DEVELOPING DEEP LEARNING-BASED PREDICTION MODEL FOR OBESITY IN CHILDHOOD

Project report submitted in partial fulfillment of the requirement for

the degree of Bachelor of Technology

in

Computer Science and Engineering/Information Technology

By

Nandini Singh (191413)

Under the supervision of

Dr. Pradeep Kumar Gupta

to

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology Waknaghat,
Solan-173234, Himachal Pradesh

CERTIFICATE

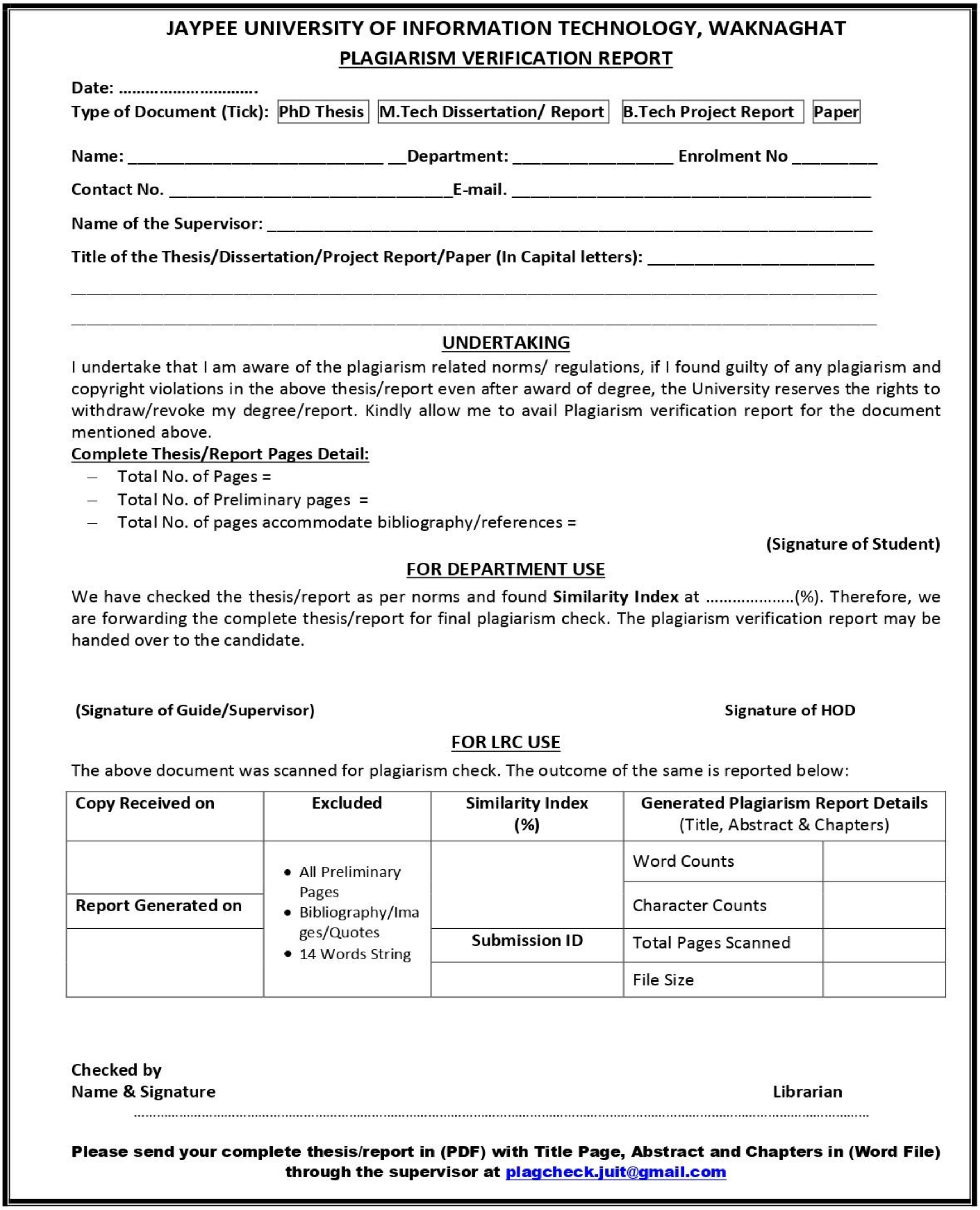
This is to certify that the work which is being presented in the project report titled "Developing deep learning based prediction model for obesity in childhood" in partial fulfillment of the requirements for the award of the degree of B. Tech in Computer Science And Engineering and submitted to the Department of Computer Science And Engineering, Jaypee University of Information Technology, Waknaghat is an authentic record of work carried out by "Nandini Singh, 191413" during the period from January 2023 to May 2023 under the supervision of Dr. Pradeep Kumar Gupta, Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat.

Nandini Singh (191413)

The above statement made is correct to the best of my knowledge.

Dr. Pradeep Kumar Gupta
Associate Professor
Computer Science & Engineering and Information Technology Jaypee University of Information Technology, Waknaghat, Solan, HP.

PLAGIARISM CERTIFICATE



JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: ……………………….

Type of Document (Tick): □ PhD Thesis □ M.Tech Dissertation/ Report □ B.Tech Project Report □ Paper

Name: ___________________________ __**Department:** _________________ **Enrolment No** _________

Contact No. ______________________________**E-mail.** _______________________________________

Name of the Supervisor: ____

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): ________________________

____

____

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

Complete Thesis/Report Pages Detail:

- Total No. of Pages =
- Total No. of Preliminary pages =
- Total No. of pages accommodate bibliography/references =

(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at ……………….(%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

(Signature of Guide/Supervisor) **Signature of HOD**

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Generated Plagiarism Report Details (Title, Abstract & Chapters)	
	• All Preliminary Pages • Bibliography/Images/Quotes • 14 Words String		Word Counts	
Report Generated on			Character Counts	
		Submission ID	Total Pages Scanned	
			File Size	

Checked by
Name & Signature **Librarian**

………

Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at plagcheck.juit@gmail.com

ACKNOWLEDGEMENT

Firstly, I express my heartiest thanks and gratefulness to Almighty God for his divine blessing that makes it possible to complete the project work successfully.

I am really grateful and wish my profound indebtedness to Supervisor **Dr. Pradeep Kumar Gupta**, **Associate Professor**, Department of CSE Jaypee University of Information Technology, Wakhnaghat. Deep Knowledge & keen interest of my supervisor in the field of **"Deep Learning"** to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Pradeep Kumar Gupta**, Department of CSE, for his kind help to finish my project.

I would also generously welcome each one of those individuals who have helped me straightforwardly or in a roundabout way in making this project a win. In this unique situation, I might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

Nandini Singh (191413)

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

1. ANN: artificial neural network
2. LSTM: long-short term memory
3. CNN: convolutional neural networks
4. RNN: recurrent neural networks
5. ROC Curve: receiver operating characteristic curve
6. AUC: Area Under the ROC Curve
7. Max: maximum
8. Min : minimum
9. EM : Elbow Method
10. MM : Matrix Multiplication
11. SW : Scalar Weighting
12. DR : Dynamic Routing
13. SF : Squashing Function
14. SVM : Support Vector Machine
15. KNN : K-Nearest Neighbour
16. DT : Decision Tree
17. MLP : Multilayer Perceptron

LIST OF FIGURES

LIST OF GRAPHS

ABSTRACT

The field of deep learning has gained momentum in recent years and is being used in many areas for building up a predictive model. Machine learning and deep learning have contributed a lot in the society for building high accuracy models and better predictive models.

Obesity is becoming a major problem in today's scenario taking into account the lifestyle the people have these days and their eating habits. Predicting obesity at an early age prevents children from becoming unhealthy adults with problems like High blood pressure (hypertension). High LDL cholesterol, low HDL cholesterol, or high levels of triglycerides (dyslipidemia), Type 2 diabetes and Coronary heart disease.

Building a prediction model for obesity in children will help them switch from obese person to a healthy one and also prevent them from chronic diseases. The objective of building this model is to warn the children and their parents who are heading towards obesity. This will help them have an ample amount of time to change their lifestyles and switch to a healthier one.

This project aims to build a deep learning based prediction model that will use using deep learning and machine learning methods to create predictive forecasts whether a child is obese or not. It will use machine learning techniques like logistic regression and deep learning techniques like LSTM to predict the same.

The purpose of building this model is to create an awareness amongst the new generation children on the importance of having a healthy lifestyle and exercising regularly and also for their parents to keep a check on their child’s growth and teach them healthy habits.

By working on this project, we got a fair understanding of CNN, LSTM and logistic regression to predict obesity. This model will further be developed for more varied data from different countries and ethnicities.

CHAPTER 1: INTRODUCTION

1.1 Introduction

Childhood obesity is a significant contributor to the non-transferable diseases that are the major public health concern of this century. In Mexico, the rates have increased thrice in the last 30 years. In Mexico today, 17% of teenagers and 16% of kids between the ages of 2 and 11 are fat. Although it was previously believed that childhood obesity was just an issue in rich nations, reports of it are now increasingly coming from middle- and low-income nations, particularly from metropolitan areas. Compared to adults, children's obesity is more difficult to measure. Body mass index is not a reliable analyser of childhood obesity[1]. Recognizing and treating childhood obesity is crucial because, if left untreated, it can lead to adult obesity and all of the associated metabolic problems. Additionally having a significant psychosocial impact, childhood obesity is consistently linked to lower academic success. Prevention is key because it's tough to shed weight after it's been gained. Due to their higher susceptibility to the constant barrage of marketing for energy-dense food, children are the target of this more so than adults.It is crucial and challenging to remove children from a setting that encourages obesity. Innovative programmes to fight childhood obesity have been investigated in a number of different countries, and each of them has something to teach public health professionals and those in charge of developing health policies. The Commission for Ending Childhood Obesity of the WHO states that this problem requires a "full -government involvement in which regulations across all areas frequently take health into account, avoid detrimental health results, and hence promote population health and health equity." We'll talk about the scope of the issue, contributing reasons, consequences of childhood obesity, and potential remedies.

This project aims to predict obesity in children using deep learning algorithms from an available dataset of Mexico. We will predict this using their body mass index, eating and drinking habits and from their backgrounds to improve this problem in the long run. The goal of this model is to survey the recent body of growing literature on Deep Learning (DL) models such as LSTM[2], Capsule Network and ML algorithms such as binary classification and logistic regression for childhood obesity prediction[8]. To do this, a taxonomy of the literature on deep learning models for predicting childhood obesity was created. Considering the outcomes of this survey, future research can create more accurate prediction models that take datasets from different fields into account. In order to create successful obesity intervention programmes, this study will provide analysis on the association and prediction of childhood obesity.

Capsule Networks- are the models that are able to retrieve data according to the space and more significant aspects in order to get over the loss of data that is seen in pooling operations- are the types of RNNs that are capable of learning order dependence in a sequence prediction problem[5]. An output vector with a direction is provided by the capsule function. Binary classification refers to tasks that require the classification of two classes[22]. The normal state is typically represented by one class in binary classification tasks, while the abnormal state is typically represented by a different class. A statistical analysis technique known as logistic regression uses former dataset results to predict a binary outcome, such as yes or no.

1.2 Problem Statement

In recent years, the problem of obesity in children has amassed a rising interest in the research field of machine learning and data science. An excessive or not natural amount of body mass is called being overweight or obese. People today are increasingly leading very unhealthy lifestyles and not exercising regularly, indulging in excessive junk food consumption, sleeping late, eating late at night and spending a lot of time sitting down[17]. It encourages the spread of complicated illnesses like liver cancer, heart disease, and stroke. In order to help hospitals and parents of the children provide them with appropriate eating habits, the project's goal is to forecast childhood obesity[1]. This product is class-oriented since it uses several bodily measurements of a child, including age, weight, height, sleep duration, outdoor activities, family diseases, past diseases, etc. The classes of this system that best describe a child's physical strength are weight, height, and outdoor activities.

1.3 Objectives

Investigating risk profiles is the aim of this study for childhood overweight and obesity in order to educate parents and schools about the problems associated with obesity and how it may stunt their children's development. The project's goal is to use DL and ML to analyze EHR data[6] to predict the risk of obesity in children between the ages of 5 and 10. In order to stop the progression of chronic diseases like diabetes and hypertension, this will aid in the early detection and prevention of obesity.

1.4 Methodology

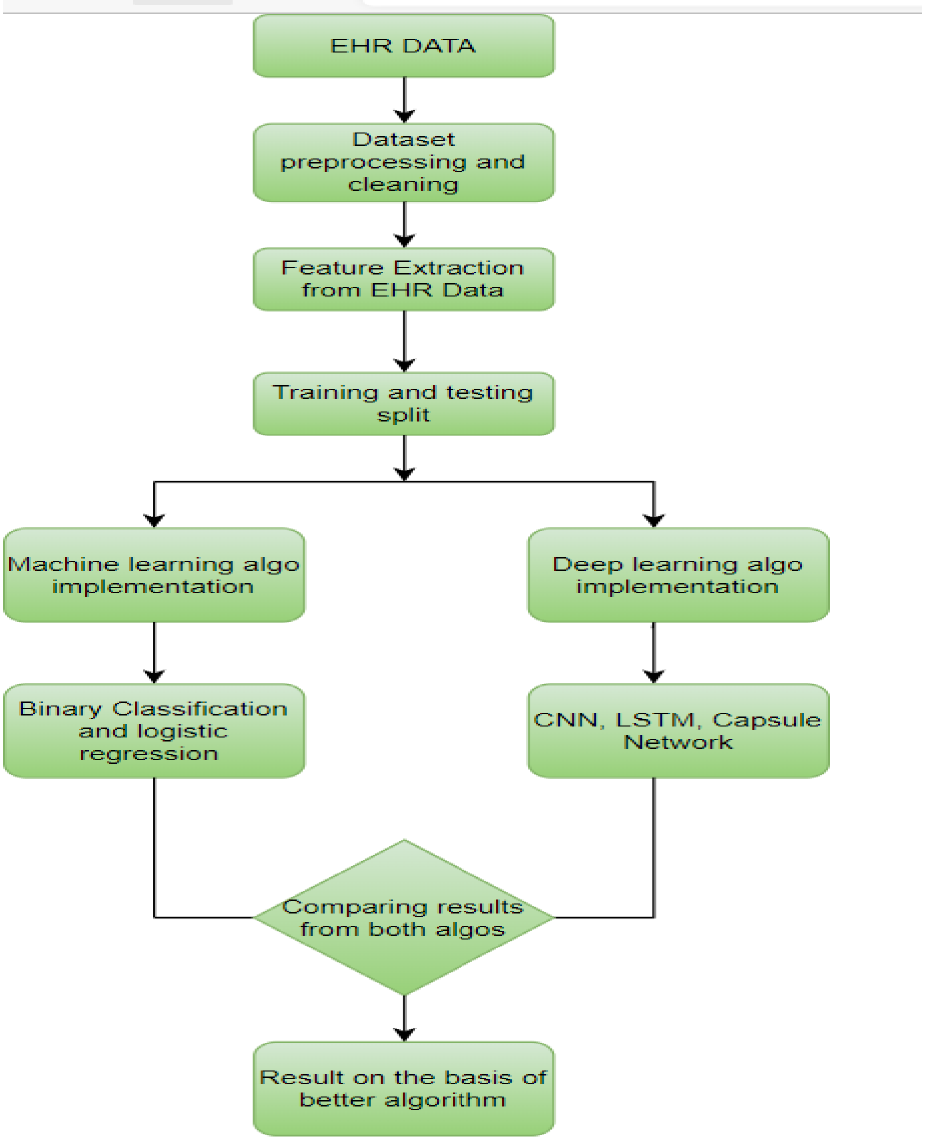

EHR DATA
Dataset preprocessing and cleaning
Feature Extraction from EHR Data
Training and testing split
Machine learning algo implementation
Deep learning algo implementation
Binary Classification and logistic regression
CNN, LSTM, Capsule Network
Comparing results from both algos
Result on the basis of better algorithm

Fig 1. Project Design/ Plan

The above is the project design/ plan for our work.

As shown in Fig.1, we have taken Electronic Health Record (EHR) data for our work which includes the height, weight, eating habits and physical activities of the people[9]. We visualized the data and established a relationship between being overweight and all the columns present in the dataset and after that we preprocessed data including label encoding , changing data types and scaling of the features. Pre-processing step is taken forward to a training and testing split where we have split our data and then we have built our model using a decision tree classifier[14]. We used BMI as the key feature in determining whether a person is obese or not. After testing the model on testing data we used various evaluation parameters to find our accuracy and the best came out to be around 95%.

K-Means Clustering

Using the Means Clustering algorithm[8], the unlabeled dataset is partitioned into diverse clusters. Here, the K value establishes the number of predetermined clusters that need to be generated; for instance, if K=3, then three clusters will be made after applying the clustering algorithm, if K=4, then four clusters will be created, and so on. The iterative method employed in the K-Means Clustering algorithm ensures that every data point belongs to a unique group that shares similar features[9].

Figure 2.a presents an effective approach for rapidly and automatically recognizing the clusters in the unlabeled dataset without the requirement of training the dataset as it is an unsupervised form of machine learning. Since the algorithm is centroid-based(the central point of the cluster is called the centroid), each cluster is assigned a centroid[9]. The primary objective of Figure 2.b is to reduce the total distances between each data point and the cluster to which it belongs.

To begin, an unlabeled dataset serves as the input for the algorithm, which subsequently divides it into k clusters. The algorithm then keeps on doing this process until no further clusters are available and thus no more clusters can be formed. It is important to establish the value of k beforehand in this method.

k-means clustering technique primarily achieves two tasks:

- uses an iterative method to choose the ideal value for K centroids or center points.
- A match is made between each dataset point and the closest k-center[3]. The dataset points that are in close proximity to a certain k-center group together to form a cluster.

Because of this, each cluster differs from the others and has some shared data points.

The figure given below explains the working of the K-means Clustering Algorithm:

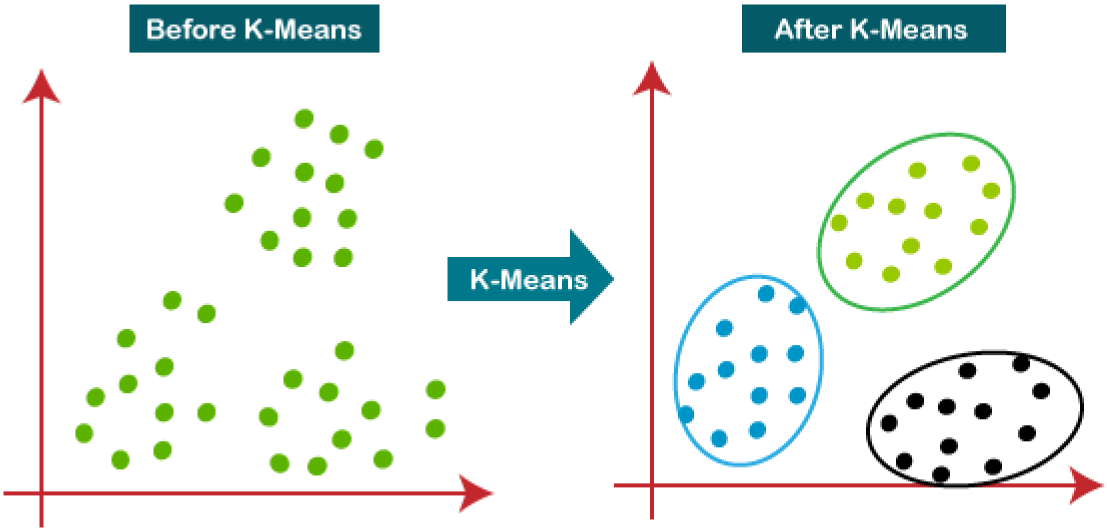


Fig 2.a. before K-Means

Fig 2.b. Clusters after K-Means

The EM [7] is majorly used methods for figuring out the optimal number of clusters. This technique makes use of the WCSS value concept. Total variations inside a cluster are denoted by the acronym "WCSS," which stands for Within Cluster Sum of Squares.

We can use any advent, including the Euclidean distance or Manhattan distance[4], to calculate the distance between dataset points and the central dataset point.

The following steps are followed by the EM to get the best cluster value:

- 1) It performs K-means clustering for different K values on a given dataset.
- 2) Determine the WCSS score for each K value.
- 3) depicts a curve between the K-cluster count with the estimated WCSS values.
- 4) The greatest K value is assigned to a bend's steep turn or a plot point that looks similar to an arm.

Since Fig.3 the elbow method because it shows a sharp curve that looks like an elbow. The graph using the elbow approach looks like the figure below:

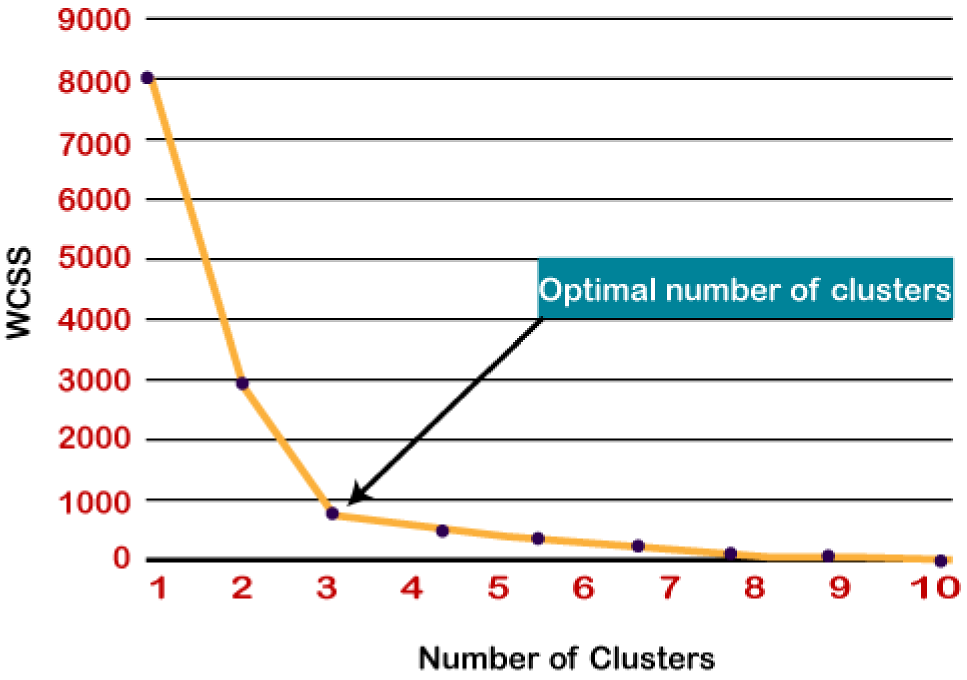


Fig 3. Optimal number of clusters

Capsule Network

In order to overcome the information loss experienced during pooling operations, Capsule Networks (CapsNet) are networks that can retrieve spatial information as well as more significant features[19]. As an output, Capsule provides us with a vector with a direction. While a neuron's output is a scalar quantity that lacks direction information, if the orientation of the image is altered, the vector will move in the same direction.

To achieve equivariance amongst capsules is the key goal. Accordingly, like shown in Fig.4 The position of a feature within an image will also alter how it is represented vectorically in the capsules, but not whether or not it is likely to exist[11]. Lower level capsules identify features, and then send this information to higher level capsules that fit well with it.

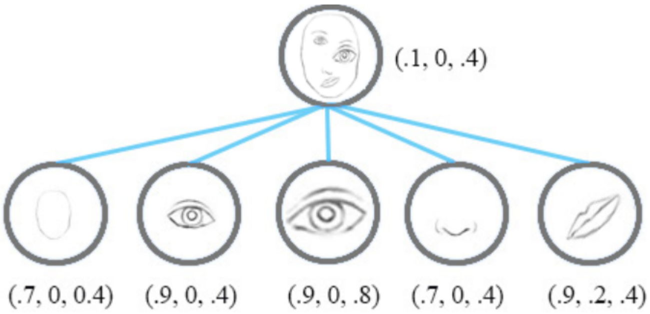


Fig 4. Capsule network Example

The following list of the CapsNet's four primary components is in no particular order:

1. Matrix Multiplication
2. Scalar Weighting of the input
3. Dynamic Routing Algorithm
4. Squashing Function

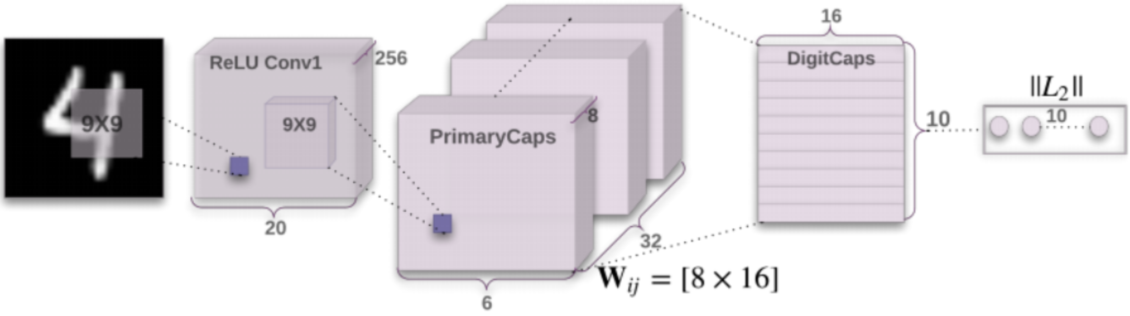


Fig 5. Components of Capsule network

As shown in Fig.5 Matrix multiplication of input vectors with weight matrices. Between the high-level characteristics and low-level features of the image, this encodes highly substantial spatial relationships. The input vector weighting determines what higher level capsule the current capsule will transmit its output to[4]. Through the use of dynamic routing, this is achieved. Using the "squash" function, nonlinearity "squashes" a vector to give it a maximum length if 1, a minimum length if 0, and to keep its direction.

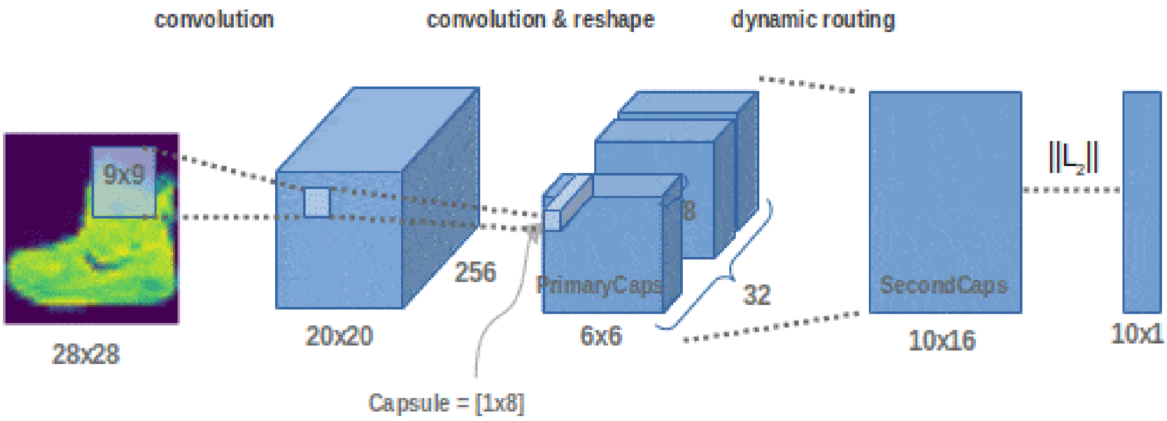


Fig 6. Capsule Network Encoder

Like in Fig.6 When an image is entered, the encoder learns how to encode it as a 17-dimensional vector that contains all of the necessary data to create the image.

Convolution Layer — Finds features that the capsules will later evaluate. 257 kernels of size 10x10c2, as suggested in the study.

a. Primary(Lower) CL— The lowest level capsule layer I previously mentioned is this layer. Each of the 32 different capsules produces a 4D vector output by applying an ninth 10x10x453 convolutional kernel to the result of the preceding convolutional layer.
b. Digit(Higher) CLr — The first Capsules would go to this layer, which is the higher level CL (using DR Algorithm)[13]. All of the instantiation parameters needed to recreate the object are contained in the 16D vectors produced by this layer.The decoder finds the 23D vector from the Digit

c. understands how to turn the object's initialization given features into a pictorial representation after finding it. To assess how accurate the reconstructed characteristics are to the new concepts that it was trained on, the decoder is linked with a loss function using Euclidean distance.

By doing this, it is ensured that the Capsules only store data that will aid in the recognition of digits within its vectors[9]. As shown in Fig.7 the decoder is a very straightforward feed-forward neural network, as explained below.

1. Layer 1 is the layer in which all neurons are connected to each other and thus it is Fully Connected (Dense)
2. Layer 2 with Full Connectivity (Dense)
3. Final Result with 10 classes from Fully Connected (Dense) Layer 3.

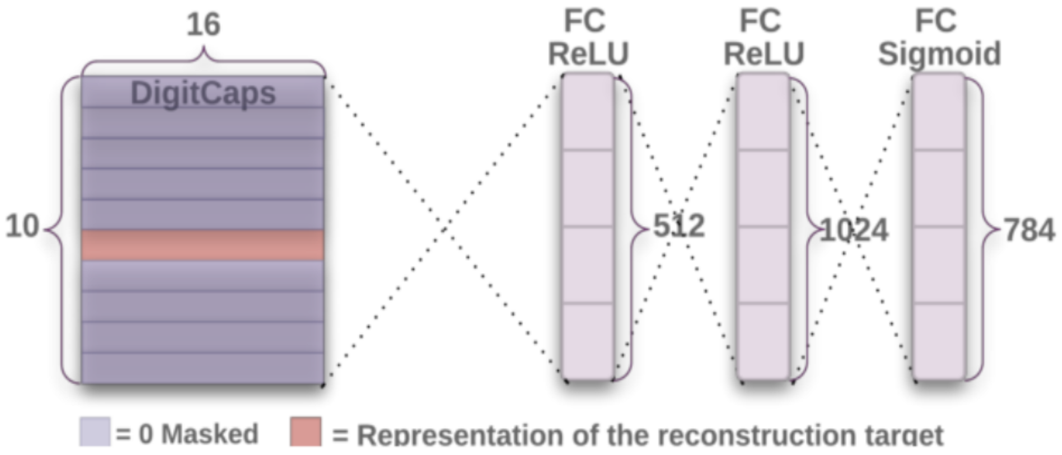


Fig 7. Feed Forward Neural Network

LSTM

The LSTM is a type of RNN architecture that piqued the interest of the natural language processing community due to its remarkable aptitude in handling sequence data. LSTMs, unlike traditional RNNs, are endowed with the ability to selectively recall or forget specific pieces of information based on their relevance to the task at hand, thanks to the network's sophisticated gating mechanisms that govern the flow of information. As a result of these techniques, LSTMs are capable of circumventing the vanishing gradient predicament that is often encountered with regular RNNs, thus making them an exemplary choice for undertaking complex jobs such as sentiment analysis, machine translation, and speech recognition. The use of LSTMs has been shown to improve the performance of various NLP tasks, and research in this area is ongoing to further improve the capabilities of these powerful networks.

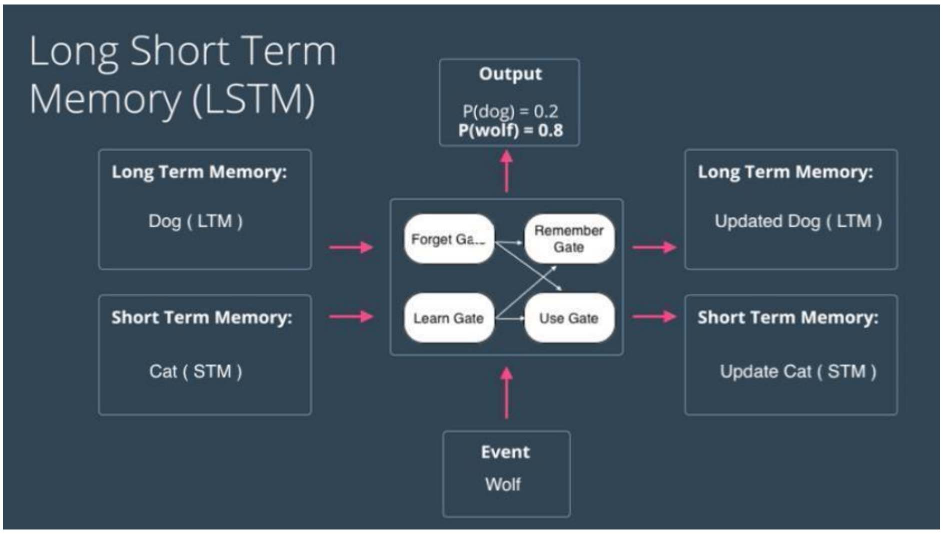


Fig 8. LSTM

Additionally, LSTMs can be used for tasks such as text classification, sentiment analysis, and machine translation by incorporating an embedding layer to transform input text data into continuous vectors. LSTMs can also be stacked to form deep networks, allowing for more complex representations of input sequences. Training Given that they demand a significant quantity of data and computer power, LSTMs can be a challenging undertaking. The model's performance can also be significantly impacted by hyperparameter adjustment, including the amount of hidden layers, the number of units per layer, and the learning rate. However, after being trained, LSTMs can perform at the cutting edge on a variety of tasks.

As shown in Fig.8, LSTMs comprise both LTM and STM and make use of the concept of gates for making the calculations simple and effective. [7]

1. Forget Gate: After entering the Forget Gate, LTM forgets useless information.
2. Learn Gate: In order to apply recent knowledge from STM to an event (or current input), STM and the event are integrated.
3. Remember Gate: Remember gate, which serves as a modernized LTM, combines LTM data that hasn't been forgotten with STM and Event data.
4. Use Gate: This gate makes use of LTM, STM, and Event to forecast the event's outcome and functions as an updated STM.

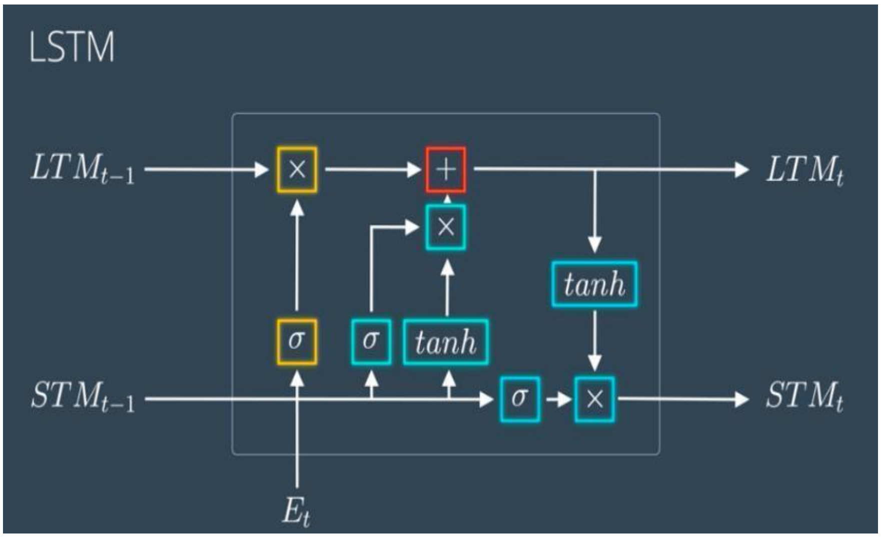


Fig 9. LSTM architecture

In Fig.9, we can see the architecture on how LSTM works.
Usage of LSTMs:

Although it resolves (or eliminates) the Vanishing Gradient problem (too-small weights that under-fit the model), it still has to deal with the Exploding Gradient problem (weights become too large that over-fits the model).

LSTMs are frequently employed in tasks like language generation, voice recognition, image OCR models, object detection, etc. because they take care of the long-term dependencies.

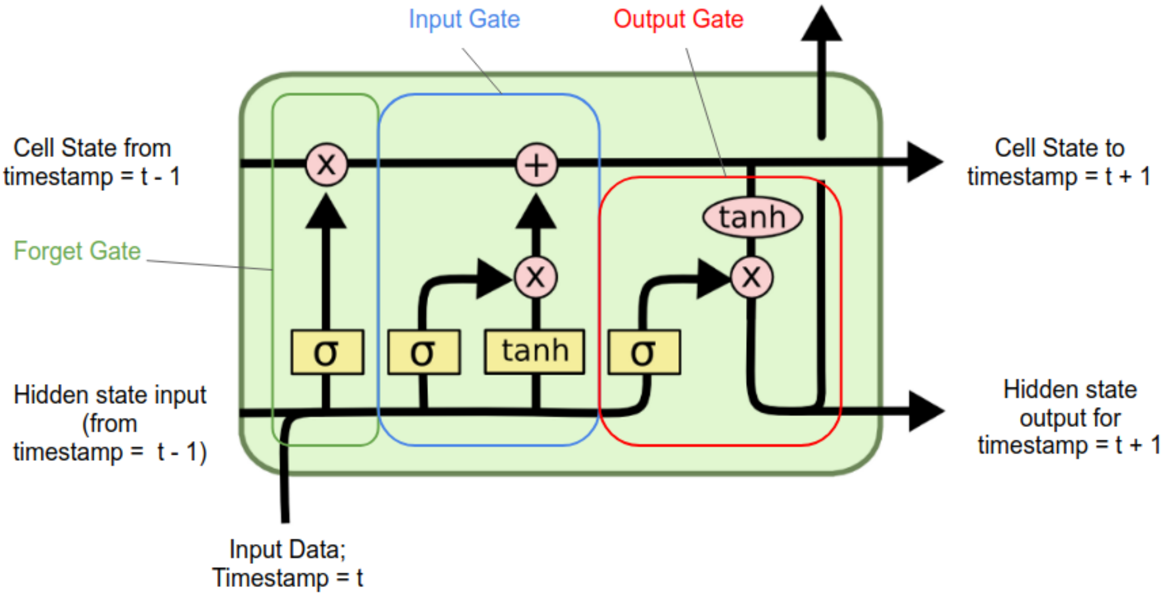

Input Gate
Output Gate
Cell State from timestamp = t - 1
Forget Gate
Hidden state input (from timestamp = t - 1)
Input Data; Timestamp = t
Cell State to timestamp = t + 1
Hidden state output for timestamp = t + 1
X
+
X
X
tanh
σ
σ
tanh
σ

Fig 10. Complete Working of LSTM

With a singular focus on every input sequence, LSTM models labor through a sequential process of analysis. Per iteration, the model performs intricate calculations that involve the current input, also the previous hidden layer and cell layer. These calculations are then used to determine the now hidden layer and cell layer, which are subsequently handed over to the following time step as the next hidden layer and cell layer, respectively. This process is repeatedly executed until the conclusion of the sequence is reached. Ultimately, the LSTM model's output is the long-anticipated output sequence that encapsulates its analytical efforts.

Binary Classification

TP is a term used to describe a situation when the model predicts that the patients will test positively[10]. When the model predicts the patient's outcomes to be negative, it is said to be TN. The binary classifier may provide false diagnosis for some cases. FN are mistakes that happen when a sick person is mistakenly classified as healthy because of a test outcome which turned out to be negative[18]. Similar to this, a FP occurs when a healthy patient receives an inaccurate disease diagnosis as a result of a positive test result.

A binary classifier can be assessed using the following criteria:

1. TP: The patient has an illness, as predicted by the model.
2. FP: The individual is in good health, despite the model's "diseased" forecast.
3. TN: The patient is in good health, as predicted by the model.
4. FN: The model predicts "fit" even when the patient is ill.

Following the acquisition of these values, the accuracy score of the binary classifier can be calculated as follows:

The following Fig.10 is a confusion matrix representing the aforementioned criteria

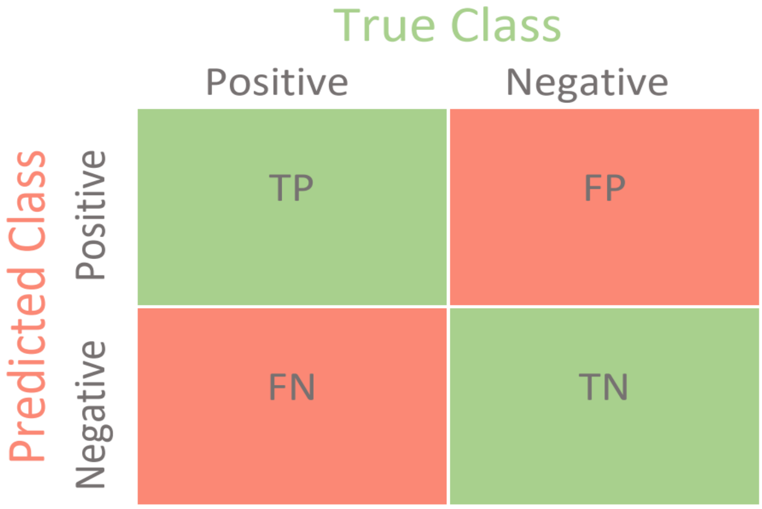


Fig 10. Confusion matrix

Many techniques in machine learning use binary classification. The majority are:

- SVM
- Naive Bayes Algorithm
- KNN Algorithm
- Decision Tree Classification
- Logistic Regression

Logistic Regression

Logistic regression, a statistical technique designed for modeling and analyzing data with binary or categorical outcomes, has gained widespread adoption in diverse fields. By utilizing the logistic function, this approach involves estimating the likelihood of an event's occurrence based on the input variable values. Fundamentally, the logistic regression model alters the linear regression equation, ultimately producing a binary classification outcome. The technique has proven valuable in various fields, including but not limited to risk management, medical diagnosis, and credit scoring. Using maximum likelihood estimation, the method computes the input variable coefficients, while the classification threshold may be adjusted to optimize the balance between sensitivity and specificity, further highlighting its versatility. The model is vulnerable to outliers and multicollinearity among the input variables, and it may be overfitted or underfitted. To prevent these problems, it is crucial to thoroughly preprocess the data and choose the right features.

Early in the 20th century, logistic regression became popular in the biological sciences. Later, it served a number of functions in social science. When the target (dependent variable) is categorical, logistic regression is used.

For instance,

To determine whether a mail received is spam or not

Whether the tumor is cancerous or not

Consider a situation where we must determine whether or not an email is spam[3]. To achieve classification if we utilize linear regression to solve this issue, a threshold must be specified. Suppose the data point is labeled as non-malignant even though the actual class is malignant and the projected continuous value is 0.4 and the threshold value is 0.5. This might have a big impact right away.

According to Fig. 11, categorization problems do not adapt easily to linear regression[21]. The unlimited character of linear regression leads to the introduction of logistic regression. They only range in value from 0 to 1.

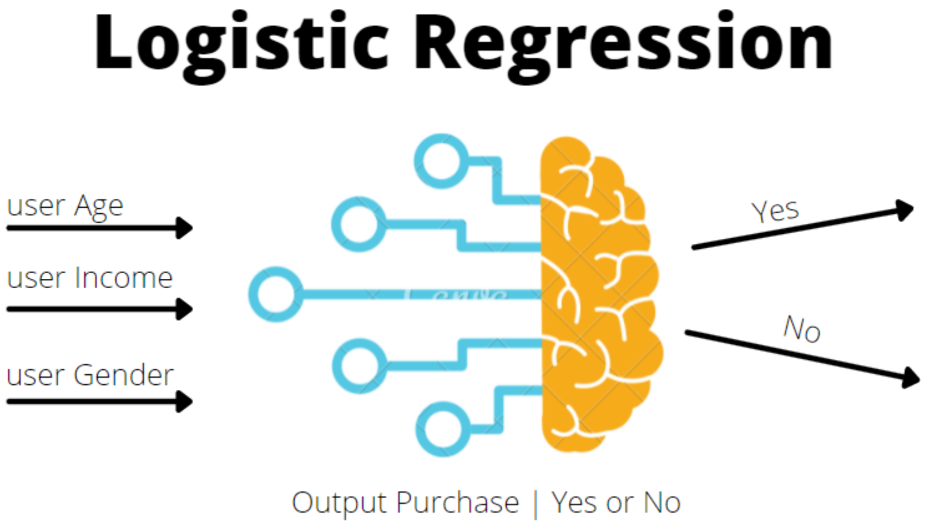


Fig 11. Logistic Regression Model

Model:

Result: 0 or 1

Z is assumed to equal WX plus B.

sigmoid (Z) = h(x)

Sigmoid Function

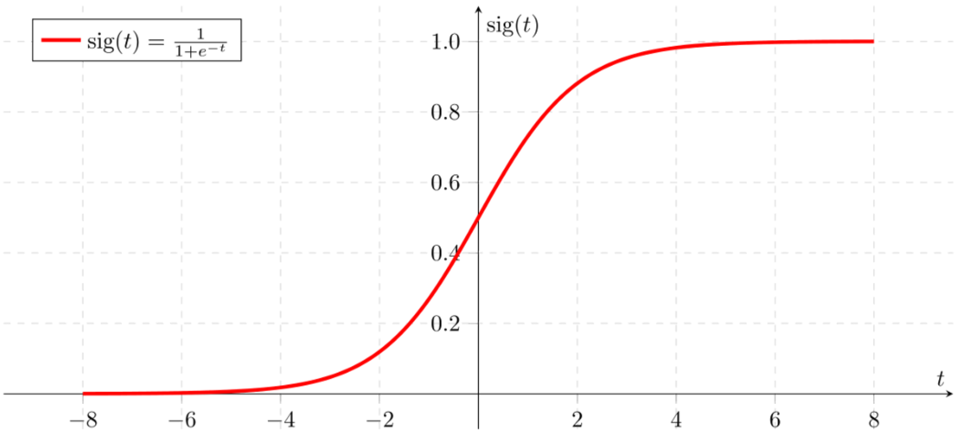


Fig 12. Sigmoid Activation Function

As in Fig.12, Y(predicted) would become 1 if Z reaches infinity, and Y(predicted) would become 0 if Z reaches negative infinity.

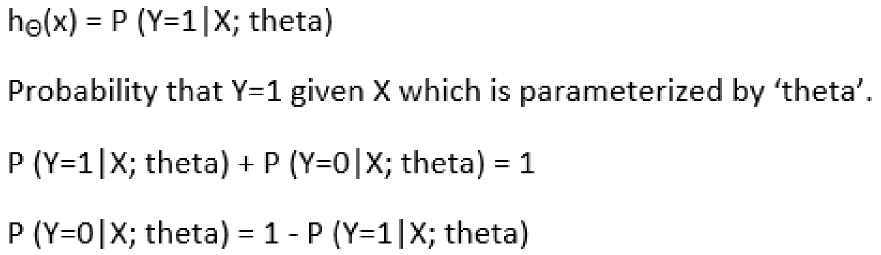
The hypothesis is analyzed

The calculated probability is the hypothesis's output[11]. This is used to determine the degree of confidence with which actual value will match anticipated value given input X.

X equals [x0, x1] and [1 IP-Address].

Let's say we calculated the predicted chance to be 0.8 based on the x1 value. This indicates that an email has an 80% chance of being spam.

The mathematical form of this is,



$h_{\Theta}(x) = P\ (Y=1|X;\ theta)$

Probability that Y=1 given X which is parameterized by ‘theta’.

$P\ (Y=1|X;\ theta) + P\ (Y=0|X;\ theta) = 1$

$P\ (Y=0|X;\ theta) = 1 - P\ (Y=1|X;\ theta)$

This explains why it is called "logistic regression." Using information that has been fitted into a linear regression model, a logistic function predicts the dependent variable for the target category.

1.5 Organization

The remainder of the paper is structured as follows: In chapter 2 we have presented the literature survey which depicts the various approaches used by authors to create an Image Captioning model[7]. Chapter 3 highlights the methodology and system development of the project. It represents various computational, experimental and mathematical concepts of the project. Also, we have focused on the software and hardware platforms needed for implementing the model. In chapter 4 we have presented the performance analysis of the project which specifies the accuracy of the project. Also, we have shown the required dataset and its related information. Chapter 5 presents the conclusions of the project and the observations seen in the results. It also provides the applications of the project and the future scope of the same.

CHAPTER 2: LITERATURE SURVEY

The exigency of effective measures to counteract the onset of this condition during the nascent stages of development is an indisputable fact, as postulated by the scholarly article [1] composed by a cadre of distinguished researchers. This corpus of work delves into the highly intricate linkage between childhood obesity and its morbidities in adulthood, and the arduous task of mitigating body mass index (BMI) later in life. The prospect of prophesying obesity prior to the age of five could represent a salient tool for pinpointing at-risk children and implementing preventative measures. However, the limited availability of data has thwarted most of the present childhood obesity prediction models, which rely heavily on information gleaned from longitudinal cohort studies. Conversely, our study employs raw and not preprocessed EHR data from the first two years of life, an innovative and unprecedented methodology in pediatric obesity research. Through a diverse array of ML algorithms that can be used for classification and can come under both supervised and unsupervised category, binary classification and regression were executed in order to discern which factors may contribute to obesity in boys and girls, thus leading to the creation of separate models for both sexes. Our pioneering study found that EHR data can be utilized to accurately forecast childhood obesity, with an AUC on par with that of cohort-based research, thus avoiding the expenses and complexities associated with additional data collection. Furthermore, the results suggest that machine learning techniques hold tremendous promise for predicting future cases of juvenile obesity via EHR data, which could catalyze significant changes in legislation, intervention planning, and clinical decision-making. Our proposed hybrid approach combining categorization and regression represents a novel and potentially transformative tool for prognosticating childhood obesity. However, it is important to note that our study is not without limitations, such as the lack of demographic diversity in the cohort, the relatively small sample size, and the inherently noisy and unreliable nature of EHR databases.

The investigation conducted by distinguished researchers under the vast topic of obesity prediction examined the risk profiles that are related with overweight and obesity among kindergarten and school children in China. To achieve this, they harnessed the power of ML and DL techniques and algorithms for model building. The survey, which took place between September and December 2020, utilized a stratified cluster random sampling strategy, and 30 kindergartens' students aged 3-6 were enrolled in Beijing and Tangshan.

The analysis was conducted using the PyCharm software, a free and open-source integrated development environment (IDE) tool. The results indicated that 1250 of the 9478 children who met the inclusion criteria were overweight or obese and were randomly split into a training group and a testing group in a 6:4 ratio. After intensive training and testing, the SVM and GBM methods emerged as the top two models, with a consistency of 0.9457 and 0.9454, respectively. Furthermore, the GBM outperformed the SVM with the highest F1 score (0.7748) while the SVM came in second with an F1 score of 0.7731.

The study revealed the top five factors related to both children and parents that were instrumental in differentiating overweight or obese children from all other children. These include child age, eating rate, the number of relatives who are obese, sweet-drinking, and paternal education, which were ranked in order of relative relevance. Remarkably, these factors were sufficient to produce satisfactory results under the GBM algorithm, as corroborated by other performance metrics.

This investigation's novelty stems from comparing deep learning methods such as LGBM and XGBoost with machine learning methods such as GBM and SVM, thus bringing the best findings from both to the forefront. However, the cross-sectional design's disadvantage is complex in nature, as only 31 components were considered in this sample, making it difficult to address the potential causal effect of childhood overweight or obesity.

In their groundbreaking research, Eom, Gayeong & Byeon, Haewon [3] employed an extensive web crawl to obtain a vast corpus of South Korea's news big data, which involved an exhaustive keyword search for "obesity." They proposed a model that could demonstrate how the issues and keywords related to obesity in South Korean society were transformed before and after the COVID-19 pandemic. Moreover, they engineered sophisticated predictive models for keyword trends based on state-of-the-art RNN and LSTM algorithms. The researchers collected an enormous dataset of 12,418 text data, of which 3,588 duplicates or irrelevant records were excised from the study. The investigators performed meticulous data preprocessing and morpheme extraction using BIGKinds, a cutting-edge data refinement source.

The study conducted latent Dirichlet allocation (LDA) topic modeling to identify obesity-related keywords and differences in participants during the COVID-19 pandemic. Text clustering was employed to sort out the themes and to gain an insight into how the clusters were dispersed based on similarity. Deep learning techniques, particularly LSTM and RNN, were leveraged to build and compare topic prediction models. The study also anticipated the timeframe of the COVID-19 pandemic using novel keyword-based algorithms. The outcome indicated that both LSTM and RNN yielded impressive accuracy, with RNN being marginally more accurate. Nonetheless, the study had limitations as it only scrutinized South Korean news data, and the findings cannot be generalized to other cultural or ethnic groups. Overall, the study underscores the pressing need to consistently monitor obesity-related factors after the COVID-19 pandemic and to develop socially responsible countermeasures. Additional NLP studies on social media are urgently needed to apprehend how individuals perceive obesity in the aftermath of the pandemic.

As posited in another research paper on the topic of obesity prediction, obesity is defined as the incommensurability of energy amid calorie consumption and growth. This was scrutinized in [4] by the said authors. Stats reveal that the rate of obesity has nearly tripled since 1975, per the Organization for World Health. In 2016, more than 650 million individuals were obese, whereof 42% of adults aged 22 or older were overweight, and 16% were obese (World Health Organisation). The same year, the number of obese children was in excess of 340,000, which ballooned to 34 million by 2019, for children under five (World Health Organisation). By the dint of the facts presented, it may be inferred that obesity could soon pose a formidable threat to the planet's inhabitants. The crux of this paper is to recognize overweight people and enlighten them about the risk elements for obesity. This research aims to foretell the risk of obesity. It is divided into two stages: first, reading the data, then inspecting the factors associated with obesity to see if they match, and finally presenting the findings. Before utilizing nine machine learning-inspired techniques to scrutinize the outcomes using the available performance metrics using confusion matrix, AUC, ROC, we first preprocessed the data as our inquiry relies on numerous variables. The most effective algorithm that can accurately predict the outcome is then determined.

We gathered data for this study with the aim of predicting the incidence of obesity in Bangladesh. A thorough study has been conducted using various ML algorithms to predict the likelihood of obesity. The risk estimate for obesity includes nine novel, distinct categories. The efficiency of the classifiers has been assessed using six important performance measures. The relative properties of the obtained outcomes have been associated by scrutinizing the outcomes of identical works. By employing logistic regression, we achieved an accuracy score of 97.09%. The intention is to encompass a more extensive range of low-, medium-, and high-obese individuals by making this study more comprehensive and utilizing a wider dataset.

The project's client must provide details about their daily routines, dietary habits, height, weight, and so on. Several methodologies, including gradient boosting machine, random forest, multilayer perceptron, adaptive boosting, and classification and regression trees (CART) (GBM), are employed. One of the limitations of the dataset is its size in comparison to other datasets. As the results primarily apply to Bangladeshis, they are not diverse.

When predicting obesity in early stages of life and related effects, a researcher [5] conducted thorough and critical analyses of machine learning models, including the most recent ones that mix deep learning with electronic health data. These models are in contrast to other well-known statistical ones that mostly utilized logistic regression. The essential traits and applications coming from these models are highlighted along with future potential.

In order to stop obesity at an early age, prediction algorithms must be made to check first which category of people fall under high risk category. Preventive measures can then be directed at the high-risk population, allowing for a more customized and economical approach to weight loss programmes.Moreover, by ranking the numerous risk factors in terms of significance through their study, predictive models let us determine which ones will be more helpful when developing these solutions. The models can also be applied as simulation tools for "what-if" analysis, in which researchers alter one or more predictor variables to evaluate how doing so would impact obesity in particular subpopulations.

The creation of prediction algorithms to identify those who are at high risk is extremely beneficial for the prevention of childhood and adolescent obesity. As a result, preventive efforts can concentrate on the high-risk group, allowing for a more tailored and practical approach to weight loss programmes. Additionally, we can rank the various risk factors according to importance and figure out which ones will be most useful when creating these interventions by studying the data from the prediction models. The models can also be used as simulation tools for "what-if" analysis, in which researchers change one or more predictor variables and then examine the effects on obesity in specific subgroups.

It is intended that this Review would help the wide range of experts in the field, including pediatricians, nurses, nutritionists, statisticians, data scientists, engineers, and epidemiologists, to approach the prevention of childhood and teenage obesity in a more efficient and creative way.

Researchers investigated the predictive capabilities of deep learning techniques for major postpartum issues from bariatric surgery, as documented in a national quality registry, in [6]. The study included patients listed in the SoReg between 2003 and 2007, with people who had undergone a surgery for decreasing their weight in 2003-2006 used for training data and those in 2007 used for test data. Complications that occur after the operation were categorized using the CD classification, with the serious ones being those requiring general anesthesia for medical intervention, resulting in organ failure, or fatal. The study compared three supervised deep learning neural networks, namely the MLP, CNN and RNN, using accuracy, sensitivity, specificity, MCC, and AUC. The artificial technique for minority oversampling was used to enhance the number of patients with catastrophic effects.

With a total of 39,831 and 6210 people used for training data and test data, respectively, the incidence rates of severe problems were 1.4% (1160/39,831) and 5.0% (136/6210). The MLP, when trained on the SMOTER data, showed promising performance with an AUC of 0.21(98% CI 0.43-0.78) but performed poorly with an AUC of 0.43 (46% CI 0.35-0.76) for the test data. The CNN performed similarly to the MLP, generating AUCs of 0.26 (63% CI 0.49-0.11) and 0.49 (15% CI 0.53-0.20), respectively, for the test data and SMOTER data. The RNN fared worse than the MLP and CNN, with AUCs of 0.35 (25% CI 0.24-0.36) and 0.13 (33% CI 0.13-0.174) for the SMOTER data and test data, respectively.

In summary, the study revealed that MLP and CNN had a better ability to predict postoperative major issues following surgery in the Soreg dataset. However, the issue that the training data is overfitting is still to be resolved , and this requires the use of intraoperative and postoperative data.

In [7], researchers sought to discover clinical thresholds for detecting high-risk children prior to the onset of obesity and explain the habits and outcomes of processes that lead to risk of children developing obesity by the age of six. Two fit groups (BMI 7th to 72th percentile) and two highly obese groups (BMI 99th percentile) were selected from the populations seen in paediatric referral and primary care clinics. Utilizing a cohort drawn from the general community, the usefulness of the identified risk thresholds was confirmed. For analysis, repeated-measures mixed modelling and logistic regression were used.

480 persons with extreme obesity and 783 people who were normal weight participated in the original trial. A significant BMI difference between those people who were at the peak of obesity and those who were at their normal weight was visible at age 4 months, or one year before the median age at the development of obesity (P .001). A threshold of the WHO 95th percentile for BMI at 3, 13, and 38 months robustly predicted severe obesity by the age of 7 years (sensitivity, 23%57%; specificity, 96%). When tested on a second independent cohort (x = 2468), this BMI threshold showed a sensitivity of 56%-89% and a specificity of 65%-89%.

The patient's birth date, birth weight, visit date, sex, and self-reported race and ethnicity were all acquired from theEHR and/or medical records. Except for the Young Child Clinic cohort, all cohorts had their insurance status verified (YCC). The initial study's execution was approved by the Institutional Review Board of the medical center(CCHMC) . The validation study received approval from the Multiple Institutional Review Board, as well as a waiver from the insurance act of consent.

CHAPTER 3: SYSTEM DEVELOPMENT

Computational

To efficiently train our model, high configuration GPUs play a crucial role, which are readily available both online and on a user's computer. The training time of the model is significantly influenced by the GPU used. To ensure optimum performance, GPUs with higher memory, typically within the range of 4-16 GB, are suggested for such applications. Numerous Python libraries, including NumPy, Keras, and Tensorflow, are available to help facilitate the development process, and software like Jupyter notebook, PyCharm, and VScode can be employed to program the model.

For our project, we utilized Google Colaboratory, an online platform that enables users to write and execute code in Python using a browser, making it an excellent choice for machine learning, deep learning, data analysis, and education. We conducted our experiments on Google Colaboratory using an Intel® Xeon® processor running at 2.20GHz and 12.72 GB of RAM, coupled with a Nvidia Tesla T4, ensuring that we had the necessary resources to complete our work efficiently.

Mathematical

1) BMI (Body Mass Index)

$$BMI = \frac{Weight(kg)}{\{Height(m)\}^2}$$

Fig 13. BMI Mathematical Equation

The Body Mass Index (BMI), illustrated in Fig. 13, is a mathematical formula that compares an individual's weight to the square of their height. Determining whether someone is obese or not heavily relies on their BMI, which categorizes individuals into various groups based on their BMI score. These groups are as follows:

- Underweight: BMI < 18.5
- Normal: BMI 18.5 - 24.9
- Overweight: BMI 25.0 - 29.9
- Obesity I: BMI 30.0 - 34.9
- Obesity II: BMI 35.0 to 39.9
- Obesity III: BMI > 40.0

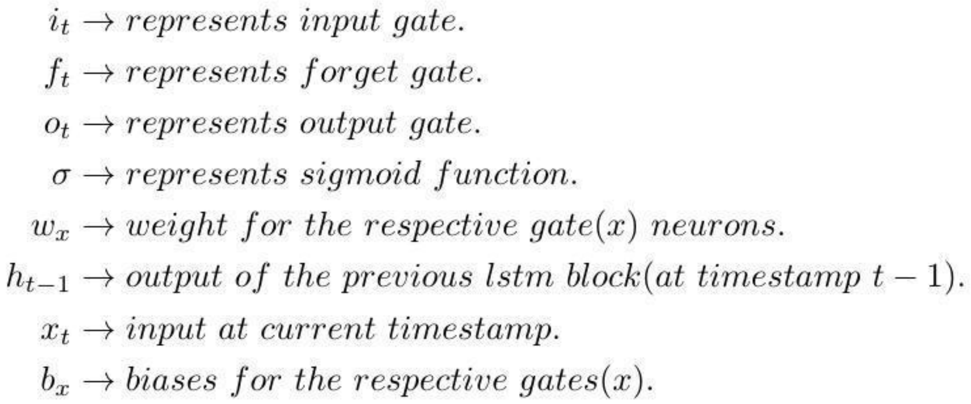
As an essential parameter, BMI provides a quick and straightforward way to assess an individual's health status and to identify potential risks associated with their weight.

(2) LSTM

$$\begin{aligned} i_t &= \sigma(w_i[h_{t-1}, x_t] + b_i) \\ f_t &= \sigma(w_f[h_{t-1}, x_t] + b_f) \\ o_t &= \sigma(w_o[h_{t-1}, x_t] + b_o) \end{aligned}$$

Fig 14. Mathematical equation for LSTM

Here in Fig.14,



$i_t \rightarrow represents\ input\ gate.$

$f_t \rightarrow represents\ forget\ gate.$

$o_t \rightarrow represents\ output\ gate.$

$\sigma \rightarrow represents\ sigmoid\ function.$

$w_x \rightarrow weight\ for\ the\ respective\ gate(x)\ neurons.$

$h_{t-1} \rightarrow output\ of\ the\ previous\ lstm\ block(at\ timestamp\ t-1).$

$x_t \rightarrow input\ at\ current\ timestamp.$

$b_x \rightarrow biases\ for\ the\ respective\ gates(x).$

The LSTM formula was used in deep learning model building for prediction and was later compared with the results of machine learning model using logistic regression.

2) LOGISTIC REGRESSION

$$y = \frac{e^{(bo + b1X)}}{1 + e^{(bo + b1X)}}$$

Fig 15. Logistic Regression Mathematical Equation

Here in Fig.15,

x = input value

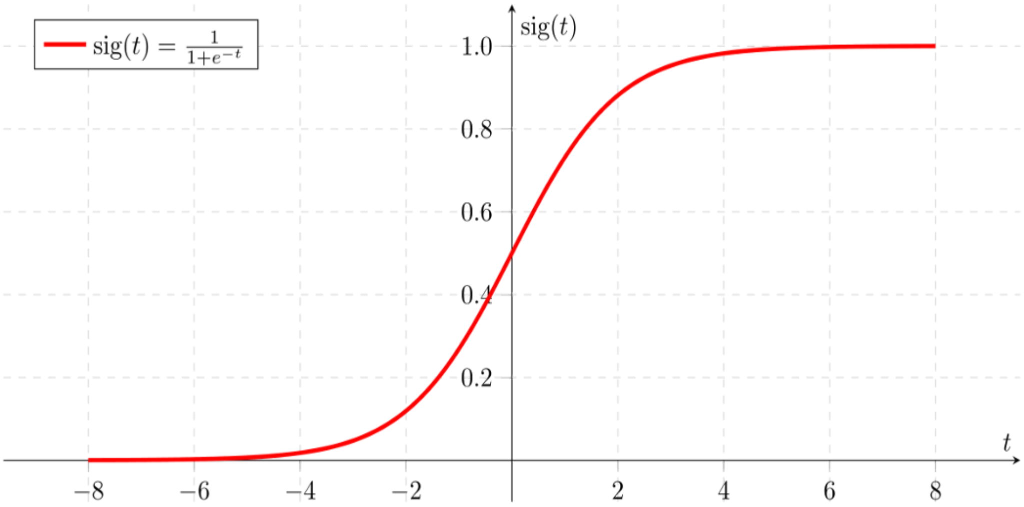
y = predicted output

b0 = bias or intercept term

b1 = coefficient for input (x)

Logistic regression formula was used in machine learning model building and the major features used were height , weight and BMI.

3) SIGMOID ACTIVATION FUNCTION



4) K-means CLUSTERING

$$J(V) = \sum_{i=1}^{C} \sum_{j=1}^{Ci} (||x_i - v_j||)^2$$

Fig 16. K-means Clustering Mathematical Equation

Where in Fig.16,

‘||xi - vj||’ is the ED between xi and vj.

‘ci’ is the number of dataset points in ith cluster.

‘c’ is the number of centroids or cluster centers.

K-means clustering was an experiment on the dataset as originally accuracy was measured using regression analysis and clustering was an experiment to check whether clustering can also be used to classify the data and predict whether a child belongs in the category of obese or not.

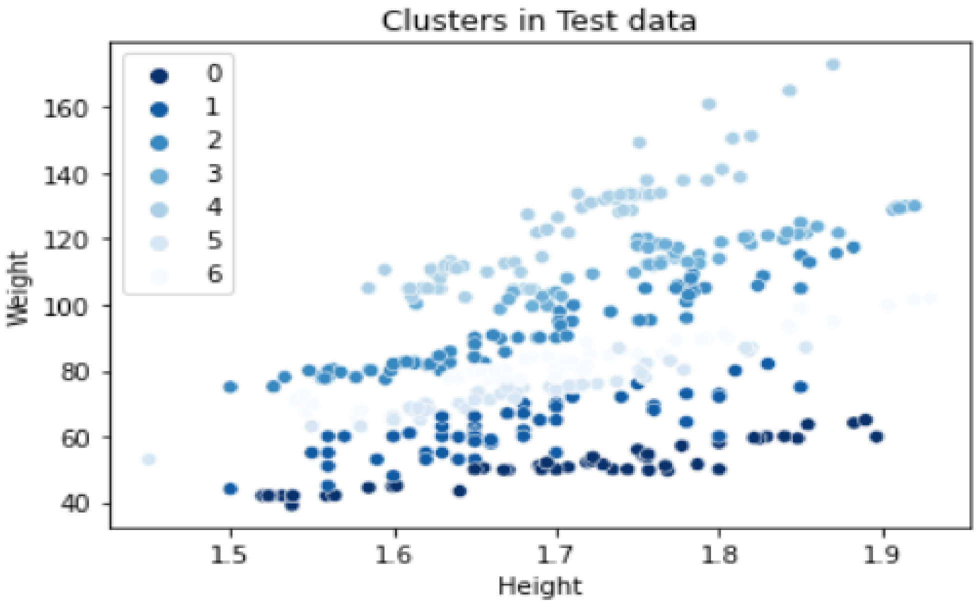


Fig.17 Clustering experimentation results

Experimental

The project was implemented as follows-

- First we studied the relationship between a person being obese and other factors like eating habits, height, weight, physical activity etc.
- We then did data preprocessing and training and testing split.
- We used Decision Tree Classifier to train and evaluate the model
- Finally we used evaluation parameters to find the accuracy of our model

Model Design

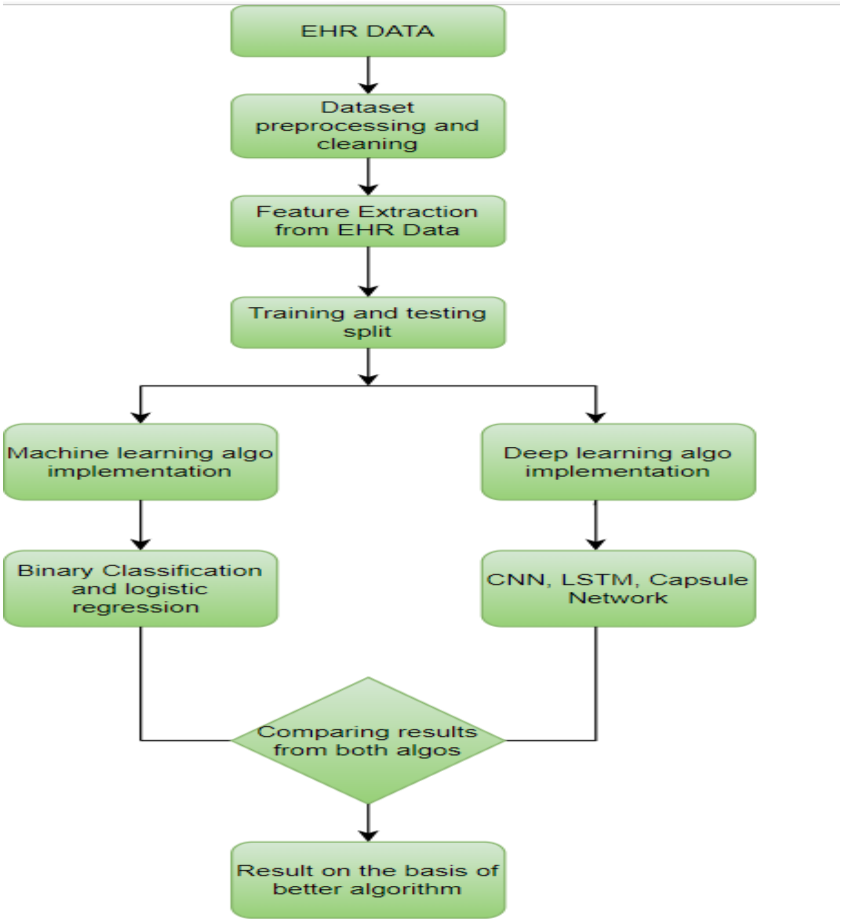
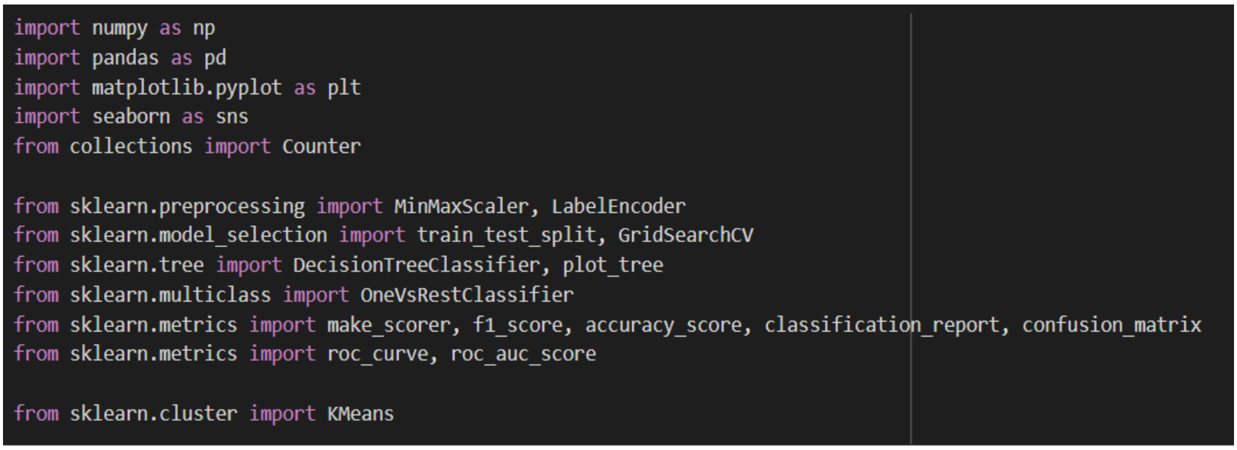

EHR DATA
Dataset preprocessing and cleaning
Feature Extraction from EHR Data
Training and testing split
Machine learning algo implementation
Deep learning algo implementation
Binary Classification and logistic regression
CNN, LSTM, Capsule Network
Comparing results from both algos
Result on the basis of better algorithm

Fig 18. Model Design

As shown in Fig.17, we have taken Electronic Health Record (EHR) data for our work which includes the height, weight, eating habits and physical activities of the people[9]. We visualized the data and established a relationship between being overweight and all the columns present in the dataset and after that we preprocessed data including label encoding , changing data types and scaling of the features. Pre-processing step is taken forward to a training and testing split where we have split our data and then we have built our model using a decision tree classifier[14].

Model Development



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from collections import Counter

from sklearn.preprocessing import MinMaxScaler, LabelEncoder
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.tree import DecisionTreeClassifier, plot_tree
from sklearn.multiclass import OneVsRestClassifier
from sklearn.metrics import make_scorer, f1_score, accuracy_score, classification_report, confusion_matrix
from sklearn.metrics import roc_curve, roc_auc_score

from sklearn.cluster import KMeans
```

Fig 18. Importing Libraries

In Fig.19, We are checking the number of items in each Gender category to know whether the dataset is biased or not. We got the number of females as 1043 and number of males as 1068 which proves the dataset is not biased.

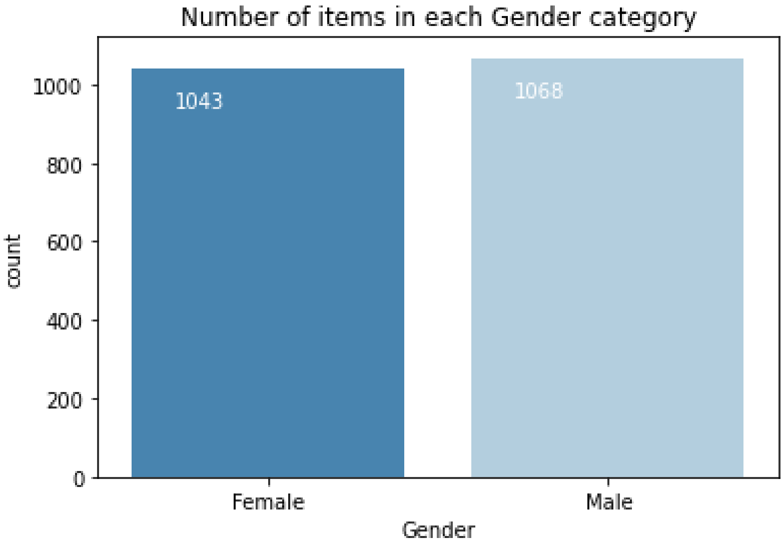


Fig 19. Checking Gender Ratio

In Fig.20, We checked the correlation between the primary factors affecting obesity, that is, height and weight. Height and weight are used to compute BMI and we came to the conclusion that height and weight are directly related to each other.

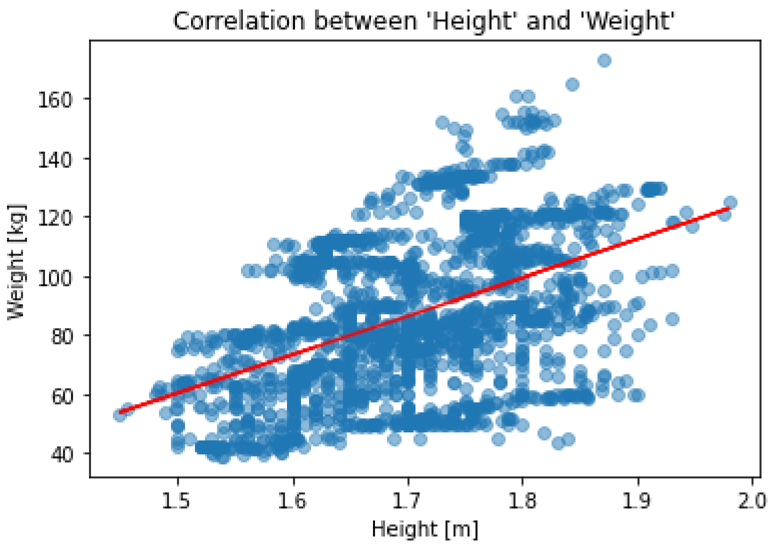


Fig 20. Determining the relationship between height and weight

In Fig.21 We computed the number of items in each category from insufficient weight to obesity type III in our dataset for understanding the dataset better.

People are classified into the following groups based on their BMI:

BMI 18.5 indicates underweight.

You are overweight if your BMI is between 25.0 and 29.9. You are obese I if your BMI is between 30.0 and 34.9. You are classified as obese II if your BMI is between 35.0 and 39.9. BMI > 40 Indicates Type III Obesity.
The number of people by category is displayed below.

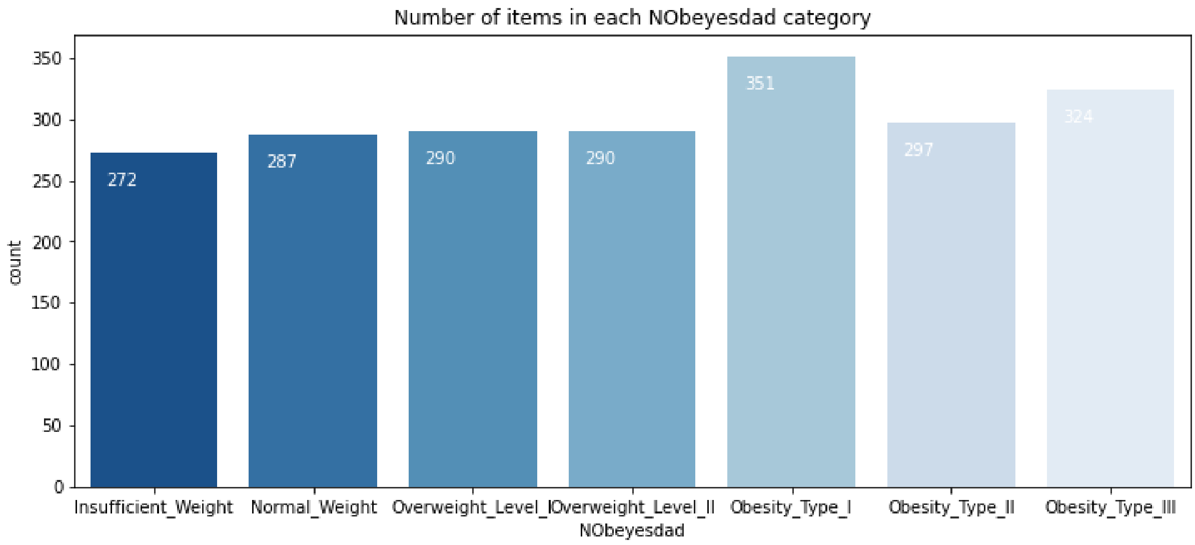


Fig 21. Data Categorization

It's fascinating to see how different category variables affect overweight and obesity. We found that "Insufficient weight" is more regular in females than in males. However, there are more fat men than women, with the exception of the ultimate category of extreme obesity. In Fig. 22, this is depicted.

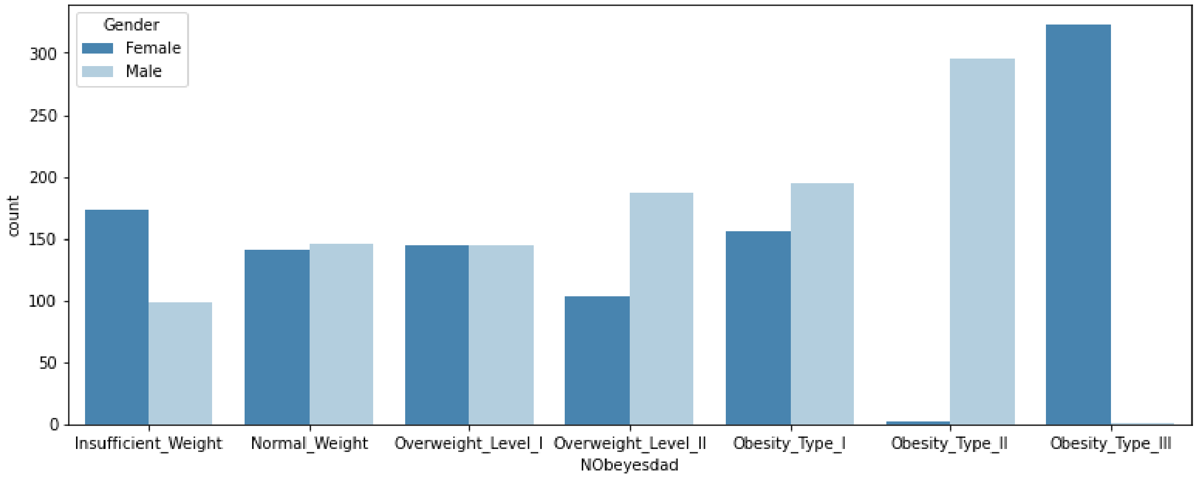


Fig 22. Relationship between being overweight and gender

It appears that obesity is a family trait. All of the people who were classified as overweight or obese had relatives who also struggled with weight issues. This can be viewed using Fig.23.

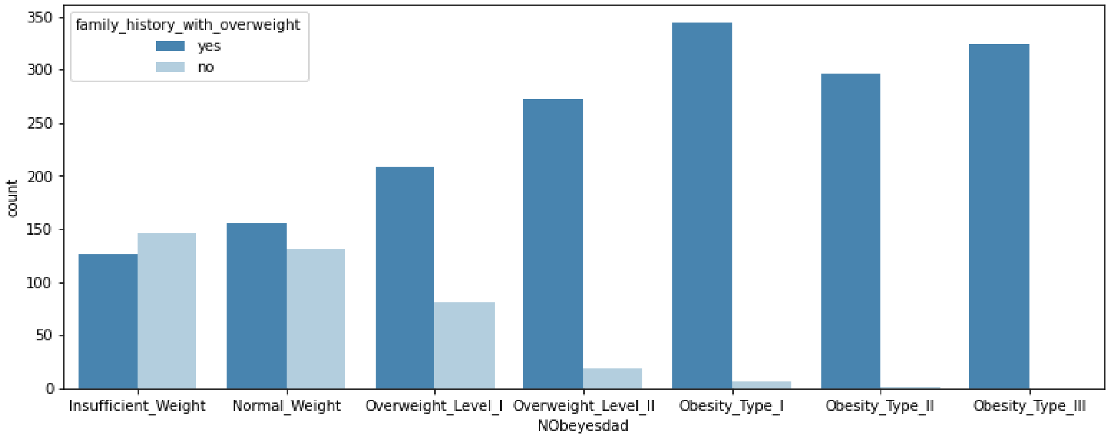


Fig 23. Relationship between being overweight and family history

People with normal weights and those who are overweight or obese eat foods high in calories. Perhaps the amount of food consumed makes a difference and influences body fats.The relationship can be seen in Fig.24

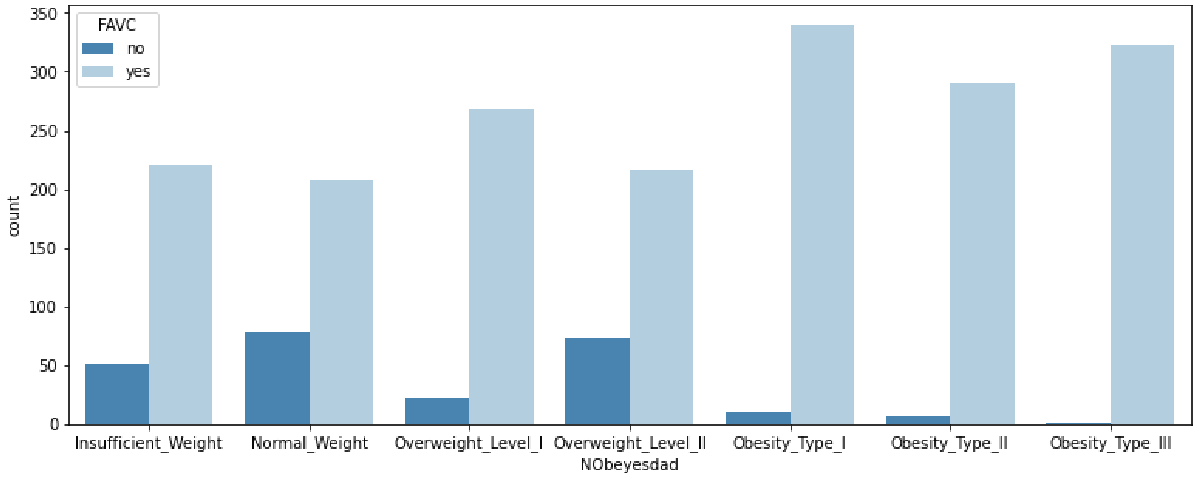


Fig 24. Relationship between being overweight and high calories food

People who "often" or "always" eat between meals do not appear to have excessive weight. Only occasional snackers fall into the "Overweight" or "Obesity" categories. This can be viewed in Fig.25.

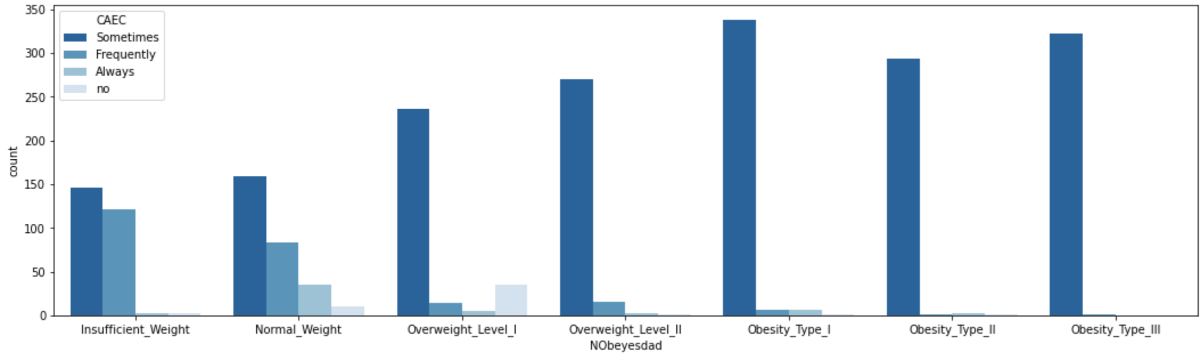


Fig 25. Relationship between overweight and food between meals

People who typically keep an eye on their caloric intake are less likely to gain weight as shown in Fig.26.

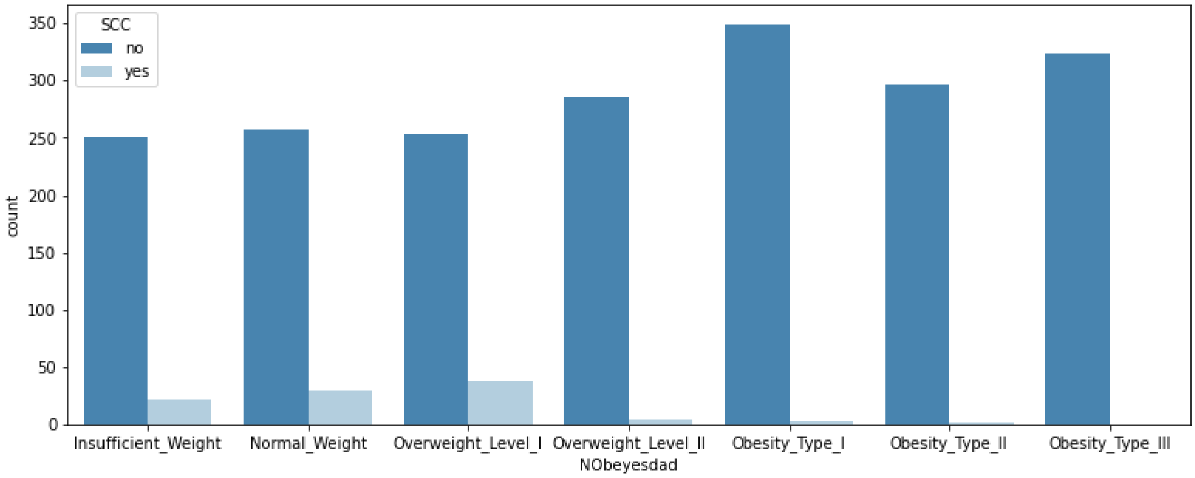


Fig 26. Relationship between being overweight and monitoring calories

The weight of a person does not appear to be (significantly) affected by transportation. All three groups—the normal, the overweight, and the slim—use cars and public transportation. This is shown in Fig.27

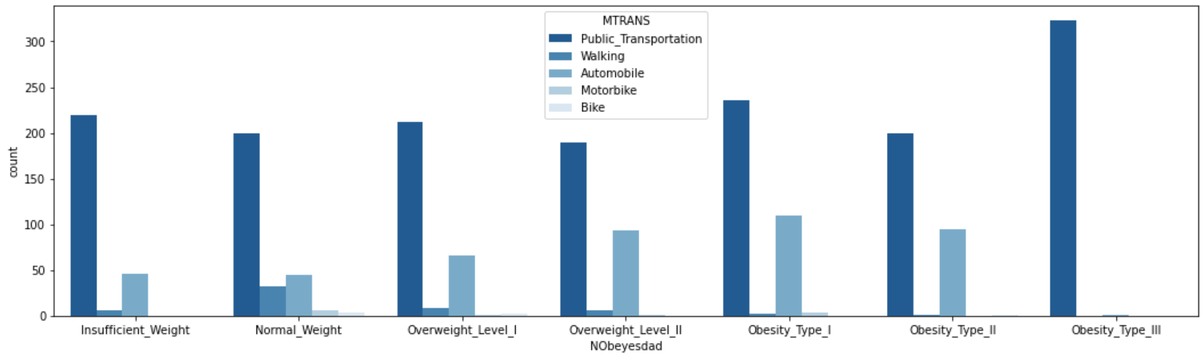


Fig 27. Relationship between being overweight and means of transport used

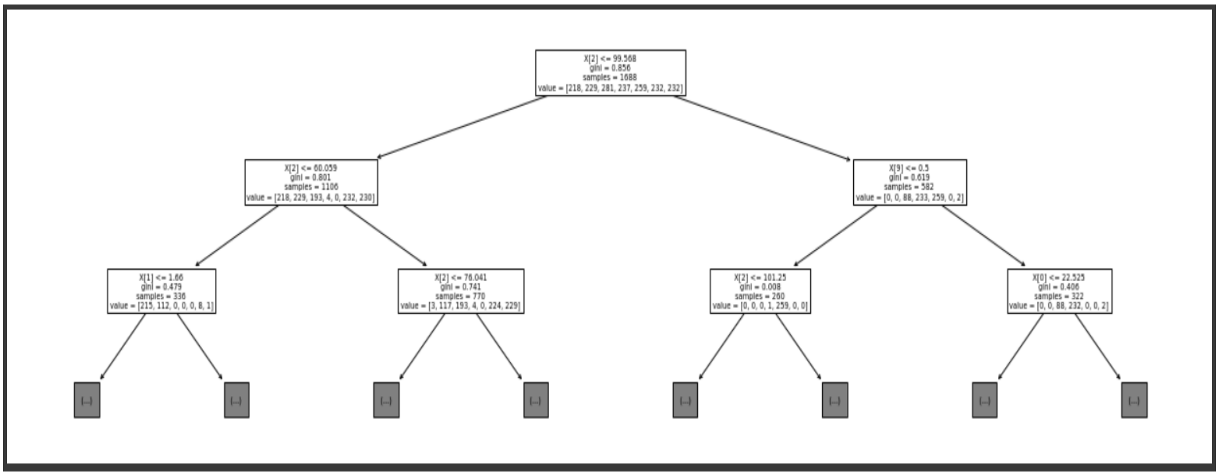
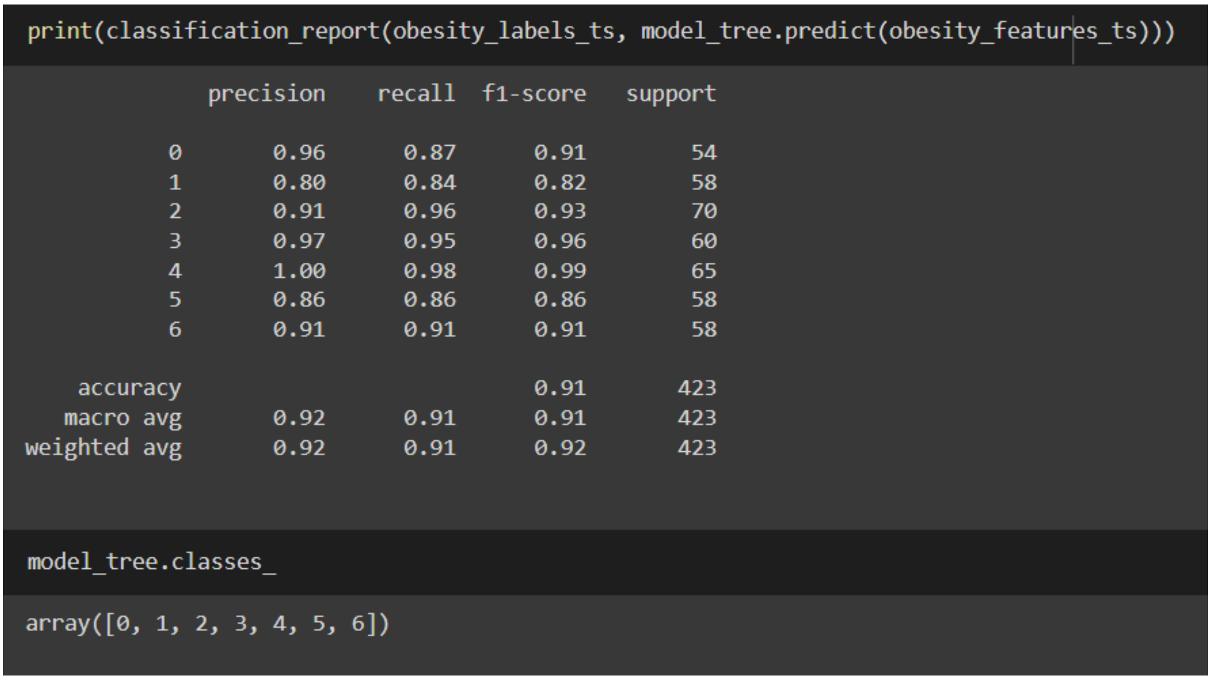


Fig 28. Decision Tree

Decision trees in Fig. 29 are easier to understand. Creating a story out of them demonstrates how choices are formed. Each node asks a question, and depending on whether it receives a "True" or "False" response, it passes the information to either the leftmost or rightmost child node. Up until there are no more questions permitted or the "max depth" restriction has been reached, this process continues.



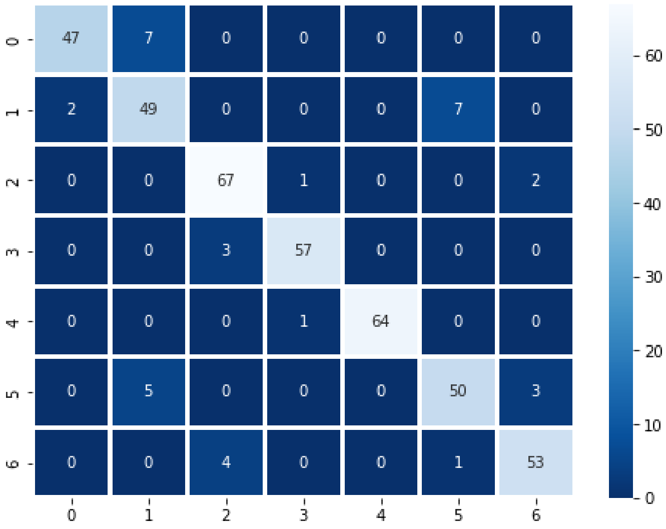
```
print(classification_report(obesity_labels_ts, model_tree.predict(obesity_features_ts)))
```

```
              precision    recall  f1-score   support

           0       0.96      0.87      0.91        54
           1       0.80      0.84      0.82        58
           2       0.91      0.96      0.93        70
           3       0.97      0.95      0.96        60
           4       1.00      0.98      0.99        65
           5       0.86      0.86      0.86        58
           6       0.91      0.91      0.91        58

    accuracy                           0.91       423
   macro avg       0.92      0.91      0.91       423
weighted avg       0.92      0.91      0.92       423
```

```
model_tree.classes_
```

```
array([0, 1, 2, 3, 4, 5, 6])
```

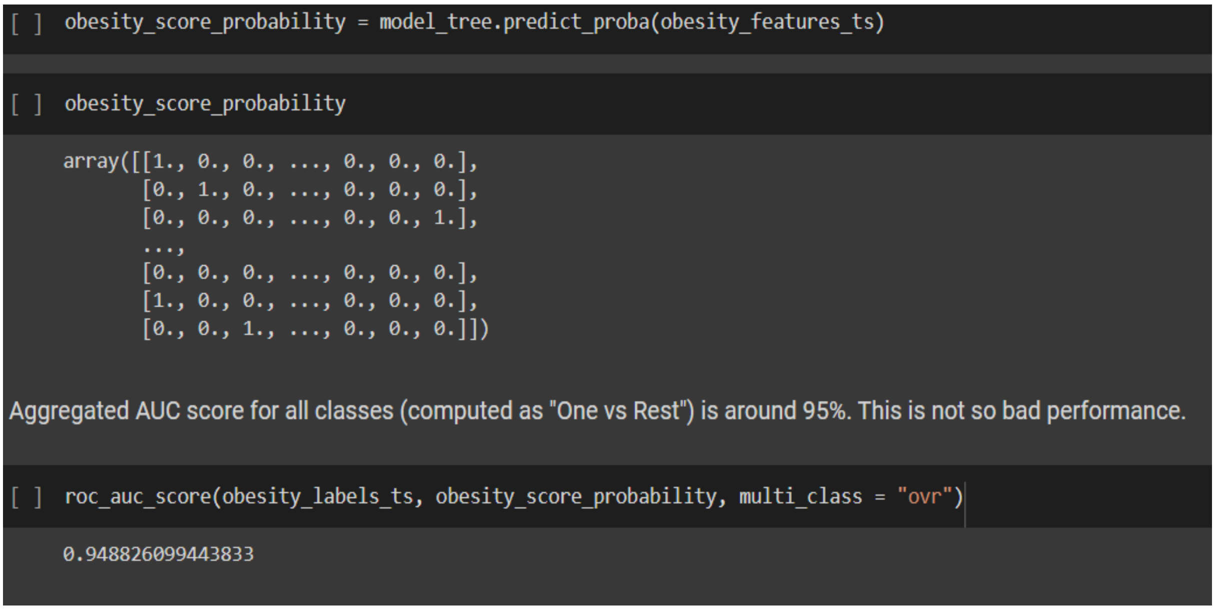
Fig 29. Classification Report



	0	1	2	3	4	5	6
0	47	7	0	0	0	0	0
1	2	49	0	0	0	7	0
2	0	0	67	1	0	0	2
3	0	0	3	57	0	0	0
4	0	0	0	1	64	0	0
5	0	5	0	0	0	50	3
6	0	0	4	0	0	1	53

Fig 30. Confusion matrix

Labels that are actual vs. predicted are shown in the Confusion matrix as in Fig.30. Actual classes are represented by rows, whereas anticipated classes are represented by columns. As an illustration, 47 samples were appropriately placed in class 0, but 7 samples were mistakenly placed in class 1. There was only ever one class 4 sample that had a class 3 labeling error.



```
[ ]  obesity_score_probability = model_tree.predict_proba(obesity_features_ts)
```

```
[ ]  obesity_score_probability

     array([[1., 0., 0., ..., 0., 0., 0.],
            [0., 1., 0., ..., 0., 0., 0.],
            [0., 0., 0., ..., 0., 0., 1.],
            ...,
            [0., 0., 0., ..., 0., 0., 0.],
            [1., 0., 0., ..., 0., 0., 0.],
            [0., 0., 1., ..., 0., 0., 0.]])
```

Aggregated AUC score for all classes (computed as "One vs Rest") is around 95%. This is not so bad performance.

```
[ ]  roc_auc_score(obesity_labels_ts, obesity_score_probability, multi_class = "ovr")

     0.948826099443833
```

Fig 31. ROC Score and Curve

In Fig. 31, we see the ROC curve, which is widely used as a classification matrix. This curve indicates the accuracy of a ML model at each given level. It is a plot of TPR versus FPR. The AUC shows the chances that a randomly chosen positive dataset point will rank higher than a randomly chosen negative dataset point . AUC values range from 0 to 1, where a model with all wrong outcomes or the wrongs predictions has an AUC of 0, and a model with all correct predictions has an AUC of 1.

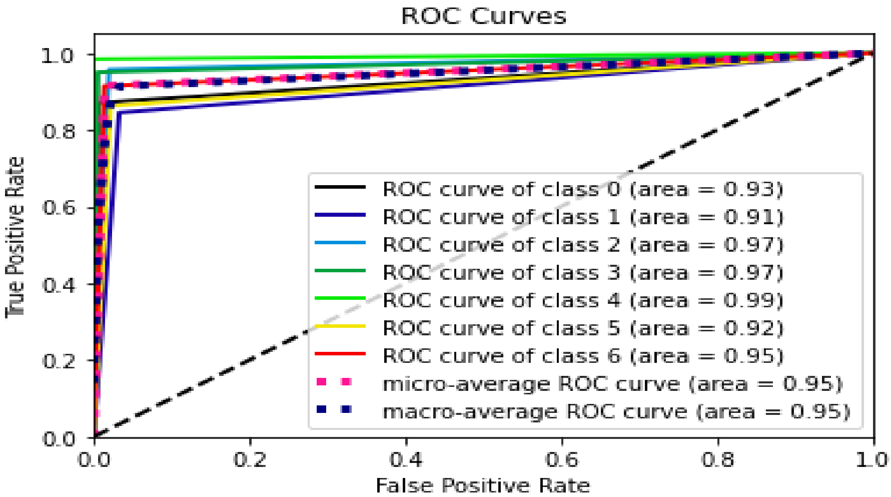


Fig 32. ROC Curve plotting

Below(Fig.33) is a plot of the clusters (made by "Height" and "Weight" attributes) in the testing data based on their actual labels.

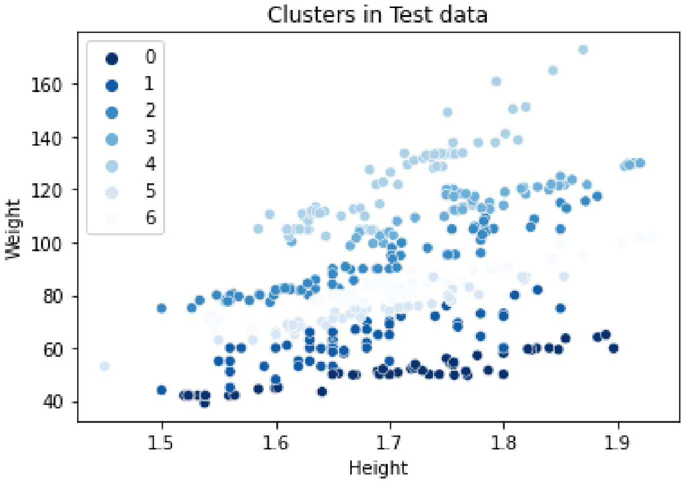


Fig 33. Clusters in Test Data

Each point (specified by feature values) is given a cluster using KMeans, which calculates the distances between each point. Clustering might therefore be viewed as a method for unsupervised learning in classification (algorithm). However, as there are no evaluation indicators for unrevised training, its performance could not be evaluated. Due to the clustering algorithm's inability to determine their order, the classes on the two plots are different (i.e., which predicted values correspond to class 0, which - to class 1, etc.). Nevertheless, KMeans was able to combine the "Height" and "Weight" data into seven groups that closely matched the test labels. This is shown in Fig.34.

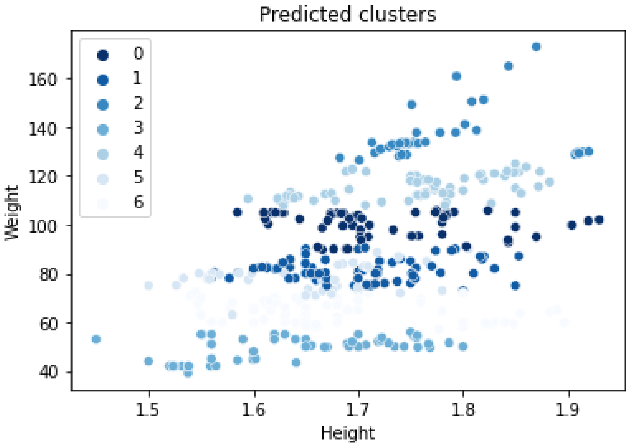
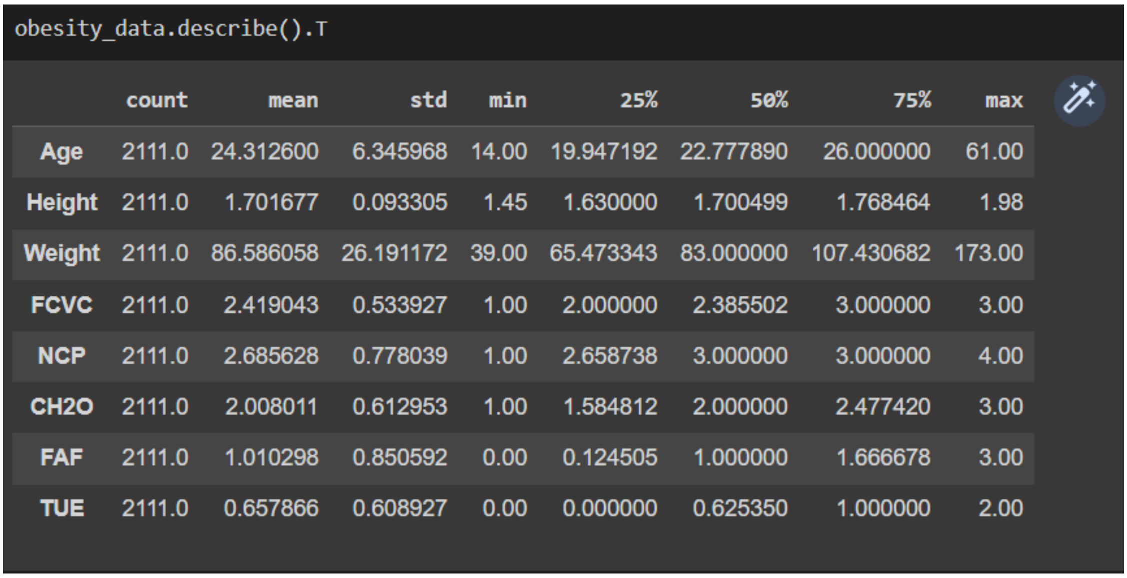


Fig 34. Predicted Clusters

Statistical

Statistics with five numbers do not provide much insight into features with numerical values. Except for age, height, and weight, most columns' data cannot be interpreted Fig.35.



```
obesity_data.describe().T
```

```
         count       mean        std    min        25%        50%         75%     max
Age     2111.0  24.312600   6.345968  14.00  19.947192  22.777890   26.000000   61.00
Height  2111.0   1.701677   0.093305   1.45   1.630000   1.700499    1.768464    1.98
Weight  2111.0  86.586058  26.191172  39.00  65.473343  83.000000  107.430682  173.00
FCVC    2111.0   2.419043   0.533927   1.00   2.000000   2.385502    3.000000    3.00
NCP     2111.0   2.685628   0.778039   1.00   2.658738   3.000000    3.000000    4.00
CH2O    2111.0   2.008011   0.612953   1.00   1.584812   2.000000    2.477420    3.00
FAF     2111.0   1.010298   0.850592   0.00   0.124505   1.000000    1.666678    3.00
TUE     2111.0   0.657866   0.608927   0.00   0.000000   0.625350    1.000000    2.00
```

Fig 35. Statistical Analysis

Outliers and quartiles are displayed in the boxplots below. The last five columns' distributions are not taken into consideration.

The first boxplot implies that the "Age" column contains outliers. However, 40, 50, or 60 years old are not eliminated because they are normal numbers (they are not outliers or errors). There don't appear to be any outliers for "Height" and only a few for "Weight." These also receive no treatment. This is shown in Fig.36.

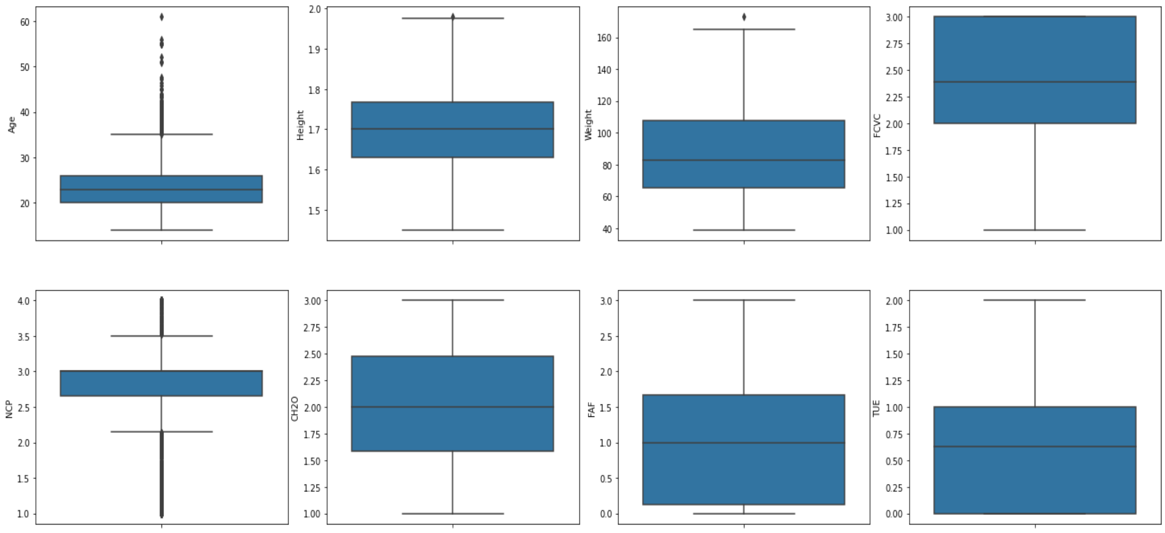


Fig 36. Statistical Representation of Dataset

There is little to no (linear) association between numerical features. Thus, the table still has all of the features. This is shown in Fig.37

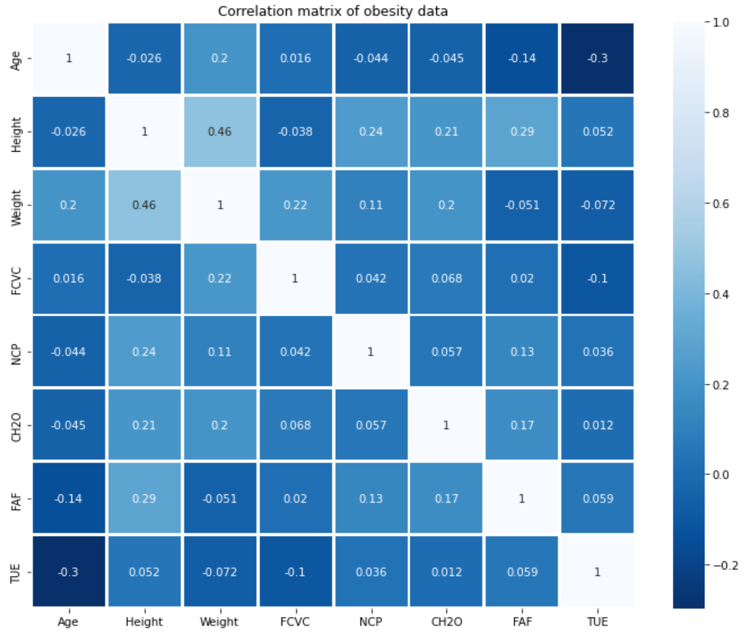


Fig 37. Correlation Matrix of Obesity Data

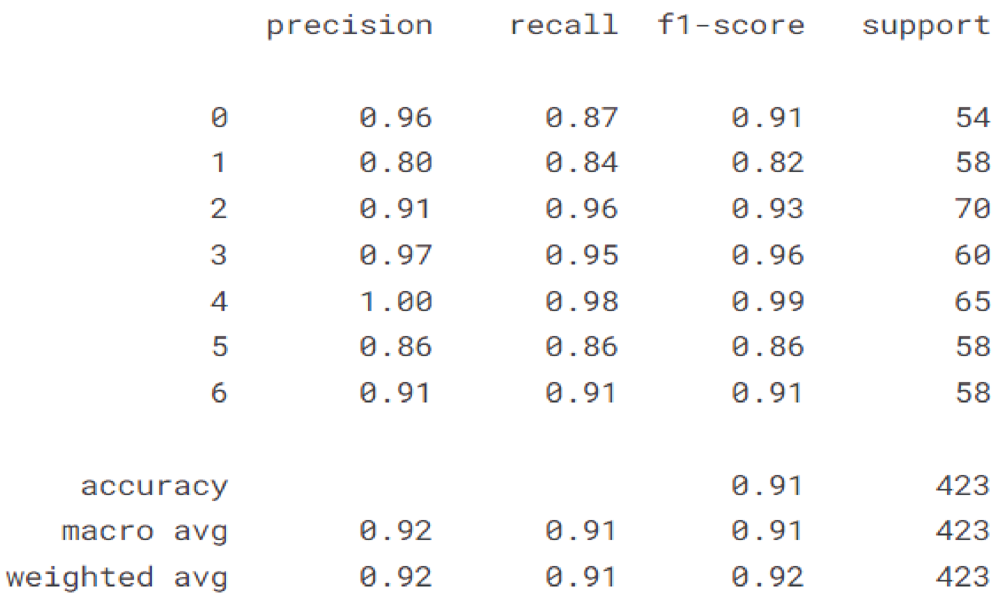

```
              precision    recall  f1-score   support

           0       0.96      0.87      0.91        54
           1       0.80      0.84      0.82        58
           2       0.91      0.96      0.93        70
           3       0.97      0.95      0.96        60
           4       1.00      0.98      0.99        65
           5       0.86      0.86      0.86        58
           6       0.91      0.91      0.91        58

    accuracy                           0.91       423
   macro avg       0.92      0.91      0.91       423
weighted avg       0.92      0.91      0.92       423
```

Fig 38. Accuracy of classification model

CHAPTER 4: PERFORMANCE ANALYSIS

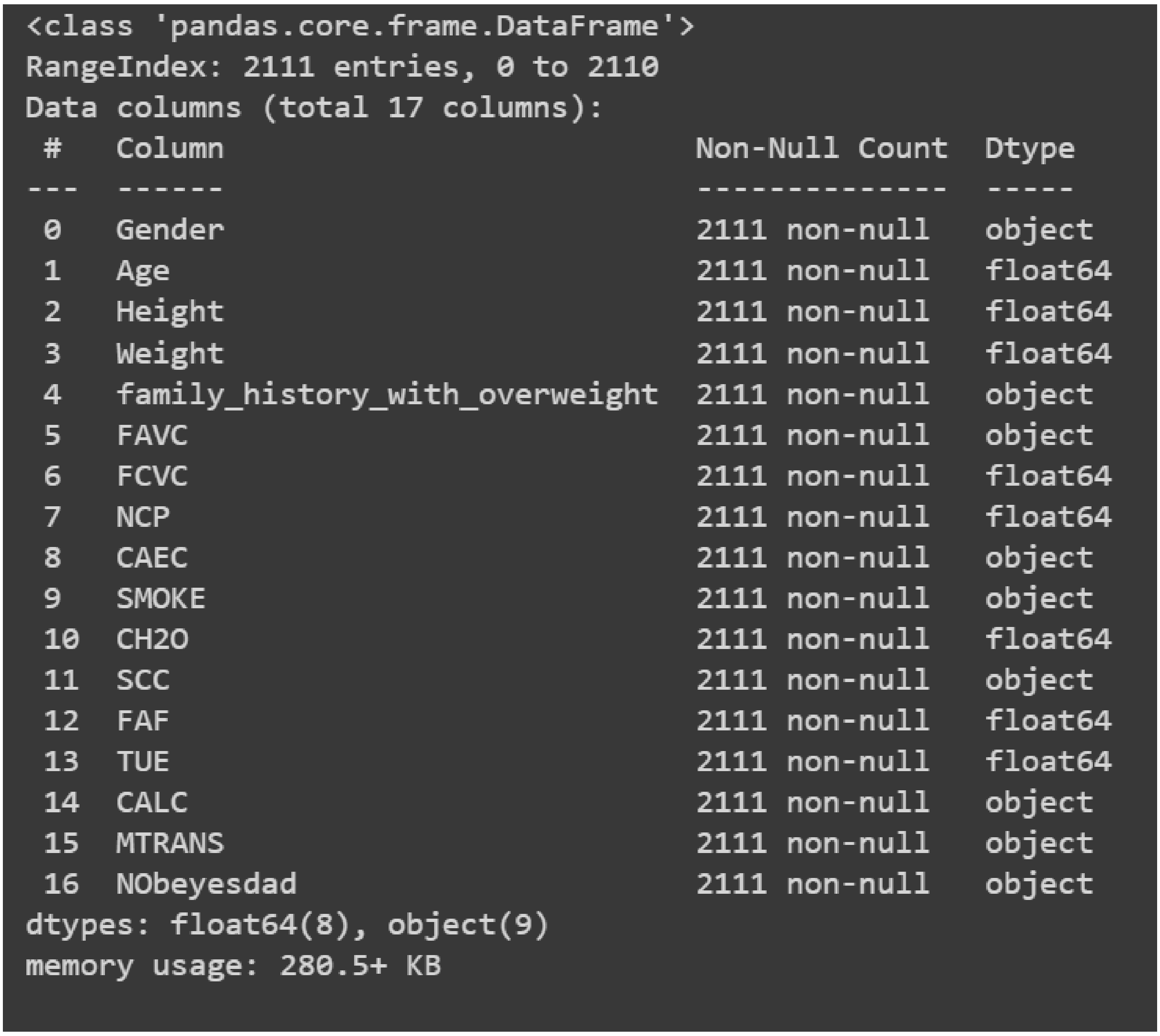
Dataset

The University of California Irvine Machine Learning Repository has a dataset for estimating obesity levels based on eating habits and physical condition in people from Colombia, Peru, and Mexico which we used for our prediction.

The collection consists of 2111 records and 17 attributes. The dataset's authors note that while 23% of the records were received directly from users via a web platform, 77% of the records were artificially manufactured using the Weka tool and SMOTE filter. The records can be categorized into seven groups according to a class attribute named "NObesity" (Obesity Level): "Insufficient".

The likelihood of being obese is impacted by a person's eating patterns, level of physical activity, and genes. After poring through the data, we came across a solid model that could tell whether a person is underweight, overweight, or whether their body falls within the normal (healthy) range. Instead, an attempt was made to cluster the data based on all features (predictors). Both the classification and clustering tasks are discussed after data exploration.

After finding a relationship between being overweight and other features we also came across the results that height and weight are the primary reasons for being obese or overweight but are also supported by other features like physical activities, eating habits and family history. We also found the relationship between height and weight using a correlation matrix as these are the primary factors affecting obesity.(Fig.38)



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2111 entries, 0 to 2110
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   Gender                          2111 non-null   object 
 1   Age                             2111 non-null   float64
 2   Height                          2111 non-null   float64
 3   Weight                          2111 non-null   float64
 4   family_history_with_overweight  2111 non-null   object 
 5   FAVC                            2111 non-null   object 
 6   FCVC                            2111 non-null   float64
 7   NCP                             2111 non-null   float64
 8   CAEC                            2111 non-null   object 
 9   SMOKE                           2111 non-null   object 
 10  CH2O                            2111 non-null   float64
 11  SCC                             2111 non-null   object 
 12  FAF                             2111 non-null   float64
 13  TUE                             2111 non-null   float64
 14  CALC                            2111 non-null   object 
 15  MTRANS                          2111 non-null   object 
 16  NObeyesdad                      2111 non-null   object 
dtypes: float64(8), object(9)
memory usage: 280.5+ KB
```

Fig 38. Dataset

Implementation

Machine learning implementation :-

- Loading the data and doing exploratory data analysis to study and know the dataset better.

- Pre-processing the data was the next step in which we did data cleaning due to the presence of noise in the dataset and we also did scaling, encoding, conversion of data type etc.

- Training and testing split was the next step which was followed by building the model using a decision tree classifier.

- The last step was using evaluation parameters to find the accuracy of the model and our best accuracy came out to be 95%.

For evaluation of performance we have used various methods such as confusion matrix, ROC Curve and clustering(training and predicted clusters).

In this situation, only "accuracy" and "f1 score" will be used to describe the Decision Tree's performance on the two sets. On the training data, accuracy and "f1 score" are both 100%, however on the testing data, they are 91%–92%. The latter demonstrates how overfit the model is. The algorithm's performance can be enhanced by regularization (e.g., shallower trees, minimum samples per leaf), feature selection (e.g., removing irrelevant columns), feature selection with feature selection, or increasing the sample sizes in both sets. Neither of these strategies is further studied because "accuracy" and "f1 score" > 90% are reliable indicators of the model's accuracy.

.Decision trees in Fig. 39 are easier to understand. Creating a story out of them demonstrates how choices are formed. Each node asks a question, and depending on whether it receives a "True" or "False" response, it passes the information to either the leftmost or rightmost child node. Up until there are no more questions permitted or the "max depth" restriction has been reached, this process continues.

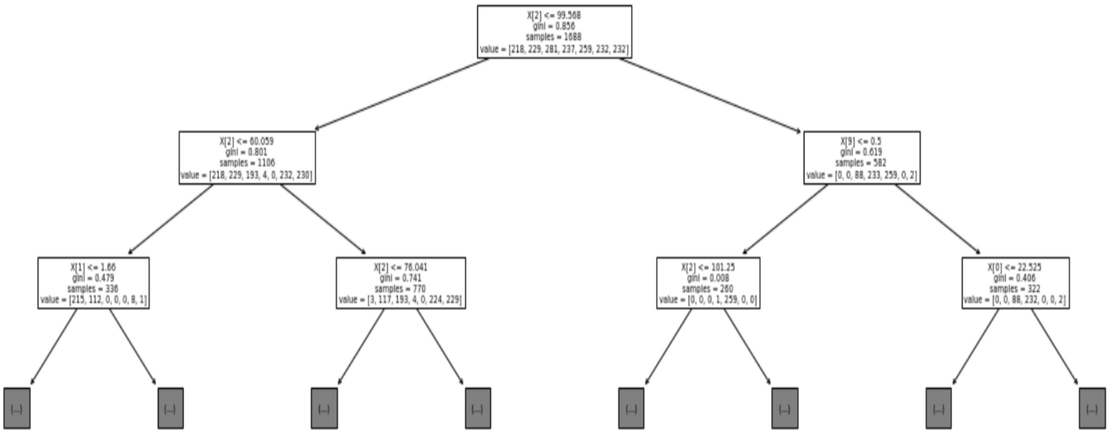
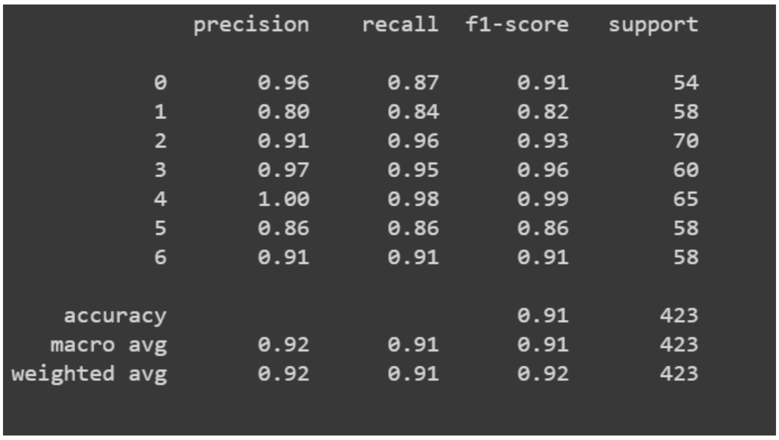


Fig 39. Decision tree

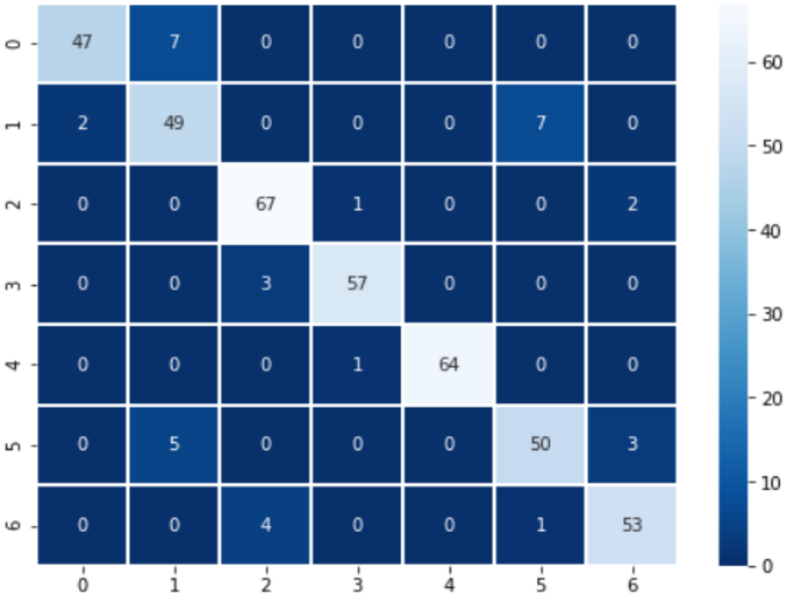
Scikit Learn's classification report function displays the classification success (metrics) for each class. For instance, "Obesity Type III" (class 4) had the majority of samples that were correctly categorized. The model achieved 99% "f1 score" and 100% "precision". However, features that indicated "Normal Weight" (class 1) were misinterpreted and received 80% or less on "accuracy" and "f1 score."(Fig.40)



```
              precision    recall  f1-score   support

           0       0.96      0.87      0.91        54
           1       0.80      0.84      0.82        58
           2       0.91      0.96      0.93        70
           3       0.97      0.95      0.96        60
           4       1.00      0.98      0.99        65
           5       0.86      0.86      0.86        58
           6       0.91      0.91      0.91        58

    accuracy                           0.91       423
   macro avg       0.92      0.91      0.91       423
weighted avg       0.92      0.91      0.92       423
```

Fig 40. Accuracy, Macro average and Weighted Average measure

Actual vs. anticipated labels are displayed in the Confusion matrix(Fig.41). Rows are used to represent actual classes, whereas columns are used to represent anticipated classes. For instance, 47 samples were correctly classified in class 0, whereas 7 samples were classified wrongly in class 1. One class 4 sample only ever received a class 3 labeling error.



	0	1	2	3	4	5	6
0	47	7	0	0	0	0	0
1	2	49	0	0	0	7	0
2	0	0	67	1	0	0	2
3	0	0	3	57	0	0	0
4	0	0	0	1	64	0	0
5	0	5	0	0	0	50	3
6	0	0	4	0	0	1	53

Fig 41. Confusion matrix

ROC curve is another popular categorization metric (Receiver Operating Characteristic curve). It is a graph that shows the effectiveness of a classification model at every level of categorization. These two parameters, True Positive Rate and False Positive Rate, are plotted on the curve. The area under the curve is a measure of how likely it is that a random positive example will be positioned next to a random negative example (AUC). AUC's value ranges from 0 to 1. A model that predicts 100% incorrectly has an AUC of 0, while a model that predicts 100% correctly has an AUC of 1. For the AUC and ROC curve, probability prediction scores must be calculated. These show the probability that a particular sample belongs to a particular class. The ROC Curves are plotted in Fig. 42 below. They ascend and move to the left, demonstrating strong model performance. As previously discovered, the model performs best for classes 4 (light green line), 6 (red line), and 0. (black line). On the legend, AUCs for all classes are shown.

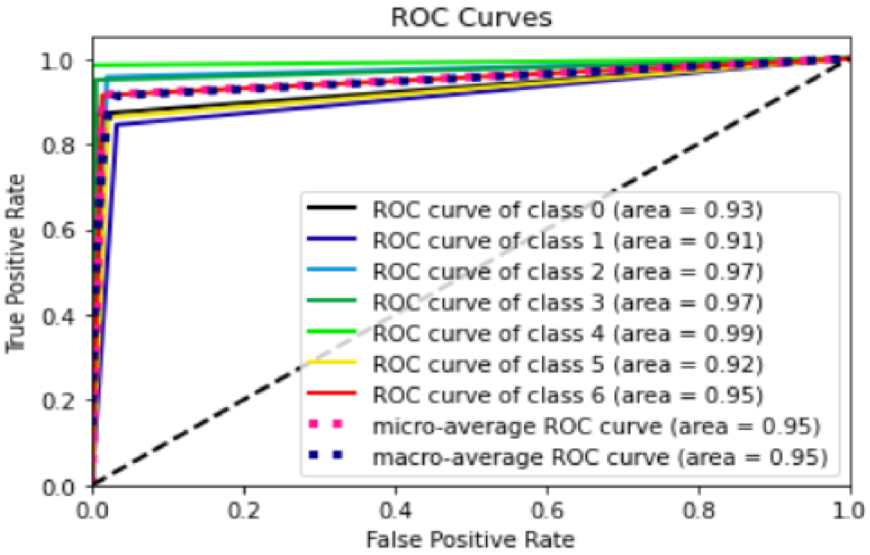


Fig 42. ROC Curve

We conducted an experiment using the features clustering of the dataset. If different clusters form, it means that certain types of overweight/obesity have particular values for the provided features. The simplest clustering algorithm, "KMeans," is used to complete the task.

Features and their projection should be shown visually to show how clustering functions. More than three dimensions cannot be represented on a 2D surface. Therefore, just those features that are necessary and contain the most important data are utilized. According to DecisionTreeClassifier, the most important columns are the second ("Height") and third ("Weight") rows (). They each provide between 21.9% and 47.85% of the data's information. The outcome in the table below also shows that some values in some columns might have been dropped since it was difficult to show how they related to obesity.

Based on their real labels, the "Height" and "Weight" attributes, which indicate the clusters in the testing data, are depicted in the following figure (Fig. 43). A cluster is assigned to each point using KMeans, which calculates the separations between each point (defined by feature values). Clustering could be viewed as a method for unsupervised learning in classification as a result (algorithm). Unsupervised training's effectiveness, however, cannot be measured because there aren't any evaluation criteria for it.

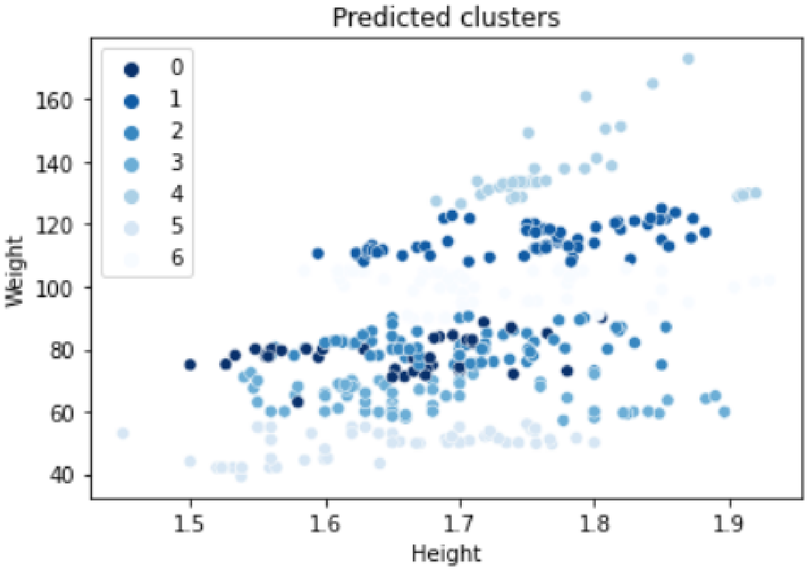


Fig 43. Predicted Clusters after K-Means Clustering

Deep learning implementation :-

After acquiring all the sub-cohorts, we initiated a complex data transformation process to effectively prepare the input data for our LSTM predictive model. Given the inherently irregular time intervals of medical visits and the distinct quantity of visits per patient, the initial data representation could not be used as is. In order to rectify this, we designed a comprehensive methodology to merge visits within a defined fixed time window, thus producing equidistant intervals. To achieve this, we employed a concatenation strategy over 30-day intervals throughout a 2-year observation window, enabling us to harmonize the sequence lengths for each patient. The process required the generation of approximately 25 sequences per patient. Our Fig. 3 illustrates the intricate details involved in producing these new sequences. Notably, every variable of interest, including the condition, medication, and procedure data that were recorded at least once over the 30-day interval, were represented as a '1' in the new sequences. Additionally, continuous variables were averaged over each 30-day period. Should there be any instances of no visits within the 30-day interval, we leveraged zero vectors to pad the sequence, ensuring equal sequence lengths for all patients. In addition to the conditions, procedures, medications, and measurements, we also incorporated the time interval between each visit sequence (Δt) to the end of each visit vector. This additional layer of complexity is highly beneficial as it has been shown to enhance the quality of time-series inputs in comparable research.

The visits for a patient monitored over time steps T = (1, 2,..., t) constitute the multivariate time-series input V=v1,v2,...,vt in our design. Each medical feature for a visit is represented as a d-dimensional vector called a visit vi. Our architecture's initial layer, an embedding layer, is utilized to convert the input dimension space from a d-dimensional space to an n-dimensional space (n d). X=x1,x2,...xtRn is a multivariate time-series that is produced by the embedding layer. We employed n = 700 and d = 1737. We go into more detail later, but this embedding layer is also utilized to increase interpretability at the feature-level. We used LSTM cells in two recurrent layers following the embedding layer to train our model using sequential visit-level (dynamic) data. The multivariate time-series X is the input to the LSTM layers. With a hidden layer dimension of size 700 for each LSTM layer, there are 3,922,800 trainable parameters. Another multivariate time-series in the n-dimensional space H=h1,h2,...,htRtn is the output of the LSTM layers. Following the LSTM layers, the softmax layer produces attention scores [a11,a12,...,a1n;...;at1,at2,...,atn]Rtn, where aij = [0, 1]. The average of each vector's values (ai1, ai2,..., ain) is then determined, with i equal to [1, 2,..., t], resulting in ai .

Time level interpretability and feature level interpretability were used in our architecture.

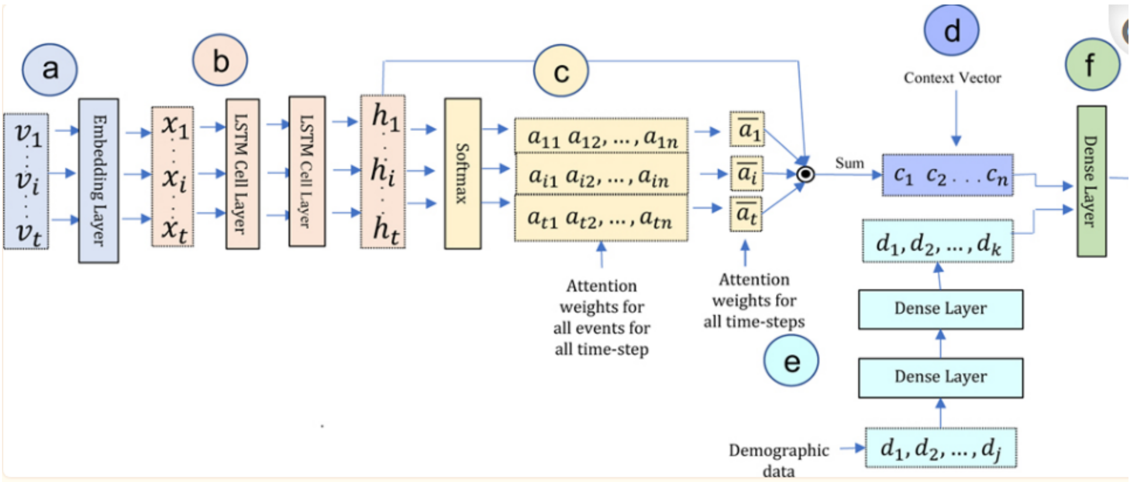


Fig 44. Overview of the deep learning model

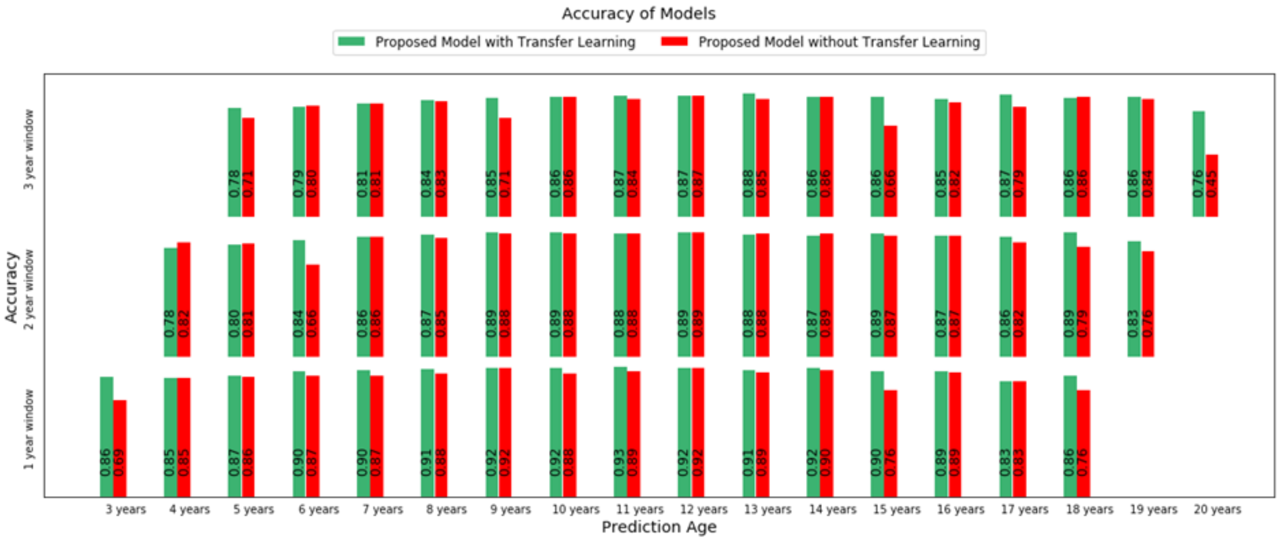


Fig 45. Accuracy of Models

CHAPTER 5: CONCLUSION

5.1 Conclusions

The conclusion that can be drawn from the project is-

- The most crucial elements affecting a person's obesity status are their height and weight.

- Other elements, including eating patterns and physical activity, may also be at play.

- It is possible to employ dataset features for classification and clustering tasks, but it is important to keep in mind that most samples are artificially generated, meaning they do not accurately represent the real world.

- Comparing the results of machine learning and deep learning models, deep learning models show more accurate results on the dataset available.

5.2 Future Scope

The future scope of the project is as follows-

- The accuracy of the model can be increased by increasing the number of attributes and including images for training.

- The comparison with other models such as the capsule network will give more accuracy and better results due to image data present in the dataset.

5.3 Applications Contributions

The primary effect of childhood obesity prediction is to assist parents and medical professionals in preventing and early detecting childhood obesity, thereby assisting the child in leading a healthy and happy life. This is because preventing childhood obesity is much simpler than treating adult illnesses because we all know that "Prevention is better than cure."

1. This project aims at predicting obesity among the young generation and thus is useful for hospitals to get an overview about the eating habits among children and can also be used by the parents to keep track of how healthy their children are.

2. This project will create an awareness amongst parents and schools regarding the diseases that come along with obesity and how it may affect their child’s growth.

3. Would serve as a huge help for parents to predict the disease early.

REFERENCES

[1] Pandita A, Sharma D, Pandita D, Pawar S, Tariq M, Kaul A, Childhood obesity: prevention is better than cure. Diabetes Metab Syndr Obes 9(83-89 (2016). https://www.ncbi.nlm.nih.gov/pubmed/27042133

[2] Qiong Wang, Min Yang, Bo Pang, Mei Xue, Yicheng Zhang, Xiangling Deng, Zhixin Zhang (2022) Machine/Deep Learning-based Approaches to Predict Overweight or Obesity in Chinese Preschool Aged Children

[3]Eom Gayeong, Byeon Haewon (2022). Development of Keyword Trend Prediction Models for Obesity Before and After the COVID-19 Pandemic Using RNN and LSTM: Analyzing the News Big Data of South Korea.

[4] Faria Ferdowsy, Kazi Samsul Alam Rahi, Md. Ismail Zabiullah, Md. Tarek Habib(2021) A Machine Learning Approach for Obesity Risk Prediction

[5] Gonzalo Colmenarejo (2020) Machine Learning Models to Predict Childhood and Adolescent Obesity

[6] Yang Cao, Scott Montgomery, Johan Ottoson, Eric Naslund, Erik Stenberg (2020) Deep Learning Neural Networks to Predict Serious Complications After Bariatric Surgery: Analysis of Scandinavian Obesity Surgery Registry Data

[7] Smego A, Woo JG, Klein J, Suh C, Bansal D, Bliss S, Daniels SR, Bolling C, Crimmins NA, High Body Mass Index in Infancy May Predict Severe Obesity in Early Childhood. J Pediatr 183(87-93.e81 (2017).

[8] Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults. Lancet 390(10113): 2627-2642 (2017).

[9] Adnan MHBM, Husain W, Damanhoori F (2010) A survey on utilization of data mining for childhood obesity prediction. In: 8th Asia-Pacific symposium on information and telecommunication technologies (Kuching, 2010). IEEE, pp 1–6

[10] Lobstein T, Jackson-Leach R, Moodie ML, Hall KD, Gortmaker SL, Swinburn BA, James WP, Wang Y, McPherson K, Child and adolescent obesity: part of a bigger picture. Lancet 385(9986): 2510-2520 (2015).

[11] Jia P, Xue H, Zhang J, Wang Y, Time Trend and Demographic and Geographic Disparities in Childhood Obesity Prevalence in China-Evidence from Twenty Years of Longitudinal Data. Int J Environ Res Public Health 14(4): (2017).

[12]World Health Organization .Obesity and Overweight. World Health Organization (WHO)(http://www.who.int/en/news-room/fact-sheets/detail/obesity-and-overweight.Accessed May 28,2018).

[13] Hassink SG, Early Child Care and Education: A Key Component of Obesity Prevention in Infancy. Pediatrics 140(6): (2017).

[14] Benjamin Neelon SE, Østbye T, Hales D, Vaughn A, Ward DS, Preventing childhood obesity in early care and education settings: lessons from two intervention studies. Child Care Health Dev 42(3): 351- 358 (2016).

[15] Yan J, Liu L, Zhu Y, Huang G, Wang PP, The association between breastfeeding and childhood obesity: a meta-analysis. BMC Public Health 14(1267 (2014).
https://www.ncbi.nlm.nih.gov/pubmed/25495402

[16] Padez C, Mourao I, Moreira P, Rosado V, Long sleep duration and childhood overweight/obesity and body fat. Am J Hum Biol 21(3): 371-376 (2009).
https://www.ncbi.nlm.nih.gov/pubmed/19189418

[17] Y. Manios et al., "Childhood Obesity Risk Evaluation based on perinatal factors and family sociodemographic characteristics: CORE index," European journal of pediatrics, vol.172, no. 4, pp. 551-555, 2013 2013.

[18] Y. Manios et al., "Utility and applicability of the "Childhood Obesity Risk Evaluation"(CORE)-index in predicting obesity in childhood and adolescence in Greece from early life: the "National Action Plan for Public Health"," European journal of pediatrics, vol. 175, no. 12, pp. 1989-1996, 2016.

[19] Z. Pei et al., "Early life risk factors of being overweight at 10 years of age: results of the German birth cohorts GINIplus and LISAplus," European journal of clinical nutrition, vol. 67, no. 8, p. 855, 2013 2013.

[20] L. Graversen et al., "Prediction of adolescent and adult adiposity outcomes from early life anthropometrics," Obesity, vol. 23, no. 1, pp. 162-169, 2015.

[21] Palechor, F. M., & de la Hoz Manotas, A. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. Data in Brief, 104344.

[22] Ilkka Rautiainen, Sami Äyrämö(2021) Predicting Overweight and Obesity in Later Life from Childhood Data: A Review of Predictive Modeling Approaches