REPORT ON MACHINE LEARNING AND NEURAL COMPUTING DATA CLASSIFICATION COURSEWORK

TASK 1

a. The Panda standard library, ***pd.read_csv*** method with default settings was used to read the data into the panda frame..

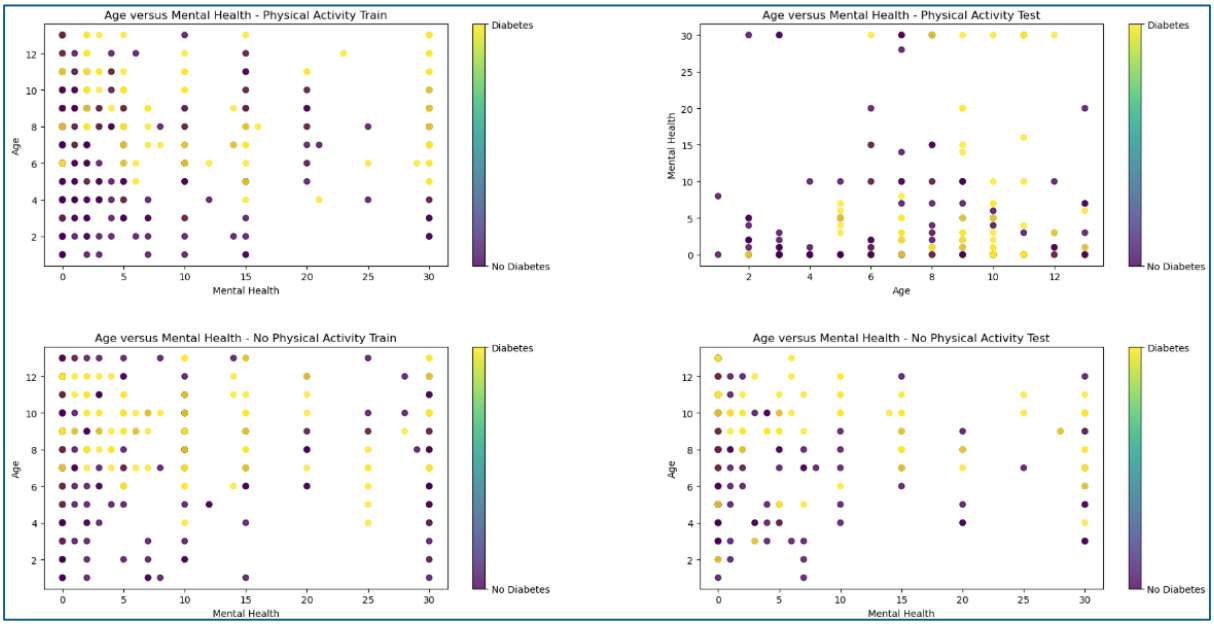

Figure1 shows the plot of age vs mental health for each dataset

b. The plot does not show any useful information in terms of a pattern or correlation. It is just a plain scatter plot of data points.

c. The data was then normalized using the **StandardScaler()** function from **sklearn** library. To do this, the numerical values for each train set (no activity and physical activity) excluding the label were fetched, and the **StandardScaler.fit()** function was applied on the values to obtain the statistics. Then, the **transform**() function was applied to normalize all values in corresponding train sets and the test sets. The mean result for the physical activity group was **-0.04,** and the standard deviation was **0.87.** The mean result for the no physical test set was 0.05, and the standard deviation was 0.93.

d.

	Principal component	NoActivity Variance (%)	PhysActivity Variance (%)
0	PC1	31.26	30.45
1	PC2	17.24	17.53
2	PC3	15.17	14.49
3	PC4	13.72	13.85
4	PC5	8.88	9.30
5	PC6	8.31	8.27
6	PC7	5.41	6.11

Figure 2: showing the PCA table showing the percentage variance by each principal component.

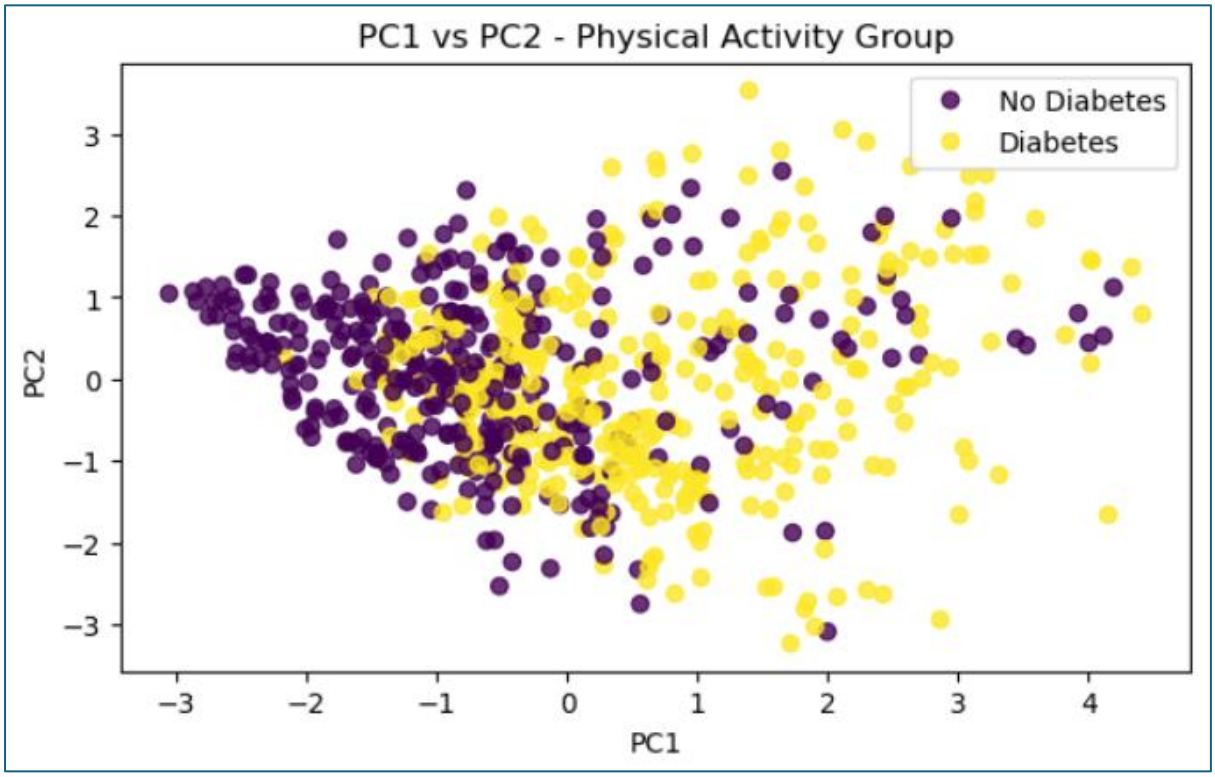

Figure 3: showing the pc1 vs pc2 -physical activity scatter plot.

The figure above shows the PCA analysis for the Physical Activity dataset. This graph shows the variance on how principal components of the labels (Diabetes, No Diabetes) of the dataset have strong discriminatory features. Here the components have features that are good discriminators.

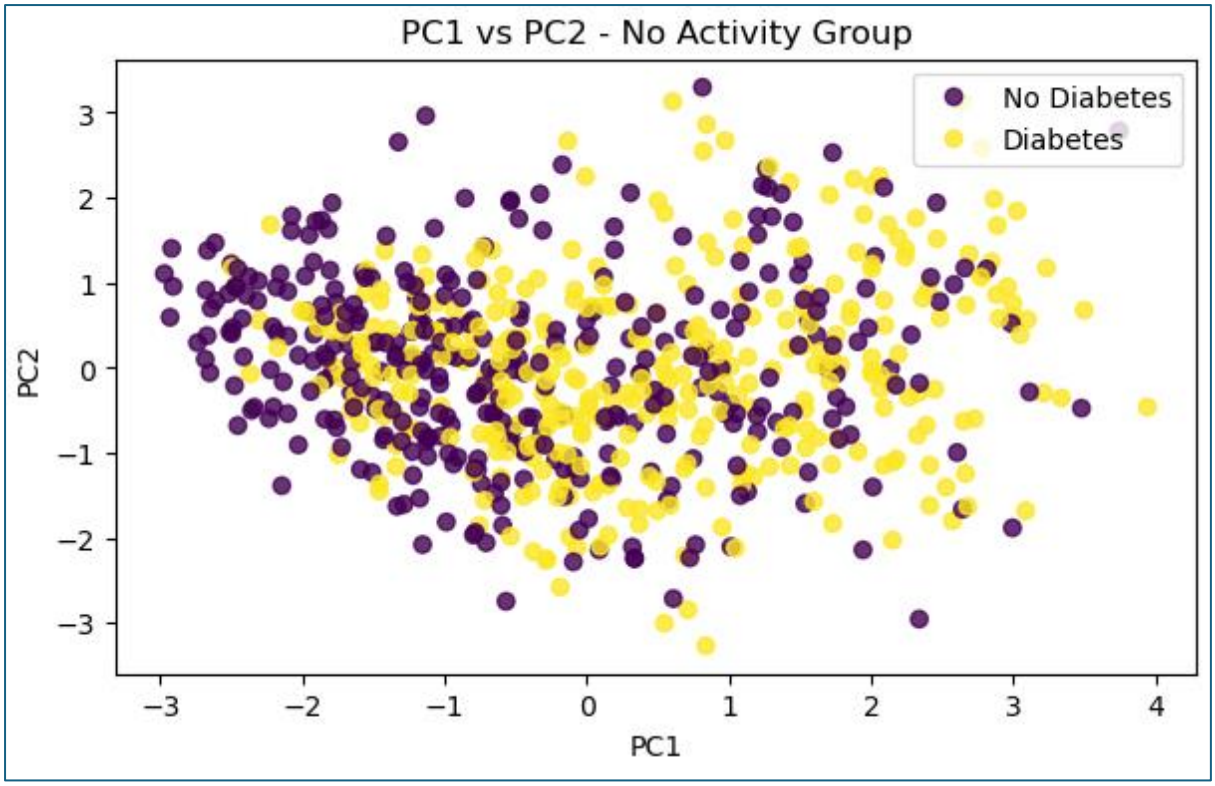

Figure 4: showing the pc1 vs pc2 no physical activity scatter plot

Figure 4 shows the pc1 vs pc2 scatter plot for the no-activity group dataset. From the plot, the features in the dataset do not have very strong differential features for discrimination between the two labels (diabetes, no diabetes) in the dataset as compared to the physical activity group.

Task 2

No Activity Dataset

a. After splitting the dataset, the smaller training set (II) has 490 rows. The validation dataset has 210 rows. I use 30% to test and 70% to train. To normalize my dataset, I use the **standardscaler()** function on the dataset to get the statistics. The statistics obtained from the train dataset for each class were used to normalise both the training dataset and the test dataset. The mean value of the train dataset of the two sets (small train set and test set) for each feature is not the same. The standard deviation of the train and test sets are approximately the same.

b. To determine the best combination of parameter values, the combination with the highest performance is deemed fit. Which is the [C, Y]: [0.5, 0.05].

c. The C and gamma values were used to train different models, including the model or the validation set and the small II train set. In the performance report, the [C, Y]: [0.5, 0.05] values had the highest values on the validation dataset with label 0 having a precision of 0.71, recall of 0.67 and f1-score of 0.69, and label 1 having a precision of 0.74, recall of 0.77 and f1-score of 0.76.

d. A model was trained on the whole train dataset with the [C, gamma]: [0.5,0.05] values and tested with the test dataset. The results obtained were a precision of 0.73, recall of 0.64,

f1-score of 0.68 for label 0, and a precision of 0.67, recall of 0.76, and f1-score of 0.71 for label 1. The overall accuracy was 0.70.

e.

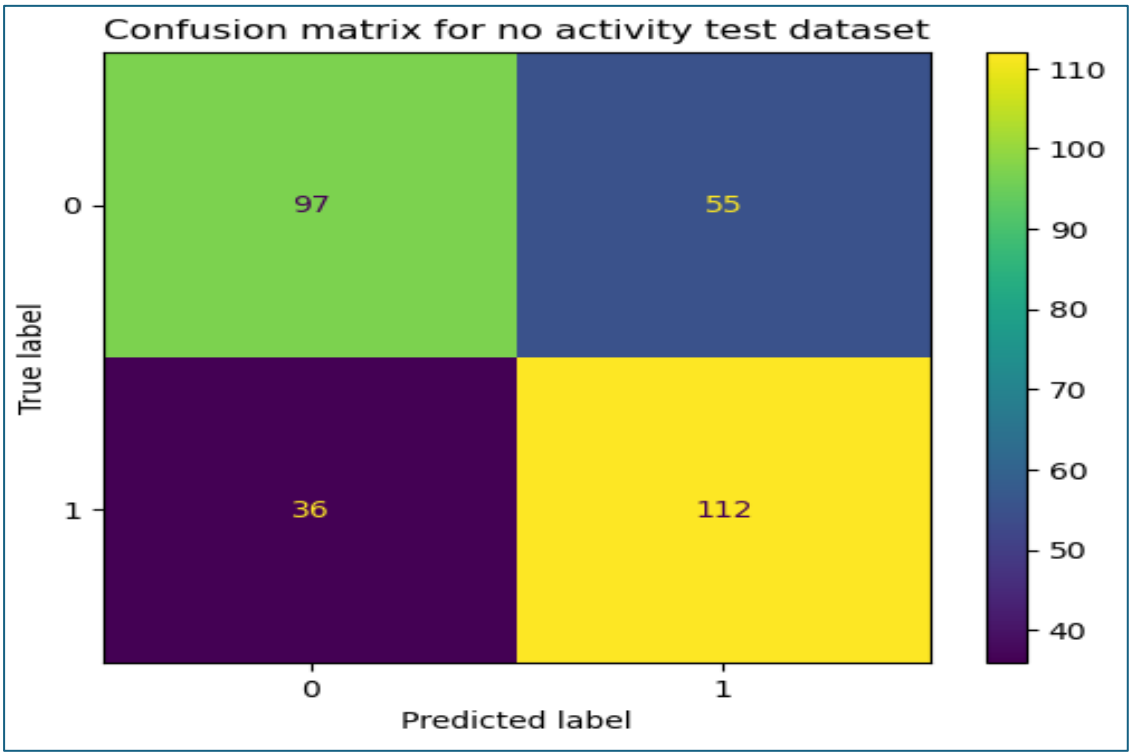

Figure 5; showing confusion matrix for the no activity dataset

From the confusion matrix, the model is better in predicting true positive values compared to the true negative values. The model will make more false positive classifications than false negative classifications.

Task 3

Physical Activity Dataset

a. After splitting the dataset, the smaller training set (II) has 490 rows. The validation dataset has 210 rows. I used 30% to test and 70% to train. To normalize my dataset, I use the **standardscaler()** function on the dataset to get the statistics. The statistics obtained from the train dataset for each class were used to normalise both the training dataset and the test dataset. The mean value of the train dataset of the two sets (train and test sets) for each feature is different. The standard deviation of the train dataset and of the two sets (small train dataset and validation dataset) are different.

b. To determine the best combination of parameter values, the combination with the highest performance is deemed fit. Which is the [C, Y]: [0.5, 0.05].

c. The C and gamma values were used to train different models, including the model or the validation set and the small I train set. In the performance report, the [C, Y]: [0.5, 0.05] values had the highest values on the validation dataset with label 0 having the precision of 0.74, recall of 0.65 and f1-score of 0.69 and label 1 having the precision of 0.74, recall of 0.81, and f1-score of 0.77. The accuracy of the model was 0.74.

d. A model was trained on the whole train dataset with the [C, gamma]: [0.5,0.05] values and tested with the test dataset. The results obtained were a precision of 0.81, recall of 0.69, f1-score of 0.74 for label 0, and a precision of 0.72, recall of 0.83, and f1-score of 0.77 for label 1. The overall accuracy was 0.76.

s

The confusion matrix is shown below:

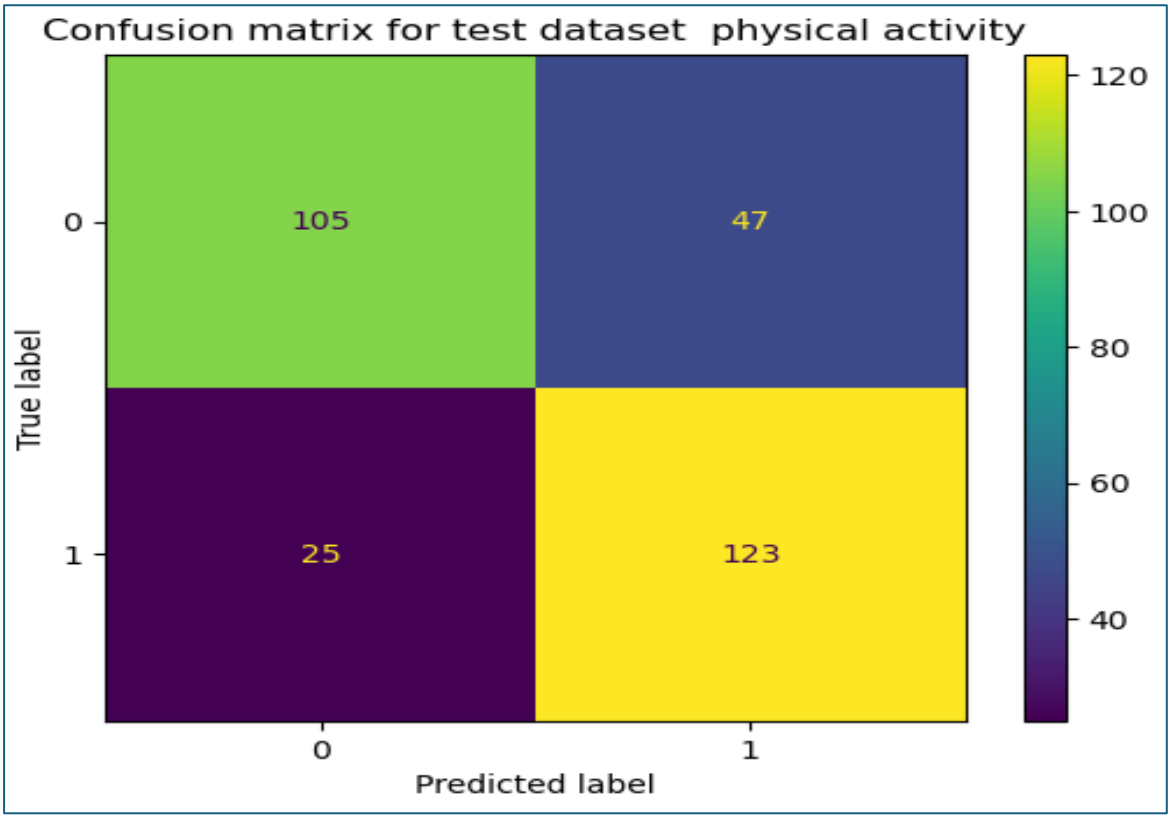

Figure 6; showing confusion matrix for the physical activity test dataset

From the confusion matrix, the model is better in predicting true positive values compared to the true negative values. The model made more false positive predictions than false negatives.

Comparing the model of task 2 and task 3. The results of the physical activity model had more performance results than Task 2. I propose this to be in conjunction with the analysis of the PCA in both train sets. The physical activity group had more discriminatory features from the analysis, and could be the reason for high performance.

Task 4

i. The physical test dataset was normalized using the statistics of the no-activity train set. While the no activity test set was normalized using the statistics of the physical activity train set. The reason was to maintain consistency between datasets.

ii. The trained models used for the cross-model testing were the model trained on the whole no activity group train set and the whole physical activity group train set. These models were trained with the C and gamma parameters that produced the best performance results.

iii. The accuracy of the cross-model testing for the no-activity model with the physical activity test set was 0.73, and that of the physical activity model with no activity test set was 0.68.

The confusion matrix is shown below:

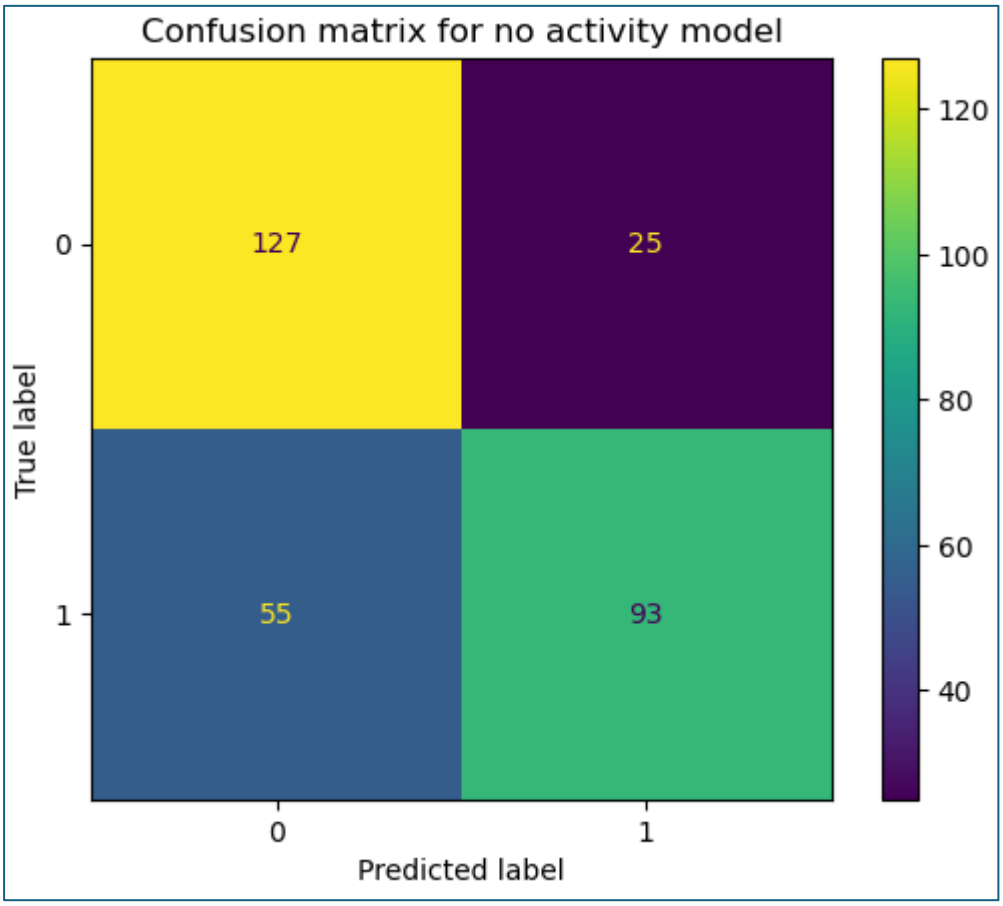

Figure 7: showing confusion matrix for no activity cross model testing on physical activity test set.

From the confusion matrix, the model has a high true negative prediction rate over true positives. The model has more false negative misses than false positive misses. The model will make better true negative classifications and good true positive classifications. However, the model will have a higher false negative classification misses and low false positive classification misses

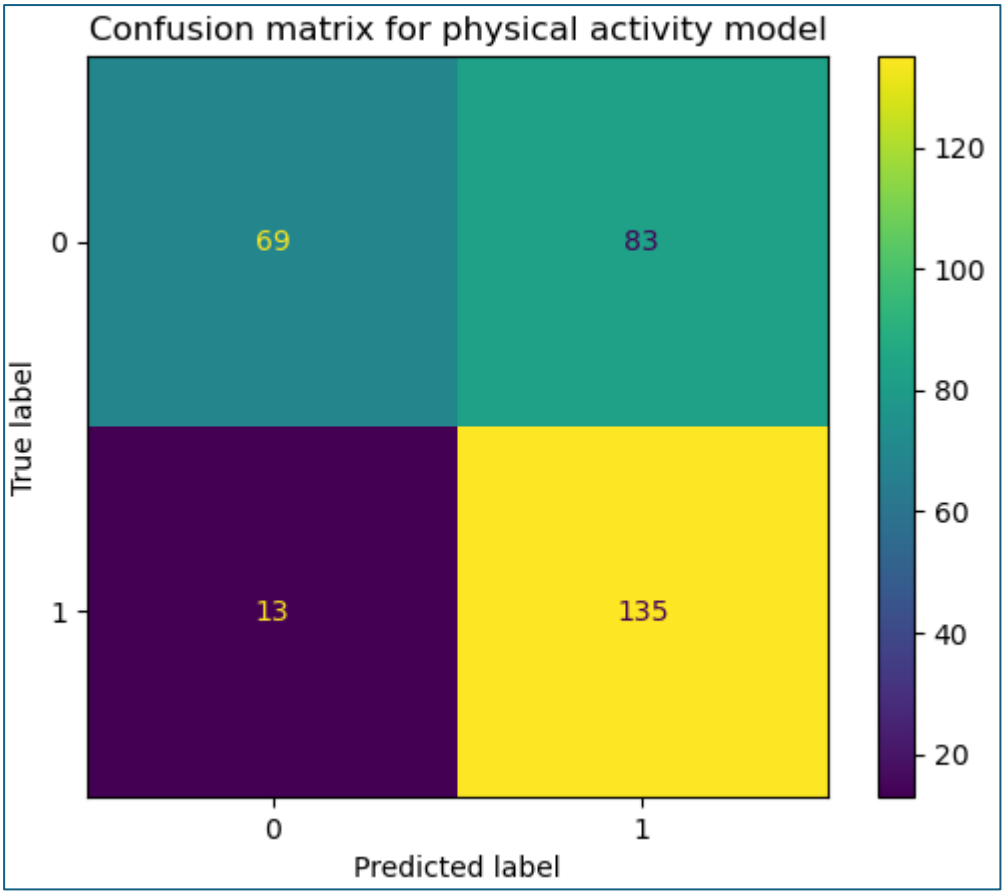

Figure 8 shows the confusion matrix for no activity cross model testing on the physical activity test set.

From the confusion matrix, the model has a high true positive prediction rate over false positive predictions. The model has more false positive misses than false negative misses. The false positive rates are higher than the true negative prediction rate. The model will always make the mistake of false positive classification. It will have low accuracy in true negative classification and high accuracy in true positive classification. The model will also make fewer false negative misses.

Conclusion

i. The no-activity model has better performance results compared to the physical activity model during cross-model validation. I suggest this is due to the mean and standard deviation differences for the train sets of both groups as analysed in task 2 and task 3 or potentially from the mean and standard deviation differences from the test sets of both groups in task 1. The difference in the accuracy result of both models is 5%. I think the potential reason could be the unbalanced dataset in both the training and test data set having four more diabetes points than the non-diabetes data points.

ii. The no-activity model has higher true negatives than the true positives. In physical activity, the model has high true positives compared to true negatives. One of the causes is that the physical health features might have been challenging for both models to generalize and might have influenced the numerical properties of other

features like BMI, Mental health, and General Health but not Education and Income.