データサイエンティスト育成コース
ベーシックステップ
統計学 Day1　宿題

今回の課題では、Kaggleの問題に挑戦してもらいます！
※Kaggleに参加するに当たって、Kaggleのアカウントが必要になります。Kaggleのアカウントを持っていない人は、事前に登録をしておいてください。

改めてKaggleとは…
世界中のデータサイエンティストがデータ分析力や機械学習のモデリング力を競い合うコンペティションサイトです。
https://www.kaggle.com/

今回は、Kaggleのチュートリアル的存在である ”Titanic : Machine Learning from Disaster” をやっていただきます。こちらは、1912年4月15日に沈没したタイタニック号の生存者を予測するという内容になっております。

【データの準備】
Titanicで扱うデータは、以下のURLからダウンロードができます。
https://www.kaggle.com/c/titanic/data
（Canvas上でも公開していますので、そちらからダウンロードしても大丈夫です。）

【データの概要】
データは、train(学習データ）とtest（検証データ）があり、以下の変数で構成されています。

PassengerId	：	乗客ID
Survived ※trainデータのみ	：	生存したかどうか（０：死亡、 １：生存）
Pclass	：	チケットのクラス（1,2,3の３種類）
Name	：	乗客の名前
Sex	：	性別
Age	：	年齢
SibSp	：	タイタニック号に同乗している兄弟 / 配偶者の数
Parch	：	タイタニック号に同乗している親 / 子供の数
Ticket	：	チケットの番号
Fare	：	運賃
Cabin	：	客室番号
Embarked	：	乗船した港（Cherbourg、Queenstown、Southamptonの3種類）

これらのことを踏まえて、以下の問いに答えなさい。ただし、今回は、’Name’, ’Ticket’, ’Cabin’の3つは、変数として考えないことにする。

問１.（前処理）

1) trainデータの'Age','Embarkedの'欠損を補完しなさい
2) testデータの'Age','Fare'の欠損を補完しなさい
3) 'Pclass','Sex','Embarked'をダミー変数にしなさい

問２.（変数の作成）

4) 'Pclass', 'Sex', 'Embarked', 'Age', 'Fare', 'SibSp', 'Parch'のいずれかを使って、新たな変数を
1つ作りなさい。

問3.（ロジスティック回帰モデルによる予測）

5) 4)で作った変数と、'Pclass', 'Sex', 'Embarked', 'Age', 'Fare', 'SibSp', 'Parch'を説明変数に’Survived’を目的変数にしてロジスティック回帰モデルを作りなさい
6) 作成したロジスティック回帰の解釈を行いなさい
7) 作成したロジスティック回帰モデルを用いて、testデータのSurvivedを予測しなさい
8) 最後に予測した結果をKaggleに提出し、算出されたスコアを記述しなさい