AFEKA TEL-AVIV ACADEMIC COLLEGE OF ENGINEERING אפקה המכללה האקדמית להנדסה בתל־אביב

Machine Learning 2022 Exercise 3

For this exercises you will participate in a Kaggle competition – “Titanic Machine Learning from Disaster”.

You will perform a classification task.

You are required to classify between to labels: survived or not.

The completion dataset provides a small amount of features to play with. You can remove features or calculate new ones based on the given features.

The data description is available at: https://www.kaggle.com/c/titanic/data . Please make sure to read it carefully, make sure you understand the meaning of each column and how each features is likely to effect the survival chance of a passenger.

In case you have question about the competition or about the data, you are advised to refer to the discussion section of the competition:
https://www.kaggle.com/c/titanic/discussion .

You may take inspiration from the work of other participator in the notebook section of the competition: https://www.kaggle.com/c/titanic/notebooks.

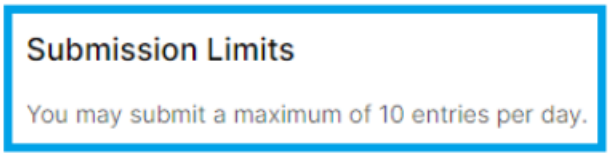
Your predictions to the competition will be measured using the accuracy metric, which is the percentage of passengers you correctly predicted rather the will survive or not.

Written by: Ido Meroz, IdoMe@afeka.ac.il
Based on the work of Mr. Aviad Malachi

AFEKA המכללה האקדמית להנדסה בתל-אביב TEL-AVIV ACADEMIC COLLEGE OF ENGINEERING אפקה

Requirements:

1. Your entire solution has to be written in a single python notebook (ipynb file)
2. Your submission to the Moodle has to include your solution file as an ipynb and as a html notebook (please refer to the attached guidance how to download your jupyter notebook as html) zipped into a single zip file.
3. How to download your ipynb as html: https://stackoverflow.com/a/64487858
4. The name of your files should begin with your full name and ID number.
5. The first MD –Mark Down cell of your submitted notebook has to include your full name, id number and a link to your Kaggle account.
6. You don't have to repeat the EDA- Essential Data Analysis phase. Use the EDA from exercise 1.
7. This exercise should be a continuation of exercise 1. Add a level 1 title (using a single #): "Exercise 3" and present your work in the cells underneath this title.
8. You may rearrange and improve work you have done in exercise 1.
9. Use random_state = 42.
10. You may use at least one (or more) of the following algorithms: KNN, NBC and or LDA.
11. Explain your choice, if you used more than one algorithm explain the differences and analyze the results.
12. Use CV (LPO, Kfold), try various feature selection algorithm. Preform hyper parameters search (Grid Search and or Random Search). Explain your choices and analyze them.
13. Please notice that you are limited to up to 10 submissions a day

    Therefore you should test your theories on the validation set prior to making a submission.
14. Try different hyper parameters, different sub groups of features.
15. Show the confusion matrix over CV, show KPIs calculated from the confusion matrix, explain the meaning of those values.
16. Plot graphs of Loss train verses validation and accuracy of train verses validation (for each fold separately). (You may compere additional statistics). Analyze your results and show your conclusions.

Written by: Ido Meroz, IdoMe@afeka.ac.il
Based on the work of Mr. Aviad Malachi

17. Use a MD cell to explain your comparisons between the train and the validation And your conclusion
18. Attach a screenshot of up to your 10 most recent submissions. Emphasize your best submission.
19. Attach a screenshot showing your score and place in the leaderboard.
20. Write a short TL;DR (under the title Exercise 3) – To Long Didn't Read (5-10 lines) at the second MD cell of your notebook (under your full name and your ID number) and a short summary on the bottom of your notebook. Use the TL;DR and the conclusion to explain your work.
21. Your last cell should hold references to resources you used including : notebooks you took inspiration from, links, books etc.

Notebook Structure:

1. Your name ID Number and link to your Kaggle account.
2. Follow the Structure from exercise 1.
3. "#" title: "Exercise 3"
4. TL;DR for Exercise 3
5. Experiments you made with feature selection, different models, ensembles and hyper parameters search.
6. Confusion matrix and KPIs , Graphs and results analysis.
7. Screenshots of submission and place in the leaderboard as mentioned above.
8. Summary
9. References

Grade Structure

1. Simple, organized explained and clean code 10%
2. Organized, understandable, explained notebook 10%
3. Effort and self-learning 10%
4. Correct implementation of the requirements and valid notebook structure 70%
5. 10% bonus for extreme effort.

Remarks

Written by: Ido Meroz, IdoMe@afeka.ac.il
Based on the work of Mr. Aviad Malachi

1. Chose a model preform feature selection, use different Ensembles methods (Bagging and or Boosting). Preform hyper parameters search using (Random Search and or Grid Search).
2. Explain your choices.
3. You should implement function in order to avoid code duplication.
4. Functions you implement should contain explanations of what they do.
5. You are advised to use meaningful names for variables and functions.
6. Show all of your work, even if it didn't work, and explain why (either why it worked or why it didn't work)

GOOD LUCK!

Written by: Ido Meroz, IdoMe@afeka.ac.il
Based on the work of Mr. Aviad Malachi