Machine Learning for Biomedical Data

A complete ML application pipeline

2021-2022

https://www.kaggle.com/ronitf/heart-disease-uci

by Fernando Caraza

First of all... Open Rstudio and install the following packages

- install.packages("dplyr") # for data manipulation
- install.packages(c("ggplot2", "ggpubr")) # for awesome graphics
- install.packages("visdat") # for additional visualizations
- install.packages("rpart.plot") # for additional visualizations
- install.packages(c("tidyverse", "titanic", "ggpubr"))
- install.packages("skimr")

- # Feature engineering packages
- install.packages("caret") # for various ML tasks
- install.packages("recipes") # for feature engineering tasks

CHALLENGE – HEART DISEASE PREDICTION

Predict whether patients will have a heart attack or not

Coronary heart disease is a type of heart disease that develops when the arteries of the heart cannot deliver enough oxygen-rich blood to the heart. It is the leading cause of death in the United States.

The “goal” of this Challenge is to predict the presence of heart disease in the patient using ECG information and other clinical featureres.

Patient Variables

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fbs: fasting blood sugar > 120 mg/dl

Hereon, variables are related to a nuclear stress test. That is, a stress test where a radioactive dye is also injected to the patient to see the blood flow:

7. restecg: resting electrocardiographic results (values 0,1,2)
8. thalach: maximum heart rate achieved
9. exang: exercise induced angina 10.oldpeak: ST depression induced by exercise relative to rest
10. slope: the slope of the peak exercise ST segment
11. ca: number of major vessels (0-3) colored by flourosopy
12. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

Patient Features

# age	# sex	# cp	# trestbps	# chol	# fbs
age in years	(1 = male; 0 = female)	chest pain type	resting blood pressure (in mm Hg on admission to the hospital)	serum cholestoral in mg/dl	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
29 – 77	0 – 1	0 – 3	94 – 200	126 – 564	0 – 1
63	1	3	145	233	1
37	1	2	130	250	0
41	0	1	130	204	0
56	1	1	120	236	0
57	0	0	120	354	0
57	1	0	140	192	0
56	0	1	140	294	0
44	1	1	120	263	0
52	1	2	172	199	1
57	1	2	150	168	0
54	1	0	140	239	0

STEPS BEFORE MODELLING

STEPS BEFORE MODELLING

1. **Define the problem**: What do we want to predict? Which data is available?

   → Make your hypotheses

2. **Explore** and understand the data that will be used to create the model.

   → Create new features?

3. **Preprocess the data**: define the necessary transformations so that the data can be interpreted by the selected machine learning algorithm.

STEPS FOR MODELLING

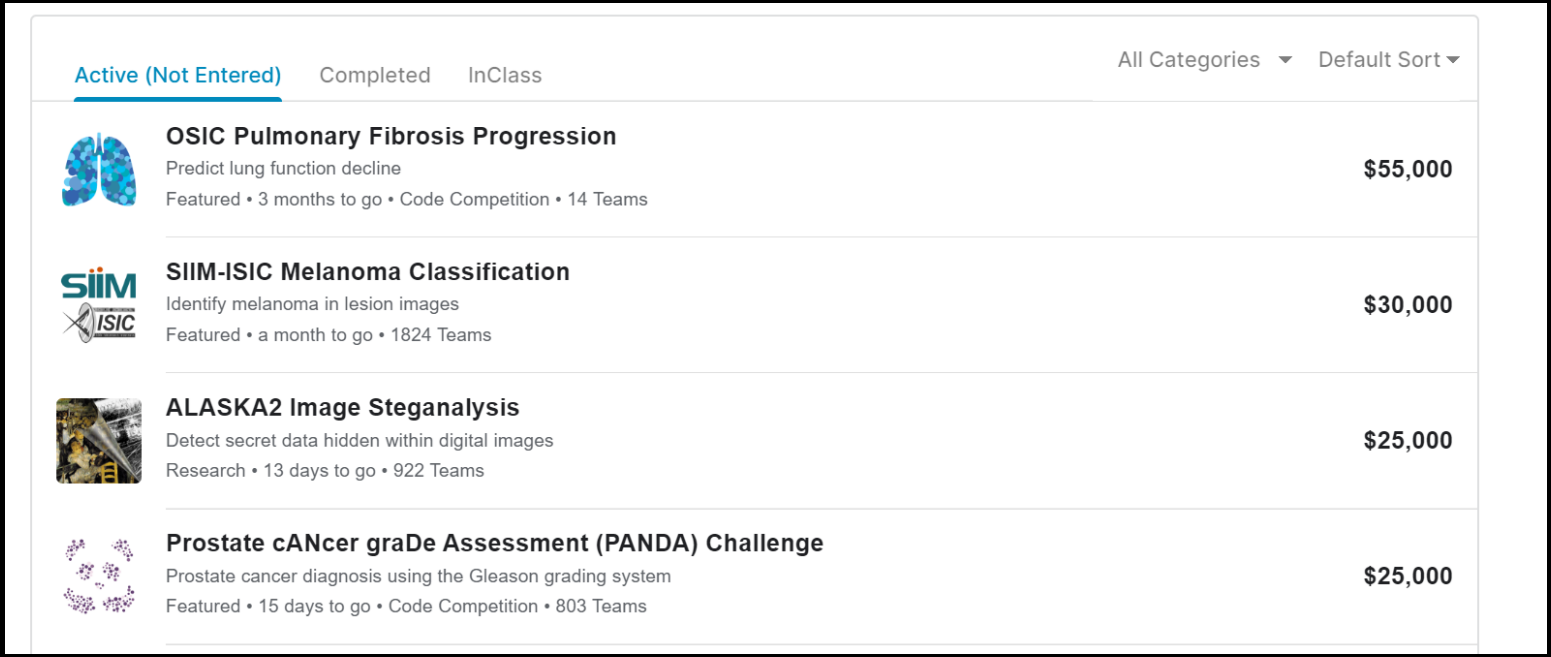
STEPS FOR MODELLING

1. **Prepare** the strategy to **evaluate the model**: separate the observations in a training set, a validation set (the latter is usually a subset of the training set) and a test set. No information from the test set should participate in the model training process.
2. **Preprocess the data**: apply the necessary transformations
3. **Select** a model
4. **Cross-validation** and Model Evaluation
5. **Hyperparameter** optimization
6. **Make the prediction** and error in the Test set

Kaggle

www.kaggle.com

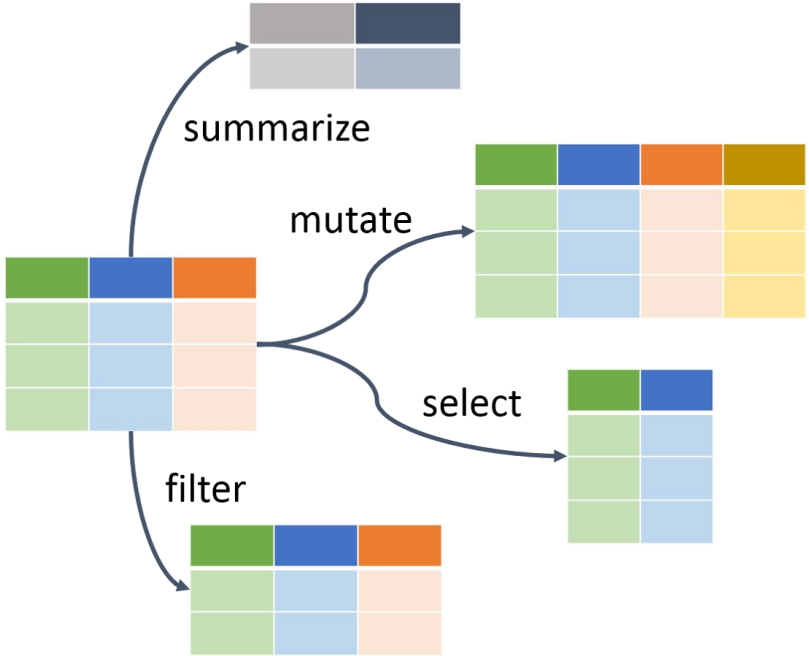
Kaggle is the world's largest data science community with powerful tools and resources to help you achieve your data science goals.



Trabajo…


ggplot2
www.rstudio.com


dplyr
www.rstudio.com


summarize
mutate
select
filter

Machine Learning for Biomedical Data

A complete ML application pipeline

2021-2022

https://www.kaggle.com/ronitf/heart-disease-uci

by Fernando Caraza