CSE 4334

DATA MINING

KAGGLE MACHINE LEARNING COMPETITION(TITANIC)

Introduction

The RMS Titanic sank in the early morning hours of 15 April 1912 in the North Atlantic Ocean, four days into her maiden voyage from Southampton to New York City. Kaggle, an online community of data scientists and machine learning practitioners created a beginner friendly ML competition inspired by the tragic story of the Titanic. The goal of the competition is to create a model capable of accurately predicting whether a passenger had survived the accident or not.

Data

In order to train and develop a somewhat accurate model, Kaggle provided three sets of data. The first one **train.csv** contains the details of a subset of the passengers on board (891 passengers, to be exact -- where each passenger gets a different row in the table). This data set indicates if each passenger survived or not.

```
train_data = pd.read_csv("/kaggle/input/titanic/train.csv")
train_data.head()
```

	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

The second one **test.csv** dataset contains similar information but does not disclose the faith for each passenger.

```
test_data = pd.read_csv("/kaggle/input/titanic/test.csv")
test_data.head()
```

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S

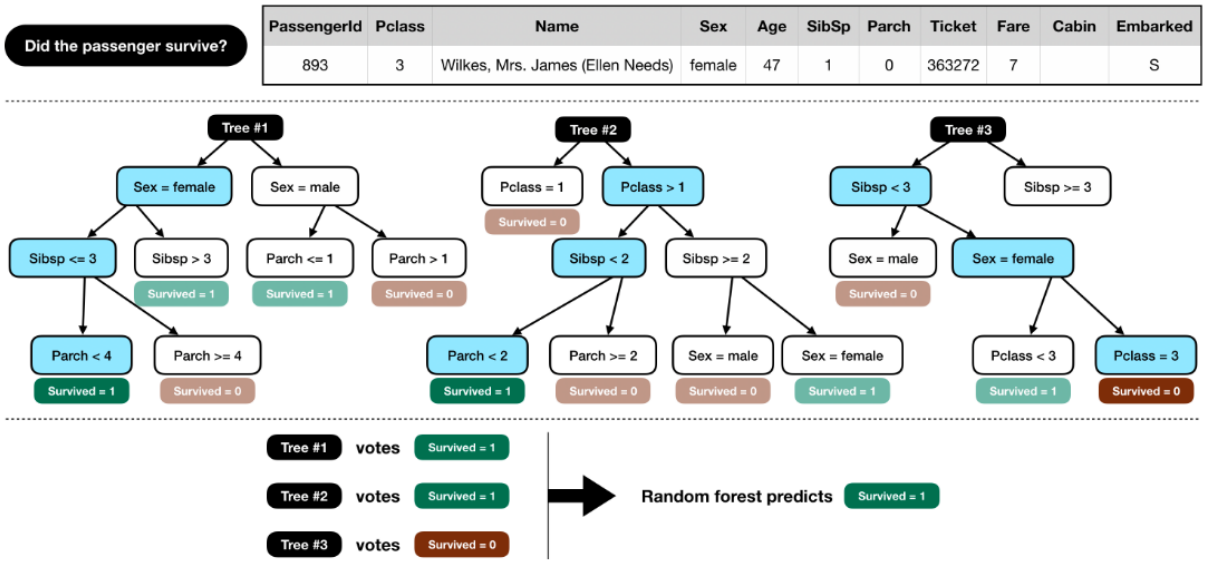
Lastly, the **gender_submission.csv** file is provided as an example that shows how you should structure your predictions. It predicts that all female passengers survived, and all male passengers died.

Methodology

The methodology suggested to approach this beginner ML competition was to use the training dataset (train.csv) to find variables and patterns to implement on the test dataset (test.csv). Inspiring ourselves from how the **gender_submission.csv** file produced a somewhat accurate result leveraging only a single attribute. We built a random forest model based on the four following attributes: socio economic status, siblings/spouse on board, parent/children on board and finally gender.

Implementation

1- Random forest model



2- Code

```
from sklearn.ensemble import RandomForestClassifier

y = train_data["Survived"]

features = ["Pclass", "Sex", "SibSp", "Parch"]
X = pd.get_dummies(train_data[features])
X_test = pd.get_dummies(test_data[features])

model = RandomForestClassifier(n_estimators=100, max_depth=5, random_state=1)
model.fit(X, y)
predictions = model.predict(X_test)

output = pd.DataFrame({'PassengerId': test_data.PassengerId, 'Survived': prediction
s})
output.to_csv('submission.csv', index=False)
print("Your submission was successfully saved!")
```

Results

Model predictions