The dataset graduation_data.csv consists of 1311 rows with 43 features. Similarly. The dataset class_data.csv consists of 12513 rows with 37 features. Considering only the unique id data, class_data.csv had 707 rows and 37 features (as each subject taken by students is recorded in column). The steps taken for this project are:

1. Explore data and its features.
2. Data Preprocessing: Join the two tables, remove null values using KMeans clustering. and analyse data.
3. Get a true label for whether the students graduated or not by using One class SVM clustering method.
4. Select important features and apply normalisation.
5. Use a neural network and get the prediction whether the student graduated or not.
6. Get insight by exploring prediction of neural networks and compare it with true labels obtained from step 3.

Step 1: Explore data and its features

While exploring the data, it is found that:

1. Target label 'Graduated In Major' has many null values. There is no label for 'No'.
2. All the students are of under-graduated level as shown in column "DEG_1" of graduation_data.csv, however, some students have also taken graduate level courses as shown by 'SUBJECT' and 'LEVEL' column of 'class_data.csv'.
3. In 'class_data.csv', the classes taken by the students are recorded. So, there are multiple similar records of a particular student like age, citizenship_status, location, act score.
4. There are 604 additional students who have graduated whose class records are not in the class data.
5. From the graduate data, it is seen that the average GPA of students is 3.1 and average ACT score is 22.66.

Step 2: Data Preprocessing

- **Join two tables:** The first step was to merge the table. However, as in class_data.csv, there are multiple records of a particular student, only a single record was selected to merge. Then, the two dataset 'graduation_data.csv' and 'class_data.csv' (with removal of duplicate records) were left joined together on 'ID'.
- **Remove null values using KMeans clustering:** k-means clustering was used to remove the null values of each of the columns except our target label "Graduated In Major". After using clustering, 6 clusters were formed from which null values were calculated.
- **Analyze data:** After the data preprocessing step, from the joined table, it was able to analyse the dataset further. Some of them are listed below:

1. Out of 12513 total courses, 12478 is an undergraduate level course, and only 35 courses are graduate level courses.
2. There are 1003 male students and 308 female students in our preprocessed joined table we used for training neural networks.
3. From the joined table, one can observe that the majority of students enrolled are in range 18-28.

Step 3: Get a true label for whether the students graduated or not by using One class SVM clustering method.

One Class SVM was used because we had only data of the positive class "Yes" and for most of the cases, labels were not recorded. So, for that, we constructed one class SVM model to learn the features of the data in the class "Yes", and use it to label the unlabelled records. Originally, there were 376 records labelled "Yes" in our joined dataset. After applying clustering, additional 43 yes labels in the non-recorded labels were obtained, and other records were classified as they did not have features of class Yes.

Step 4: Select important features and apply normalization

Label Encoder was used to transform the non-numeric data to numeric form, min-max normalization was used to normalize the data, and PCA was used to select 5 important features.

Step 5: Use a neural network and get the prediction whether the student graduated or not.

Before using neural networks, the dataset was splitted to train (60%), validation (20%) and test set (20%) with stratified sampling to maintain uniform records of target class in each of the set. Then, a decision tree classifier was applied using scikit-learn on the data set in which test accuracy was obtained to be 0.97338. After applying the decision tree, a neural network was used to check if accuracy can be improved further. Early stopping was used with a tolerance value of 30. Different values of tolerance were used to determine the value that gives good results with less overfitting. I played with different tolerance values for the model, and tolerance value of 30 gave goodresult with less overfitting. The epoch was stopped at 171 using early stopping. BinaryCrossEntropy Loss and SGD with batch-size 32 and learning rate 0.001 was used.

The loss plot obtained was:

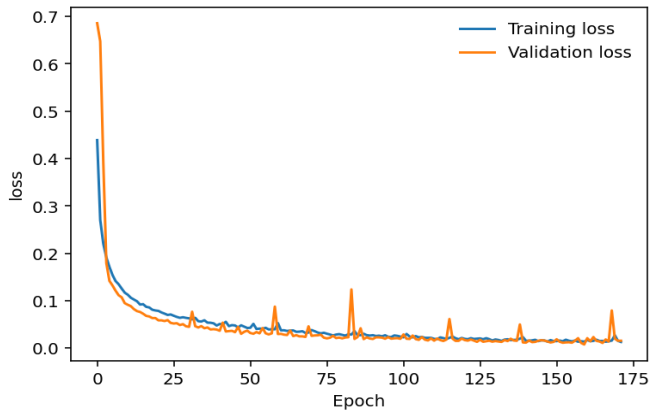
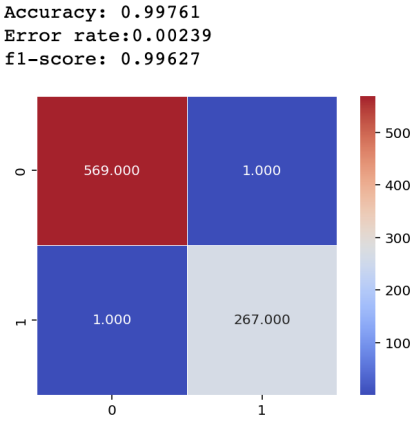
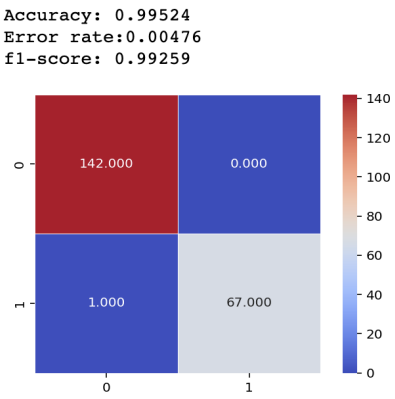


Fig. 1. Loss plot

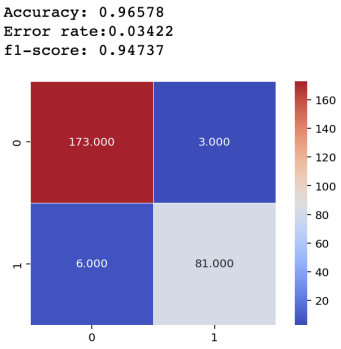
The confusion matrix obtained from neural network is as follows:



a) Train



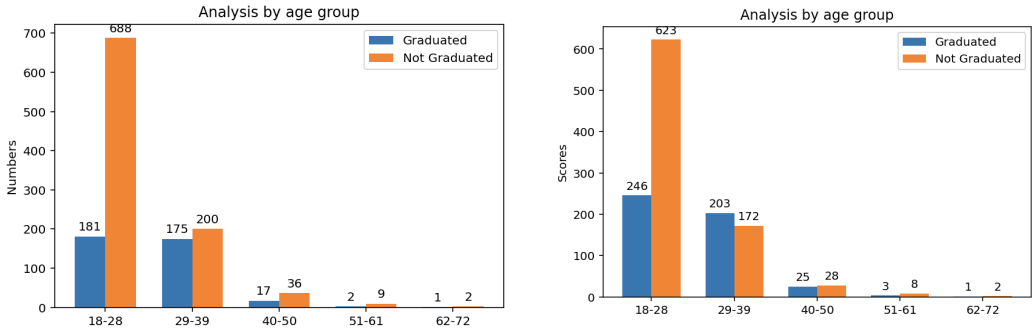
b) validation



c) Test

Fig. 2. Confusion Matrix

Analysis of Predicted and True Label and Insights Obtained from Neural Network



True label (Age Group) Predicted Label(Age Group)

Fig. .3 Graph of Age Group (True vs. Predicted label)

For the age-group graph as shown in fig. 3., we have divided the data to 5 age groups and calculated the total number of graduated and not-graduated students for each group.

Age Group	Total	% graduated	% not graduated
18-28	869	20.8	79.87
29-39	375	46.67	53.33
40-50	53	32.07	67.93
51-61	11	18.18	81.82
62-72	3	33.33	66.67

Predicted label

Age Group	Total	% graduated	% not graduated
18-28	869	27.8	72.2
29-39	375	54.6	45.4
40-50	53	45.3	54.7
51-61	11	27.27	72.73
62-72	3	33.33	66.67

From both the true label and predicted label, it can be observed that there are more students in the age-group 18-28, however, students in the age group 29-29 have graduated most of the times and looks promising.

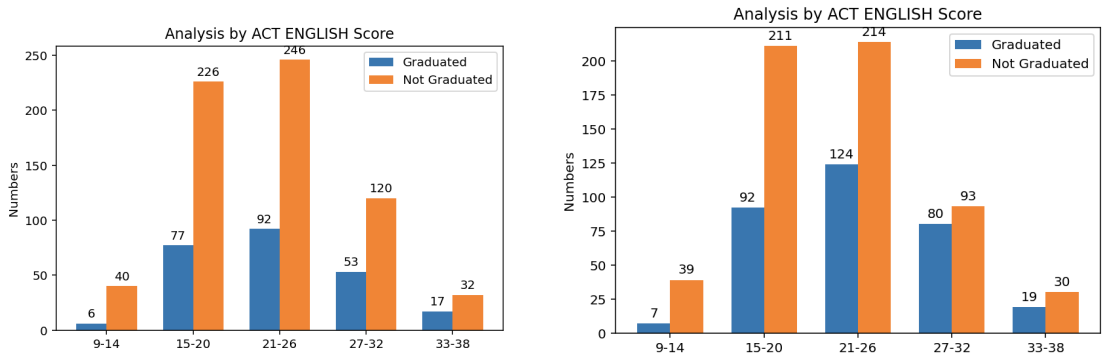


Fig. .4 Graph of ACT English score (True vs. Predicted label)

For the English ACT score graph, we have divided the data into 5 groups, 9-14, 15-20, 21-26, 27-32 and 33-38, and calculated the total number of graduated and not-graduated students for each group. We found the following observation:

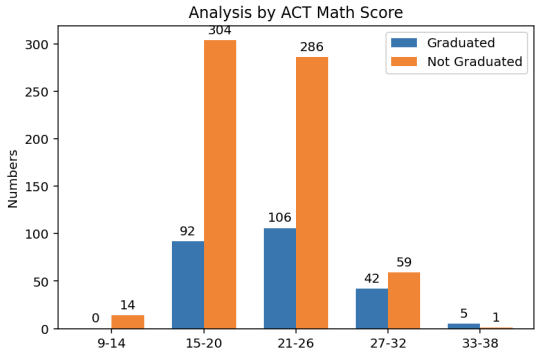
Observation from True label

ACT English	Total	% graduated	% not graduated
9-14	46	13.043	86.95
15-20	303	25.41	74.58
21-26	338	27.21	72.78
27-32	173	30.63	69.34
33-38	49	34.6	65.3

Predicted label

ACT English	Total	% graduated	% not graduated
9-14	46	15.21	84.79
15-20	303	30.36	69.64
21-26	338	35.7	64.3
27-32	173	45.6	54.4
33-38	49	38.7	61.3

From both the true label and predicted label, it can be observed that there are more students with the ACT English score 21-26, however, students with ACT English score 27-32 have graduated most of the time and look promising.



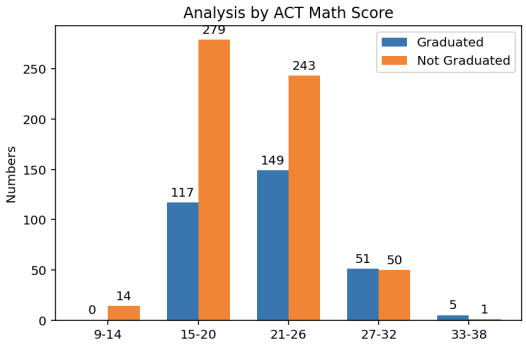


Fig. .5 Graph of ACT Math score (True vs. Predicted label)

Observation from True label

ACT Math	Total	% graduated	% not graduated
9-14	14	0	100
15-20	396	23.23	76.77
21-26	392	27	72
27-32	101	41.58	58.41
33-38	6	83.33	16.67

Predicted label

ACT Math	Total	% graduated	% not graduated
9-14	14	0	100
15-20	396	29.3	70.7
21-26	392	37.2	62.8
27-32	101	50.49	49.5
33-38	6	83.34	16.66

From both the true label and predicted label, it can be observed that there are more students with the ACT math score 15-20, however, students with ACT math score 33-38 have graduated most of the time and look promising. There are less samples of students in the range 33-38. The second range 27-32 looks promising as well.

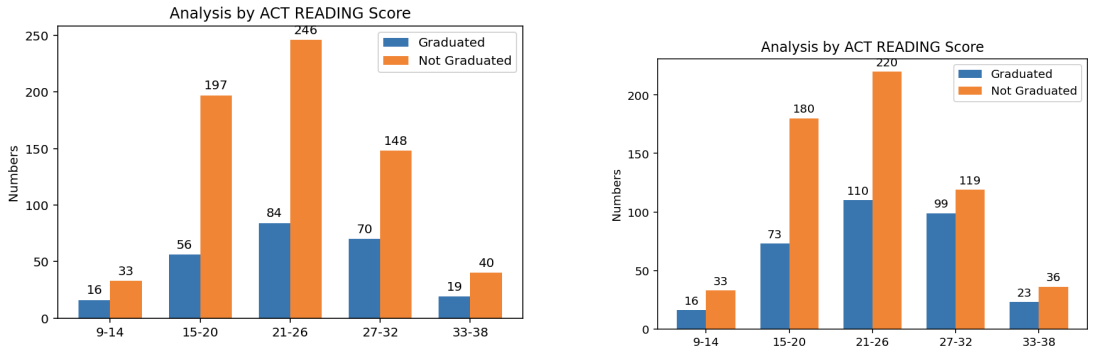


Fig. .6 Graph of ACT Reading score (True vs. Predicted label)

Observation from True label

ACT Reading	Total	% graduated	% not graduated
9-14	49	32.6	67.4
15-20	253	22.1	77.8
21-26	330	25.45	74.5
27-32	218	32.11	67.8
33-38	59	32	67.79

Predicted label

ACT Reading	Total	% graduated	% not graduated
9-14	49	32.6	67.3
15-20	253	27.6	72.4
21-26	330	33.33	66.67
27-32	218	45.4	54.6
33-38	59	37.28	62.72

From both the true label and predicted label, it can be observed that there are more students with the act reading score 21-26, however, students with ACT english score 27-32 have graduated most of the times and looks promising.

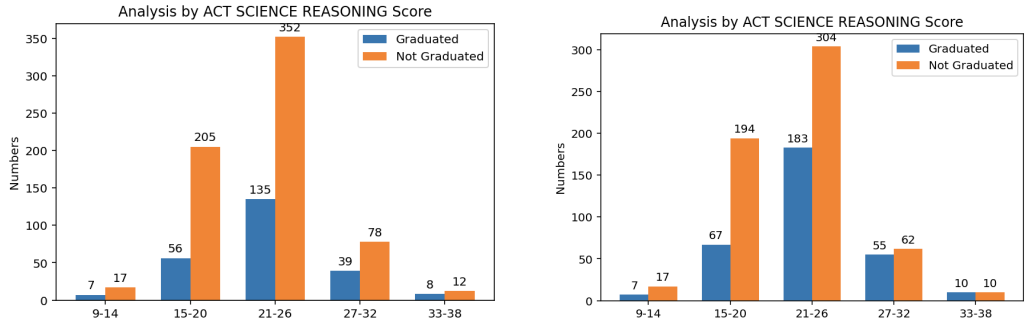


Fig. .7 Graph of ACT Science Reasoning score (True vs. Predicted label)

Observation from True label

ACT Science	Total	% graduated	% not graduated
9-14	24	29.16	70.84
15-20	261	21.4	78.6
21-26	487	27.7	72.3
27-32	117	33.33	66.67
33-38	20	40	60

Predicted label

ACT Science	Total	% graduated	% not graduated
9-14	24	29.16	70.84
15-20	261	25.3	74.7
21-26	487	37.16	62.84
27-32	117	46.15	53.8
33-38	20	50	50

From both the true label and predicted label, it can be observed that there are more students with the act science score 21-26, however, students with ACT science score 27-32 have graduated most of the times and looks promising.

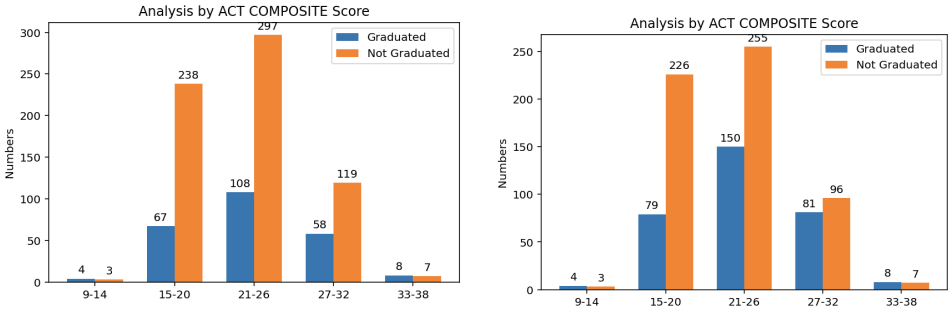


Fig. .8 Graph of ACT Composite score (True vs. Predicted label)

Observation from True label

ACT Composite	Total	% graduated	% not graduated
9-14	7	57.14	42.85
15-20	305	21.96	78.03
21-26	405	26.6	73.3
27-32	177	32.7	67.3
33-38	15	53.4	46.6

Predicted label

ACT Composite	Total	% graduated	% not graduated
9-14	7	57.14	42.86
15-20	305	25.9	74.1
21-26	405	36.29	63.7
27-32	177	45.19	54.81
33-38	15	53.34	46.66

From both the true label and predicted label, it can be observed that there are more students with the act composite score 21-26, however, students with ACT composite score 33-38 have graduated most of the times and looks promising.

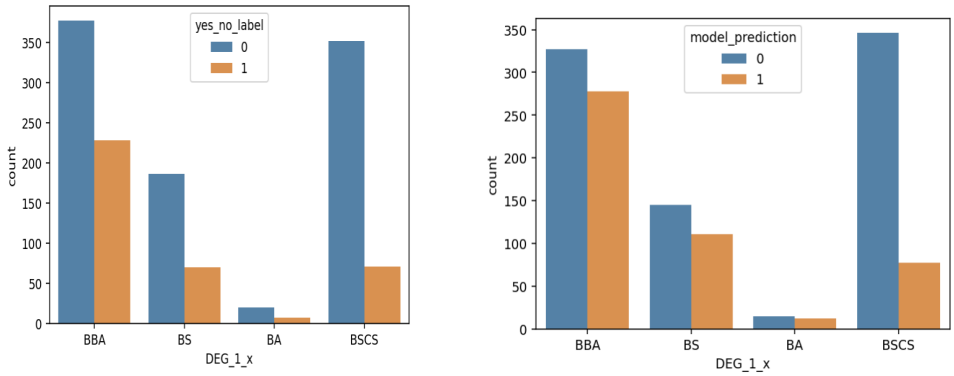


Fig. .9 Graph based on Degree (True vs. Predicted label)

From both the true and predicted label analysis of the graduation by subject, we can see that most of the students take BBA, students taking BA have high graduation rate. Students taking BSCS have a low graduation rate.