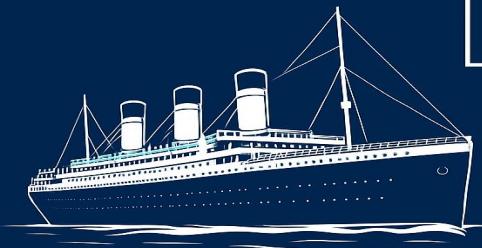


# *Titanic Survival Model*

## **TITANIC SURVIVAL MODEL**

**SURVIVED = 1  
NOT SURVIVED = 0**



درس مدل‌سازی مقدماتی ریاضی

استاد راهنما: دکتر کوشکی

دانشجویان: ارغوان آژیر، سارینا غفوری، آروین عزتی

# چرا تایتانیک؟



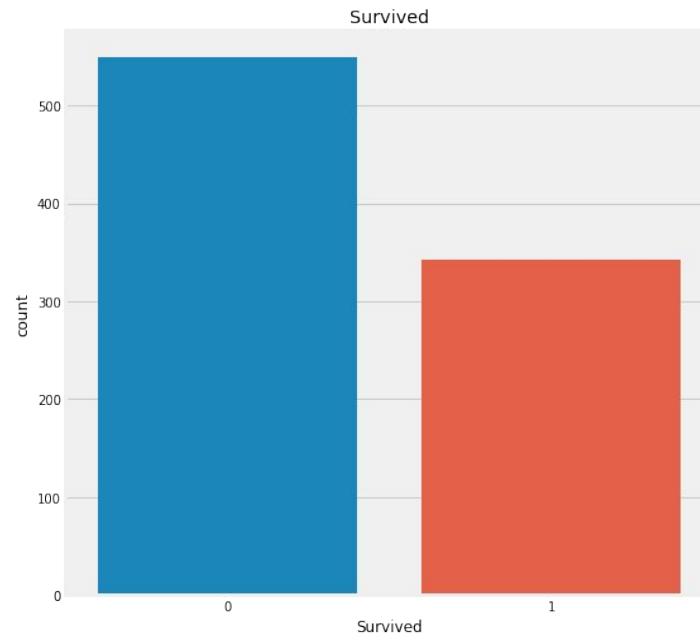
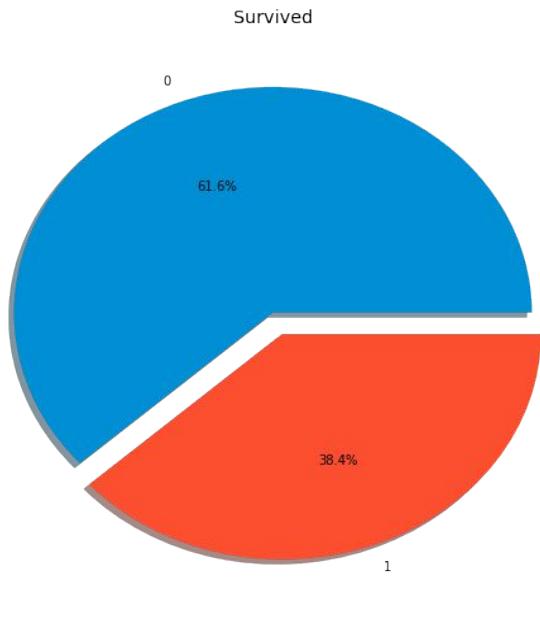
در این پروژه با استفاده از داده‌های تاریخی مسافران و الگوریتم‌های یادگیری ماشین، تلاش می‌کنیم تا مدل‌هایی برای پیش‌بینی احتمال نجات مسافران طراحی و ارزیابی کنیم.

**هدف:** تحلیل داده‌های مسافران و مقایسه عملکرد مدل‌های یادگیری ماشین از جمله رگرسیون لجستیک، درخت تصمیم و جنگل تصادفی در پیش‌بینی بقا.

**متغیر هدف:** پیش‌بینی اینکه شخص زنده مانده یا نمانده (*survived or not survived*)

فوت شده = ۰ ، زنده مانده = ۱

**مهمترین ویژگی:** کلاس مسافری (*Pclass*) ، جنسیت (*Gender*) ، سن (*Age*) ، تعداد همراهان (*SibSp, Parch*) ، قیمت بلیط (*Fare*) ، بندر سوار شدن (*Embarked*)



با توجه به نمودارهای فوق، مشخص است که افراد زیادی از این سانحه جان سالمی به در نبردند. تنها حدود ۳۵۰ نفر از ۸۹۱ مسافر نجات یافتند، یعنی تقریباً ۳۸.۴٪ از کل مسافران تایتانیک نجات یافتند.

اما میزان بقا باید با استفاده از ویژگی‌های مختلفی بررسی شود تا بتوان مدل ریاضی‌ای با کمترین خطا از داده‌های در دسترس به دست آورد.

این ویژگی‌ها به ۳ دسته تقسیم‌بندی می‌شوند:

۱. **اسمی**: به متغیرهایی گفته می‌شود که دارای دو یا چند دسته هستند مثلاً جنسیت یک متغیر اسمی است که شامل زن و مرد می‌شود و هیچ ترتیبی ندارد.
۲. **ترتیبی**: بر عکس مورد قبل این نوع متغیرها دارای ترتیب‌اند اما مقدار عددی نمی‌گیرند مانند کلاس مسافری.
۳. **پیوسته**: یک ویژگی زمانی پیوسته در نظر گرفته می‌شود که بتواند مقداری بین دو نقطه مشخص را پذیرد، مثل سن.

تحلیل تک متغیره هر شاخص (درصدهای مذکور درصد زنده ماندن مسافران تایتانیک هستند):

۱- کلاس مسافری ( $Pclass$ ):

کلاس ۳:٪۲۴.۲

کلاس ۲:٪۴۷.۳

کلاس ۱:٪۶۲.۹

کلاس ۱ | ٪۶۲.۹

کلاس ۲ | ٪۴۷.۳

کلاس ۳ | ٪۲۴.۲

نتیجه: هرچه کلاس مسافری پایین‌تر بوده، دسترسی به قایق‌های نجات نیز کمتر شده.

## ۲- جنسیت (Gender):

زنان: ٪ ۷۴.۲ (۳۱۴ از ۴۳۳)

مردان: ٪ ۱۸.۹ (۵۷۷ از ۳۱۰)

زنان | ٪ ۷۴.۲

مردان | ٪ ۱۸.۹

**نتیجه:** سیاست «نجات زنان اول» باعث به وجود آمدن چنین تفاوت فاحشی شد.

## ۳- سن (Age): کودکان(< ۱۲ سال): ٪ ۵۹.۱

بزرگسالان(> ۱۲ سال): ٪ ۳۸.۴

کودکان | ٪ ۵۹.۱

بزرگسالان | ٪ ۳۸.۴

**نتیجه:** با اینکه نجات کودکان اولویت داشته، اما تفاوت چندان چشمگیر نبوده.

۴- تعداد همراهان (*SibSp, Parch*):

۱-۳ همراه:٪۵۵.۴

≤۴ همراه:٪۲۵.۰

نها:٪۳۰.۴

۱-۳ نفر:٪۵۵.۴

≤۴ نفر:٪۲۵.۰

نتیجه: خانواده‌های کوچکتر شанс بیشتری برای زنده ماندن داشتند (خانواده‌های بزرگ احتمالاً در کلاس ۳ بودند)

۵- قیمت بلیط (*Fare*): میانگین بلیط نجات‌یافتنگان: £۴۸ میانگین بلیط غرقشدگان: £۲۲

نجات‌یافتنگان: (£۴۸ ~)

غرقشدگان: (£۲۲ ~)

درصد نجات	حدوده قیمت بلیط (پوند)	کلاس
۶۲.۹ %	۳۰-۵۱۲	۱
۴۷.۳ %	۱۳-۷۲	۲
۲۴.۲ %	۱-۶۹	۳

**نتیجه:** قیمت بالاتر-> کلاس بالاتر -> دسترسی بهتر به قایق‌های نجات.

مسافران کلاس ۱ که بیشترین قیمت را پرداخت کرده بودند، بالاترین درصد نجات را داشتند.

**رابطه‌ی غیرخطی قیمت بلیط و نجات:**

مسافران با بلیط رایگان (خدمه): اگرچه قیمت بلیط تقریباً صفر بود، اما ۲۴٪ از آنها نجات یافتند. (به دلیل دسترسی به عرضه)

مسافران با گران‌ترین بلیط (£۵۱۲): ۱۰۰٪ نجات یافتند. (همگی از کلاس ۱ و نزدیک به قایق‌ها)

## ۶- بندر سوار شدن *:(Embarked)*

نقطه سوار شدن	تعداد مسافران	نجات یافته‌ها	درصد نجات
<i>Cherbourg (C)</i>	۱۶۸	۹۳	۵۵.۴ %
<i>Southampton (S)</i>	۶۴۴	۲۱۷	۳۳.۷ %
<i>Queenstown (Q)</i>	۷۷	۳۰	۳۹.۰ %

**تفسیر:** بندر C بالاترین درصد نجات را داشته چرا که مسافران این بندر عمدتاً از کلاس ۱ بودند.

بندر S کمترین نجات را داشته چرا که مسافران این بندر عمدتاً از کلاس ۳ (کارگران و مهاجران) بودند.

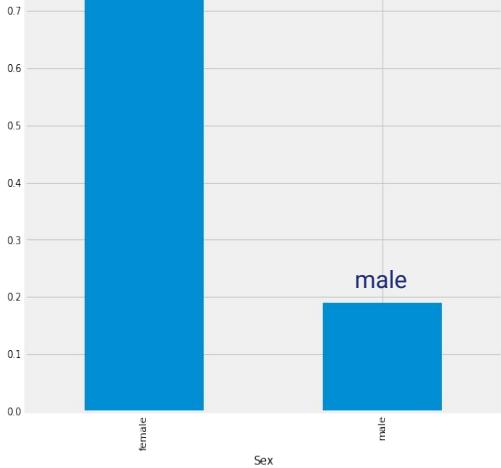
**نکته:** این شاخص به تنها یی علت مستقیم نجات نبود، بلکه پروکسی (نماینده) برای کلاس اجتماعی و اقتصادی مسافران است.

## جنسیت :(Categorical/Nominal Feature)

female

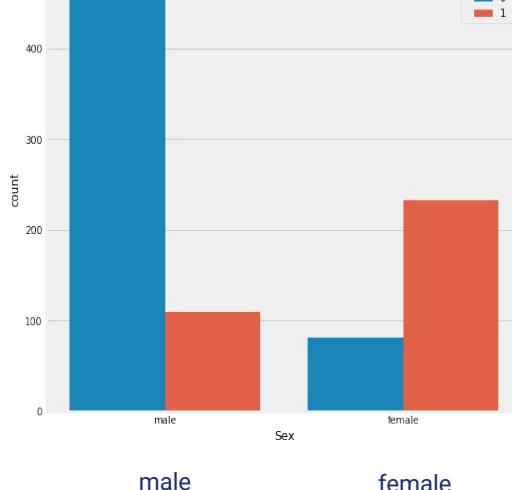
Survived vs Sex

Survived



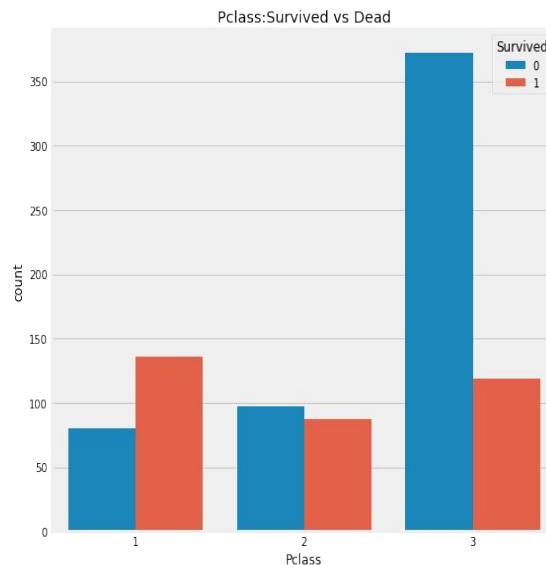
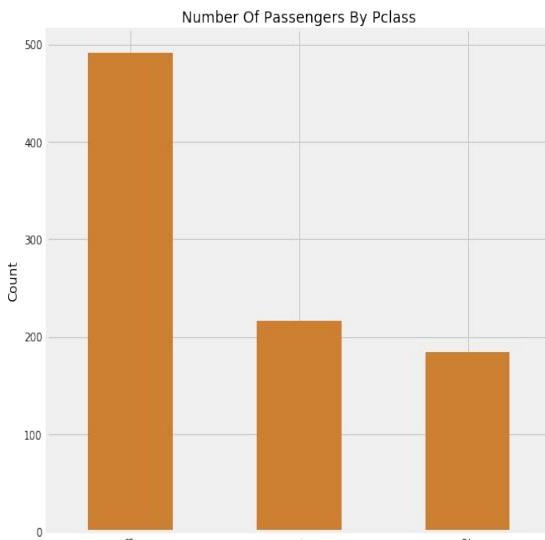
Sex:Survived vs Dead

Survived



با اینکه تعداد مردانی که در کشتی حضور داشتند بیشتر از زنان بوده، اما تعداد زنان نجات یافته تقریباً دو برابر مردان نجات یافته است. احتمال زنده ماندن زنان در کشتی چیزی حدود ۷۵٪ است در حالی که این احتمال برای مردان تنها بین ۱۸ تا ۱۹ درصد است. پس به نظر می‌رسد جنسیت می‌تواند یکی از مهمترین ویژگی‌ها برای مدلسازی باشد.

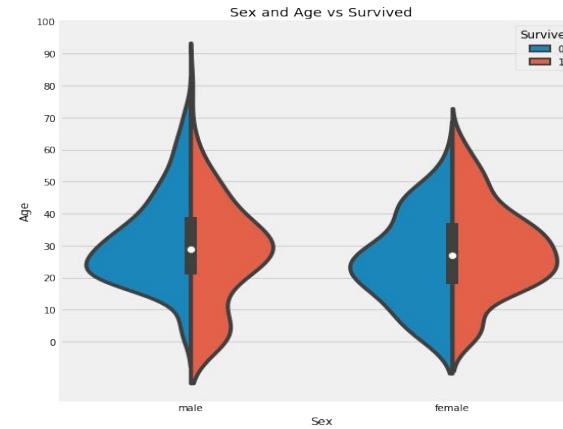
## کلاس مسافری(Pclass):(Ordinal Feature)



مردم می‌گویند با پول نمی‌توان همه چیز را خرید  
اما واضح‌ا در اینجا می‌بینیم که مسافران کلاس ۱  
برای نجات اولویت بودند، اگرچه تعداد مسافران  
کلاس ۳ خیلی بیشتر از آنها بوده اما تعداد نجات  
یافتگان این کلاس، بسیار پایین‌تر هستند، چیزی  
حدود ۲۵٪.

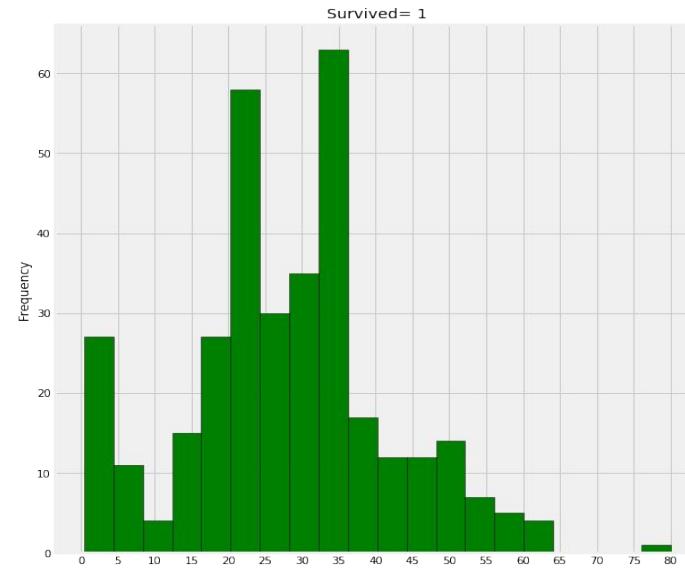
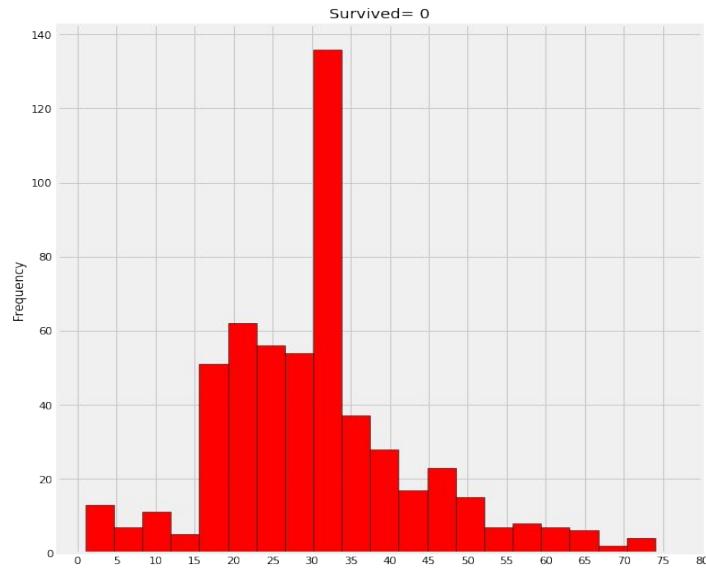
نجات‌یافتگان کلاس ۱، ۶۳٪ و برای کلاس ۲،  
۴۸٪ بوده است.

## سن : (Continuous Feature)



تعداد کودکان در کلاس‌های پایین‌تر بیشتر است و نرخ نجات برای مسافران زیر ۱۰ سال (کودکان) بهم‌طور کلی بالا است،  
صرف‌نظر از اینکه در کدام کلاس قرار دارند.

احتمال نجات برای مسافران ۲۰ تا ۵ ساله در کلاس اول بالا است و این احتمال برای زنان حتی بیشتر هم هست. اما برای مردان با افزایش سن، احتمال نجات کاهش می‌یابد.



تعداد زیادی از کودکان خردسال (زیر ۵ سال) نجات یافته‌اند که نشان‌دهنده اجرای سیاست «زنان و کودکان در اولویت» است. مسن‌ترین مسافر که ۸۰ سال سن داشته، نجات یافته‌است. بیشترین تعداد فوت شدگان مربوط به گروه سنی ۳۰ تا ۴۰ سال بوده است.

## تعداد همراهان (SibSp)

نمودارهای میله‌ای و دسته‌ای (barplot و

(factorplot) نشان می‌دهند مسافرانی که تنها بودند (بدون خواهر، برادر یا همسر) حدود ۴۲.۵٪ احتمال زنده ماندن داشتند.

هرچقدر تعداد اعضای خانواده (SibSp) افزایش پیدا کند، نرخ زنده ماندن کاهش پیدا می‌کند.

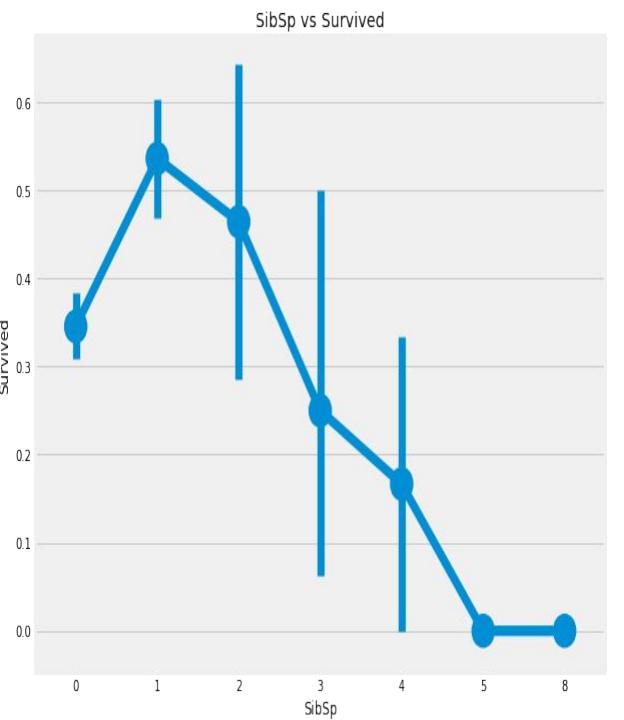
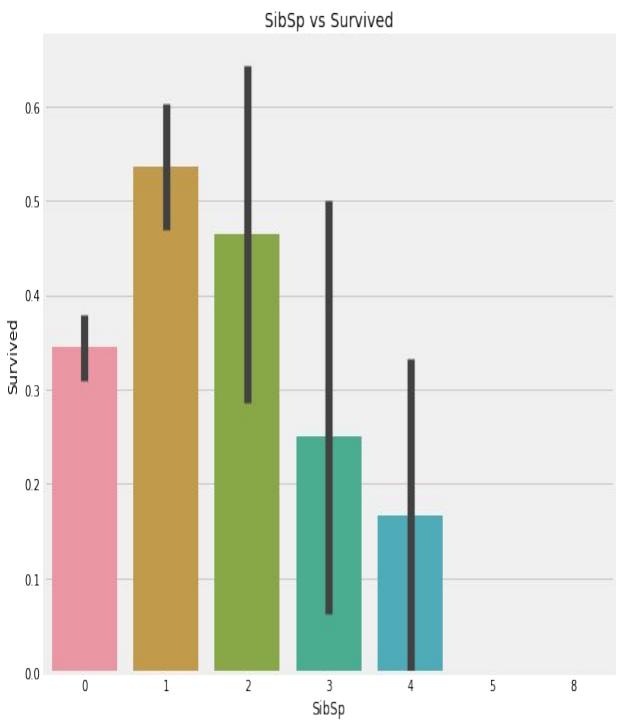
بی‌تردید، این موضوع منطقی به نظر می‌رسد؛ چرا که در شرایط بحرانی، نخستین واکنش انسان نجات عزیزانش است، نه خود.

اما نکته‌ای تلخ و تأملبرانگیز در میان داده‌ها پنهان است: در میان خانواده‌هایی که پنج تا هشت عضو داشتند، هیچ‌کدام زنده نماندند.

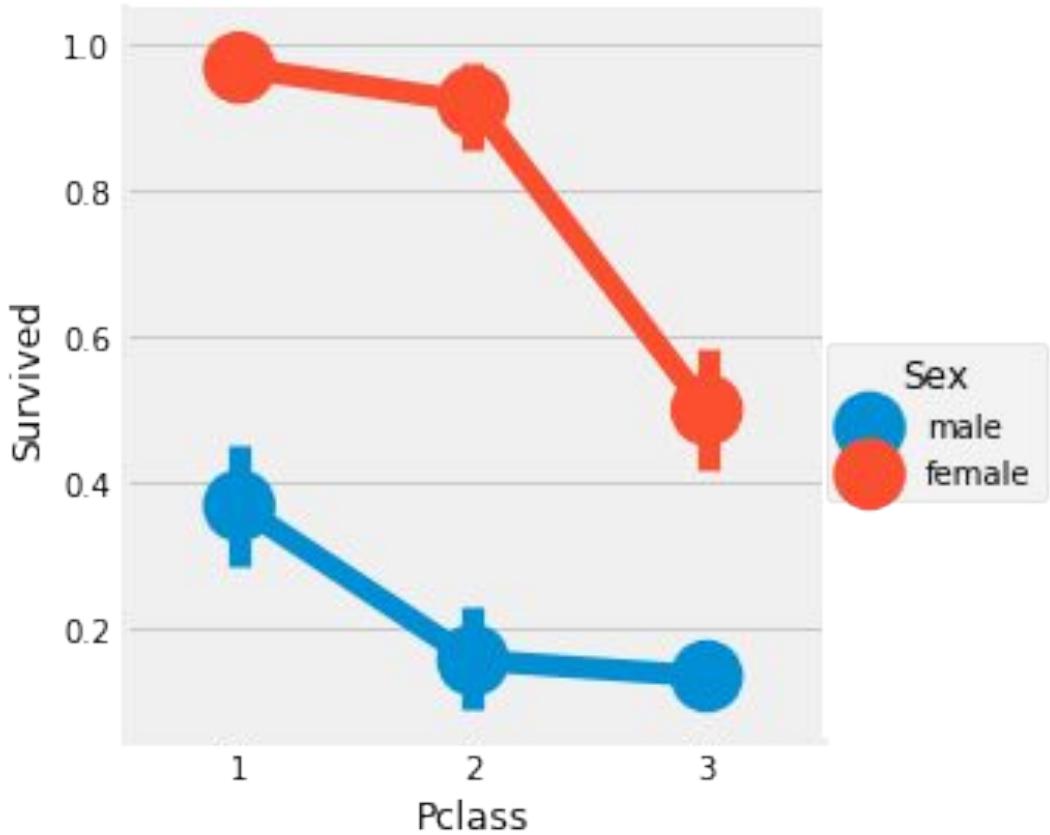
علت این فاجعه را شاید بتوان در جایگاه اجتماعی و طبقه‌ی سفر آنان جستجو کرد. بررسی‌ها نشان می‌دهد که تمام مسافرانی که شمار بستگان

همراهان (SibSp) بیش از سه نفر بوده، در کلاس سوم سفر می‌کردند؛ همان طبقه‌ای که کمترین بخت را برای نجات داشت.

پس می‌توان گفت که خانواده‌های پرجمعیت، به ویژه آنان که در طبقه‌ی سوم جای داشتند، تقریباً همگی در آن شب شوم، جان خود را از دست دادند.



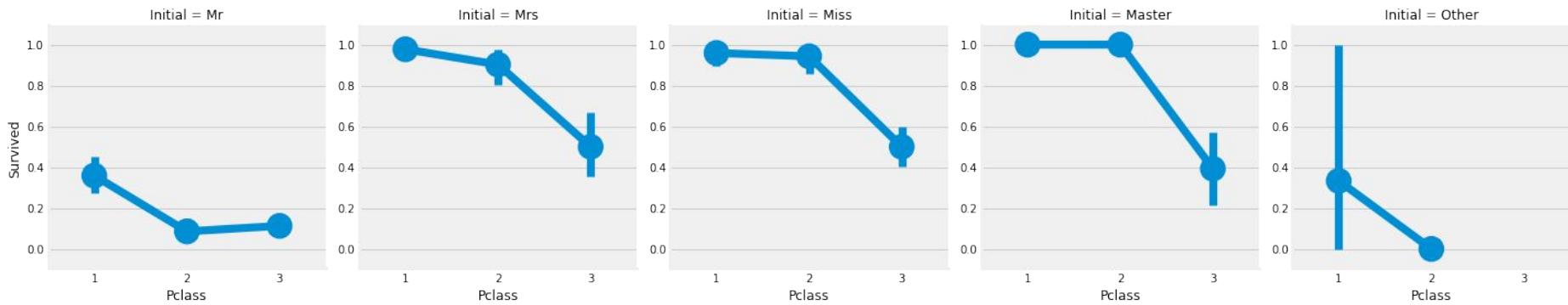
## جنسيت و کلاس مسافري(Sex And Pclass)



با مشاهده این نمودار به سادگی پی می بریم که نرخ زنده ماندن زنان در کلاس مسافری ۱ چیزی در حدود ۹۵ تا ۹۶ درصد بوده، یعنی تنها ۳ مسافر زن کلاس یک از ۹۴ تای آنها جان باختند.

مشخص است که علاوه بر اولویت کلاس های مسافری، زنان نیز اولویت اول را در نجات داشتند در صورتی که مردان کلاس ۱ نرخ نجات یافتنشان بسیار پایین است.

## سن و جنسیت و کلاس مسافری (Survival Rate With Initial And Pclass)



این نمودارها نشان می‌دهند که نرخ زنده ماندن به طور واضح تحت تأثیر سن و جنسیت (Initial) و کلاس سفر است.

زنان (Mrs) و Miss در تمام کلاس‌ها نرخ زنده ماندن بالایی دارند، مخصوصاً در کلاس‌های ۱ و ۲. این موضوع مؤید اجرای سیاست "زنان و کودکان اول" است.

کودکان (Master) نیز نرخ نجات بالایی دارند، حتی در کلاس‌های پایین.

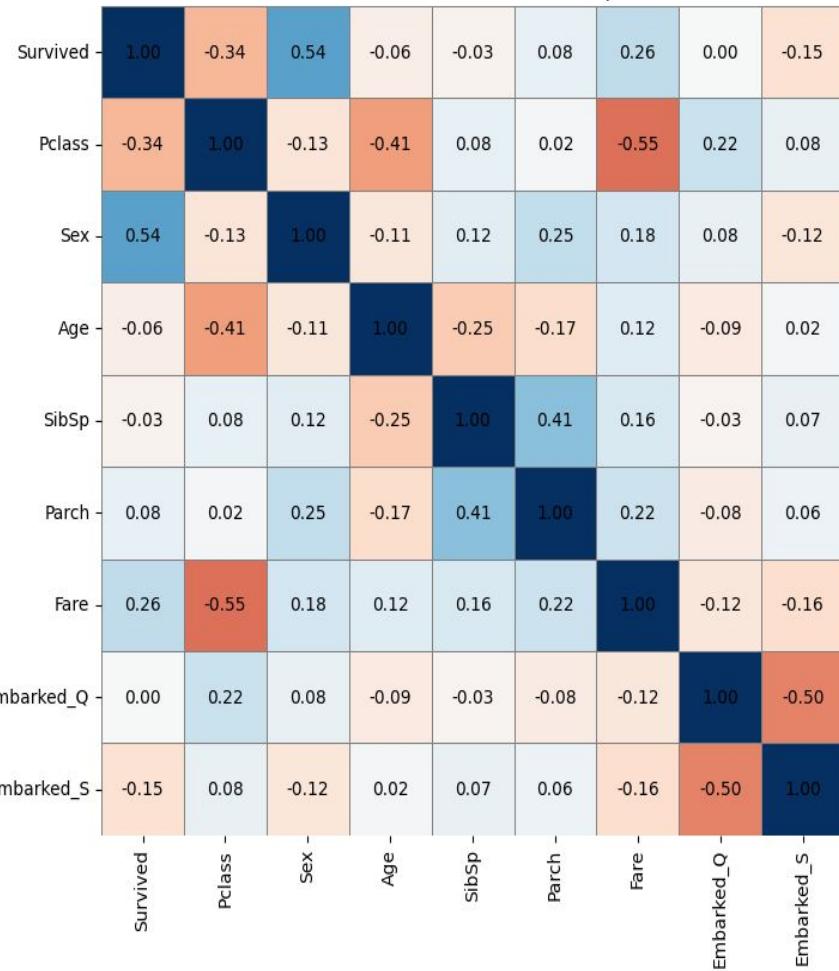
مردان (Mr) نرخ نجات بسیار پایینی دارند، به خصوص در کلاس‌های ۲ و ۳.

گروه‌های دیگر (Other) به دلیل تعداد کم داده، نوسان زیادی دارند اما به طور کلی نرخ نجات پایینی دارند.

با کاهش کلاس (از ۱ به ۳)، نرخ زنده ماندن در همه گروه‌ها کاهش می‌یابد، اما این کاهش برای مردان محسوس‌تر است.

سیاست نجات زنان و کودکان اول، مستقل از کلاس اجتماعی، تا حد زیادی رعایت شده. با این حال، کلاس سفر نیز همچنان نقش مهمی در شанс زنده ماندن ایفا کرده است.

Feature Correlation Heatmap



## تحلیل همبستگی ویژگی‌ها (Heatmap)

این هیتمپ (Heatmap) میزان همبستگی خطی بین ویژگی‌های مختلف دیتاست تایتانیک را نشان می‌دهد. رنگ‌های تیره‌تر (آبی پررنگ یا قرمز پررنگ) بیانگر همبستگی قوی‌تر هستند.

مقادیر نزدیک به +1 نشان‌دهنده همبستگی مثبت قوی، مقادیر نزدیک به -1 همبستگی منفی قوی و مقادیر نزدیک به 0 نشان‌دهنده عدم همبستگی هستند.

جنسیت (Sex) بیشترین تأثیر مثبت بر نجات (همبستگی ۰.۵۴) دارد.

کلاس بلیط (*Pclass*) همبستگی منفی با بقا دارد (-۰.۳۴)، یعنی افراد در کلاس‌های بالاتر شанс بیشتری برای نجات داشته‌اند.

مقدار کرایه (*Fare*) نیز با بقا رابطه مثبت دارد (۰.۲۶).

ویژگی‌های *SibSp* و *Parch* با یکدیگر همبستگی مثبت دارند (۰.۴۱)، که نشان‌دهنده ارتباط بین تعداد اعضای خانواده همراه است.

**نتیجه:** ویژگی‌های جنسیت، کلاس سفر و کرایه بلیط بیشترین قدرت پیش‌بینی نجات را دارند.

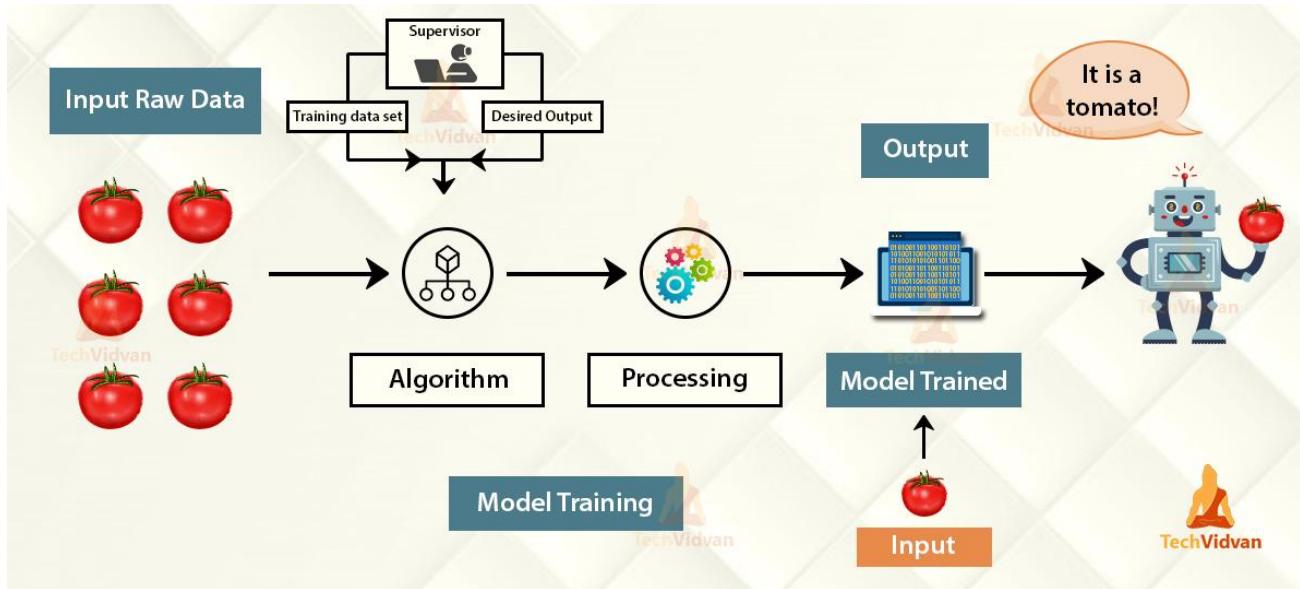
# Survival Rate Table:

شاسنخ	کروه	تعداد کل	نجات یافته	درصد نجات	تفسیر
جنسیت	زن (Female)	314	233	74.2	اولویت نجات زنان و کودکان
	مرد (Male)	577	109	18.9	کمترین شанс به دلیل سیاست نجات
کلاس مسافری	کلاس ۱ (1st)	216	136	63.0	نزدیکی به عرش و قایق‌ها
	کلاس ۲ (2nd)	184	87	47.3	
	کلاس ۳ (3rd)	491	119	24.2	دوری از قایقهای نجات
سن	کودکان (≤۱۲ سال)	88	52	59.1	اولویت نجات کودکان
	بزرگسالان (>۱۲ سال)	755	290	38.4	
همراهی خانواده	تنها	537	163	30.4	
	۱-۳ همراه	294	163	55.4	کمک متقابل در نجات
	≤ ۴ همراه	60	15	25.0	احتمالاً از کلاس ۳ بودند
بندر سوار شدن	C	168	93	55.4	مسافران پردرآمدتر
	S	644	217	33.7	
	Q	77	30	39.0	

- Decision Tree
- Random Forest
- Logistic Regression

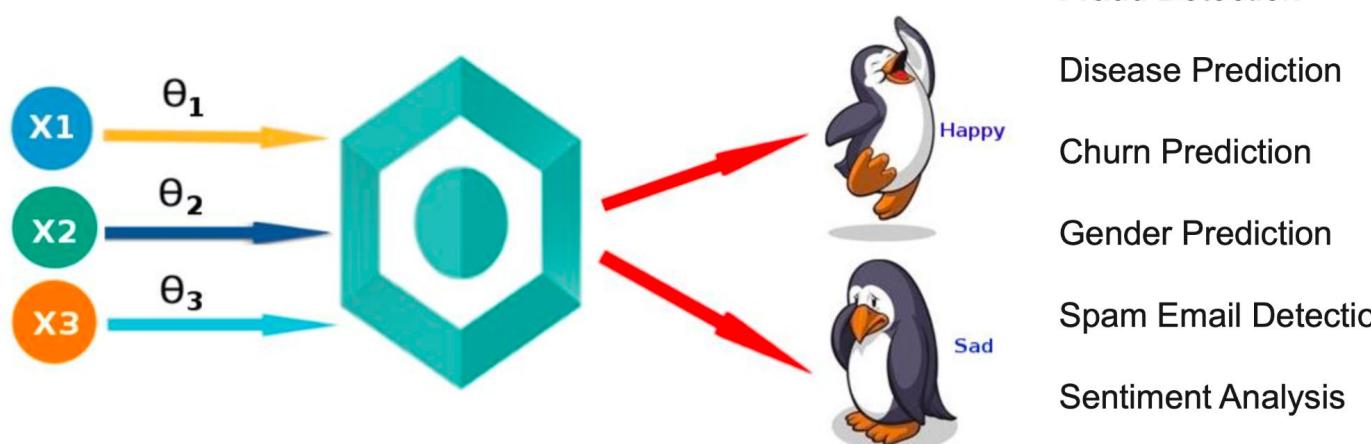
چرا انتخاب شدن و چطوری کار می‌کنن؟

# Supervised Learning



ما در اینجا یک مسئله‌ی طبقه‌بندی دودویی (*binary classification*) داریم، یعنی دو حالت ۰ و ۱ را برای مسئله در نظر می‌گیریم.

## Binary Classification



# معیارهای ارزیابی مدل‌های classification

**TP (True Positive):**

مدل درست تشخیص داده که نمونه، مثبت بوده

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

**TN (True Negative):**

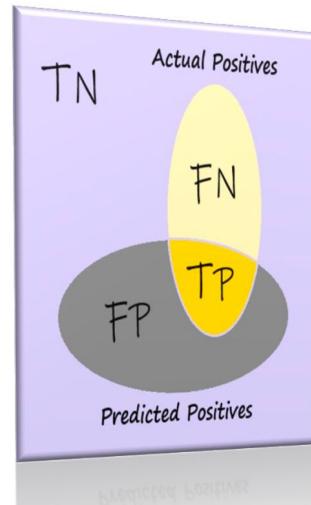
مدل درست تشخیص داده که نمونه، منفی بوده

**FP (False Positive):**

مدل اشتباهی مثبت تشخیص داده

**FN (False Negative):**

مدل اشتباهی منفی تشخیص داده



		Actual Values	
		Positive	Negative
Predicted Values	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

# معیارهای ارزیابی مدل‌های classification

$$\text{Recall} = \frac{TP}{TP + FN}$$

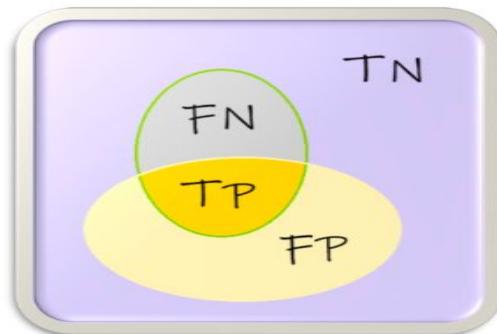
معنی چقدر از نمونه‌های مثبت واقعی درست پیش‌بینی شده.

معنی از بین پیش‌بینی‌های مثبت مدل، چند درصد واقعاً درست بودند.

$$\text{Precision} = \frac{TP}{TP + FP}$$

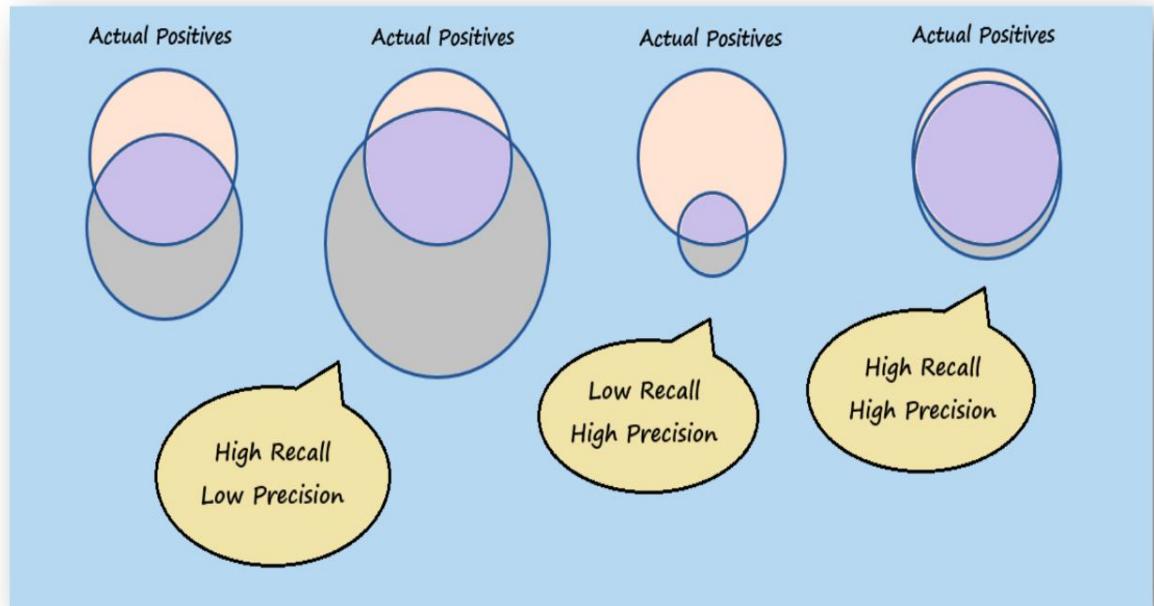
یه میانگین هارمونیک بین دقت و بازخوانی، برای بررسی توازن بین این دو تا.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$



# Precision - Recall - F1 score

Precision	Recall	F1
50	50	50
60	40	48
70	30	42
80	20	32
90	10	18



# معیارهای ارزیابی مدل‌های classification

## False Positive Rate (FPR):

یعنی از بین همه نمونه‌های منفی واقعی، مدل چقدر اشتباهی مثبت پیش‌بینی کرده

## True Positive Rate (TPR) یا Sensitivity:

درست یعنی از بین همه نمونه‌های مثبت واقعی، مدل چقدر پیش‌بینی کرده

$$FPR = \frac{FP}{FP + TN}$$

$$TPR = \frac{TP}{TP + FN}$$

# ROC Curve , AUC

## :ROC Curve

یک نمودار است که محور x آن FPR و محور y آن TPR است.  
مدل هر چقدر بهتر باشد، نمودارش به گوشه بالا سمت چپ نزدیکتر است.

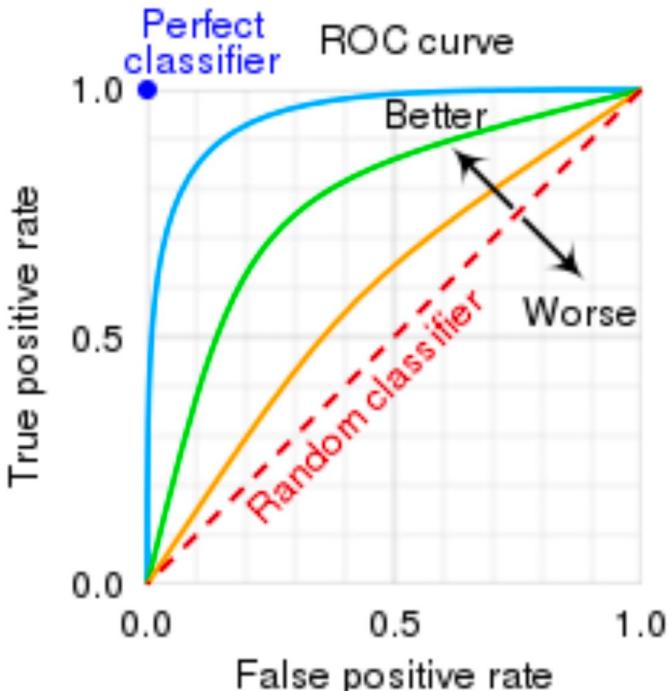
## :AUC (Area Under Curve)

مساحت زیر نمودار ROC است.

اگر **AUC = 1** باشد → مدل عالی است.

اگر **AUC = 0.5** باشد → مدل مثل حدس زدن شانسی است.

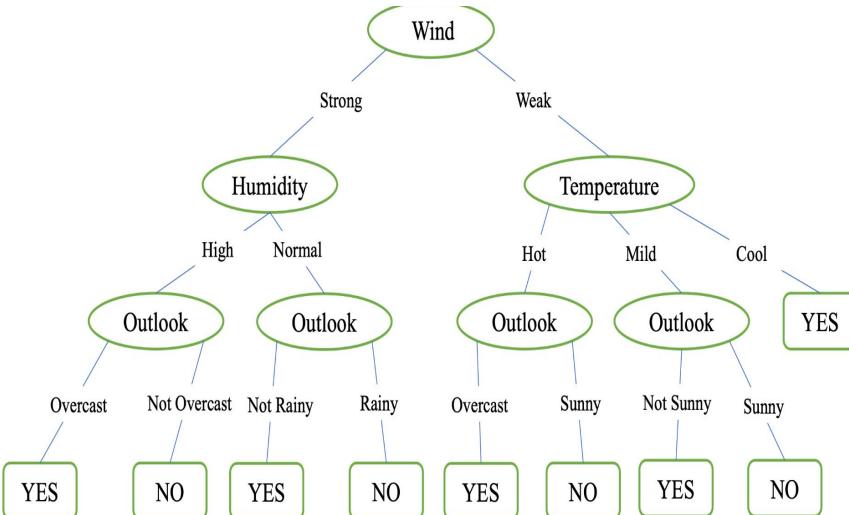
هر چقدر AUC به 1 نزدیکتر باشد، مدل بهتر است.



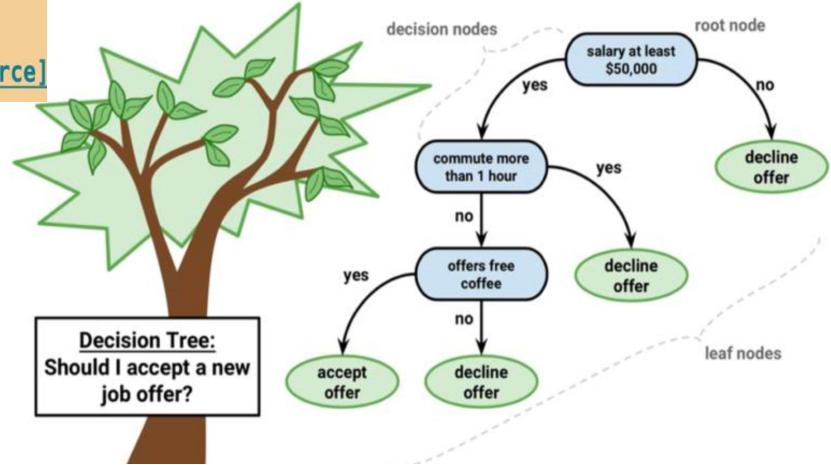
# Decision Tree

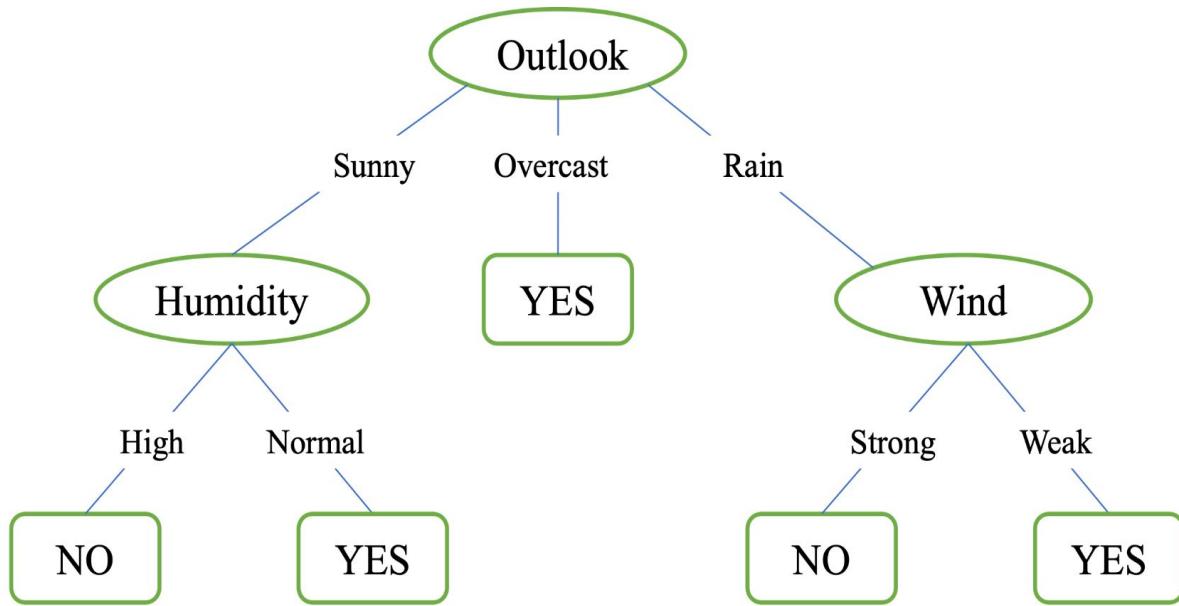
# درخت تصمیم

```
class sklearn.tree.DecisionTreeClassifier(*, criterion='gini',
splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1,
min_weight_fraction_leaf=0.0, max_features=None, random_state=None,
max_leaf_nodes=None, min_impurity_decrease=0.0, class_weight=None,
ccp_alpha=0.0, monotonic_cst=None)
```

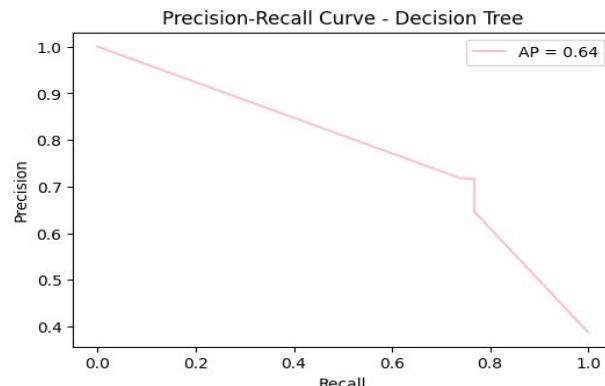
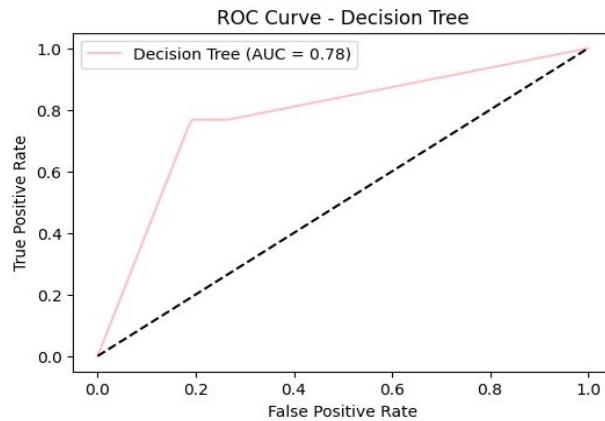
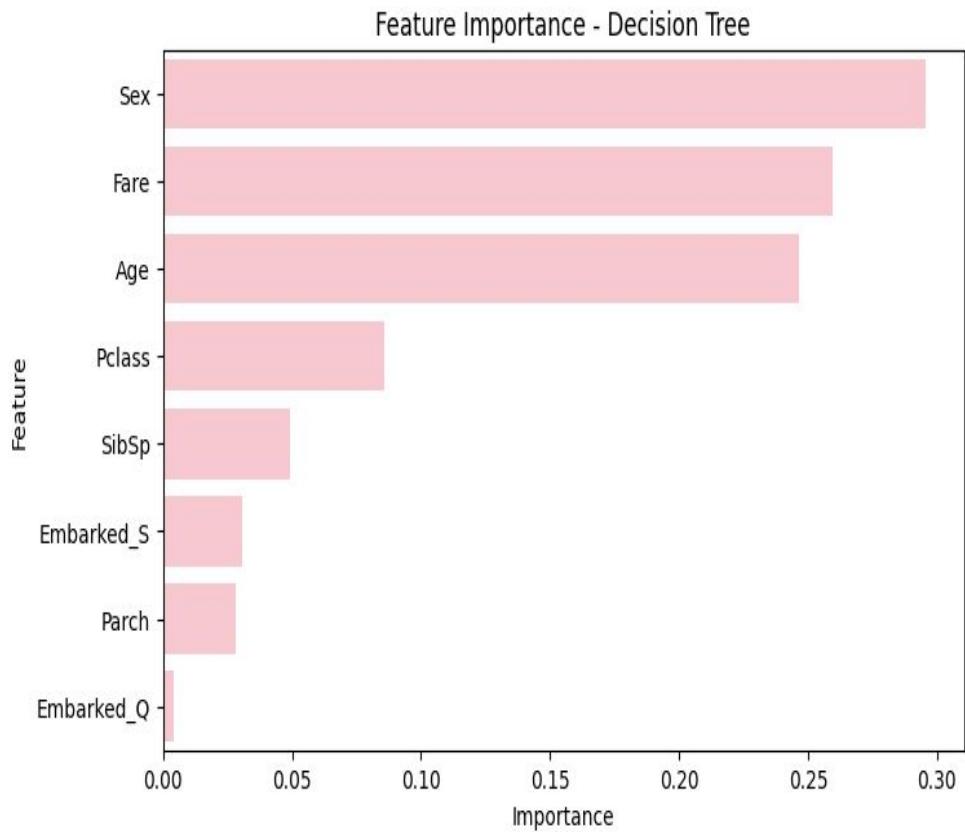


[source]



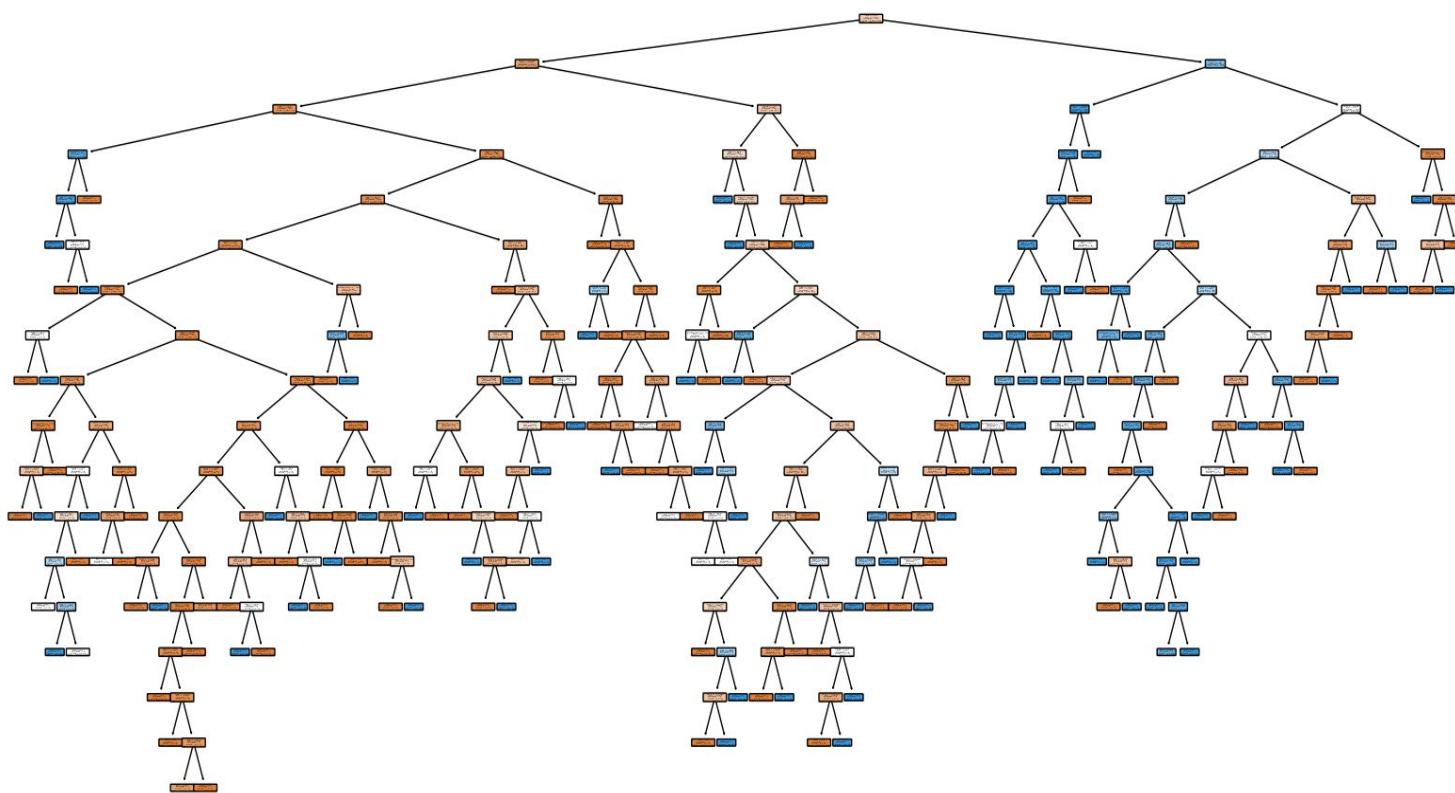


- فیلتر کردن داده‌ها
- سریع‌تر و بهتر
- زودتر به نتیجه رسیدن



# درخت تصمیم تایتانیک

Decision Tree Structure

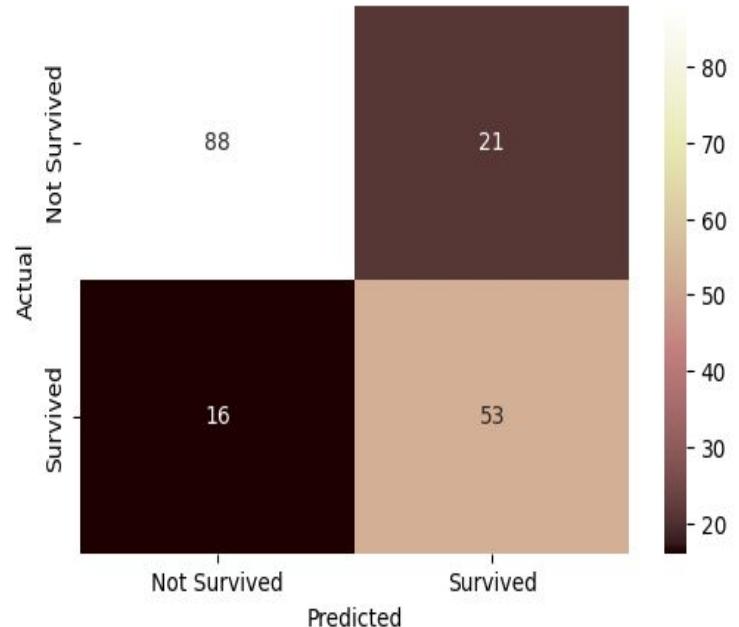


Accuracy: 0.7921348314606742

Classification Report:

	precision	recall	f1-score	support
0	0.85	0.81	0.83	109
1	0.72	0.77	0.74	69
accuracy			0.79	178
macro avg	0.78	0.79	0.78	178
weighted avg	0.80	0.79	0.79	178

Decision Tree - Confusion Matrix

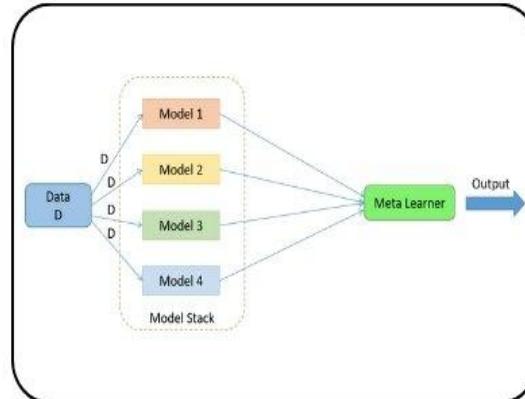
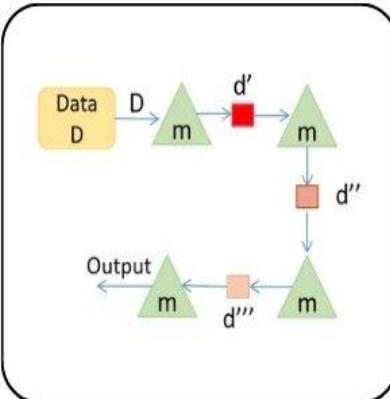
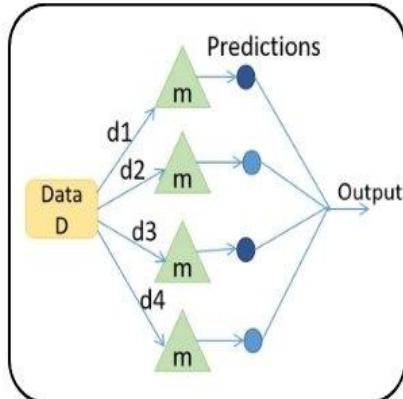


## Ensemble Methods

Bagging

Boosting

Stacking



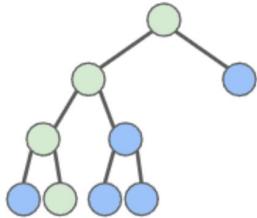
# RandomForest

```
class sklearn.ensemble.RandomForestClassifier(n_estimators=100, *,  
criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1,  
min_weight_fraction_leaf=0.0, max_features='sqrt', max_leaf_nodes=None,  
min_impurity_decrease=0.0, bootstrap=True, oob_score=False, n_jobs=None,  
random_state=None, verbose=0, warm_start=False, class_weight=None,  
ccp_alpha=0.0, max_samples=None, monotonic_cst=None)
```

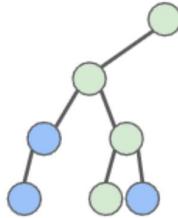
[\[source\]](#)

جنگل تصادفی:

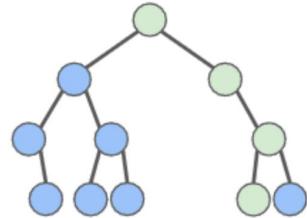
مجموعه‌ای از درختان تصمیم که هر کدام روی نمونه‌های تصادفی آموزش دیده‌اند و پیش‌بینی نهایی از طریق رأی‌گیری (*Voting*) یا میانگین‌گیری صورت می‌گیرد.



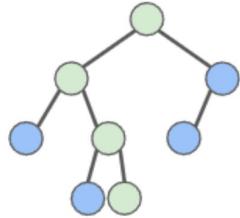
Tree 1: Cat



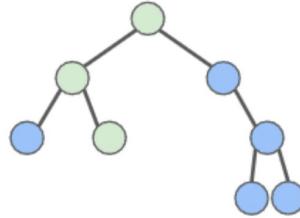
Tree 2: Dog



Tree 3: Cat

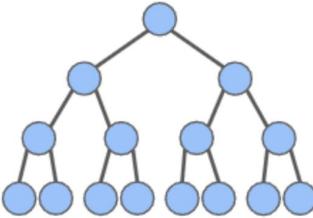


Tree 4: Cat



Tree 5: Cat

...

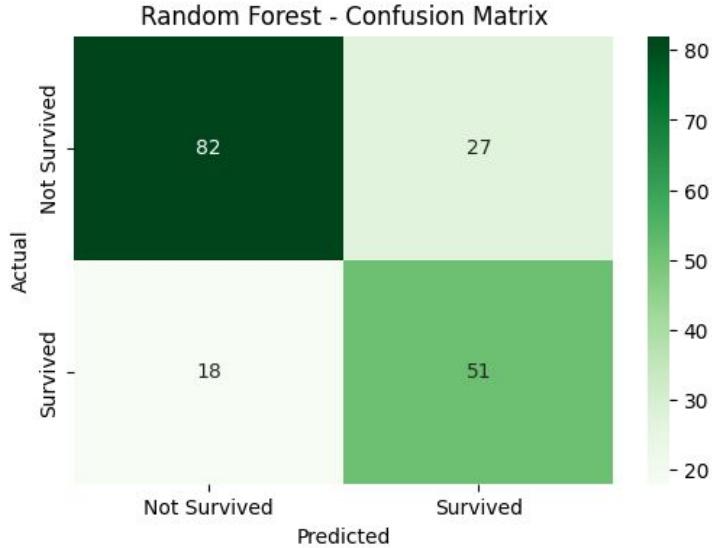


Tree n



- Accuracy: 0.7471910112359551
- Classification Report:

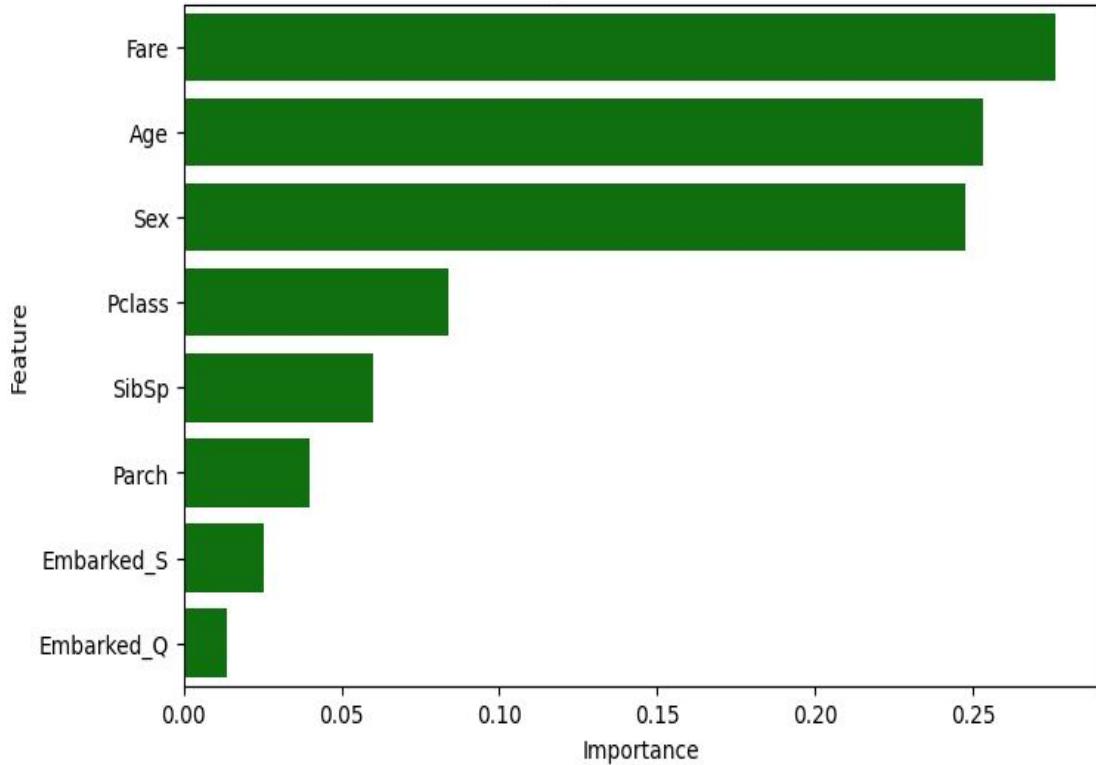
	precision	recall	f1-score	support
0	0.82	0.75	0.78	109
1	0.65	0.74	0.69	69
accuracy			0.75	178
macro avg	0.74	0.75	0.74	178
weighted avg	0.76	0.75	0.75	178

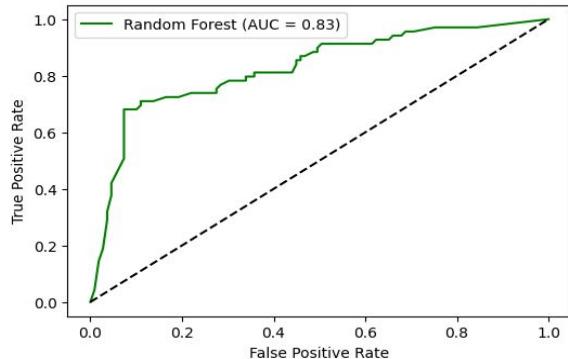
مدل *Random Forest* با دقت حدود ۷۵٪ عملکرد مناسبی در پیش‌بینی بقای مسافران داشته است.

دقت مدل در شناسایی نجات‌یافته‌ها کمی پایین‌تر است که با بهبود داده‌ها یا تنظیم پارامترها قابل بهتر شدن است.

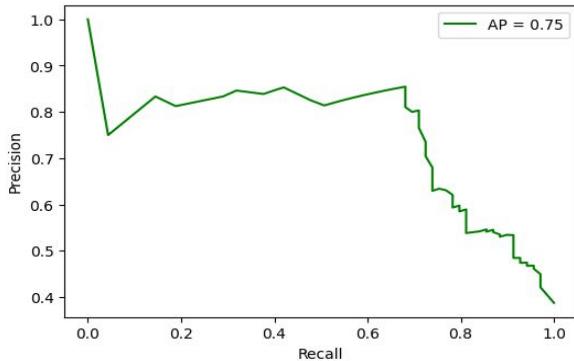
### Feature Importance - Random Forest



### ROC Curve - Random Forest



### Precision-Recall Curve - Random Forest



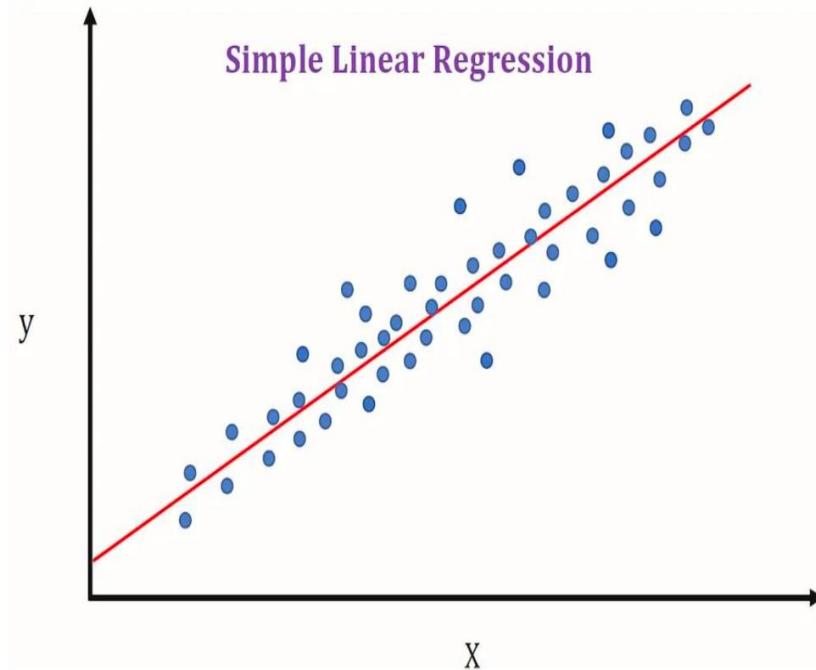
# Simple Linear Regression

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Annotations for the equation:

- Dependent Variable →  $Y_i$
- Population Y intercept →  $\beta_0$
- Population Slope Coefficient →  $\beta_1$
- Independent Variable →  $X_i$
- Random Error term →  $\varepsilon_i$

Brackets indicate:  
Linear component:  $\beta_0 + \beta_1 X_i$   
Random Error component:  $\varepsilon_i$



# Simple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \varepsilon$$

Dependent Variable (Response Variable)

Independent Variables (Predictors)

احتمال زنده ماندن مسافر:  $p(x)$

عرض از مبدأ:  $\beta_0$  (intercept)

ضرایب ویژگی‌ها:  $\beta_1, \beta_2, \dots, \beta_i$

ویژگی‌های پیش‌بین (متغیر مستقل) (مانند سن، جنسیت، کلاس مسافری و...):  $X_1, X_2, \dots, X_i$

Y intercept

Slope Coefficient

Error Term

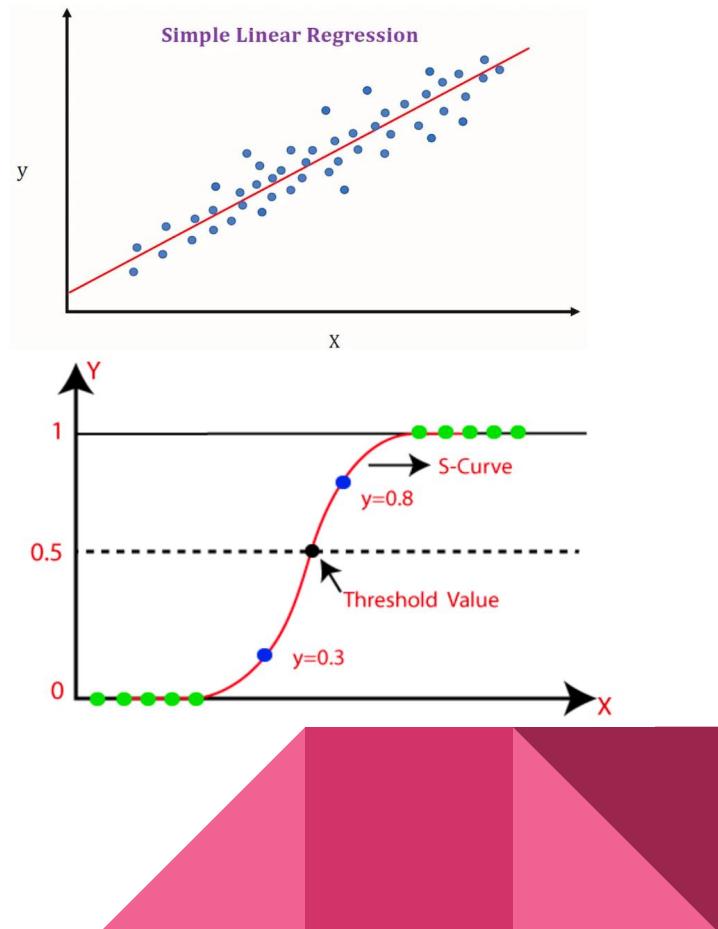
$$p(X) = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\dots+\beta_kX_k)}}$$

# Logistic Linear Regression

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

```
class sklearn.linear_model.LogisticRegression(penalty='l2', *, dual=False,  
tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None,  
random_state=None, solver='lbfgs', max_iter=100, multi_class='deprecated',  
verbose=0, warm_start=False, n_jobs=None, l1_ratio=None)
```

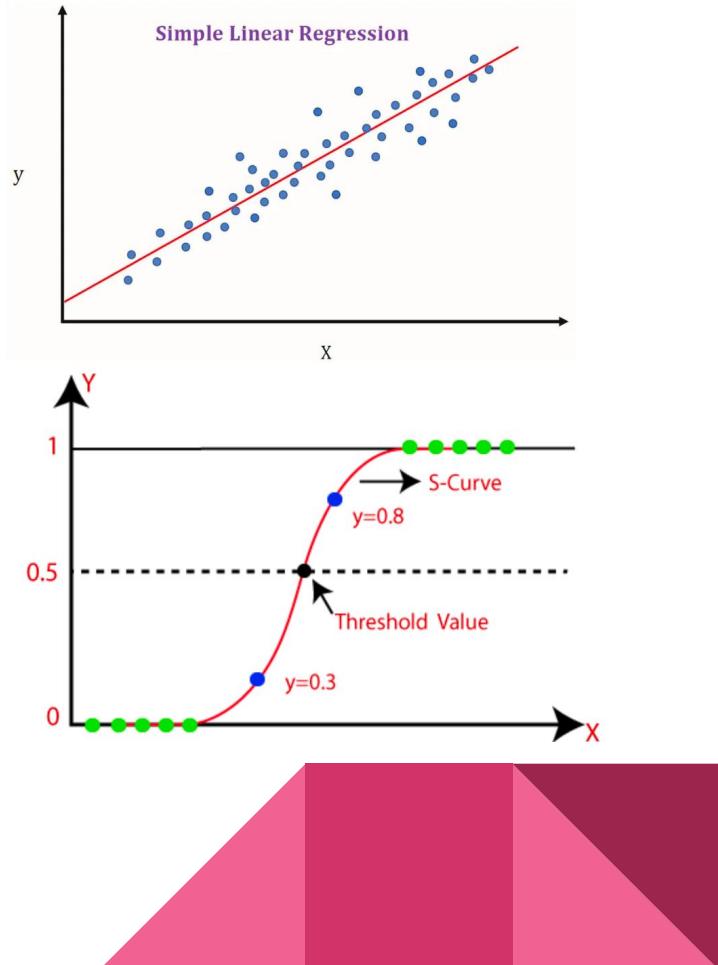
[\[source\]](#)



# Logistic Linear Regression

$$\hat{y} = P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}}$$

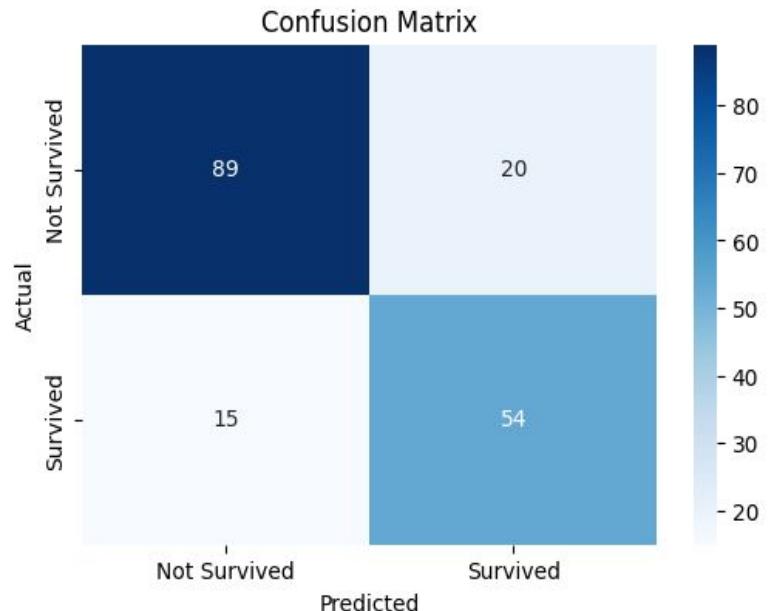
$$\text{Log Loss} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

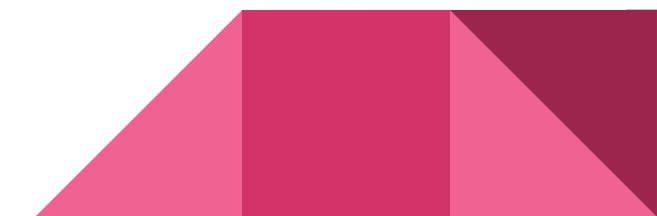
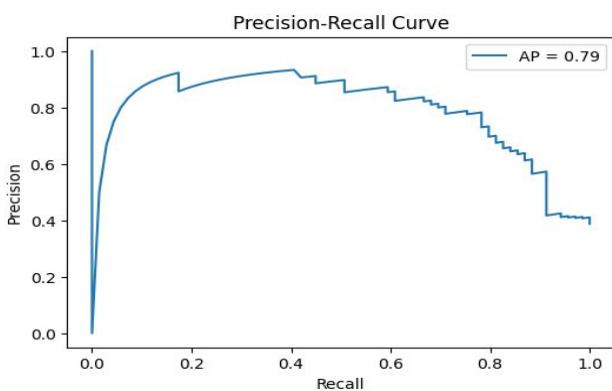
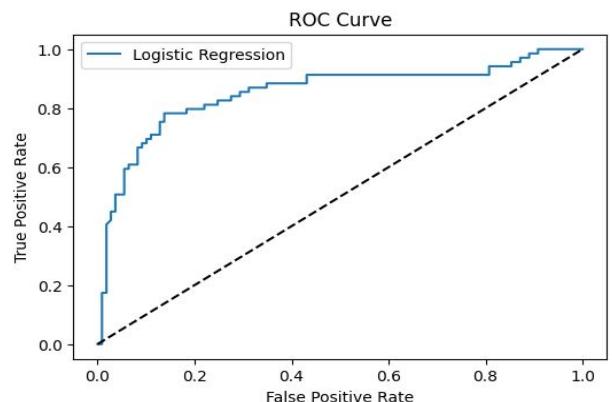
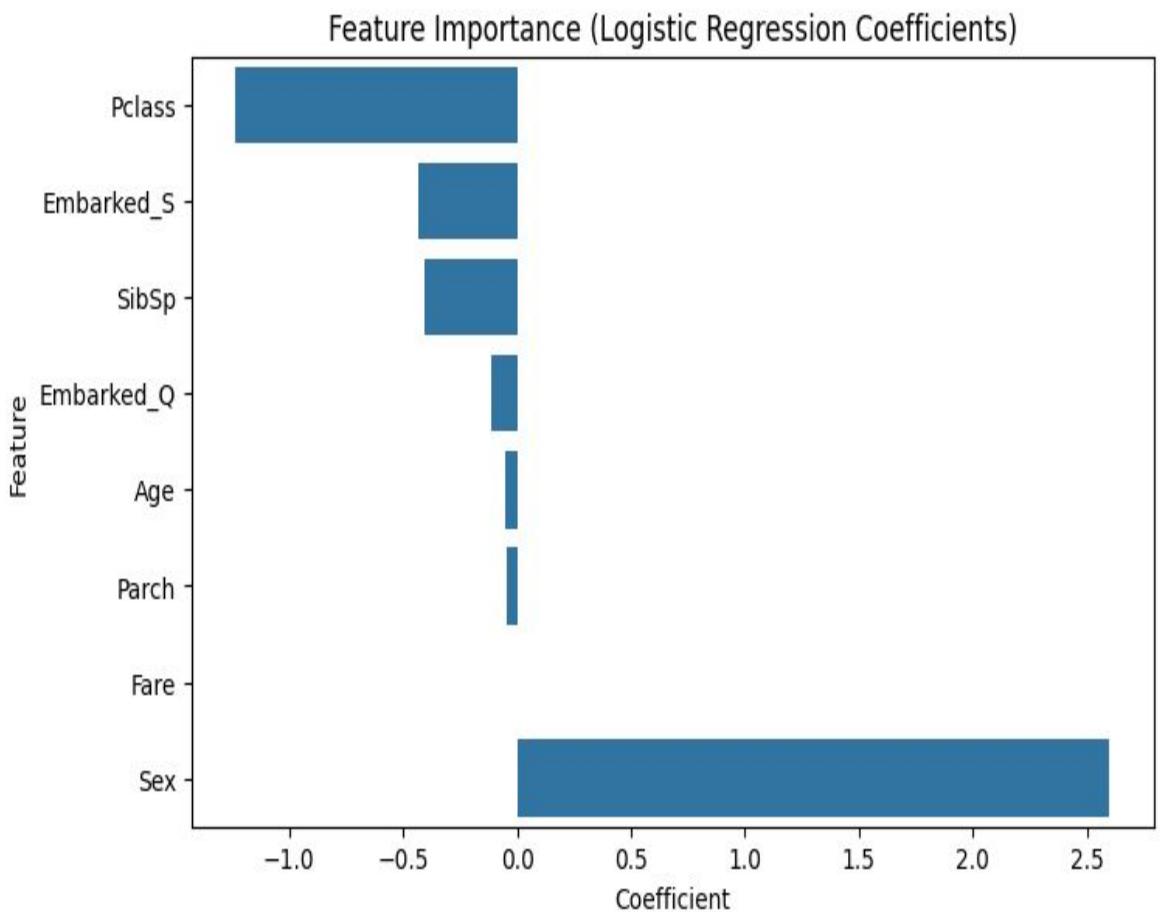


Accuracy: 0.8033707865168539

Classification Report:

	precision	recall	f1-score	support
0	0.86	0.82	0.84	109
1	0.73	0.78	0.76	69
accuracy			0.80	178
macro avg	0.79	0.80	0.80	178
weighted avg	0.81	0.80	0.80	178

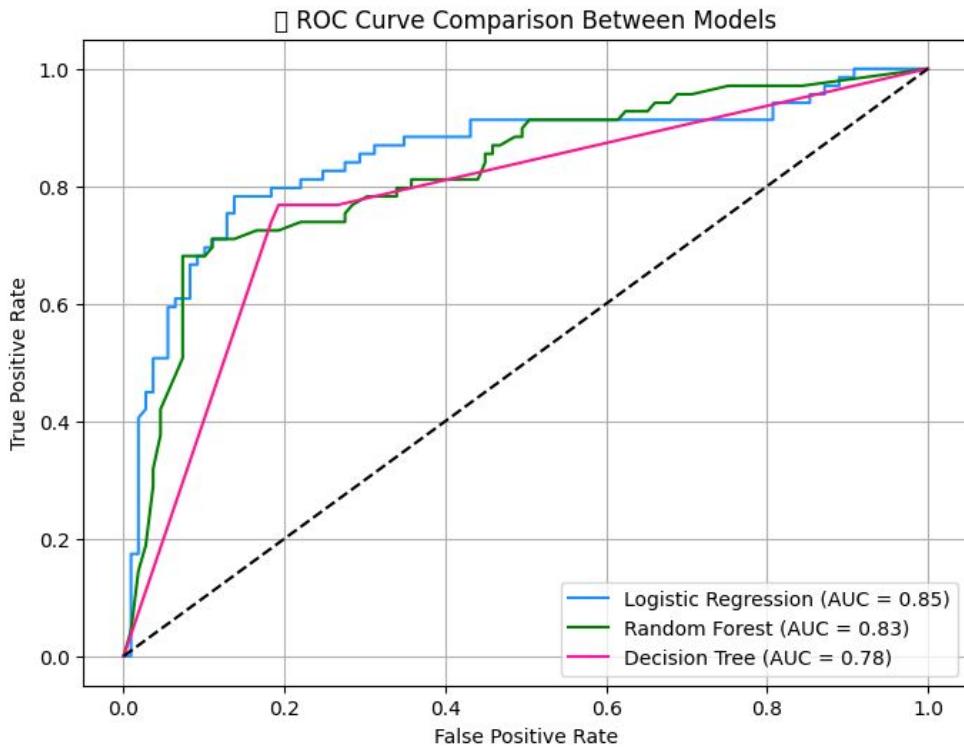




مقایسه

مدل‌ها

# Roc-Accuracy-...



Model	Accuracy	AUC
Logistic Regression	0.803371	0.852214
Random Forest	0.747191	0.826951
Decision Tree	0.792135	0.779949

اگر بخوایم دقت بالا و سادگی تفسیر داشته باشیم،  
**بهترین گزینه Logistic** روی بهینه‌سازی بیشتر سرمایه‌گذاری کنیم،  
ترکیب مدل‌ها یا بهبود **Random Forest** هم  
میتوانه موثر باشه

این همه تحقیق و مطالعه؛  
برای چه؟

# کاربرد مدل در دنیای واقعی

2. به یک شرکت بیمه یا نهاد نجات کمک می‌کند در لحظه حادثه، تصمیم بگیرد چه کسی در اولویت باشد یا چه کسی ارزش بیمه‌گذاری بیشتری دارد. در واقع هدف شرکت بیمه گرفتن حق بیمه بیشتر از کسانی که احتمال مردن اونها بالاتر و بر عکس؛ با استفاده از مدل تایتانیک به عنوان یک سناریوی واقعی.

3. ساده‌بودن داده‌ها و ساختار مدل، این مدل را به گزینه‌ای مناسب برای آموزش مفاهیم پایه‌ای مانند پیش‌پردازش داده‌ها، مهندسی ویژگی‌ها و ارزیابی عملکرد در ماشین لرنینگ تبدیل کرده است.

1. بعد از این حادثه، فشار مردم نسبت به نابرابری سطح ایمنی بخاطر مادیات، بالا گرفت. همونطور که تو نتایج مدل دیدیم، در حادثه‌های این چنینی قیمت و کلاس بلیط نقش زیادی تو احتمال زنده موندن داره که این اصل برابری جان انسان‌ها رو نقض میکنه. همین بحث منجر به ایجاد کنوانسیون بین‌المللی ایمنی جان اشخاص در دریا (SOLAS) در سال 1914 شد. نتایج مدل کاملاً لزوم ایجاد همچین مجموعی رو تایید میکنه.

# یه بررسی جذاب: احتمال زنده موندن شما تو همچین سوانحی چقدر؟

جنسیت	سن	کلاس بلیط	احتمال زنده ماندن
خانم	20	2	70%
آقا	20	2	28%
آقا	30 به بالا	2	کمتر از 19%



و در نهایت: جک و رز، نمونه از درستی مدل ما:)

- <https://www.kaggle.com/competitions/titanic>
- <https://www.youtube.com/watch?v=PjYS-U0NRWo>
- <https://encord.com/blog/classification-metrics-accuracy-precision-recall/>
- <https://www.statlearning.com/>
- <https://link.springer.com/book/9780387310732>
- <https://hastie.su.domains/ElemStatLearn/>
- [https://en.wikipedia.org/wiki/Logistic\\_regression](https://en.wikipedia.org/wiki/Logistic_regression)
- <https://www.khanacademy.org/math/statistics-probability>
- <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier>
- <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
- [https://scikit-learn.org/stable/modules/generated/sklearn.linear\\_model.LogisticRegression.html](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)
- <https://viso.ai/deep-learning/ensemble-learning/>