DESIGN ANALYSIS AND MODELING TECHNIQUES PROJECT REPORT

(2242-CSE-5301-001)

Project Group: 15

Project Name: Predicting Obesity Level Based on Different Factors

Professor: Prof. Ramakrishna Koganti

Team Members:

Syed Saboor Ahmed (1002153699)

Siva Ganesh Kolusu (1002161883)

Nahin Dey (1002204142)

Shruti Bhau Borge (1002122976)

Predicting Obesity Level Based on Different Factors

Introduction

Obesity is a leading cause for many different health concerns. Around 31.6 percent of the adult population in U. S. A. is found to be overweight and 7.7 percent are severely obese[1]. It is caused due to accumulation of body fat in large quantity. To handle this, it is important to understand different aspects of human life that may lead to obesity. These aspects include a person's eating habits, how much exercise they do, their family history with obesity, if they smoke or not etc. By understand these aspects we can understand which factors contribute more to obesity and potentially suggest some prevention strategies to tackle this problem.

To address this, we have employed different machine learning and Data Mining techniques on a data set taken from online source. To achieve our goal, we first performed data preprocessing and then employed different Classification models on the pre-processed data like KNN, Decision Trees etc.

These models were implemented to find which model is best suitable to classify this type of data. The target classification variable is name 'NObeyesdad'. Finally, based on the result we select the best model considering factors like accuracy and efficiency.

Dataset Data Source

This dataset is taken from an online source:

(https://archive.ics.uci.edu/dataset/544/estimation%2Bof%2Bobesity%2Blevels%2Bbased%2Bon%2Beating%2Bhabits%2Band%2Bphysical%2Bcondition)

This is a UC Irvine machine learning repository which includes obesity level of individuals from different countries like Mexico, Peru, and Colombia. It has different attributes like eating habits, smoking, height, weight, family history with obesity etc. In total of 17 columns including the target attribute (NObeyesdad) which is classified into 6 classes based on Mass Body Index with over 2111 record out of which 77% are synthetic and 23% are taken from users.

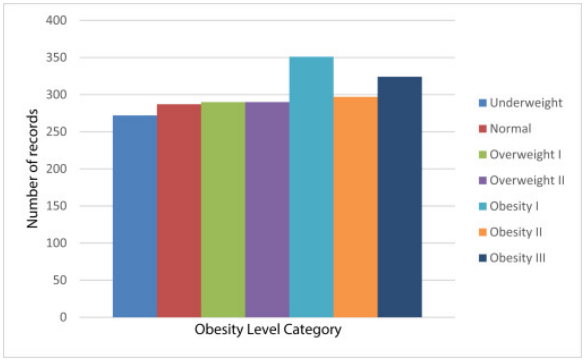

Fig. 1 Distribution of data regarding the obesity levels category

Problem Statement

Obesity, a significant global health issue, results from an excessive accumulation of body fat, leading to adverse health consequences. Therefore, understanding the interplay of genetic and behavioural factors is crucial for effective prevention and treatment strategies.

Project Goal

Implementing a prediction model, such as linear regression, to analyse the impact of various factors such as age, eating habits, family history, gender etc., on individual's obesity levels

Business Case:

To understand different factors that affect obesity level of a person and employ different preventive measure to make sure the person is healthy.

Benefits:

By implementing prediction model and understanding different factors that affect obesity we can try and reduce the obesity level by educating people on better life style.

Prediction Model:

1. Decision Tree: Performing Decision Tree was a good option to perform on this data as they tend to be more interpretable and best in handling both categorical and numerical features making them suitable for dataset with mixed data types. They can automatically handle feature selection and identify interactions between several types of features. As it was required to check the key features, it helped identify them. Also, Decision tree can be applied to both binary and multiclass classification it becomes a versatile tool in data mining.

2. Random Forest: Random Forest and Decision Tree algorithms are both popular choices for data mining tasks, but they have distinct characteristics and are suited for different scenarios. For example, Random Forest is particularly effective when dealing with datasets with many features or dimensions. It can handle high-dimensional data without overfitting, making it suitable for complex datasets.

   If the dataset contains complex relationships or interactions between features, Random Forest can capture these intricacies better than a single Decision Tree.

   Random Forest provides a feature importance measure, allowing you to assess the contribution of each feature to the model's predictive power. This can aid in feature selection and identifying the most relevant variables in the dataset.

3. Support Vector Machine (SVM): SVMs (Support Vector Machine) perform well in high-dimensional spaces, making them suitable for datasets with multiple features related to eating habits and physical conditions. SVMs can effectively handle non-linear relationships

through kernel functions. This is beneficial when dealing with intricate connections between variables that may not be linearly separable.

4. K-Nearest Neighbors (KNN): KNN is based on the idea that similar instances in the feature space tend to have similar outcomes. This is relevant for estimating obesity levels based on eating habits and physical conditions, where similar individuals may share common characteristics. KNN adapts well to local patterns in the data, making it suitable for capturing localized variations in obesity levels that may be influenced by specific eating habits or physical conditions.

Analysis:

On all the models we used 80-20 split for training and testing, also we used to stratify which is a data mining approach for making sure the variance in result is as low as possible. The target attribute is 'NObeyesdad' and all the other attribute as features.

Note: all the analysis is done online on Google Collab using python language.

Decision Tree:

Decision Tree Accuracy: 0.9432624113475178
Classification Report for Decision Tree:

	precision	recall	f1-score	support
Insufficient_Weight	0.92	0.96	0.94	56
Normal_Weight	0.87	0.89	0.88	62
Obesity_Type_I	0.97	0.94	0.95	78
Obesity_Type_II	0.95	0.95	0.95	58
Obesity_Type_III	1.00	1.00	1.00	63
Overweight_Level_I	0.93	0.91	0.92	56
Overweight_Level_II	0.96	0.96	0.96	50
accuracy			0.94	423
macro avg	0.94	0.94	0.94	423
weighted avg	0.94	0.94	0.94	423

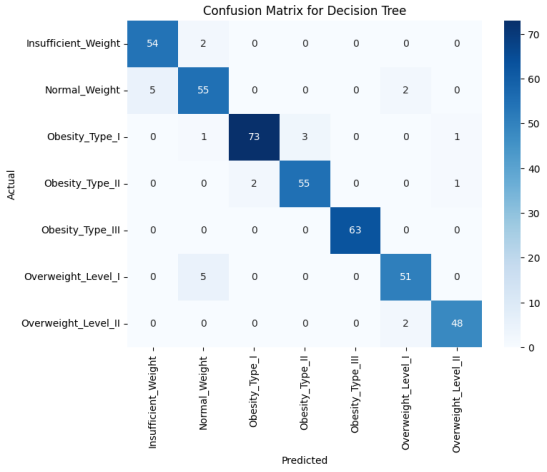

Fig. 2. Classification report and confusion matrix for Decision tree

We observe that the decision tree has good accuracy which might be due to its effective classification on discrete variable as most of our data set consists of discrete variable. We can see that the accuracy of the decision tree is around 94.3% which is quite high and can be considered as good model. The time taken by decision tree is approximately 0.02 seconds.

Random Forest:

Random Forest Accuracy: 0.9574468085106383
Classification Report for Random Forest:

	precision	recall	f1-score	support
0	1.00	0.94	0.97	54
1	0.84	0.98	0.90	58
2	0.97	0.97	0.97	70
3	0.98	0.98	0.98	60
4	1.00	0.98	0.99	65
5	0.96	0.88	0.92	58
6	0.96	0.95	0.96	58
accuracy			0.96	423
macro avg	0.96	0.96	0.96	423
weighted avg	0.96	0.96	0.96	423

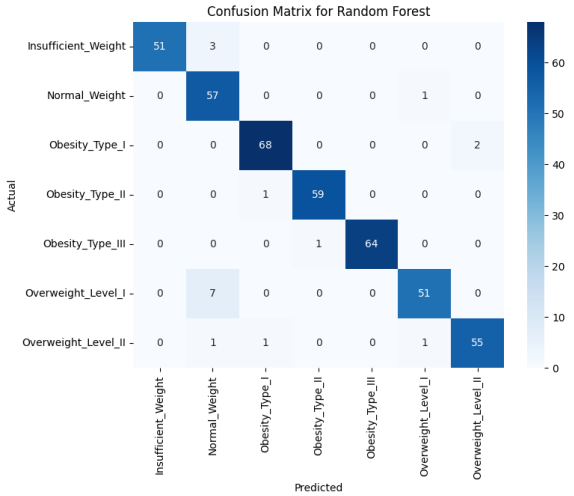

Fig. 3. Classification report and confusion matrix for Random Forest

The Random Forest accuracy is more than decision tree which is around 95.7% and the time taken to complete the task is 0.77 seconds which is quite high compared to decision tree. The confusion matrix of both decision tree and random forest is quite evenly spread out for miss classification and no inference can be found from this.

Support Vector Machines (SVM):

SVM Accuracy: 0.8628841607565012
Classification Report for SVM:

	precision	recall	f1-score	support
0	0.85	0.93	0.88	54
1	0.82	0.69	0.75	58
2	0.89	0.91	0.90	70
3	0.92	0.98	0.95	60
4	1.00	0.98	0.99	65
5	0.70	0.84	0.77	58
6	0.87	0.67	0.76	58
accuracy			0.86	423
macro avg	0.86	0.86	0.86	423
weighted avg	0.87	0.86	0.86	423

Number of Features: 16

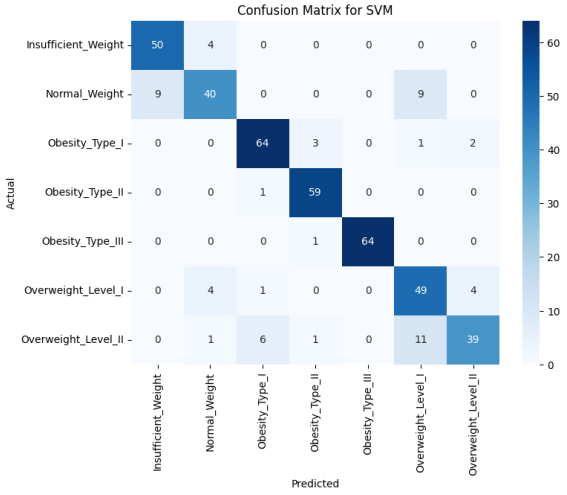

Fig. 4. Classification report and confusion matrix for SVM

SVM also give high accuracy on the test data with over 86.3% and completed the training of the model 0.33 which is also good but we can see decision tree is much more accurate. Also we can see in the confusion matrix that the mode is slightly biased between overweight level 1 and 2 and also between insufficient weight and normal weight which makes this model not so good for our prediction.

K-Nearest Neighbors (KNN):

KNN Accuracy: 0.8794326241134752
Classification Report for KNN:

	precision	recall	f1-score	support
Insufficient_Weight	0.81	0.96	0.88	56
Normal_Weight	0.82	0.45	0.58	62
Obesity_Type_I	0.89	0.97	0.93	78
Obesity_Type_II	0.98	0.97	0.97	58
Obesity_Type_III	0.97	1.00	0.98	63
Overweight_Level_I	0.77	0.89	0.83	56
Overweight_Level_II	0.90	0.90	0.90	50
accuracy			0.88	423
macro avg	0.88	0.88	0.87	423
weighted avg	0.88	0.88	0.87	423

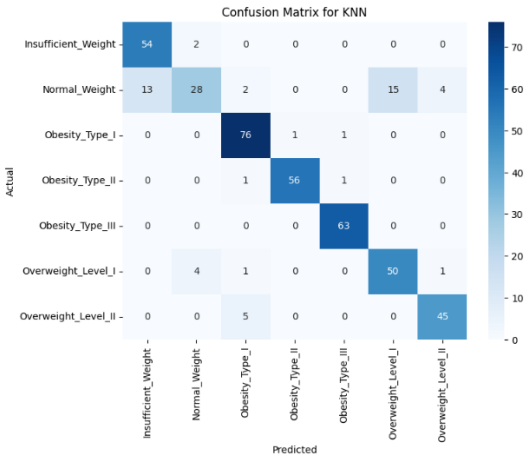

Fig. 5. Classification report and confusion matrix for KNN

We can see from the results that is almost same as SVM in terms of accuracy with around 87.9% accuracy but with higher training time around 0.87 seconds which is not desirable. Also we can see the same trend as in the case of SVM where there is some baseness between the variables.

Conclusion and Summary

Sr. No.	Model	Accuracy	Precision	F1-score	Recall	Training Time (sec)
1	Decision Tree	0.943	0.94	0.94	0.94	0.02
2	Random Forest	0.957	0.96	0.96	0.96	0.77
3	KNN	0.879	0.88	0.87	0.88	0.87
4	SVM	0.862	0.87	0.86	0.86	0.33

Based on the results we can choose model like Decision Tree or Random Forest but according to data mining techniques, Decision Tree should be selected even though Random Forest have better accuracy with almost a 1% higher than Decision Tree, the Decision Tree in general is a simpler model and simpler models are given more preference as there is less chance of overfitting. Also, the training time of the Random Forest is much high considering that we have a small data set of 2111 records. Therefore, we the proposed mode should be Decision Tree.

By using decision Tree model, we can say that a persons Obesity level can be predicted with 94% accuracy and Professionals can use this model to extract useful information from this data set and potentially figure out all the factors that contribute to obesity the most.

References

https://frac.org/obesity-health/obesity-u-s-2#:~:text=Adult%20Obesity%20in%20the%20U.S.,7.7%20percent%20are%20severely%20obese.) [1]

Fabio Mendoza Palechor, Alexis de la Hoz Manotas, Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico, Data in Brief, Volume 25, 2019, 104344, ISSN 2352-3409, *https://doi.org/10.1016/j.dib.2019.104344*.