CS 180

Filipino Student Employability Predictor using Multilayer Perceptrons

Team #14

Ralf Michael Balatibat, 2020-11685
Stephen Mary Encarnacion, 2020-02988
Michael Sean Brian Omisol, 2020-10660
Aaron Jude Tanael, 2020-04275

University of the Philippines Diliman

June 9, 2023

Presentation agenda

1. Introduction
2. Materials and methods
3. Results
4. Discussion, Q&A

Web app link

https://theredswarm-employability-predictor.streamlit.app

Presentation agenda

1. Introduction
2. Materials and methods
3. Results
4. Discussion, Q&A

Project introduction

- We tackle the project introduction into four (4) parts:
  - Background
  - Motivation and Importance
  - Objectives and Goals
  - Scope

Background

- In recent years, artificial intelligence and machine learning have demonstrated their transformative potential in addressing complex problems, optimizing processes, and uncovering valuable insights from vast amounts of data.
- Organizations across sectors, ranging from healthcare and finance to transportation and manufacturing, have recognized the immense benefits of leveraging artificial intelligence and machine learning technologies.

Background

- As graduation approaches, many fresh graduates worry about finding a job quickly. This can create immense pressure for them (Springboard Philippines, 2021). They may even experience anxiety due to the looming transition from studying at a university to becoming part of the workforce (Vallesteros et al., 2020).
- As the transition can be seen as a daunting shift in environment, the students may feel concerned and worried, as they often lack insights and the adaptability that the employers are looking for.

Motivation and importance

  - *As graduation approaches, many fresh graduates worry about finding a job quickly. This can create immense pressure for them (Springboard Philippines, 2021). They may even experience anxiety due to the looming transition from studying at a university to becoming part of the workforce (Vallesteros et al., 2020).*
  - *As the transition can be seen as a daunting shift in environment, the students may feel concerned and worried, as they often lack insights and the adaptability that the employers are looking for.*

- We see this as an opportunity to utilize the power of machine learning in order to reduce the anxiety imposed by this looming transition.

Motivation and importance

- Our project aims to capitalize on the power of machine learning to tackle the specific challenges posed by the Filipino students worries on employability.
- With this, the students may be guided on what areas they need to work on in order to improve their chances of landing a job.

Objectives and goals

- The group aims to develop and train the multilayer perceptron algorithm in order to achieve the following goals:
  - Uncover meaningful patterns and insights with regards to students' employability.
  - Predict student employability based on the given traits a student has.
  - Develop a user-friendly web application that intuitively applies the algorithm into general use.

Objectives and goals

- To accomplish the project goals stated above, the group has to satisfy the project's key objectives as follows:
    - Preprocess the acquired "Students' Employability Dataset - Philippines" dataset for the model to use.
    - Train the Multilayer Perceptron model and rigorously evaluate their performance and generalization capabilities.
    - Apply the trained models for predictive analytics, enabling sufficiently accurate predictions, and actionable recommendations.
    - Deploy a web application that future employees can use to identify whether they can be employed based on the following features defined in the model.

Scope

- This project seeks to predict whether a Filipino student is employable based on their assessed traits.
- This means that the algorithm "judges" whether the student is employable or less employable (i.e., prone to rejections in job applications) given the student's traits.
- However, this does not mean the algorithm serves as the hiring manager that outrightly accepts or rejects that student on the spot.

Scope

- Note that the basis of classification, aside from the student's input traits, is also limited by the dataset itself.
- Nevertheless, it is still upon the student's overall performance in the job application process whether or not that student gets accepted in the applied job position, which is out of the project's scope.

Presentation agenda

1. Introduction
2. Materials and methods
3. Results
4. Discussion, Q&A

Project materials and methods

- We tackle the project materials and methods into four (4) parts:
  - Data set and sources
  - Preprocessing steps
  - Methodology
    - Model input/features
    - Model output/target
    - How the model works
  - Why is the model appropriate for the problem?

Data set and sources

- As previously stated in the project background, the data set used for the implementation of the project is the *Students' Employability Dataset - Philippines* readily available from *Kaggle*.
- `https://www.kaggle.com/datasets/anashamoutni/students-employability-dataset`
- Owned by *Anas Hamoutni*.

Data set and sources

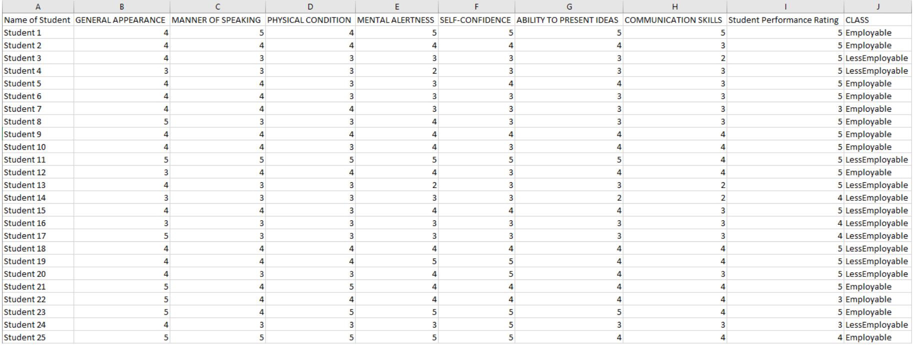
- Notes about the data set (as shown from Kaggle):
  - The data set was collected from different university agencies in the Philippines
  - The data set consists of mock job interview results of 2,982 observations.
  - However, the dataset collected requires normalization and cleaning.
  - The data set that was collected is compliant with the Data Privacy Act of the Philippines.

Data set format

- The data set is in .xlsx (MS Excel Spreadsheet) format.
- Can be imported using read_excel from pandas.

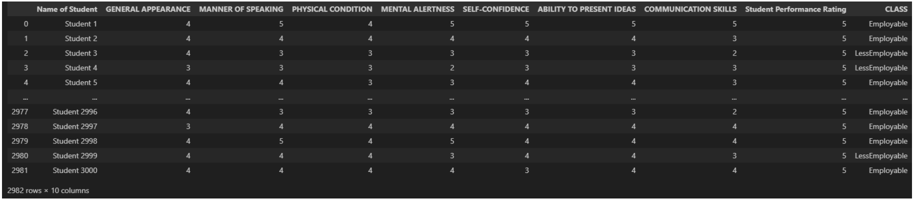
```
data = pd.read_excel('../data/Student-Employability-Da
 ↪   tasets.xlsx')
```

Data set as shown in MS Excel



A	B	C	D	E	F	G	H	I	J
Name of Student	GENERAL APPEARANCE	MANNER OF SPEAKING	PHYSICAL CONDITION	MENTAL ALERTNESS	SELF-CONFIDENCE	ABILITY TO PRESENT IDEAS	COMMUNICATION SKILLS	Student Performance Rating	CLASS
Student 1	4	5	4	5	5	5	5	5	Employable
Student 2	4	4	4	4	4	4	3	5	Employable
Student 3	4	3	3	3	3	3	2	5	LessEmployable
Student 4	3	3	3	2	3	3	3	5	LessEmployable
Student 5	4	4	3	3	4	4	3	5	Employable
Student 6	4	4	3	3	3	3	3	5	Employable
Student 7	4	4	4	3	3	3	3	3	Employable
Student 8	5	3	3	4	3	3	3	5	Employable
Student 9	4	4	4	4	4	4	4	5	Employable
Student 10	4	4	3	4	3	4	4	5	Employable
Student 11	5	5	5	5	5	5	4	5	LessEmployable
Student 12	3	4	4	4	3	4	4	5	Employable
Student 13	4	3	3	2	3	3	2	5	LessEmployable
Student 14	3	3	3	3	3	2	2	4	LessEmployable
Student 15	4	4	3	4	4	4	3	5	LessEmployable
Student 16	3	3	3	3	3	3	3	4	LessEmployable
Student 17	5	3	3	3	3	3	3	4	LessEmployable
Student 18	4	4	4	4	4	4	4	5	LessEmployable
Student 19	4	4	4	5	5	4	4	5	LessEmployable
Student 20	4	3	3	4	5	4	3	5	LessEmployable
Student 21	5	4	5	4	4	4	4	5	Employable
Student 22	5	4	4	4	4	4	4	3	Employable
Student 23	5	4	5	5	5	5	4	5	Employable
Student 24	4	3	3	3	5	3	3	3	LessEmployable
Student 25	5	5	5	5	5	4	4	4	Employable

Data set as shown in Python Notebook



	Name of Student	GENERAL APPEARANCE	MANNER OF SPEAKING	PHYSICAL CONDITION	MENTAL ALERTNESS	SELF-CONFIDENCE	ABILITY TO PRESENT IDEAS	COMMUNICATION SKILLS	Student Performance Rating	CLASS
0	Student 1	4	5	4	5	5	5	5	5	Employable
1	Student 2	4	4	4	4	4	4	3	5	Employable
2	Student 3	4	3	3	3	3	3	2	5	LessEmployable
3	Student 4	3	3	3	2	3	3	3	5	LessEmployable
4	Student 5	4	4	3	3	4	4	3	5	Employable
...	...	...	...	...	...	...	...	...	...	...
2977	Student 2996	4	3	3	3	3	3	2	5	Employable
2978	Student 2997	3	4	4	4	4	4	4	5	Employable
2979	Student 2998	4	5	4	5	4	4	4	5	Employable
2980	Student 2999	4	4	4	3	4	4	3	5	LessEmployable
2981	Student 3000	4	4	4	4	3	4	4	5	Employable

2982 rows × 10 columns

Preprocessing steps

- Before performing the MLP algorithm into our data set, we should consider normalizing and cleaning the data set first to achieve best results.

Data cleaning

- Since the data set is well-structured from the start (i.e., no null entries), all we need to do is to remove the Name of Student and CLASS columns in our data set prior to splitting the data set into train and test sets.

```
data = data.drop("Name of Student", axis=1)
x = data.drop('CLASS', axis = 1)
```

Data normalization

- Since we are using the MLP algorithm, we use MinMaxScaler with tanh activation to normalize our data set.

```
#Normalize train and test sets for MLP
from sklearn.preprocessing import MinMaxScaler

#TANH activation
scaler = MinMaxScaler(feature_range=(-1,1))
```

- We will use scaler later on the transformation of our train and test sets.

Model input/features

- GENERAL APPEARANCE
- MANNER OF SPEAKING
- PHYSICAL CONDITION
- MENTAL ALERTNESS
- SELF-CONFIDENCE
- ABILITY TO PRESENT IDEAS
- COMMUNICATION SKILLS
- Student Performance Rating

```
data = pd.read_excel('../data/Student-Employability-Datasets.xlsx')
data = data.drop("Name of Student", axis=1)
x = data.drop('CLASS', axis = 1)
```

Model output/target

- CLASS
  - Employable
  - LessEmployable

```
y = data['CLASS']
```

How the model works

- Splitting the data set into train and test sets using train_test_split

```
#70% training and 30% testing
x_train, x_test, y_train, y_test =
↪  train_test_split(x,y,train_size=0.7)
```

How the model works

- Fit and transform the x_train set using scaler from MinMaxScaler

```
x_train = scaler.fit_transform(x_train)
```

How the model works

- Transform the x_test set using scaler from MinMaxScaler

```
x_test = scaler.transform(x_test)
```

How the model works

- Fit the MLPClassifier model with tanh activation for MLP using the x_train and y_train sets

```
from sklearn.neural_network import MLPClassifier
mlp = MLPClassifier(activation='tanh')
mlp.fit(x_train,y_train)
```

How the model works

- Perform hyperparameter tuning on the model for better accuracy

```
param_grid = {'hidden_layer_sizes': [8], 'max_iter':
 ↪  [20000], 'learning_rate_init':
 ↪  [0.01],'alpha':[0.001], 'activation': ['tanh']}
grid = GridSearchCV(mlp,param_grid,refit=True)
grid.fit(x_train,y_train)
```

How the model works

- Obtain predictions from using the model

```
grid_predictions = grid.predict(x_test)
```

Why is the model appropriate for the problem?

- The model is appropriate for the problem since the data set is not particularly huge and the problem pretty much begs for a model solution with better accuracy.
- Although sacrificing processing speed, the use of MLP for this problem helps yield better results even though the data set used is not particularly huge as compared to the ones used for deep learning.
    - Accuracy is more important than speed since we aim to predict student employability based on the given traits a student has as accurate as possible.

Presentation agenda

1. Introduction
2. Materials and methods
3. Results
4. Discussion, Q&A

Project Results

- We tackle the project results into two (2) parts:
  - Summary of the results
  - Comparison to other models

Summary of the results

- For the result summary, we will talk about the:
  - Accuracy of the model
  - View of the confusion matrix
  - Classification Report

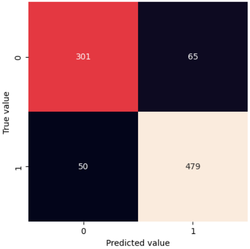
Accuracy of the model

- Upon using the Multilayer Perceptron to create predictions, we observed that the accuracy is around 87% (exactly 0.8715083798882681), as shown in the provided Python Notebook file.

```
# Display accuracy
from sklearn.metrics import accuracy_score
grid_predictions = grid.predict(x_test)
mlp_score = (accuracy_score(y_test, grid_predictions))
print(mlp_score)
```

View of the confusion matrix

- We view the confusion matrix as follows:



- We can see that there are some edge cases that the model did not handle well, causing false negative and false positive classifications.
- In this confusion matrix, the model has a bit more difficulty in classifying `LessEmployables` than of `Employables`.

View of the confusion matrix

```
# Display the confusion matrix
from sklearn.metrics import confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt

mat = confusion_matrix(y_test, grid_predictions)
sns.heatmap(mat, square=True, annot=True, fmt='d',
 ↪  cbar=False)
plt.xlabel('Predicted value')
plt.ylabel('True value')

plt.show()
```

Classification Report

- We view the classification report as follows:

```
              precision    recall  f1-score   support

           0       0.86      0.82      0.84       366
           1       0.88      0.91      0.89       529

    accuracy                           0.87       895
   macro avg       0.87      0.86      0.87       895
weighted avg       0.87      0.87      0.87       895
```

- We can see in the classification report that there is not that much of the difference between the two, although we can see in the `f1-scores` that classifying `Employable` students (presented as `1`) is a bit easier for the model to do than `LessEmployable` ones.

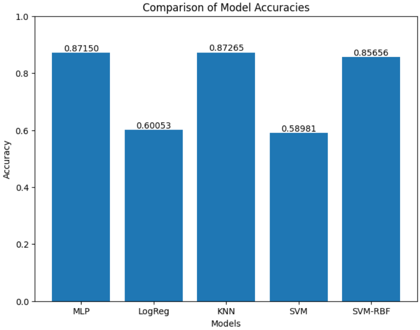
Classification Report

```
# Display classification report
from sklearn.metrics import classification_report
print(classification_report(y_test,grid_predictions))
```

Comparison to other models

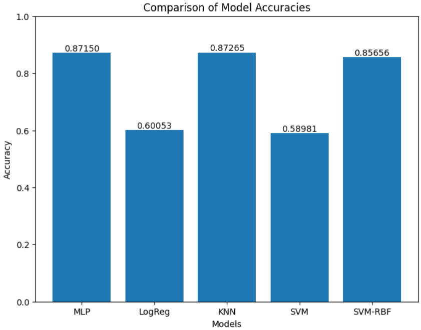
- *The model is appropriate for the problem since the data set is not particularly huge and the problem pretty much begs for a model solution with better accuracy.*
- *Although sacrificing processing speed, the use of MLP for this problem helps yield better results even though the data set used is not particularly huge as compared to the ones used for deep learning.*
    - *Accuracy is more important than speed since we aim to predict student employability based on the given traits a student has as accurate as possible.*

Comparison to other models



- Upon comparing the accuracy of the MLP model among different models utilized by a similar research shown above, it can be said that the MLP model is among the top performers when it comes to predicting student employability.

Comparison to other models



- It has a better accuracy score compared to logistic regression and SVM (linear), and is slightly better than SVM-RBF. Our model's accuracy is only beaten by the KNN model by a small margin.

Presentation agenda

1 Introduction

2 Materials and methods

3 Results

4 Discussion, Q&A

Discussion, Q&A

- We tackle the discussion part into four (4) parts:
  - Result interpretation
  - Conclusion
  - Improvements

What did we realize?

- We realize the power of Multilayer Perceptrons in solving classification problems in terms of accuracy.
- We realize that there can also be so many ways of solving a particular problem, may it be a mathematical solution or a machine learning algorithm.
- Upon this, with so many methods already given to us, there is always a chance we can use such methods to solve more problems in the future.

Conclusion

- We conclude in this project that using Multilayer Perceptrons as machine learning algorithm is not necessarily a bad idea even though it sacrifices performance for sake of accuracy.
    - After all, there are use cases wherein a problem demands for a solution with better accuracy than a faster one.
- We also conclude that we can find ways on solving interesting problems using machine learning. We can even use it to check if we are even fit to get employed in a job!
- We realized that machine learning can still do so many ways despite it being on top of artificial intelligence.

Improvements

- We are actually limited to the provided dataset in Kaggle. It would be much nicer if we can obtain larger datasets containing these kinds of evaluations from different companies in the Philippines.
- More features would be nice for a much accurate assessment of the student with regards to employability.
- If we, by chance, obtain a much larger dataset regarding employability, the use of deep learning might be an option. We may even work with other types of inputs such as a video sample of student speaking so that we may implement a sort of AI-based interviewer to evaluate if the student is `Employable` or `LessEmployable` in the future.

Q&A portion

Thank you!