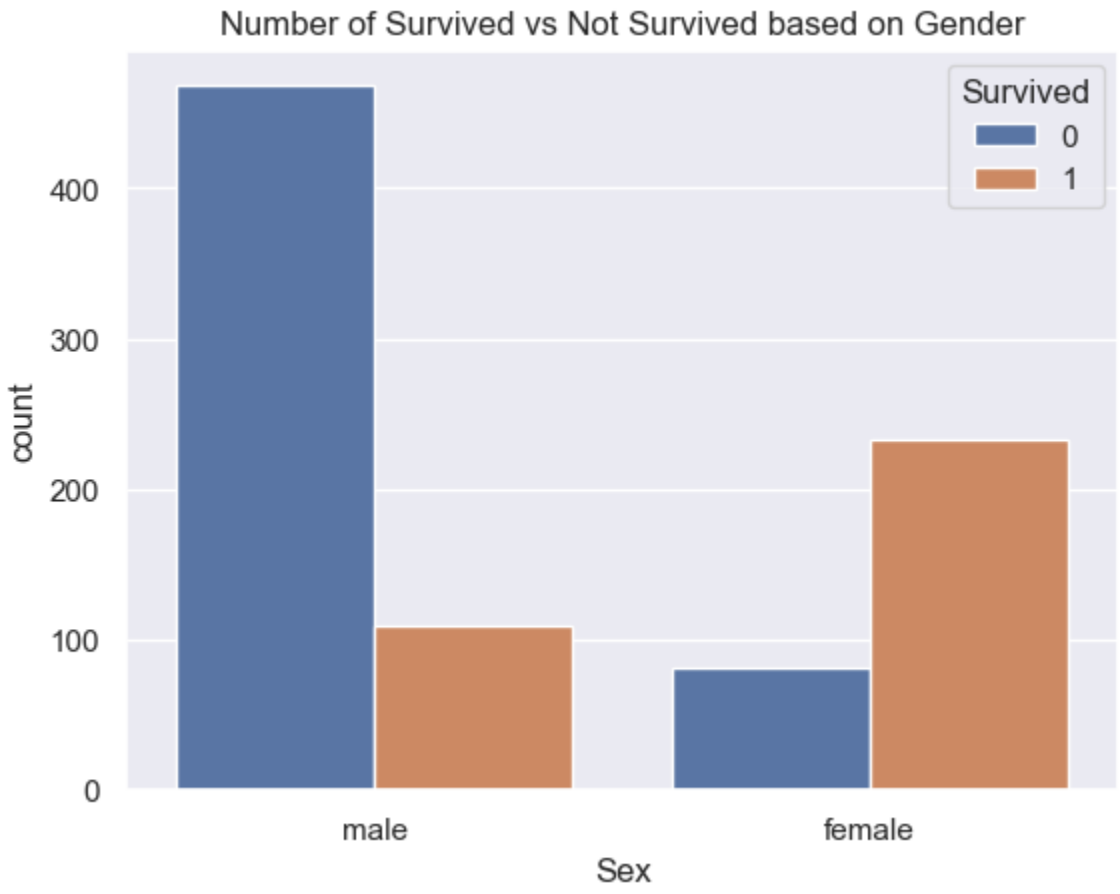
توزیع داده‌ها : نمودارهای زیر ارتباط میان متغیرها و وضعیت بقا را نشان می‌دهند.

نمودار ۱ : ماتریس پراکندگی - (Pairplot) شناسایی روابط
این نمودار روابط جفتی بین تمام ویژگی‌های عددی را نمایش می‌دهد. برای مثال، می‌توان دید که مسافران با Pclass پایین‌تر (کلاس بالاتر) شانس بقای بیشتری داشته‌اند .



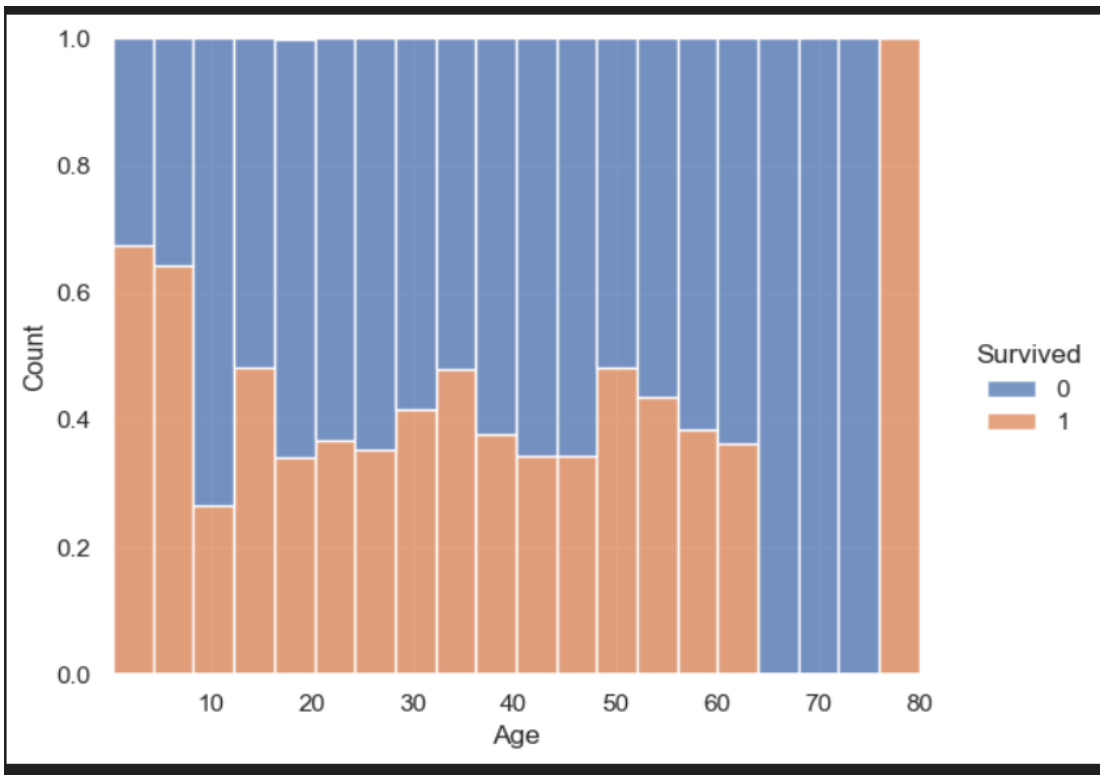
نمودار ۲: توزیع بقا بر اساس جنسیت

این نمودار نشان می‌دهد که زنان احتمال بقای به‌مراتب بیشتری نسبت به مردان داشته‌اند .



نمودار ۳: توزیع بقا بر اساس سن

پس از حذف مقادیر گمشده سن، مشاهده می‌شود که کودکان شانس بقای بالاتری داشته‌اند .



۳ .پیش‌پردازش داده‌ها

مراحل پیش‌پردازش به شرح زیر انجام شد :

1. حذف ستون‌های غیرضروری : ستون‌های PassengerId، Name، Ticket و Cabin حذف شدند.
2. مدیریت مقادیر گمشده : با توجه به تعداد اندک داده‌های گمشده در Embarked و جلوگیری از کاهش حجم داده‌ها، این ستون حذف شد.
3. کدگذاری (Encoding) : ستون Sex از مقادیر متنی به مقادیر عددی (۰ و ۱) تبدیل شد.
4. مقیاس‌بندی(Feature Scaling) : ویژگی‌های عددی با استفاده از StandardScaler نرمال‌سازی شدند.
5. تفکیک ویژگی‌ها و متغیر هدف:
    - Xشامل : Pclass ، Sex، SibSp، Parch، Fare
    - Y شامل : Survived
6. تقسیم داده‌ها : داده‌ها با نسبت ٪۸۰ آموزش و ٪۲۰ آزمون تفکیک شدند (test_size = 0.2).

پس از نرمال‌سازی، ماتریس پراکندگی داده‌ها به شکل زیر تغییر یافت :


Pclass
Sex
SibSp
Parch
Fare
Survived
0
1

داده نرمال‌شده به شکل زیر درآمد یعنی میانگین‌ها نزدیک صفر و انحراف معیار نزدیک 1 خواهند بود .

	Pclass	Sex	SibSp	Parch	Fare	Survived
count	8.910000e+02	8.910000e+02	8.910000e+02	8.910000e+02	8.910000e+02	891.000000
mean	-8.772133e-17	3.987333e-17	4.386066e-17	5.382900e-17	3.987333e-18	0.383838
std	1.000562e+00	1.000562e+00	1.000562e+00	1.000562e+00	1.000562e+00	0.486592
min	-1.566107e+00	-7.376951e-01	-4.745452e-01	-4.736736e-01	-6.484217e-01	0.000000
25%	-3.693648e-01	-7.376951e-01	-4.745452e-01	-4.736736e-01	-4.891482e-01	0.000000
50%	8.273772e-01	-7.376951e-01	-4.745452e-01	-4.736736e-01	-3.573909e-01	0.000000
75%	8.273772e-01	1.355574e+00	4.327934e-01	-4.736736e-01	-2.424635e-02	1.000000
max	8.273772e-01	1.355574e+00	6.784163e+00	6.974147e+00	9.667167e+00	1.000000

در حالیکه پیش از preprocessing به شکل زیر بود.

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200