Obesity Level Prediction Machine Learning Report

(By :- Suraj and Shivam)

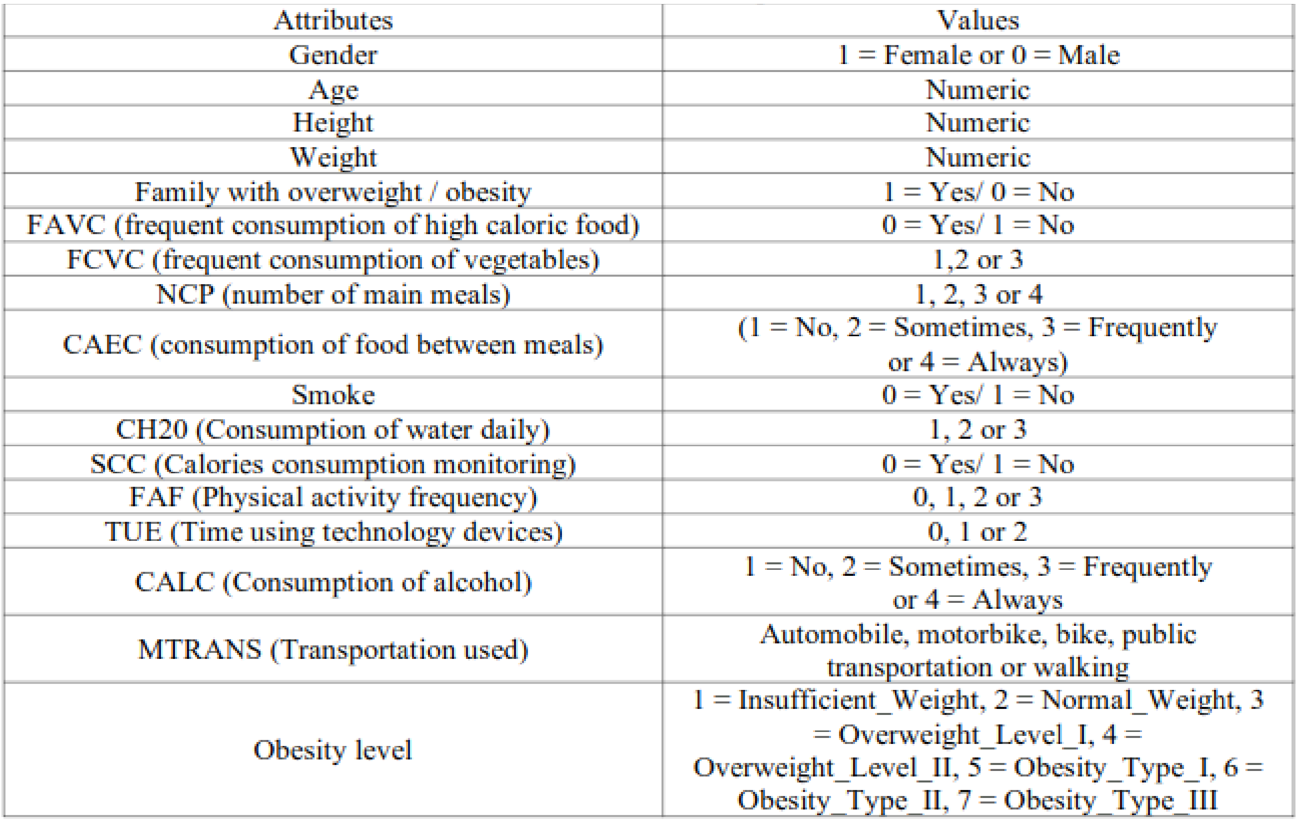
Introduction

Welcome to the future of personalized health!We have built an Obesity Level Predictor using Machine Learning which offers precise insights into individuals' obesity risks.We have applied 5 advanced algorithms(ANN,SVM,Logistic,KNN,Random forest). These machine learning models could help physicians identify overweight or obese individuals and thus accelerate the early detection, prevention, and treatment of obesity-related diseases. In this way these predictive models can help the doctors in better decision-making.

About the Data

This data comes from the UCI Machine Learning Repository This dataset includes data for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition.

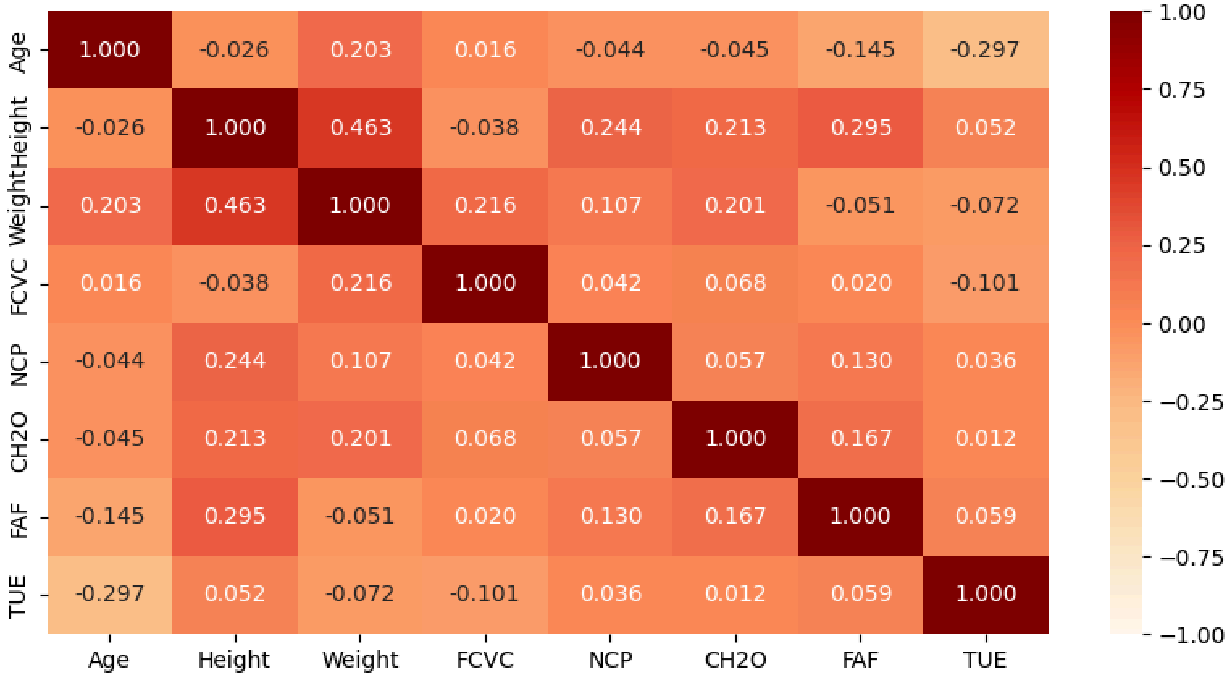
Table 1 Dataset Description

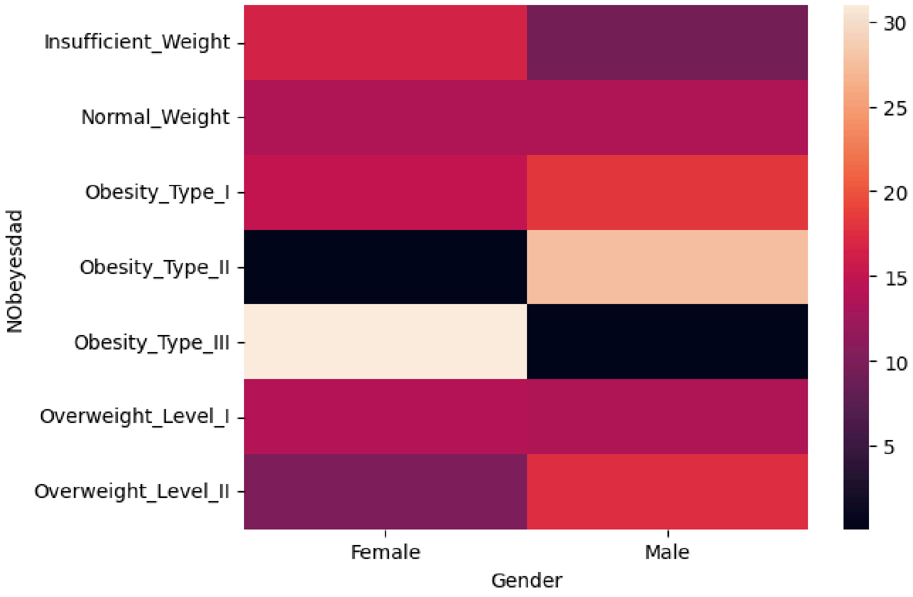


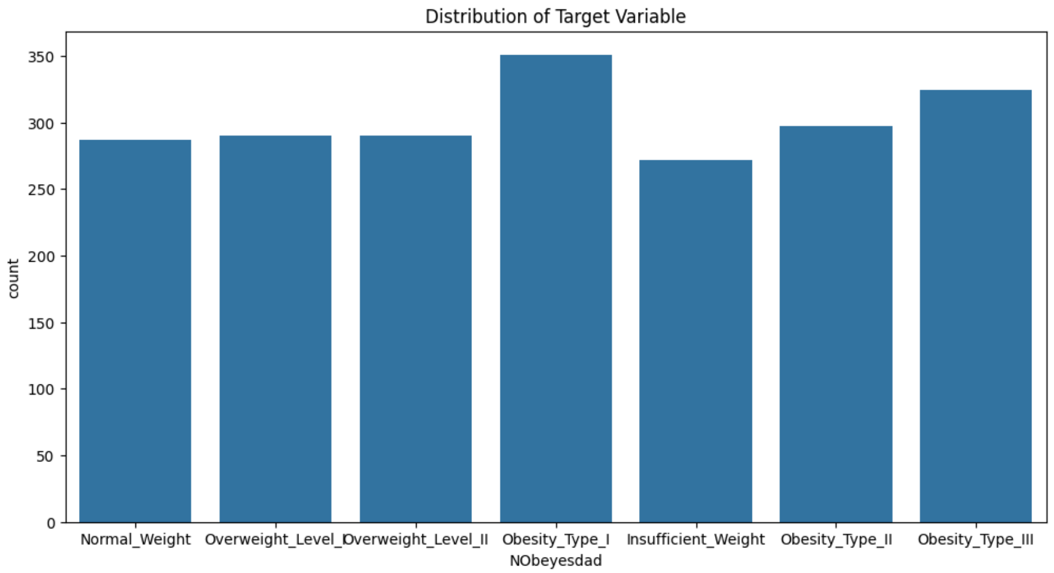
Attributes	Values
Gender	1 = Female or 0 = Male
Age	Numeric
Height	Numeric
Weight	Numeric
Family with overweight / obesity	1 = Yes/ 0 = No
FAVC (frequent consumption of high caloric food)	0 = Yes/ 1 = No
FCVC (frequent consumption of vegetables)	1,2 or 3
NCP (number of main meals)	1, 2, 3 or 4
CAEC (consumption of food between meals)	(1 = No, 2 = Sometimes, 3 = Frequently or 4 = Always)
Smoke	0 = Yes/ 1 = No
CH20 (Consumption of water daily)	1, 2 or 3
SCC (Calories consumption monitoring)	0 = Yes/ 1 = No
FAF (Physical activity frequency)	0, 1, 2 or 3
TUE (Time using technology devices)	0, 1 or 2
CALC (Consumption of alcohol)	1 = No, 2 = Sometimes, 3 = Frequently or 4 = Always
MTRANS (Transportation used)	Automobile, motorbike, bike, public transportation or walking
Obesity level	1 = Insufficient_Weight, 2 = Normal_Weight, 3 = Overweight_Level_I, 4 = Overweight_Level_II, 5 = Obesity_Type_I, 6 = Obesity_Type_II, 7 = Obesity_Type_III

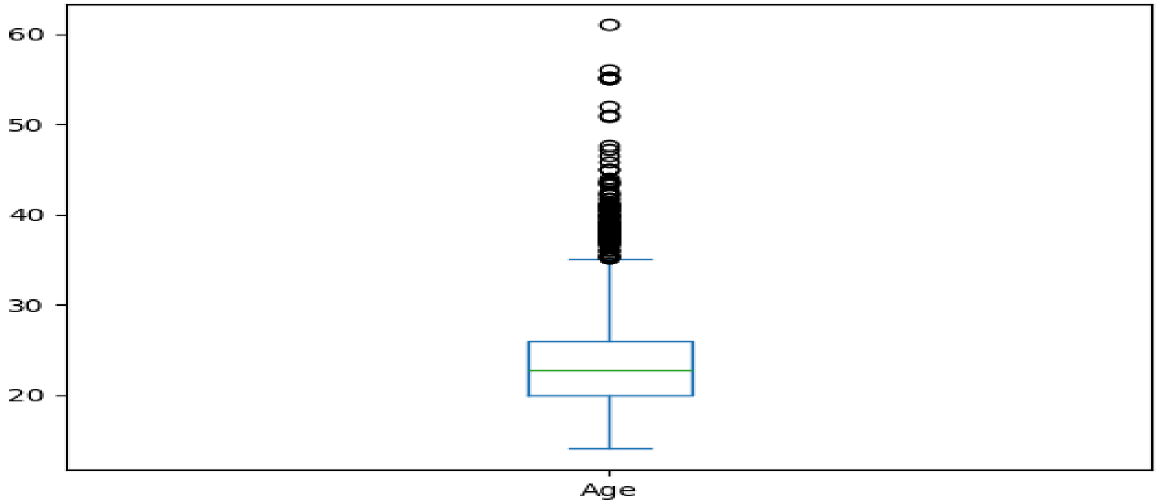
EDA

- Handle null value with mean for numerical column and most frequently occurrence for categorical column**.**
- **Correlation matrix analysis**
    - All columns except height are independent
    - Which is good for our model









Experiments and Evaluation

In the training and evaluation process, five machine learning methods were tested.
The predictions of the tested models were internally validated by 10-fold cross-validation and scores for accuracy, recall, precision and f1 are taken as the mean of these 10 fold results.. On the other hand,we have selected combinations of hyperparameters for each of the tested models. Table 2 presents the optimal hyperparameters for each model that we have tried to our best for better performance.

Table 2
Hyperparameters of the models

S.N	Models	Hyperparameters
1	Support vector machines	kernel='poly', C=0.5,
2	Random forest	n_estimators=300,criterion='entropy', min_samples_leaf=2
3	K-Nearest Neighbors	weights:uniform, n_neighbors:8, leaf_size: 30, algorithm: 'kd_tree'
4	Logistic Regression	tol: 0.0001, max_iter=1000, penalty="l2", solver="saga", shuffle: True
5	ANN	Nodes: 64,hidden layer: 2, dropout=0.12,kernel_initializer='uniform', activation='relu'

Table 3
Mean performance values of the models

Model	Train Score	Test Score	Accuracy	Precision (weighted)	Recall (weighted)	F1-score (weighted)
RF	0.95 (±0.01)	0.95 (±0.01)	0.96 (±0.01)	0.96 (±0.01)	0.96 (±0.01)	0.96 (±0.01)
Logistic	0.89 (±0.00)	0.88 (±0.01)	0.88 (±0.01)	0.87 (±0.01)	0.88 (±0.01)	0.88 (±0.01)
KNN	0.83 (±0.01)	0.79 (±0.01)	0.79 (±0.01)	0.79 (±0.01)	0.79 (±0.01)	0.78(±0.01)
SVC	0.89 (±0.00)	0.84 (±0.01)	0.84 (±0.01)	0.84 (±0.01)	0.84 (±0.01)	0.84 (±0.01)
ANN	0.99 (±0.00)	0.95 (±0.01)	0.95 (±0.01)	0.95 (±0.01)	0.95 (±0.01)	0.95 (±0.01)

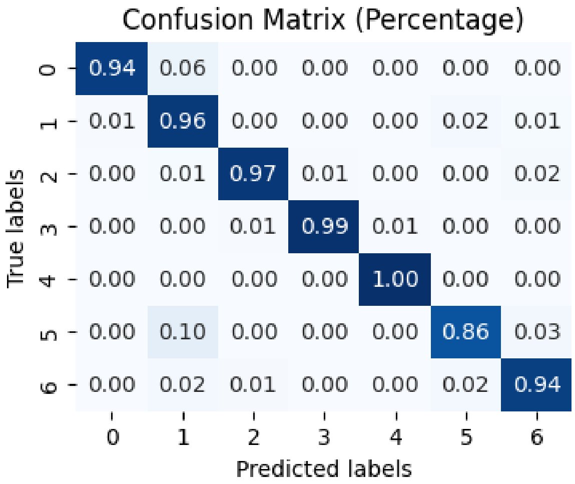
The results of Table 3 show that the K Nearest (KNN) model presented the lowest performance values, compared to the other models. On the other hand, Random Forest (RF) and Artificial Neural Network (ANN) exhibit the highest accuracy among the models, with RF slightly outperforming ANN. Logistic Regression, Support Vector Machines (SVM) show lower accuracy compared to RF and ANN.

RF and ANN models have shown robustness with narrow confidence intervals across all with high accuracy in all these parameters. Given their higher accuracy and relatively stable performance, Random Forest and Artificial Neural Network could be suitable choices for prediction.

Confusion Matrix Analysis:

In all confusion matrices, the diagonal elements are representing correctly classified instances. They are dominant, indicating that the models perform reasonably well in predicting the correct labels for most classes except in case of KNN having correctly classified about 37 percent.
Off-diagonal elements are representing misclassifications for corresponding labels. In logistic class 5 and 6 has been misclassified to 6 and 5 respectively about 14 percent each and class 1 has been up to 30 percent. Logistic classification, KNN and SVM shows similar misclassification across different classes but KNN has badly misclassified class 1.
For ANN and RF, we can see that more darker shades along diagonals indicate a lower misclassification rate compared to other modals.so on overall RF and ANN appears to be better performance.

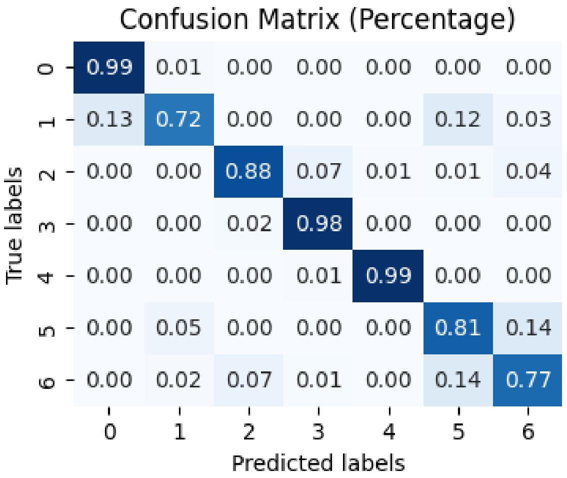
1. RF



Confusion Matrix (Percentage)

True labels \ Predicted labels	0	1	2	3	4	5	6
0	0.94	0.06	0.00	0.00	0.00	0.00	0.00
1	0.01	0.96	0.00	0.00	0.00	0.02	0.01
2	0.00	0.01	0.97	0.01	0.00	0.00	0.02
3	0.00	0.00	0.01	0.99	0.01	0.00	0.00
4	0.00	0.00	0.00	0.00	1.00	0.00	0.00
5	0.00	0.10	0.00	0.00	0.00	0.86	0.03
6	0.00	0.02	0.01	0.00	0.00	0.02	0.94

2. Logistic Regression

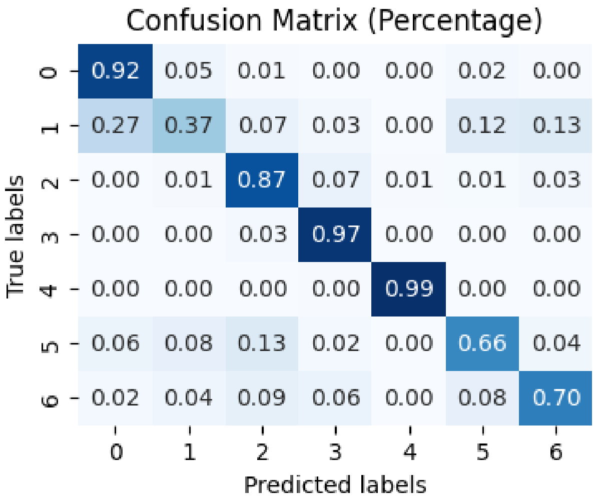


Confusion Matrix (Percentage)

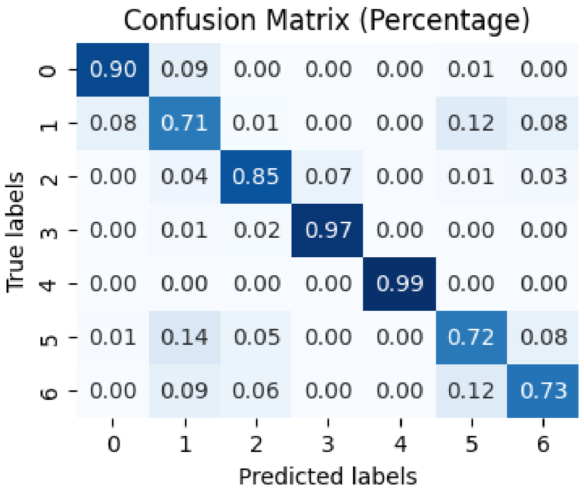
True labels \ Predicted labels	0	1	2	3	4	5	6
0	0.99	0.01	0.00	0.00	0.00	0.00	0.00
1	0.13	0.72	0.00	0.00	0.00	0.12	0.03
2	0.00	0.00	0.88	0.07	0.01	0.01	0.04
3	0.00	0.00	0.02	0.98	0.00	0.00	0.00
4	0.00	0.00	0.00	0.01	0.99	0.00	0.00
5	0.00	0.05	0.00	0.00	0.00	0.81	0.14
6	0.00	0.02	0.07	0.01	0.00	0.14	0.77

3. KNN

4. SVM

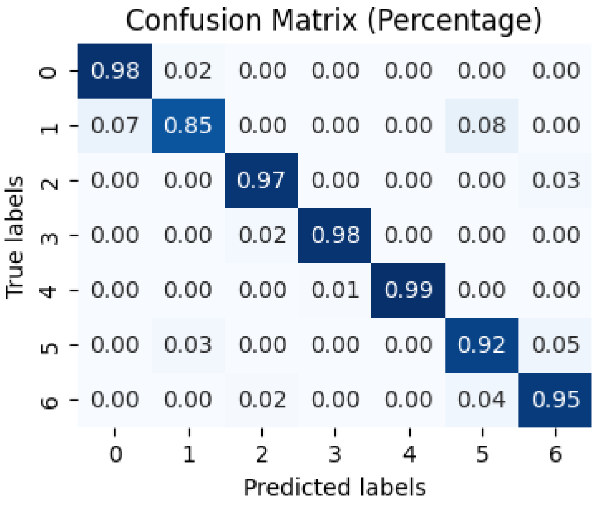

Confusion Matrix (Percentage)

	0	1	2	3	4	5	6
0	0.92	0.05	0.01	0.00	0.00	0.02	0.00
1	0.27	0.37	0.07	0.03	0.00	0.12	0.13
2	0.00	0.01	0.87	0.07	0.01	0.01	0.03
3	0.00	0.00	0.03	0.97	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.99	0.00	0.00
5	0.06	0.08	0.13	0.02	0.00	0.66	0.04
6	0.02	0.04	0.09	0.06	0.00	0.08	0.70


Confusion Matrix (Percentage)

	0	1	2	3	4	5	6
0	0.90	0.09	0.00	0.00	0.00	0.01	0.00
1	0.08	0.71	0.01	0.00	0.00	0.12	0.08
2	0.00	0.04	0.85	0.07	0.00	0.01	0.03
3	0.00	0.01	0.02	0.97	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.99	0.00	0.00
5	0.01	0.14	0.05	0.00	0.00	0.72	0.08
6	0.00	0.09	0.06	0.00	0.00	0.12	0.73

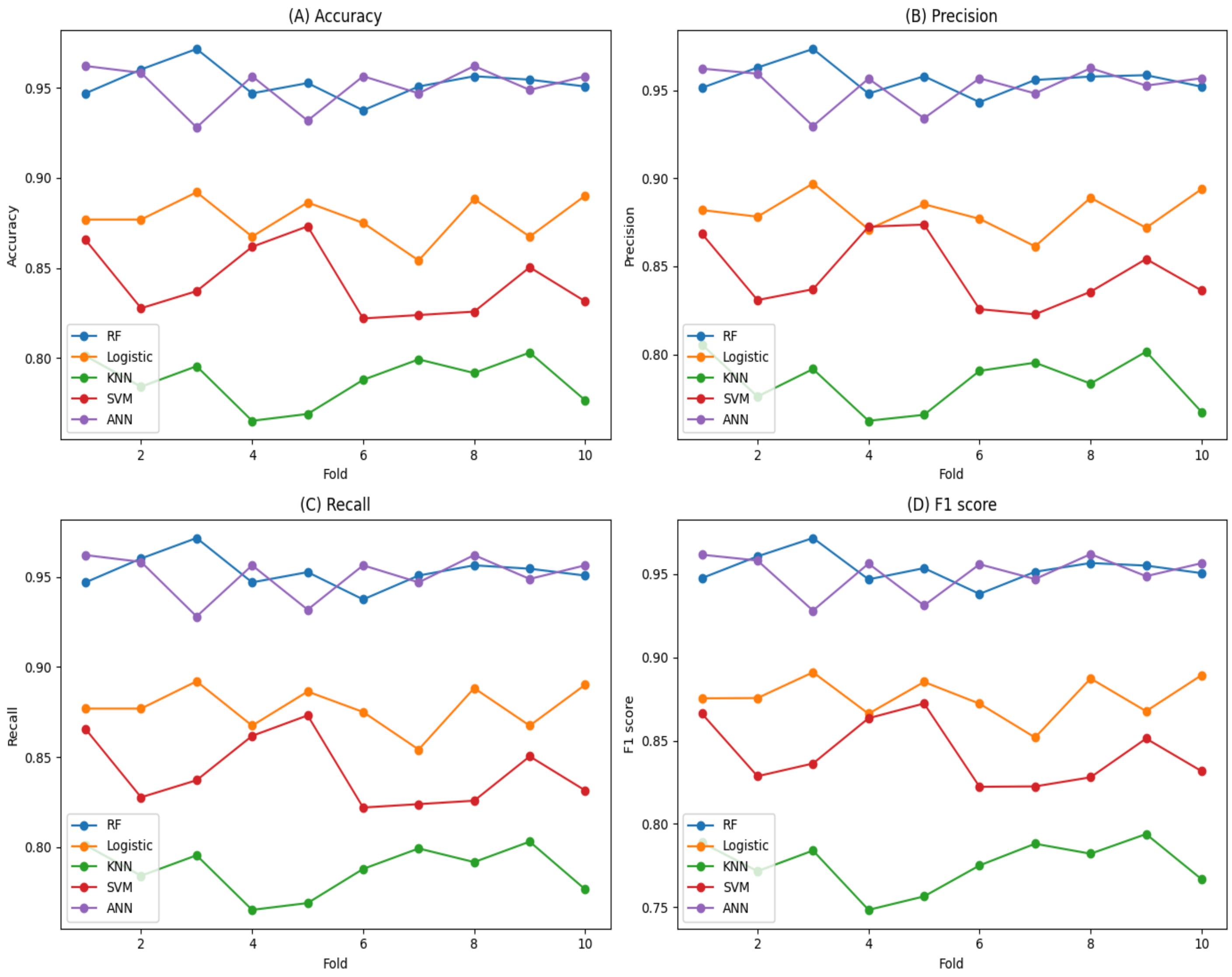
5. ANN


Confusion Matrix (Percentage)

	0	1	2	3	4	5	6
0	0.98	0.02	0.00	0.00	0.00	0.00	0.00
1	0.07	0.85	0.00	0.00	0.00	0.08	0.00
2	0.00	0.00	0.97	0.00	0.00	0.00	0.03
3	0.00	0.00	0.02	0.98	0.00	0.00	0.00
4	0.00	0.00	0.00	0.01	0.99	0.00	0.00
5	0.00	0.03	0.00	0.00	0.00	0.92	0.05
6	0.00	0.00	0.02	0.00	0.00	0.04	0.95

Accuracy, Precision, Recall and F1 Score Analysis

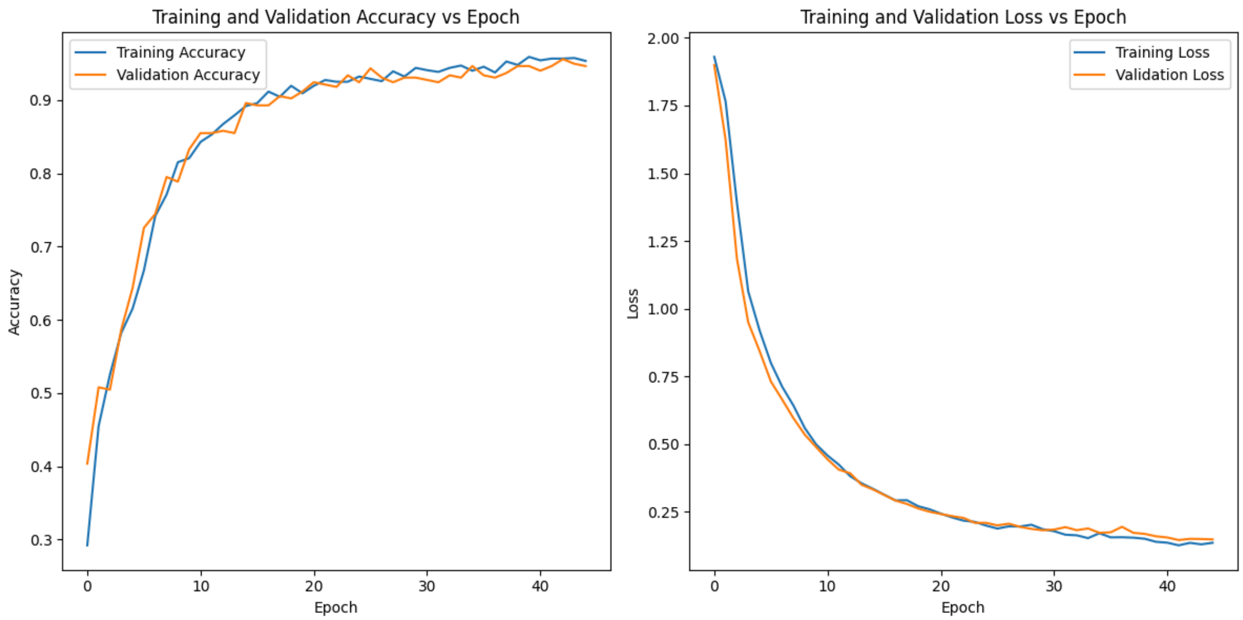
Below figure shows the performance behavior in each fold, according to the metrics of accuracy, precision, recovery, and F1-score. This shows that each model is generated with cross-validation with 10-fold and it has shown the difference behavior according to each random set of data for each fold. There are greater variations in their performance in each fold. RF and ANN appear to be more robust and consistent performance in each fold but with minor fluctuations.



Performance of models tested with cross-validation (10-folds)

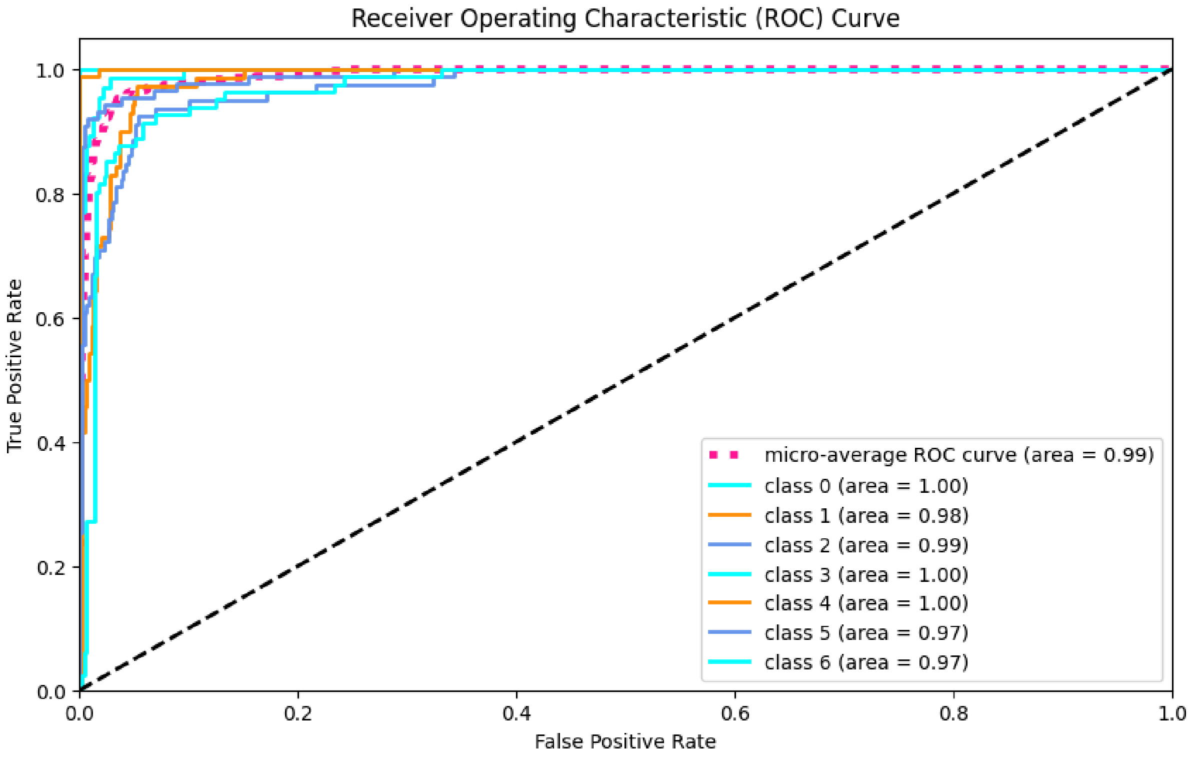
A Special Case with ANN:

Both the training and validation accuracy show an increasing trend over epochs, indicating that the model is learning and improving its performance over time.
Similarly, the training and validation loss plots show a decreasing trend over epochs, suggesting that the model is minimizing its loss function and becoming more optimized. This suggests that the model is improving its generalization ability and is learning meaningful patterns from the data.

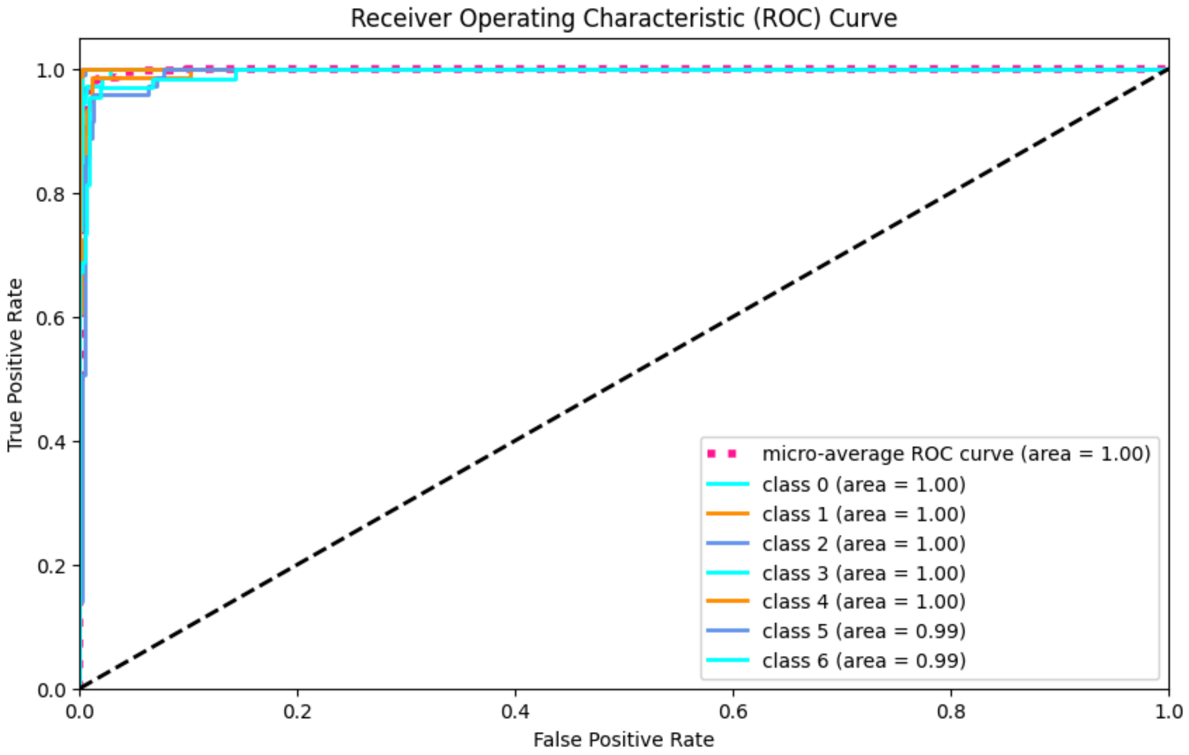


Some Other Evaluation Matrix

The ROC curves show how well the model is able to discriminate between different classes. A steeper ROC curve indicates better discrimination between positive and negative classes. In the first and second plot for logistic regression and ANN, all classes have high AUC values (0.97 - 1.00), this suggests that the modal has good discrimination ability for most of the classes.

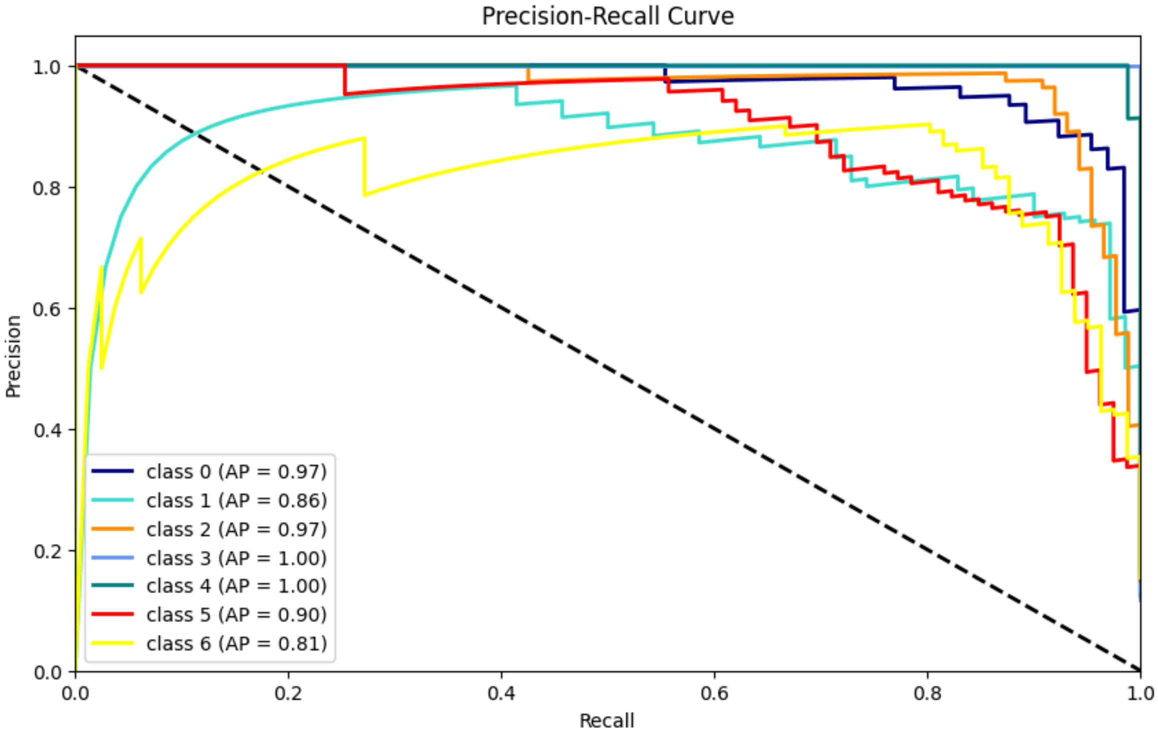


ROC for Logistic Regression curve

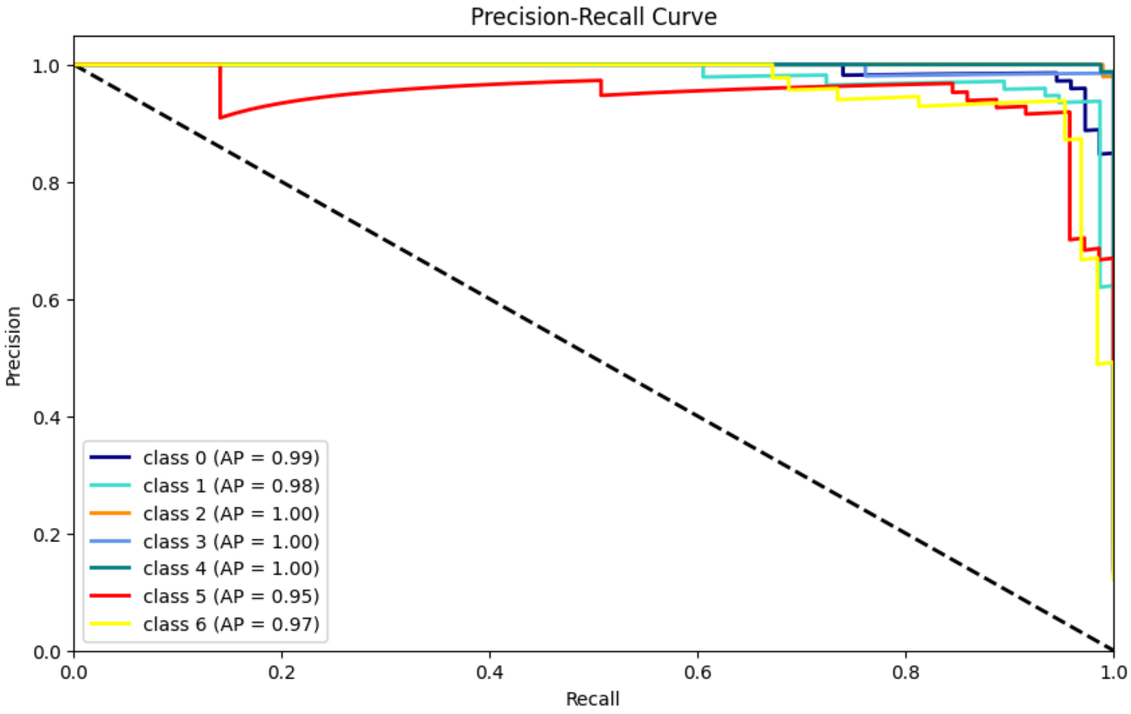


ROC for ANN

The area under each curve provides a single-value summary of the model's performance for each class in terms of precision and recall. An of 1.0 represents perfect precision and recall, while lower values indicate less effective performance. In case of ANN the area values range from approximately 0.95 to 1.00 across different classes which is an indication of excellent precision and recall performance. While in case ofLogistic regression, the area values range from approximately 0.81 to 0.97. class 1, have lower AP values compared to others.



Precision-Recall for Logistic Curve



Precision-Recall for ANN Curve

Conclusion

After studying and analyzing the five machine learning models we concluded that the results of the study show that RF and ANN have achieved the highest accuracy about 95.78 % in performance measures followed by logistic regression having around a performance measure of 88 %. The study concludes that machine learning is an effective tool in medicine that can be used to make timely treatment decisions for people at risk of obesity. Other algorithms also produced satisfactory accuracy results but could not compete with the RF and ANN model.