Random Forest Project Tutorial

Editar en Github

- So far, we have only used the Titanic dataset for reference in our learning process. It has been our guide in the exploratory data analysis, data preprocessing part and a reference in our first model: logisitc regression. In the current project and the next one, Titanic will be our project main data.
- If you have been practicing with Titanic so far, you may have some clean data ready for modeling. If not, we recommend you start doing some eda and cleaning on the Titanic dataset to have it ready for the next two models.
- In this project, practice your new Random Forest skills to try making an accurate prediction of Titanic survival. Then, as always, optimize your hyperparameters.
- If you already have a notebook for Titanic in your Github repo, feel free to use it for this project. Continue your notebook with a new modeling part, this time using Random Forest! If you haven't built a repo for Titanic yet, follow the instructions on how to start this project.

🌱 How to start this project

You will not be forking this time, please take some time to read this instructions:

1. Create a new repository based on machine learning project by clicking here.
2. Open the recently created repository on Gitpod by using the Gitpod button extension.
3. Once Gitpod VSCode has finished opening you start your project following the Instructions below.

🚛 How to deliver this project

Once you are finished creating your model, make sure to commit your changes, push to your repository and go to 4Geeks.com to upload the repository link.

📝 Instructions

Predicting Titanic survival using Random Forest

We need to build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data (ie name, age, gender, socio-economic class, etc). To be able to predict which passengers were more likely to survive we will use Random Forest to train the model.

Step 1:

The dataset can be found in this project folder as 'titanic_train.csv' file. You are welcome to load it directly from the link (`https://raw.githubusercontent.com/4GeeksAcademy/random-forest-project-`

`tutorial/main/titanic_train.csv`), or to download it and add it to your data/raw folder. In that case, don't forget to add the data folder to the .gitignore file.

Time to work on it!

Step 2:

Explore and clean the data.

Step 3:

Build a first predictive model using Random Forest. Chose an evaluation metric and then optimize your model hyperparameters.

Step 4:

Use the app.py to create your pipeline.

Step 5:

To save your model and be able to use it later use the following code:

```
import pickle

filename = 'finalized_model.sav'
pickle.dump(model, open(filename, 'wb'))
```

In your README file write a brief summary.