

گزارش تمرین اول درس داده‌کاوی

**«داده‌کاوری یک مسابقه در Kaggle: تایتانیک»**

گردآورنده: سعید دادخواه

استاد: دکتر ناظرفرد

بهمن ۱۳۹۵

# مقدمه

غرق شدن کشتی تایتانیک یکی از بزرگ‌ترین فاجعه‌های تاریخ بوده است. در این مسابقه اطلاعاتی در مورد مسافران این کشتی داده شده است و در مورد برخی از آن‌ها که داده‌های آموزش را تشکیل می‌دهند نجات یافتن آن‌ها یا غرق شدن آن‌ها نیز در این حادثه ذکر شده است. داده‌هایی که اطلاعات مربوط به نجان یافتن افراد را ندارند داده‌های سنجش هستند و دقت الگوریتم نیز با پیش‌بینی نجات یافتن یا غرق شدن آن‌ها سنجیده می‌شود. در این گزارش که مرحله به مرحله یا توجه به آموزش با زبان R موجود در توضیحات مسابقه آورده شده است مراحل و چگونگی پیش‌بینی این موضوع توضیح داده شده است.

# به دست آوردن داده‌ها

داده‌ها در صفحه مسابقه و در بخش داده‌های آن (<https://www.kaggle.com/c/titanic/data>) قابل دسترسی هستند. دو فایل csv موجود در این بخش با نام‌های train.csv و test.csv باید دانلود شوند. برای این که این داده‌ها را بتوان در سایت Kaggle خواند باید آن‌ها را در پوشه قبل از پوشه کد در پوشه جدید به نام input (“../input/train.csv” و “../input/test.csv”) قرار داد.

برای پیش‌بینی باید ساختار داده‌های آموزش و سنجش یکسان باشد به همین دلیل هرکاری که روی داده‌های آموزش انجام می‌گیرد باید روی داده‌های سنجش نیز انجام گیرد. برای سادگی کار این دو دسته را با هم ترکیب می‌کنیم. برای این کار از تابع bind\_rows استفاده می‌کنیم که یکی از توابع پکیج dplyr است.

# مرحله اول مهندسی ویژگی‌ها

## عنوان مسافران

ابتدا عنوان مسافران را با استفاده از عبارت منظم استخراج می‌کنیم سپس تعداد افراد را در یک جدول نسبت به جنسیت و عنوان قرار می‌دهیم. سپس گروه‌هایی که تعداد کمی دارند را با هم ترکیب می‌کنیم.

## نجات یا غرق شدن اعضای خانوانده با هم

ابتدا یک ویژگی جدید برای مسافران با عنوان اندازه خانواده درست می‌کنیم که بر اساس ویژگی‌های قبلی ساخته می‌شود. یک ویژگی دیگر نیز بر همین پایه می‌سازیم که به صورت فاکتور خواهد بود. افرادی که اندازه خانواده‌شان در کشتی برابر یک، دو تا چهار و بیش از چهار نفر به ترتیب به صورت تنها، کوچک و بزرگ برچسب می‌خورن.

## عرشه

با استفاده از ویژگی کابین افراد عرشه آن‌ها مشخص می‌شود. حرف اول کابین عرشه مسافر را مشخص می‌کند.

# پر کردن داده‌های ناقص

## Embarked

ابتدا ویژگی Embarked را مورد بررسی قرار می‌دهیم. در این ویژگی دو رکورد فاقد مقدار هستند. برای پر کردن این مقادیر رکوردهای دیگر را در یک نمودار جعبه‌ای نمایش می‌دهیم. در این نمودار برای هر دسته از مقادیر Embarked و Pclass یک جعبه و جعبه‌ها را بر اساس Fare رسم می‌کنیم. در نمودار به‌دست آمده کاملا قابل مشاهده است که میانگین Fare داده‌هایی که Pclass آن‌ها برابر با مقدار این ویژگی در این دو داده ناقص است مقدار Embark در آن‌ها برابر ‘C’ است پس نقص این داده‌ها را اینگونه رفع می‌کنیم.

## Fare

در این ویژگی فقط یک نقص وجود دارد. برای رفع این ویژگی هیستوگرام داده‌هایی که Pclass و Embarked برابر داده ناقص دارند را رسم می‌کنیم. با توجه به نمودار این مقدار را برابر با مقدار میانه در این داده‌ها قرار می‌دهیم.

## Age

در این ویژگی ۲۶۳ داده ناقص وجود دارد. برای رفع این نقص از پکیج MICE (Multivariate Imputation by Chained Equations) استفاده می‌کنیم. برای این کار دو ویژگی دیگر که یکی نام خانوادگی و دیگر نام خانواده است را نیز به داده‌ها اضافه می‌کنیم و روش rf برای این کار استفاده می‌شود. در نهایت ستون Age از خروجی mice جایگزین ستون اصلی Age می‌شود.

# مرحله دوم مهندسی ویژگی‌ها

## کودکان و مادرها

در این مرحله مسافرانی که سنی کمتر از ۱۸ دارند با عنوان کودک و بقیه با عنوان بزرگسال برچسب می‌خورند و زنانی که فردی با رابطه والد-فرزندی داشته باشند، سنشان بیشتر از ۱۸ باشد و عنوانشان بانو نباشد با عنوان مادر برچسب می‌خورند.

# پیش‌بینی

در این مرحله ابتدا داده‌های آموزشی از داده‌های سنجش جدا می‌شوند سپس یک randomForest آموزش داده می‌شود. برای نمایش اهمیت هرکدام از ویژگی‌ها یک نمودار رسم می‌شود که اهمیت هرکدام از ویژگی‌ها را نمایش می‌دهد. در نهایت نیز داده‌های سنجش نیز به عنوان ورودی به الگوریتم داده می‌شود و خروجی آن‌ها در یک data frame ذخیره و در یک فایل csv به عنوان خروجی قرار داده می‌شوند تا در سایت Kaggle مورد ارزیابی قرار گیرند.

# سوال‌ها

۱. DataFrame چیست؟

DataFrameها ساختمان داده‌هایی برای نگه‌داری داده‌هایی با طول برابر به صورت جدول هستند. در این ساختمان داده هر ستون می‌تواند header و هر سطر یک نام داشته باشد.

۲. مقصود از Imputation چیست؟

داده‌هایی که برای پیش‌بینی استفاده می‌شوند کامل نیستند و تعدادی از ویژگی‌هایشان مقدار ندارند. برای پیش‌بینی بهتر ابتدا باید این نقص‌ها را برطرف کرد. این کار را Imputation می‌نامند.

۳. Normalize کردن داده‌ها به چه منظوری انجام می‌گیرد؟

در صورتی که داده‌ها نرمال نشده باشند می‌توانند روی عملکرد الگوریتم تاثیر بگذارند. برای مثال ممکن است الگوریتم نیاز به زمان بیشتری برای همگرایی داشته باشد. حتی ممکن است عملکرد الگوریتم نیز تضعیف شود.

۴. آیا در درخت تصمیم، تمام ویژگی‌ها دارای اهمیت یکسانی هستند؟

خیر، اهمیت ویژگی‌ها در الگوریتم‌ها متفاوت است و در گزارش فوق نیز یک نمودار برای نمایش اهمیت ویژگی‌ها رسم شد.

۵. در الگوریتم جنگل تصادفی دو پارامتر max\_depth و min\_sample\_split بیانگر چه چیزی هستند؟