

دانشکده مهندسی

بخش مهندسی کامپیوتر

درس یادگیری ماشین

گزارش تمرین اول(تایتانیک)

استاد درس:دکتر مهدی افتخاری

دانشجو:محمد امین جهانی

پاییز ۱۴۰۳

**مقدمه**

تایتانیک، کشتی بخار عظیم و لوکسی بود که در ۱۰ آوریل ۱۹۱۲، برای اولین سفر خود از بندر ساوت‌همپتون انگلستان به نیویورک آمریکا به آب انداخته شد. این کشتی، یکی از بزرگترین و پیشرفته‌ترین کشتی‌های زمان خود بود**.**در این تمرین، هدف تحلیل داده‌های مربوط به مسافران کشتی تایتانیک است که در اولین سفر خود در سال ۱۹۱۲ دچار حادثه شد و غرق گردید. این داده‌ها شامل اطلاعاتی نظیر جنسیت، سن، کلاس بلیط، هزینه بلیط و غیره برای هر مسافر است. این تحلیل به ما امکان می‌دهد عواملی را که بیشترین تأثیر را بر نرخ بقا داشته‌اند شناسایی کنیم.

**بررسی دیتاست**

دیتاست‌های مورد استفاده در این پروژه شامل سه فایل با نام‌های train.csv، test.csv و gender\_submission.csv هستند که از داده‌های معروف تایتانیک در سایت Kaggle استخراج شده‌اند. این داده‌ها اطلاعاتی درباره‌ی مسافران کشتی تایتانیک، که در سال ۱۹۱۲ غرق شد، فراهم می‌کنند. فایل train.csv شامل ۸۹۱ نمونه از مسافران با ویژگی‌هایی چون جنسیت، سن، کلاس سفر، وضعیت بقا و غیره است که برای آموزش مدل‌ها به کار می‌رود. فایل test.csv شامل داده‌های مشابه از مسافران دیگر است که برای آزمایش مدل‌ها به کار می‌رود. فایل gender\_submission.csv به عنوان نمونه‌ای از فایل پیش‌بینی ارائه شده است و نشان می‌دهد که یک پیش‌بینی اولیه بر اساس جنسیت چگونه می‌تواند باشد. این سه دیتاست با هم به تحلیل عوامل مؤثر بر بقا در فاجعه‌ی تایتانیک کمک می‌کنند.

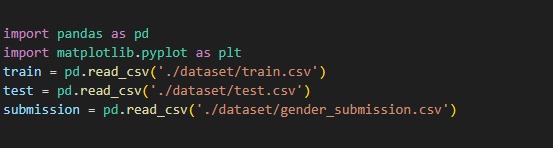
دیتاست تایتانیک شامل ۱۲ ویژگی مختلف برای هر مسافر است که اطلاعاتی درباره وضعیت بقا ، کلاس سفر ، جنسیت ، سن ، تعداد خویشاوندان و ، کرایه‌ی بلیت ، بندر سوار شدن و برخی ویژگی‌های متنی همچون نام و شماره‌ی بلیت را شامل می‌شود. در مجموع، ۸۹۱ نمونه در فایل train.csv و ۴۱۸ نمونه در فایل test.csv وجود دارد.

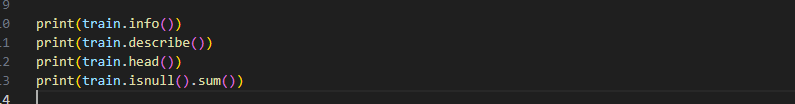
**بررسی کد**

در ادامه به بررسی کد نوشته شده می پردازیم

**۱-وارد کردن کتابخانه‌ها و بارگذاری داده‌ها**

معرفی کتابخانه‌ها:

در این کد، از سه کتابخانه اصلی برای تجزیه و تحلیل داده‌ها استفاده شده است. **Pandas** یکی از مهم‌ترین کتابخانه‌ها در زبان برنامه‌نویسی Python است که به کاربران اجازه می‌دهد تا داده‌ها را به آسانی بارگذاری، پردازش و تجزیه و تحلیل کنند. این کتابخانه به ویژه برای کار با داده‌های جدولی و ساختارهای داده‌ای مشابه، مانند دیتافریم‌ها، طراحی شده است**Matplotlip**  به عنوان یک ابزار قدرتمند برای ترسیم نمودارها و تجسم داده‌ها عمل می‌کند و به کاربر این امکان را می‌دهد تا نتایج تحلیل‌های خود را به صورت گرافیکی و بصری نمایش دهد. ترکیب این سه کتابخانه به تحلیل داده‌ها و استخراج بینش‌های ارزشمند کمک شایانی می‌کند.

**۲-بررسی ساختار داده‌ها**

در اینجا، ساختار دیتاست با استفاده از info() بررسی می‌شود که اطلاعاتی درباره نوع داده‌ها و تعداد مقادیر غیر خالی را نمایش می‌دهد. describe() آمار توصیفی از ویژگی‌های عددی دیتاست ارائه می‌دهد و head() پنج ردیف اول دیتاست را برای مشاهده محتویات آن نشان می‌دهد.

**۳-پرکردن مقادیر گم شده**

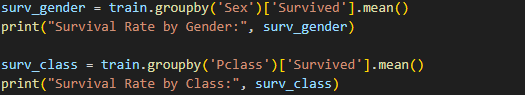


در اینجا، مقادیر گم‌شده در ستون سن (Age) با میانگین سن و مقادیر گم‌شده در ستون بندر سوار شدن (Embarked) با مد (پر تکرارترین مقدار) پر می‌شوند. این کار به حفظ یکپارچگی داده‌ها و جلوگیری از مشکلات ناشی از مقادیر گم‌شده کمک می‌کند.

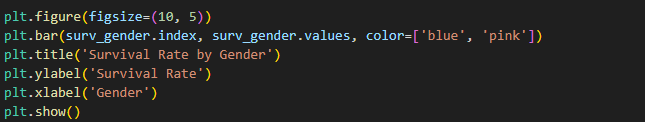
**۴-محاسبه نرخ بقا**

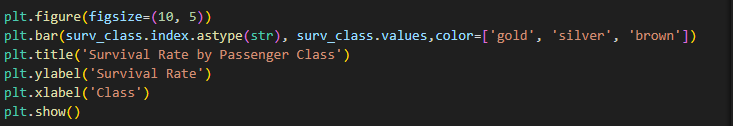
در این بخش، نرخ بقا با محاسبه میانگین ستون Survived به دست می‌آید. این عدد نشان‌دهنده درصد مسافرانی است که در حادثه زنده مانده‌اند.

**۵-تحلیل بقا بر اساس جنسیت، کلاس سفر**

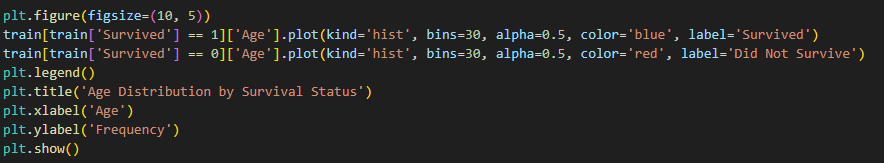
****

**۶-نمودار بقا بر اساس جنسیت**

****

**۷-نمودار بقا بر اساس کلاس سفر  
**

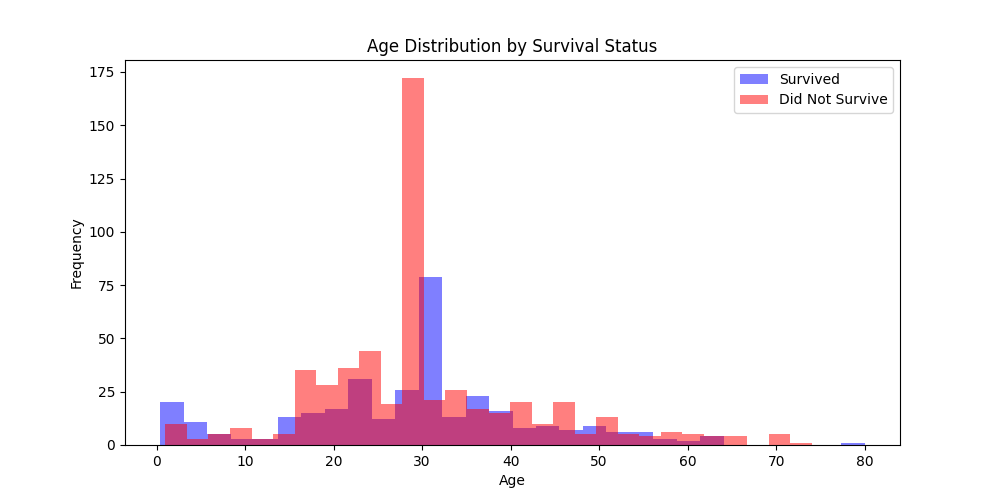
**۸-توزیع سنی بقا**

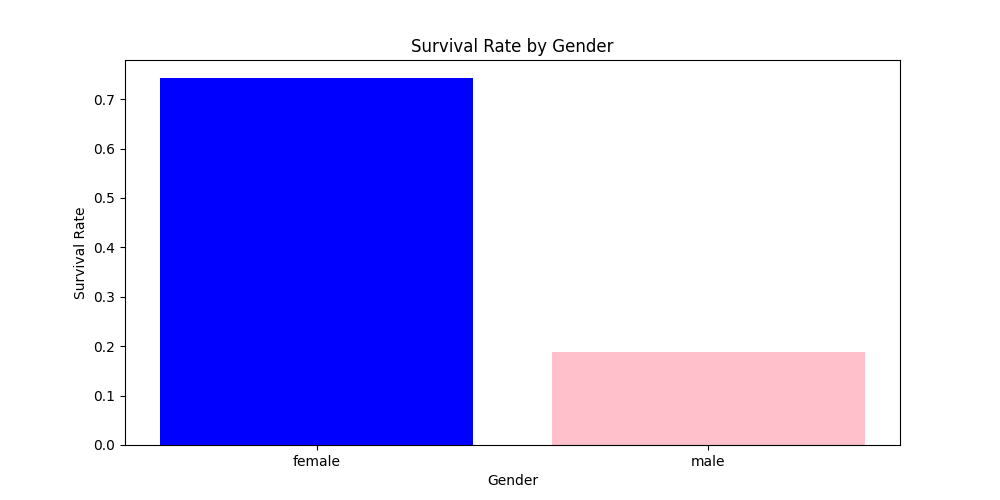
****

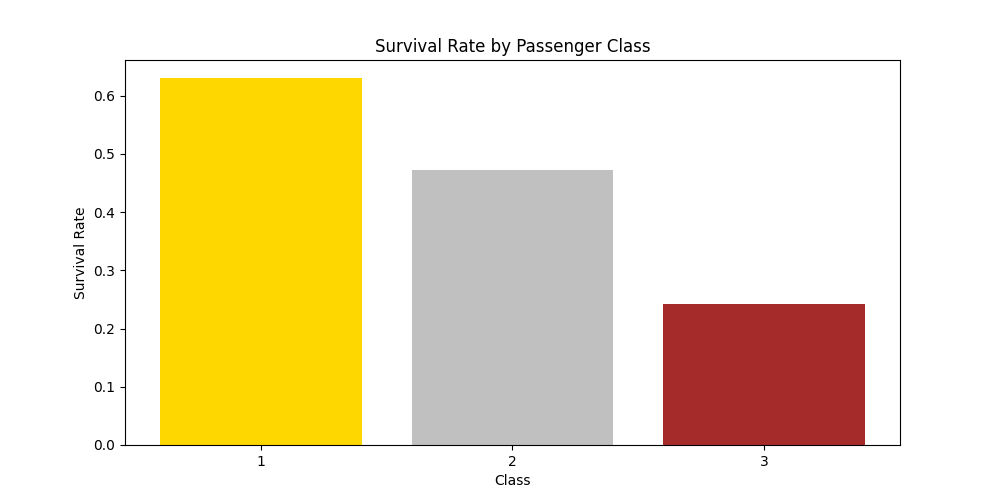
**بررسی تحلیل کد از داده‌ها**

این داده‌ها تحلیل‌های آماری جالبی ارائه می‌دهند. به عنوان مثال، نرخ کلی بقا در دیتاست train.csv حدود ۳۸٪ است که نشان‌دهنده آن است که کمتر از نیمی از مسافران توانسته‌اند زنده بمانند. به علاوه، تجزیه و تحلیل‌های اولیه نشان می‌دهد که جنسیت و کلاس سفر از جمله عوامل اصلی تأثیرگذار بر بقا هستند؛ نرخ بقا برای زنان به مراتب بالاتر از مردان است و مسافران کلاس اول نیز شانس بقا بیشتری نسبت به کلاس‌های دوم و سوم دارند. همچنین، ستون‌های Age و Embarked دارای مقادیری خالی هستند که برای کامل کردن داده‌ها نیاز به روش‌های اکتشافی مانند پرکردن با میانگین یا مد دارند.

در نهایت، بررسی توزیع سنی نشان می‌دهد که بیشتر بازماندگان زیر ۴۰ سال هستند و در میان افرادی که بقا نیافته‌اند، سنین مختلفی دیده می‌شود. این الگوها نشان می‌دهند که ترکیبی از ویژگی‌های سن، جنسیت و کلاس سفر می‌تواند به شناسایی الگوهای بقا در این فاجعه کمک کند.







**منابع**

<https://pandas.pydata.org/>

<https://matplotlib.org/>

<https://openai.com/>