TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

KHOA CÔNG NGHỆ THÔNG TIN



Final Report Machine Learning

Introduction Machine Learning

*Người hướng dẫn*: TRẦN LÊ ANH CƯỜNG

*Người thực hiện*: NGUYỄN THÀNH KHANG – 518H0372

MẠC THUẬT ĐẠT-518H

Lớp : 18H50303

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

KHOA CÔNG NGHỆ THÔNG TIN



BÀI TẬP LỚN MÔN PHÂN TÍCH

& THIẾT KẾ GIẢI THUẬT

GIỚI THIỆU VỀ PHÂN TÍCH VÀ THIẾT KẾ GIẢI THUẬT

*Người hướng dẫn*: TS NGUYỄN CHÍ THIỆN

*Người thực hiện*: NGUYỄN THÀNH KHANG – 518H0372

MẠC THUẬT ĐẠT – 518H

Lớp : 18H50303

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020

**LỜI CẢM ƠN**

Đầu tiên, chúng em xin dành lời cảm ơn chân thành đến Trần Lê Anh Cường đã nhiệt tình giúp đỡ chúng em trong môn học này. Để nghiên cứu và hoàn thành bài tập này đúng hạn và hoàn thiện, xin cảm ơn những tác giả được nhắc đến ở phần tham khảo đã cho chúng em những tài nguyên hữu ích để hoàn thành bài viết này. Cùng lúc đó, xin được cảm ơn đến những người bạn và gia đình đã cùng giúp đỡ chúng em để hoàn thành được bài viết.

Cuối cùng, chúng em xin dành lời cảm ơn đến thầy, xin trân trọng cảm ơn.

ĐỒ ÁN ĐƯỢC HOÀN THÀNH

TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan đây là sản phẩm đồ án của riêng tôi và được sự hướng dẫn của TS Nguyễn Chí Thiện;. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình. Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 22 tháng 11 năm 2020*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Nguyễn Thành Khang*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng 11 năm 2020

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng 11 năm 2020

(kí và ghi họ tên)

**TÓM TẮT**

Present machine learning issues by building machine learning models for classification

problem.

1)Standardize data: using a dataset of numeric and categorial (nominal) types (1p)

2)Processing with unbalanced data, it means there is unequal/or bias distribution of classes

in the dataset. (1p)

3) Draw a chart showing statistical information of the dataset: quantity by labels (classes).

(1p)

4) Solving the problem by using classification models: kNN, NB, SVM, Logistic Regression,

MultiLayer Perceptron (i.e FFNN). (2p)

5) Using early stopping for training process, using validation dataset. (1p)

6) Train the model using parameters: batch size and epoch number. (1p)

7) Display model evaluation results with different measures. (1p)

8) Present the overfiting issue and illustrate this issue through examples. (1p)

9) Presenting (by a program) an example of using Convolution Neural Network or Long Short Term Memory (2đ)

**MỤC LỤC**

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

CÁC KÝ HIỆU

*f*: function

CÁC CHỮ VIẾT TẮT

**f: function**

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

DANH MỤC HÌNH

DANH MỤC BẢNG

# **Predictors of mental health illness**

### **Question:**

**Can you predict whether a patient should be treated of his/her mental illness or not according to the values obtained in the dataset?**

This is my first kernel after taking several courses of ML. I'm trying to understand better every concept by practicing and Kaggle is a good place to do it. Thanks Kaggle team.

The proccess is the following:

1. [Library and data loading](https://www.kaggle.com/kairosart/machine-learning-for-mental-health-1/data?select=survey.csv#Library_and_data_loading)
2. [Data cleaning](https://www.kaggle.com/kairosart/machine-learning-for-mental-health-1/data?select=survey.csv#Data_cleaning)
3. [Encoding data](https://www.kaggle.com/kairosart/machine-learning-for-mental-health-1/data?select=survey.csv#Encoding_data)
4. [Covariance Matrix. Variability comparison between categories of variables](https://www.kaggle.com/kairosart/machine-learning-for-mental-health-1/data?select=survey.csv#Covariance_Matrix)
5. [Some charts to see data relationship](https://www.kaggle.com/kairosart/machine-learning-for-mental-health-1/data?select=survey.csv#Some_charts_to_see_data_relationship)
6. [Scaling and fitting](https://www.kaggle.com/kairosart/machine-learning-for-mental-health-1/data?select=survey.csv#Scaling_and_fitting)
7. [Tuning](https://www.kaggle.com/kairosart/machine-learning-for-mental-health-1/data?select=survey.csv#Tuning)
8. [Evaluating models](https://www.kaggle.com/kairosart/machine-learning-for-mental-health-1/data?select=survey.csv#Evaluating_models)

* [Logistic Eegression](https://www.kaggle.com/kairosart/machine-learning-for-mental-health-1/data?select=survey.csv#Logistic_regressio)
* [KNeighbors Classifier](https://www.kaggle.com/kairosart/machine-learning-for-mental-health-1/data?select=survey.csv#KNeighborsClassifier)
* Naïve Bayes
* SVM
* Multilayer Peceptron

1. [Predicting with Neural Network](https://www.kaggle.com/kairosart/machine-learning-for-mental-health-1/data?select=survey.csv#Predicting_with_Neural_Network)
2. [Library and data loading](https://www.kaggle.com/kairosart/machine-learning-for-mental-health-1/data?select=survey.csv#Library_and_data_loading):
3. import numpy as np
4. import pandas as pd
5. import matplotlib.pyplot as plt
6. import seaborn as sns
8. from scipy import stats
9. from scipy.stats import randint
11. # prep
12. from sklearn.model\_selection import train\_test\_split
13. from sklearn import preprocessing
14. from sklearn.datasets import make\_classification
15. from sklearn.preprocessing import binarize, LabelEncoder, MinMaxScaler
17. # models
18. from sklearn.linear\_model import LogisticRegression
19. from sklearn.tree import DecisionTreeClassifier
20. from sklearn.ensemble import RandomForestClassifier, ExtraTreesClassifier
22. # Validation libraries
23. from sklearn import metrics
24. from sklearn.metrics import accuracy\_score, mean\_squared\_error, precision\_recall\_curve
25. from sklearn.model\_selection import cross\_val\_score
27. #Neural Network
28. from sklearn.neural\_network import MLPClassifier
29. from sklearn.model\_selection import learning\_curve, GridSearchCV
31. #Bagging
32. from sklearn.ensemble import BaggingClassifier, AdaBoostClassifier
33. from sklearn.neighbors import KNeighborsClassifier
35. #Naive bayes
36. from sklearn.naive\_bayes import GaussianNB
38. #Stacking
39. from mlxtend.classifier import StackingClassifier
41. # Dataset is now stored in a Pandas Dataframe
42. url = 'https://raw.githubusercontent.com/nthanhkhang/Machine\_Learning/main/survey.csv'
43. train\_df = pd.read\_csv(url)
44. #print(df.values[:5])
45. #print(train\_df.columns)
46. train\_df.head(5)
47. #print(train\_df.info())
48. [Data cleaning](https://www.kaggle.com/kairosart/machine-learning-for-mental-health-1/data?select=survey.csv#Data_cleaning):