CMPE430-Assignment Report

by Öner Efe GÜNGÖR

Submission date: 13-Dec-2025 03:52PM (UTC+0300)

Submission ID: 2845235671

File name: 16732_Oner_Efe_GUNGOR_CMPE430-Assignment_Report_56049_880317807.docx (360.09K)

Word count: 1621

Character count: 9226

ATILIM UNIVERSITY
FACULTY OF ENGINEERING

DEPARTMENT OF INFORMATON SYSTEMS ENGINEERING

CMPE430 ASSIGNMENT PROGRESS REPORT

Prepared By
1. Öner Efe, Güngör, 23243610018

12.12.2025

1. Dataset Overview

Dataset Name: Student Depression Dataset

Problem Type: Binary Classification

Dataset Info:

1) There are 27901 rows in the dataset which represents unique students and some other people with different professions(will be cleaned in further steps).

2) And there are 18 feature columns that represents ['id', 'Gender', 'Age', 'City', 'Profession', 'Academic Pressure', 'Work Pressure', 'CGPA', 'Study Satisfaction', 'Job Satisfaction', 'Sleep Duration', 'Dietary Habits', 'Degree', 'Have you ever had suicidal thoughts ?', 'Work/Study Hours', 'Financial Stress', 'Family History of Mental Illness', 'Depression']

Demographic: Gender, Age, City, Profession
Academic: CGPA, Academic Pressure, Study Satisfaction, Degree
Lifestyle: Sleep Duration, Dietary Habits, Work/Study Hours
Pyschological: Have you ever had suicidal thoughts ?
Other: id, Work Pressure, Financial Stress, Family History of Mental Ilness

3) Target variable is ‘Depression’ value which represented by ‘0’ and ‘1’ [‘1’ means student most likely has depression, ‘0’ means student has most likely no depression]

2. Data Preprocessing Strategy

- **Steps Applied:**
  1) Removing unrelated/unneccessary datas on dataset;

  1.1) City, Job Satisification, Work Pressure and id cleaned on dataset. Since we are dealing with students, ‘Job Satisification and Work Pressure’ values are unneccessary for the model and since ‘City’ values are too wide it is also not important data for the model and lastly ‘id’ is not significantly important for that model so they are all cleaned.

  1.2) Some noise reduction applied such as all rows dropped except than ‘Profession: Student’ rows since we are dealing with ‘Student Depression’ and after row reduction, ‘Profession’ column also cleaned from dataset for efficiency

  2) Encoding string valued features on dataset;

  2.1) Ordinal values such that: ‘Sleep Duration and Dietary Habits’ are encoded into integers relying on the importance and effect for the targeted value. It is higher value if it affects the targeted value in the positive way (‘Depression : 0’) and it is lower value if it affects the targeted value in the negative way (‘Depression : 1’).

  2.2) Nominal values such that: ‘Gender and Degree’ are encoded using One-Hot Encoding Method and for the ‘Suicidal thoughts and Family history’ are encoded into binary values which ‘Yes’ represents ‘1’ ‘No’ represents ‘0’.

  3) Scaling the Dataset

  3.1) Binary values remained same since they are already in the boundaries of [0,1] which are already small numbers.

  3.2) Non-binary values scaled into boundaries of [0,1] using MinMaxScaler algorithm.

- **Justification for Skipped Steps:** I didn’t applied Outlier Detection and Removal step. Since the dataset consists of student records with naturally bounded ranges
- **Data Splitting:** I splitted as %70 Training %15 Validation %15 Test.

3. Baseline Model

Model Architecture: 2 Hidden Layers, 2 Droput Layer(%20) 64 neurons (1st hidden layer), 32 neurons (2nd hidden layer), activation ReLU, optimizer 'adam', loss function 'binary-crossentropy'

Test Results:

```
=== FINAL RESULTS ===
----------------------------------------
Test Set Loss:          0.3493
Test Set Accuracy:      0.8491
----------------------------------------
Validation Loss:        0.3673
Validation Accuracy:    0.8337
----------------------------------------
```

Loss Graph:

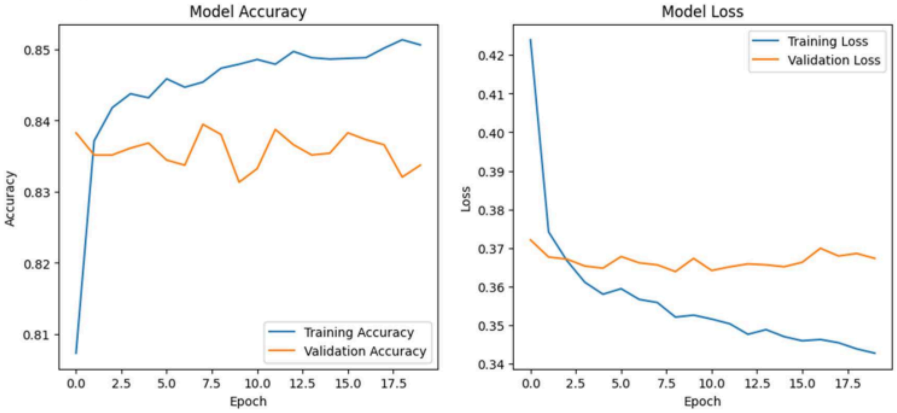

Figure 1: Baseline Model Loss and Accuracy Graph.

Initial Comment: Training is accurate but however, loss is too high and validation accuracy is lower than the training accuracy there may be an overfitting risk. Model may has too much neurons which causes memorization of patterns and overfitting.

4. Experimental Process

4.1. Experiment 1:

Hypothesis: Overfitting will be fixed with the following changes.
Change Made: 64 neurons decreased to 32 at the first layer, 32 neurons decreased to 16 at the second layer which is expected to make the model less 'clever' and make it stop to memorize the patterns.
Results:

```
=== FINAL RESULTS ===
----------------------------------------
Test Set Loss:          0.3476
Test Set Accuracy:      0.8536
----------------------------------------
Validation Loss:        0.3630
Validation Accuracy:    0.8359
----------------------------------------
```

Loss Graph:

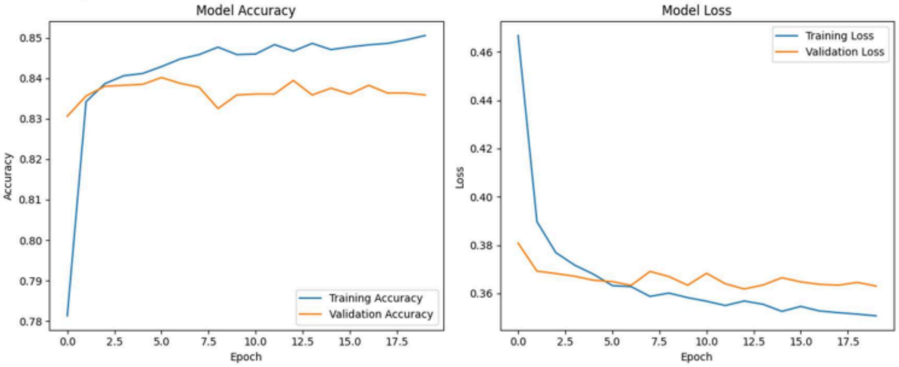

Figure 2: Experiment 1 Model Loss and Accuracy Graph.

Analysis: The difference between training and validation accuracy has been decreased due to the changes as expected so overfitting problem solved. We can say that is a good fit. But loss is still too high than expected so I need to check the hyperparameters to make sure it is the optimal one.

4.2. Experiment 2:

Hypothesis: The loss will be decrased and model accuracy will be increased by implementing optuna to the model, it will decide the best configuration for the hypermarameters which will be result with the goal.
Change Made: Implemented optuna with 4 trial to the model to set the best configuration .
Results:

```
=== FINAL RESULTS ===
----------------------------------------
Test Set Loss:          0.3479
Test Set Accuracy:      0.8515
----------------------------------------
Validation Loss:        0.3655
Validation Accuracy:    0.8335
----------------------------------------
```

Loss Graph:

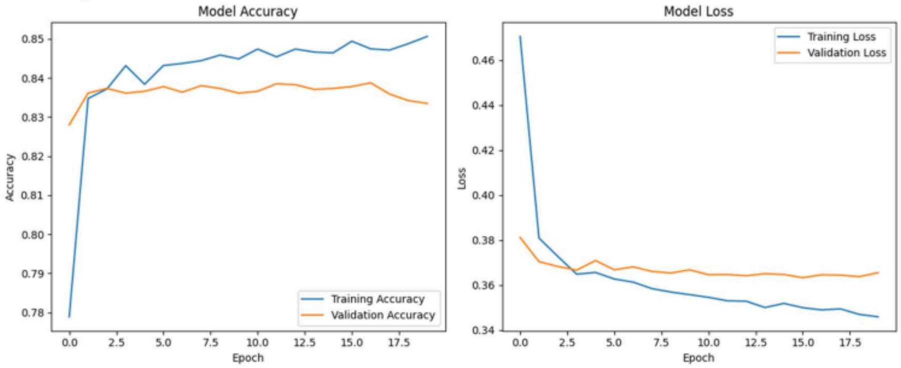

Figure 3: Experiment 2 Model Loss and Accuracy Graph.

Analysis: Optuma made some minimal changes to the results so we can say that it didn't work significantly , loss is still high and suprisingly, the difference between acurracies are increased which may be a sign of overfitting but we can assume that it is a good fit since the difference is not huge. I will try completely different setted model which includes less layer and more simple.

4.3. Experiment 3:

Hypothesis: Loss values will be decreased by configuring simpler model which may be more efficient.
Change Made: A complete simpler model has been setted which has 1 hidden layer with 8 neuron for efficiency.
Results:

```
=== FINAL RESULTS ===
----------------------------------------
Test Set Loss:          0.3432
Test Set Accuracy:      0.8543
----------------------------------------
Validation Loss:        0.3611
Validation Accuracy:    0.8366
----------------------------------------
```

Loss Graph:

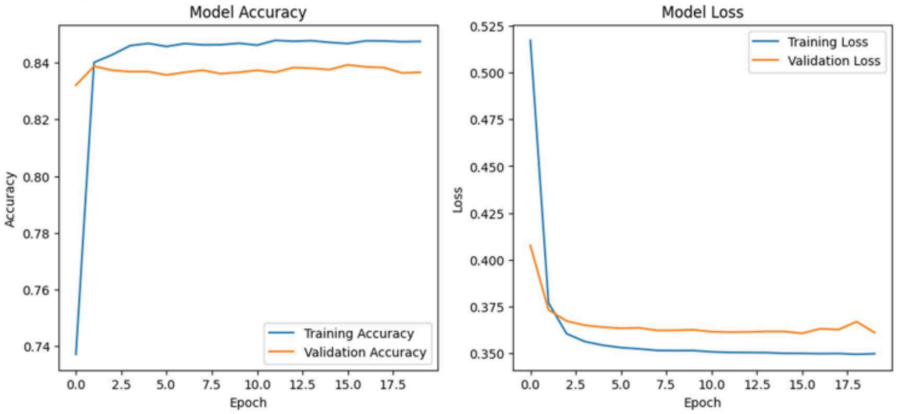

Figure 4: Experiment 4 Model Loss and Accuracy Graph.

Analysis: A simpler model worked more efficiently and more accurate as expected, the baseline model was too complex for simple binary classification problem with simple parameters. Graph is more smooth for this model and loss is smaller than the others. For furher experiments I will try to improve this 'Experiment 3 Model' as baseline model. In a nutshell , it is a good fit.

4.4. Experiment 4:

Hypothesis: Activation function change will effect the results in the postive way.
Change Made: Changed activation function ‘ReLU’ to ‘tanh’ to observe the effect of activation function.
Results:

```
=== FINAL RESULTS ===
----------------------------------------
Test Set Loss:          0.3415
Test Set Accuracy:      0.8531
----------------------------------------
Validation Loss:        0.3603
Validation Accuracy:    0.8400
----------------------------------------
```

Loss Graph:

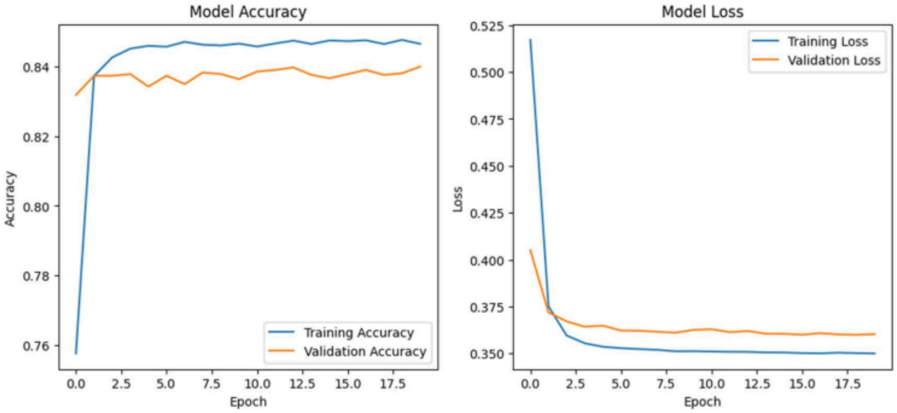

Figure 5: Experiment 4 Model Loss and Accuracy Graph.

Analysis: There is no significant difference between accuracies and model loss, on the contrary model accuracy graph is more ‘wavy’ than the ‘Experiment 3 model’ so we can clearly say that ReLU activation function is the best for that model. However, it is still good fit because accuracy is in the acceptable level.

4.5. Experiment 5:

Hypothesis: Changing optimizer to SGD with momentum and increasing epoch counts and batch size will be result with more accurate and less lossy model.
Change Made: Changed optimizer to SGD with momentum with learning rate 0.015 and momentum 0.5 and increased epoch counts 20 to 30 and batch size 32 to 64. Also activation function changed back to the ReLU.
Results:

```
=== FINAL RESULTS ===
----------------------------------------
Test Set Loss:          0.3420
Test Set Accuracy:      0.8529
----------------------------------------
Validation Loss:        0.3602
Validation Accuracy:    0.8371
----------------------------------------
```

Loss Graph:

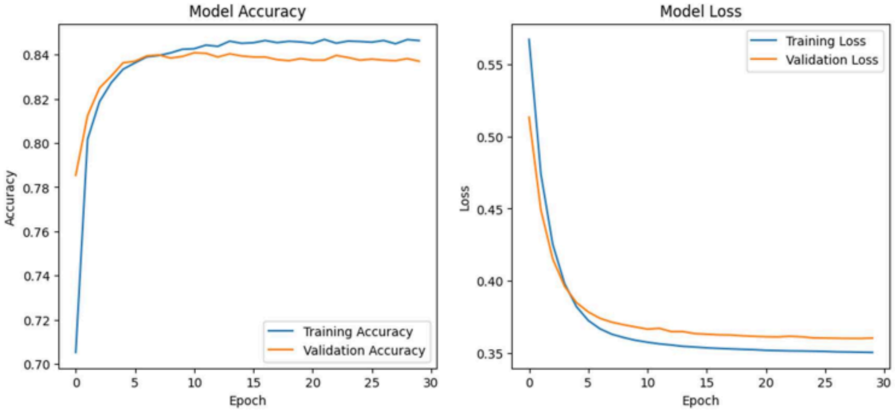

Figure 6: Experiment 5 Model Loss and Accuracy Graph.

Analysis: This graph is the smoothest one between all the models, loss is still high and accuracy is not more than 0.84 but it may be at its best. The reason may be the human answers always 'noisy' as data so it is hard to make an prediction at high accuracy. Model is in good fit.

5. Conclusion Table

Model	Change Made	Test Loss	Other Metrics	Overfitting/Underfitting?
Baseline	(Standard)	Test Set Loss: 0.3493	Test Set Accuracy: 0.8491 Validation Loss: 0.3673 Validation Accuracy: 0.8337	Overfitting

Exp 1	64 neurons decreased to 32 at the first hidden layer, 32 neurons decreased to 16 at the second hidden layer which is expected to make the model less ‘clever’ and make it stop to memorize the patterns.	Test Set Loss: 0.3476	Test Set Accuracy: 0.8536 Validation Loss: 0.3630 Validation Accuracy: 0.8359	Good Fit
Exp 2	Implemented optuna with 4 trial to the model to set the best configuration.	Test Set Loss: 0.3479	Test Set Accuracy: 0.8515 Validation Loss: 0.3655 Validation Accuracy: 0.8335	Good Fit
Exp 3	A complete simpler model has been setted which has 1 hidden layer with 8 neuron for efficiency.	Test Set Loss: 0.3432	Test Set Accuracy: 0.8543 Validation Loss: 0.3611 Validation Accuracy: 0.8366	Good Fit
Exp 4	Changed activation function ‘ReLU’ to ‘tanh’ to observe the effect of activation function.	Test Set Loss: 0.3415	Test Set Accuracy: 0.8531 Validation Loss: 0.3603 Validation Accuracy: 0.8400	Good Fit
Exp 5	Changed optimizer to SGD with momentum with learning rate 0.015 and momentum 0.5 and increased epoch counts 20 to 30 and batch size 32 to 64. Also activation function changed back to the ReLU.	Test Set Loss: 0.3420	Test Set Accuracy: 0.8529 Validation Loss: 0.3602 Validation Accuracy: 0.8371	Good Fit

Table 1: Conclusion table.

Final Conclusion

I believe that the best model is the ‘Experiment 5’ model despite the loss values are still high and accuracy values are not significantly high. I think the reason is ‘unexpected human pyshcology’ which result with high loss values that means suprising values for the model. So I believe that the graphs and the objection values are completely normal and ‘Experiment 5’ model is a good model to estimate ‘Depression’ for the students. For this work, simpler models was always more accurate than the others and clearly the reason was; in the dataset, there are very simple metrics which some of them has ‘Binary: Yes or No’ values and that metrics was the most important ones for the target value. In a nutshell, after 5 experiments I think we find the optimal model for deciding ‘Depression’ value for students and that model is the ‘Experiment 5 Model’.