BMI Prediction for 10-year-old:

Using Time Series And Demographic Features

Name: Shweta K Bhiwapurkar
Student number: 2058070
Number of words: 8622

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

Dr. Marijn van Wingerden
Dr. Peter Hendrix

Tilburg University
School of Humanities & Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
27-01-2022

Preface

Writing a thesis after more than a decade of academic gap was a real challenge. I am immensely thankful to so many people who helped me in this journey and had faith in me when I didn't have any. I would like to express my gratitude to my supervisor Dr. Marijn van Wingerden for all of the advice given during our weekly meetings. This thesis has been a challenge for me, and I could not have completed it without his guidance. I would also like to express my gratitude to my mentor at Esculine, Thijs de Bruijn for all his constructive feedback and support I have received during this study. Finally, I would like to thank my family for their encouragement and immense support, without them this thesis would have been impossible to complete.

Table of Contents

BMI Prediction for 10-Year-Old:
Using Time Series And Demographic Features

Shweta K Bhiwapurkar

Obesity is a condition wherein excessive fat accumulation in the body can impair health or can be a source of other diseases. Recent research shows that there is an alarming increase of obese children around the world. If childhood obesity is predicted at an early stage, then necessary steps can be taken to lower future disease risk. In the past, studies on this topic were mostly about classification of obesity levels and not BMI prediction in general. This study about predicting BMI values for 10-year-old Dutch children using longitudinal BMI and static demographic data with more emphasis on predicting accurately for overweight children. This is done by comparing four machine learning models to understand which performs best in predicting BMI values. ARIMA, XGBoost, LSTM and RNN models were built to predict the BMI values for children aged 10. The results show that all the models perform almost same as the baseline linear regression model with LSTM performing slightly better on predicting BMI for overweight children. The four models have MAE in the range 1.4-1.7 and R-square in the range 0.48-0.54. This study also implements two resampling techniques to obtain balanced datasets. The results also show that the models predict better with a balanced larger training set. Since all the models performed equally, future research should focus on an ensemble of models with a larger training set.

1. Legal & Ethical Considerations

This study was conducted in collaboration with EscuLine, a company specializing in business intelligence and health analytics solutions for health care institutions. Since healthcare data falls in the category of personal sensitive data, privacy is an important concern and needs to be protected under GDPR. Also, under GDPR informed consent by the data subject is a requirement in order to process the data. However, for this study, the data provided was completely anonymized hence GDPR and explicit consent does not apply (GDPR EU Recital 26). From the ethical perspective, it is essential that the models are not biased in their predictions. Poor representation of the data can cause the model to be biased towards the majority class. Hence resampling techniques are deployed to ensure that the training data is a balanced representation of the classes.

2. Introduction

According to the WHO, 13% of the adult population in the world were obese in 2016 and this number is rising across the world. Sedentary lifestyle, foods high in refined sugars, little to no physical activity are some of the important causes of obesity in adults. Obesity is a major risk factor for diseases like cardiovascular diseases, diabetes, cancers. Worldwide prevalence of obesity has increased in the last few decades. The statistics for childhood obesity are no different. It is estimated that currently 38.2 million children under the age of 5 are obese or overweight (WHO). In 2016, over 340 million children of age group 5-19 were obese or overweight. In the US, childhood obesity is a serious problem with more than 24% of children aged 2-4 classified as obese. In 1990 the Netherlands had 14% of obese children which has gradually increased to 20% by 2016 (Ritchie et al., 2017). This trend is more visible in the developing countries, especially in the mid to low-income population groups (Chung et al., 2016). Obesity is a chronic disease which begins early in the life. Once established it is quite difficult to root out, as sedentary lifestyle and unhealthy food habits are deeply rooted especially in adolescents. Hence it is important to prevent childhood obesity by implementing strategies which can effectively help in its prevention. In order to achieve this, predictive modeling can be helpful to identify high-risk children so that personalized and cost-effective strategy can be built. Obesity level is measured by Body Mass Index also known as BMI. BMI is defined as: “a simple index of weight-for-height that is commonly used to classify underweight, overweight and obesity in adults” (Ritchie et al., 2017). The higher the BMI, the higher is the risk of certain diseases like diabetes, heart disease, cancer, blood pressure. The risk of developing obesity as an adult is higher if the person is obese as a child (Singh et al., 2008).

In the Netherlands, children have regular visits to the pediatric doctor (Jeugdgezondheidszorg - JGZ) from birth till age 5. After age 5 there is no regular contact with the JGZ until the child is 10 years old. It therefore becomes more important to predict BMI in the initial life of children and understand which trajectory (obese or not obese) a child will follow so that preventive measures can be taken in time. With these preventive interventions in early childhood, future obesity related diseases and

obesity can be reduced which can further help in reducing healthcare costs and overall improvement in personal health. With the issue of shortage of trained medical personnel being at an all-time high, better predictions can also help to effectively plan the available resources.

There have been few models developed for prediction of obesity for children of different age groups using statistical methods (Morandi et al., 2012, Weng et al., 2013, Graversen et al., 2014). Recently machine learning models and neural networks have been explored for obesity predictions (Hammond et al., 2019, Pang et al., 2019, Zheng and Rugggiero, 2017). Although statistical models, machine learning and neural nets gave promising results, they were mostly used for obesity classification. Another type of prediction method is Time Series forecasting. This method is to forecast future events based on known past events. This method has been most commonly used to predict progress of disease, forecast healthcare cost and estimate mortality rate in healthcare domain. Although time-series prediction is very common in healthcare domain, it has rarely been applied on BMI prediction. As the data used for this study consists of longitudinal BMI data, four time-series models are built using this longitudinal data and static features and their performances are compared. The compared four models are ARIMA, XGBoost, RNN and LSTM. The model comparison aims to answer the following research question:

1) **Which time-series based machine learning model is best in predicting BMI values (of children aged 10 in Netherlands)?**

It is common practice to follow BMI definitions provided by WHO or CDC for adults. However, in case of children, due to high variability in the height and weight the standard WHO definitions do not apply. Instead, BMI cut-offs are determined by age and gender-based BMI distribution of the population. For example: In Netherlands, a 3-year-old male with BMI greater than 18 will fall in the overweight category, but a 3-year-old female with BMI greater than 17.5 will fall in the overweight category. As the BMI values differ between male and female gender for children, this research question is further sub-divided into following sub-question:

Sub-question 1.1: Does a gender specific time-series model outperform the general model?

In the real world, data imbalance is ubiquitous. Data often exhibits skewed distributions, where certain target values have very few observations. This skewness is not very obvious in a regression problem as compared to a classification problem. In order to mitigate this imbalance problem, two data resampling or balancing techniques are applied in this study which gives rise to the following research question.

2) **Does a balanced distribution of the target (continuous) feature help with improving the predictions?**

The main findings of this study show that although none of the models outperformed the baseline significantly, the models did perform better with a balanced and a larger training set. To obtain a balanced dataset, two resampling techniques are implemented – Random Under Sampling and SMOTEr. As expected, with a better representation of the samples the models performed better on the balanced sets, however all the models performed similar upon comparing the different evaluation metrics.

3.Related Work

Technology has now become an integral part of our lives. With digitization of hospital and clinic systems, huge volumes of data are now available for analysis. An EHR (Electronic Health Record) is a digital file providing patient's medical and treatment history that makes information available instantly which can be leveraged with machine learning (ML), deep learning (DL) and artificial intelligence (AI) to identify diseases and diagnosis, personalized care, better health record management, etc. Predictive modelling has also been quite popular in the healthcare sector. This section highlights the different applications of machine learning and artificial intelligence on previous work done on childhood obesity predictions

3.1 Predicting Childhood Obesity With Statistical Models

In the past studies, statistical models have been researched upon extensively to predict childhood obesity. Due to clinical interest in detecting conditions leading to pathological complications, most of the statistical models were for classification of obesity levels - overweight, obese (Colmenarejo, 2020). The most popular model is logistic regression, applied to predict binary results like being obese vs not obese. Most of the statistical models predicting childhood obesity had common predictor variables. The most popular ones were gender, parental BMI, parental education, birth weight, vital signs, smoking during pregnancy. A comprehensive comparison of different statistical models for predicting childhood obesity is done by Colmenarejo (2020). The major limitation in these studies was missing data at random, reducing the population size considerably. Also, in this era of Big Data, statistical models cannot handle high-dimensional data and hence they usually underperform when compared to machine learning models.

3.2 Predicting Childhood Obesity using Machine Learning models

With the popularity of machine learning models in healthcare, there has been quite a number of studies on using machine learning models for childhood obesity prediction. Recently, EHR data have been widely used in machine learning models as it has vast amount of information which can be used to learn or predict about diseases and or medical conditions (Rajkomar et al., 2018, Obermeyer et al., 2016).

There are some researches using machine learning techniques to predict childhood obesity or BMI of which two were most relevant to this study. The first one is the one done by Gupta et al (2019). This study used EHR data from Nemours Children Health System to predict BMI, which is further used to classify children as obese or non-obese. The LSTM model proposed by the study uses attention modelling to make the predictions interpretable and predicts obesity status at three different points in the future. The study compares performance of proposed LSTM model, random forest regressor and linear regression. The LSTM model performs much better as compared to the two baseline

models. For different prediction ages, the accuracy ranged from 0.75 to 0.92 for the LSTM model. The second one is by Singh et al (2020). This study evaluates seven ML algorithms using BMI values from ages 3, 5, 7 & 11 to classify a child as obese or overweight at age 14. Issue of data imbalance was handled by Synthetic Minority Oversampling Technique (SMOTE). Algorithms which performed poorly on imbalanced dataset performed much better with balanced dataset. Among the seven algorithms compared, Multi-layer Perceptron (MLP) performed the best with a balanced dataset with an accuracy of 0.96.

Another novel approach to predict BMI was to use facial images. This research done by Siddiqui et al (2020) was mainly concentrated on adult facial images. Five different deep learning models were evaluated, namely VGG19, ResNet50, DenseNet, MobileNet and lightCNN and their performances were compared. DenseNet and ResNet gave superior performances with mean absolute error (MAE) of 1.39 and 1.04 respectively as compared to other three models (VGG19-1.49, MobileNet-2.10, lightCNN-1.90).

3.3 Time-series Prediction In Healthcare

Healthcare or EHR data has abundant time series information. However, data is recorded at different time intervals for different individuals. This makes the data sparse. Also, sometimes it has missing data or wrong data entry due to human error. Hence a raw time series data needs quite a large amount of preprocessing to make it model ready for accurate predictions. Nevertheless, this method of prediction is quite popular in healthcare domain. A typical application of time series forecasting in the health domain is to predict healthcare cost (Morid et al., 2019). Some other applications of time series forecasting methods in healthcare domain are predicting patient count in emergency department (Choudhary et al., 2020), inpatient admissions (Zhou et al., 2018), expenditure on medications (Kaushik et al., 2020), healthcare waste generation (Chauhan et al., 2017), etc. A study by Monteiro et al., (2005) showed that longitudinal obesity-related measures (like weight, BP) have a high association

with the future obesity patterns. A highly accurate forecast can help to better plan and allocate economic as well as human resources.

3.4 Data Imbalance In Regression

Data imbalance is an unexplored area in a regression problem. Similar to a class imbalance in a classification setting, data imbalance in regression can cause an algorithm to overfit due to distribution bias. There have been few strategies proposed to mitigate this issue like random under and over-sampling, Synthetic Minority Oversampling Technique for Regression (SMOTEr) and Weighted Relevance-based Combination Strategy (WERCS). SMOTEr proposed by Torgo et al., (2013) is based on SMOTE algorithm. SMOTEr randomly under samples normal observations and over samples extreme or rare observations using a relevance threshold. Branco et al., (2019) proposed WERCS which is an informed combination of under and over sampling based on a relevance threshold. This user defined relevance is then used as weights for sampling. Studies implementing these data balancing strategies in regression are very rare.

As described in the sub-sections above, most of the models built using either statistical or machine learning methods were for classification of obesity and not predicting BMI values. As per the BMI chart followed in the Netherlands, a 10-year-old is considered obese if he/she has BMI greater than 19.86/19.84 respectively (GROEIDIAGRAMMEN IN PDF-FORMAAT). A child with predicted BMI of 25 might need a different care plan than one predicted with 20. As most of the existing studies are about classification, this study on predicting BMI values at age 10 can help the medical professionals to plan better intervention strategies based on the predictions. Predicting BMI values is a regression problem for which no prior suitable studies exist to be used as a baseline. Hence, a simple linear regression model is built to be used as a baseline. The data used in this study is highly imbalanced with majority of records for children with normal BMI, hence data balancing strategies are implemented to reduce distribution bias.

4.Methods

As mentioned in the introduction section, four different models are built to predict BMI values for children aged 10. These four models are selected because they are the most commonly used models in time-series prediction. This section will briefly explain about these four models. Furthermore, a brief explanation is given for feature selection, resampling strategy implemented for this study and hyperparameter optimization technique.

4.1 ARIMA

ARIMA stands for Autoregressive Integrated Moving Average. This model is very popular in forecasting future demand. The model is useful to understand past data or predict future data in a series.

An ARIMA model has three parameters: p, d and q. p is the order of the autoregression term. q is the order of the moving average term. d represents how many times the term has to be differenced in order to make the time series stationary. As the name suggests, ARIMA is an integration of Autoregression and Moving average models.

Hyndman et al., (2018) explains the autoregression and moving average as:

> "In an autoregression model, we forecast the variable of interest using a linear combination of past values of the variable. The term autoregression indicates that it is a regression of the variable against itself.
>
> Thus, an autoregressive model of order p can be written as
>
> $$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$
>
> A moving average model uses past forecast errors in a regression-like model.
>
> $$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$
>
> where ε_t is white noise."

An autocorrelation plot and partial auto-correlation plot can be used to determine the value for p and q respectively. The time series input to ARIMA must be stationary. A stationary time series can be obtained by subtracting the previous value from the current value (Prabhakaran S, 2021). For a stationary time-series, d = 0.

4.2 XGBoost

XGBoost stands for Xtreme Gradient Boosting and is an ensemble algorithm for classification and regression. It was initially developed by Tianqi Chen and described in their 2016 paper "*XGBoost: A Scalable Tree Boosting System*". The major advantages of XGBoost are execution speed, scalability and model performance (Chen et al., 2016). Real world data can be sparse or have missing data, XGBoost is designed to handle sparse matrix or missing data. In the boosting ensemble approach, the trees are built sequentially. Each subsequent tree tries to reduce the errors of the previous tree. Hence, all the trees learn from its predecessors and the next tree in sequence will learn from previous errors. Although XGBoost is more popular for regression and classification tasks, it can be used for time-series forecasting as well.

Every ML algorithm needs to be optimized to reduce error. XGBoost is trained by minimizing loss of an objective function against a dataset (Brownlee, 2021). A loss function must be selected during the optimization process to calculate the error of the model. XGBoost has various built-in loss functions for classification and regression tasks. The loss functions can also be customized to improve model performance. The goal of this study is to predict better on the extreme BMI values. The typical regression evaluation measures like mean squared error are sensitive to outliers. Hence a custom loss function is built to predict better for the extreme BMI values.

4.3 Artificial Neural Network

Neurons are fundamental units of brain which carry electrical impulses and an Artificial Neural Network (ANN) simulates this network of neurons. The network of neurons called as 'units' in ANN terminology are interconnected with each other. A typical neural network has at least 3 layers or units,

input, hidden and output. The input layer receives the input which then goes through one or more hidden layers. The output of hidden layer is determined by the weights of the connections between input and hidden layer. This output is then fed into next hidden layer or an output layer which is dependent on weights between hidden and output layer (Hinton, 1992). The network learns more about the data as it goes through each unit.

Two different forms of ANN will be implemented in this study, namely Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM). An input to any neural network is a vector of numbers. The data for this study consists of both numeric and categorical features. In order to use these categorical features as an input to the neural network, these features are transformed into a smaller denser space using user defined embedding space.

4.3.1 RNN

RNN is one of the basic forms of neural networks. Because of their internal memory feature RNN are preferred for sequential data like financial, text, time-series. A typical neural network processes the input and provides an output. However, in RNN each input is dependent on the previous one for making a decision and uses a method called backpropagation to loop information back in the network. Due to this dependency RNN's are not suitable for long sequences. This dependency is also called as vanishing gradient problem which slows the training process for RNN. For this study, the input to the RNN network was a concatenation of embedding layer and the numeric features. The number of dense layers and the size of units was determined using Bayesian Optimization.

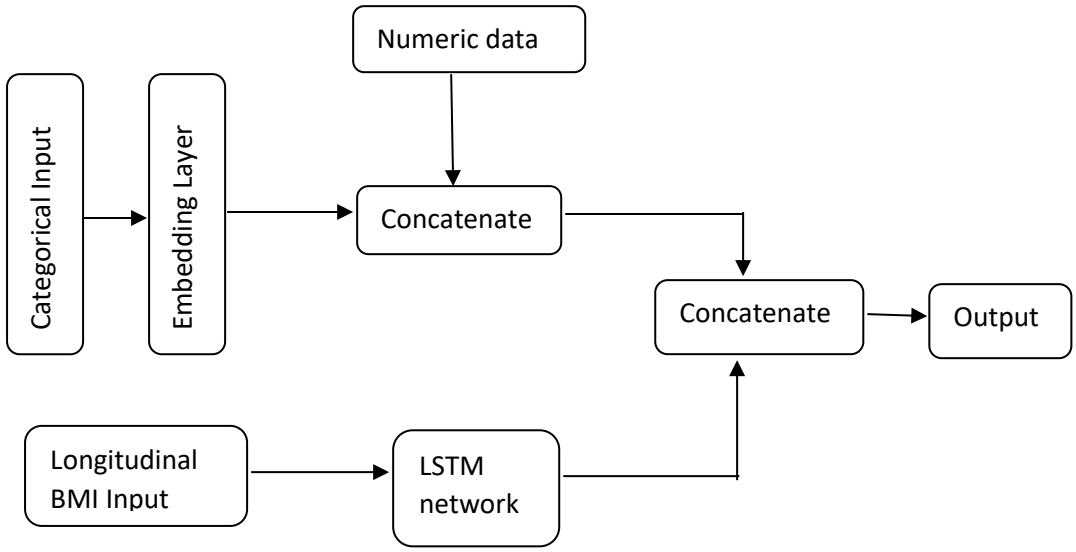
4.3.2 LSTM

As the name suggests LSTM can store information in memory for a longer time. LSTM solves the vanishing gradient problem of RNN and hence performs better. For this study, the LSTM network is built as shown in Figure 1. This was inspired by the architecture implemented in the study by Gupta et al., 2019. Categorical input layer is transformed using embeddings and concatenated with numeric input data. The LSTM model is trained with longitudinal BMI data. The output of LSTM layer is

concatenated with embedding and numeric input and passed thru a dense layer to predict the final BMI value.

Figure 1

Overview of the LSTM model



4.4 Feature Selection

The main goal of any feature selection method is to select only those features which contribute to better predict target variable. For regression, a very popular feature selection technique is correlation – which is a measure of how two variables change together. A good set of features are highly correlated with target and uncorrelated with each other. The SelectKBest class from scikit-learn library implements correlation in f_regression function, which is used in this study to select the best features which contribute in predicting the target feature.

4.5 Resampling

Imbalanced data is a term usually referred in terms of classification problem. It is when distribution of classes is not uniform within a dataset. This imbalance causes the model to be biased towards the majority class. Research shows that with a balanced dataset there is an increase in predictive accuracy,

especially for rare observations (Moniz et al., 2016). There are several ways to handle this imbalance depending on the problem at hand. In this study, two different resampling techniques are applied and models are tested on each resampled dataset as well as on the original unbalanced dataset.

4.5.1 Random Under Sampling

In this technique, the majority class is reduced to the size same as the minority class. The majority class instances are randomly selected till they are of the same size as minority class. The major drawback of this method is that useful data from the majority class is lost and the size of dataset gets reduced based on the size of minority class. Although this method is more popular with classification problems, it can be applied to regression by stratifying the continuous target feature.

4.5.2 SMOTE For Regression

Another resampling technique is SMOTE, which is an abbreviation for Synthetic Minority Over-sampling Technique. However, this technique is intended for classification problems. For regression we have SMOTEr. SMOTEr was designed specially to solve imbalanced dataset problem for regression by Torgo et al., 2013. It has the same underlying principle as SMOTE - combination of under-sampling the majority class and over-sampling the minority class. In regression, values commonly found near the mean of a normal distribution in the response variable y are in majority and the rare or minority values are typically found in the tails. SMOTEr uses a notion of relevance of target variable where the target observations are placed into majority (normal class) and are under sampled or minority (rare or interesting class) and are over sampled, based on the user specified relevance threshold. The python implementation of SMOTEr is through SMOGN library (Branco et al., 2017).

4.6 Bayesian Hyperparameter Optimization

Hyperparameter (HP) are adjustable parameters chosen by user to train a model. Performance of any machine learning model depends on the values of HP. Hence it is essential to find or tune the best HP for optimal model performance. For HP tuning, grid search and random search techniques are the most popular methods. However, both have their pros and cons. Grid search searches in all possible

user defined grid space for the optimal hyperparameters. Since it is an exhaustive search, it takes a lot of computing power and time. With random search, it randomly searches the HP space and is faster than grid search but is bound to miss better performing parameters. Another technique for HP tunning is Bayesian Optimization. It is considered to be more efficient than grid and random search for HP tuning (Bergstra et al., 2015, Putatunda et al., 2018). Bayesian optimization approach uses the past results to build a probabilistic model of the objective function. There are three components when performing Bayesian optimization. First is the HP search space. Second is to build a model to approximate the true objective function. Third is to build an acquisition or surrogate function which helps to determine the next hyperparameter to evaluate. The objective function takes in a set of hyperparameters from the user provided search space and provides a score that indicates how well a set of hyperparameters performs on the validation set. Depending on the requirement, the objective function can be minimized or maximized. The approximation of the objective function is a surrogate function which maps HPs on a probability score and suggests the best one based on this score for further evaluation. This evaluation of the HPs takes lesser time as compared to grid or random search and hence Bayesian methods can find better HPs in less time (Koehrsen Will, 2018).

5. Experimental Setup

5.1 Dataset Description

For this study the data is provided to EscuLine by their client JGZ (Jeugdgezondheidszorg). The English translation of Jeugdgezondheidszorg (JGZ) is youth health care. The JGZ provides preventive health checkups and vaccinations for children. An Electronic Health Record (EHR) collected by them usually consist of weight and height measurements, vaccination and illness record, etc. The dataset for this study consists of clients living in the western part of the Netherlands and the data dates back from current to 2011.

Figure 2

Dataset before and after preprocessing

Client id	Features	Age	BMI
1		3 months	
1		5 months	
1		9 months	
2		5 months	
2		9 months	
2		12 months	
3		2 years	

(a) Raw data

Client id	Features	3 months	5 months	9 months	12 months	2 years
1		BMI value	BMI value	BMI value		
2			BMI value	BMI value	BMI value	
3						BMI value

(b) Transformed data

The raw dataset consists of around 1 million records. For each unique client the raw dataset had records for the measurements done at different time steps (age of client). Hence the dataset had to be reorganized to show measurements done at different time steps for a unique client i.e., transform data into a format which a machine learning model can understand. A schematic representation of data is shown in Figure 2(a).

5.2 Dataset Preprocessing

An EHR consists of many features and not all features are consistently recorded for every client. As seen in Figure 2(b), the dataset after transformation has missing BMI values. These missing values can be attributed to various reasons like, missed appointment, relocation, illness, etc. which makes EHR data sparse. The actual counts of missing data for different age groups in the raw dataset is given in Table 1.

The JGZ system was digitized in 2011. Hence a client born in 2011 and is in the JGZ system will have a higher chance that the record will not have missing values till age 5. But will have no information for age 10 as the child has not reached this age yet. Age 10 being the target feature, this

record cannot be used for training. On the other hand, clients joining the system at different age groups will have missing data for the initial years of life**.** Since the target is BMI prediction at age 10, it is crucial to have non-missing BMI values measured close to target time step. In order to have enough records for model training and testing, only records with all measurements for last 2 timesteps along with target variable are considered i.e., age 3.9, age 5. As seen from the Table 1, most of the missing data is from the early age groups (3 weeks – 18 months). Due to the limitation of available data, these age groups will not be considered further. The missing values for other features are imputed using NaNImputer, which fills any missing values in every column of the given dataset using XGBoost based model. Categorical features like gender, birth country of parents (categorized into Netherlands, Western, non-Western), region (current region of stay in Netherlands) is dummy-ed for ARIMA and XGBoost models. Dummying the categorical features increases the dimensionality of the dataset. For neural nets, in order to reduce this dimensionality of the categorical features an embedding layer is implemented in the network.

Along with this main raw dataset, there were other files provided by EscuLine consisting of demographic (gender, parents birth country, parents age, current region of stay, etc.) and contact moment details. Contact moments is the number of times the child visited the JGZ segmented into four different time windows. All these files were concatenated based on the unique identifier.

Table 1

Missing BMI values in each age group

Age	Missing Values	Non-Null Values
3-4 weeks	9905	28
7-8 weeks	9892	41
3 months	9860	73
5 months	9814	119
6 months	9688	245
8 months	9933	0
11 months	9101	832
14-15 months	8703	1230
18 months	8022	1911

2 year	6616	3317
3 year	0	6691
3.9 year	0	9933
5 year	0	9933
10 year	0	9933

5.3 Feature Selection

The features used in the past on obesity or BMI prediction varied significantly. Some of the common features in the past studies are gender, birth weight, mother's weight, mother's age, race or ethnicity (Morandi et al., 2012, Redsell et al., 2016). Initially the features were handpicked based on the literature review and as suggested by the JGZ contact. In order to determine which features, apart from past BMI values help to predict the BMI at the target age of 10, SelectKBest with f_regression function is implemented on the final dataset (9070 records, 17 features). This method calculates the correlation between each feature and target which is then converted to F-score for each feature. The higher the score, the more important the feature is. The feature selection is implemented after dummy-ing the categorical features.

5.4 Final dataset

After preprocessing and concatenation of the files, the final dataset has 9070 records and 17 features. Each record is for a unique client. As it is common in medical domain, this final dataset is highly unbalanced. It has more records for clients with normal BMI than for those with overweight. In an unbalanced dataset, a model is more likely to be biased towards majority class, in this case normal BMI. Hence the prediction for extreme values will be in the range of normal values. In order to mitigate this issue, dataset resampling techniques are applied to create balanced datasets and models are then fit to the unbalanced and balanced datasets. The performance is compared accordingly.

5.5 Exploratory Data Analysis

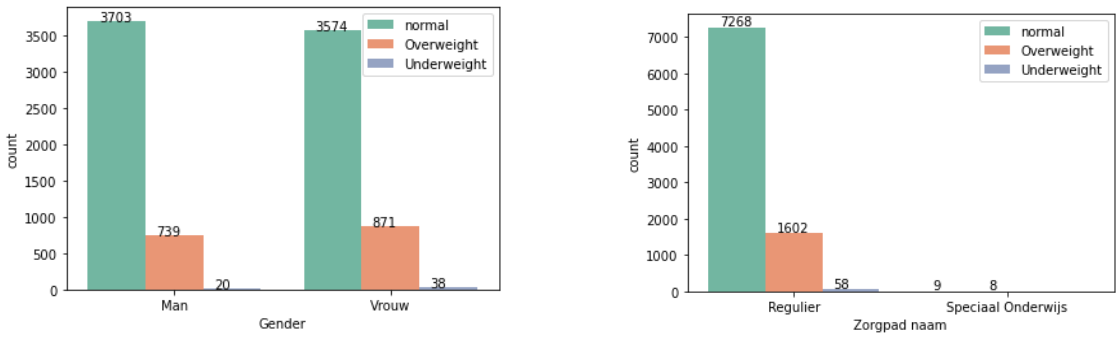
Exploratory data analysis (EDA) is performed to understand the data better. It can help to gather many insights and understand patterns within the data. EDA is performed on the unbalanced dataset obtained after preprocessing.

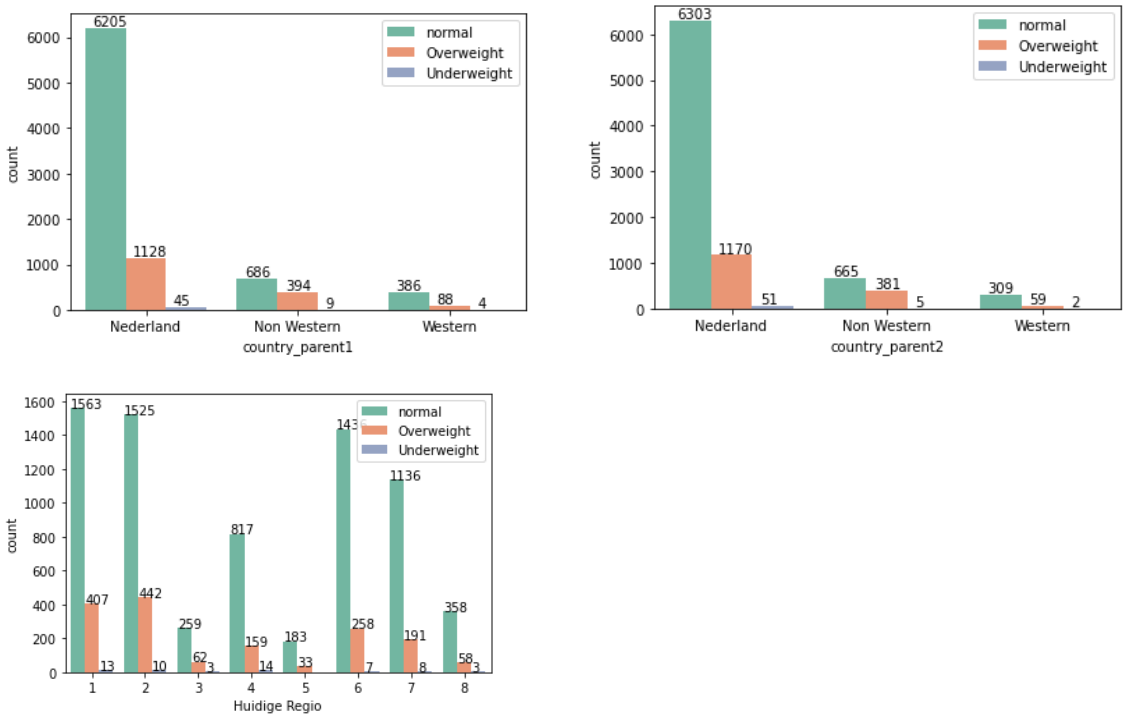
5.5.1 Categorical Predictors

The dataset has five categorical predictors. The distribution of these categorical features grouped by BMI category is shown in Figure 3. From the count plot of gender, it can be seen that there is not much difference in the counts of each BMI category for male gender vs female gender. The *Zorgpad naam* (caretype) feature shows that the data has high majority of children receiving regular care services. For majority of the children, Netherlands is the birthplace of the parents. Hence it is not a good mix of different ethnicities. Data rich with different ethnicities can be a good predictor as studies have shown that it does have an impact on BMI (Davis et al., 2013, Heymsfield et al., 2016). The regions are anonymized for client privacy purpose. The region-wise plot shows the count of children in each category. The region 8 has least number of overweight children. This can be attributed to the fact that it is an economically well-off region and hence lesser number of underprivileged families. Regions 1 and 2 have relatively higher number of overweight children. Socio-economic reasons and the fact that these two regions are geographically adjacent might be one of the contributory reasons for this higher number (The Netherlands in numbers).

Figure 3

Plots of categorical features grouped by BMI category





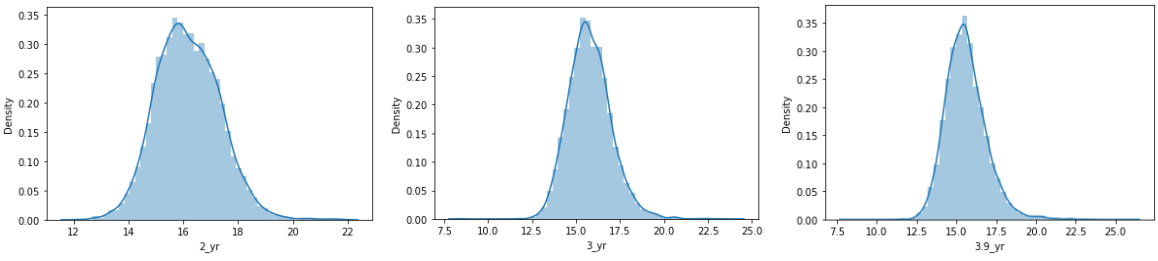
5.5.2 Continuous Predictors

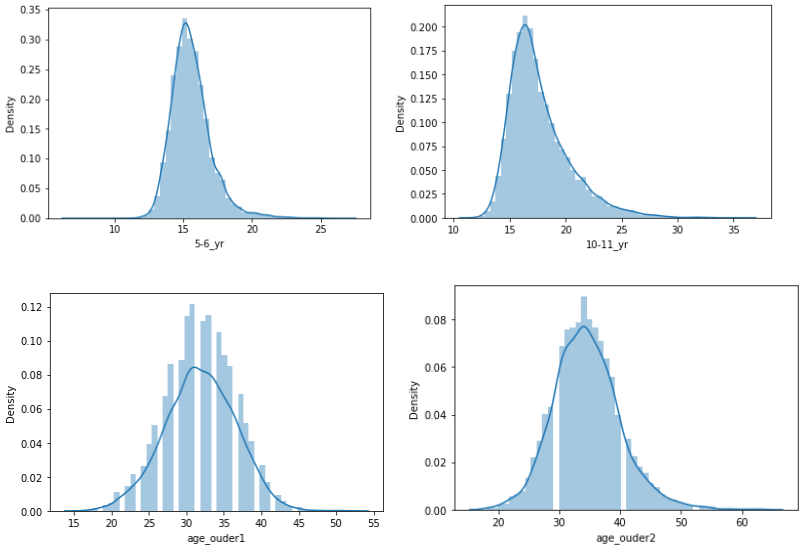
To understand the distribution of continuous features, distribution plots have been constructed. As seen in Figure 4(a), the BMI values at different ages appear to be normally distributed except for the target feature which is slightly positively skewed. Also, the age of parents appears to be normally distributed. The age of the parents is at the time of the birth of the child, not the current age. To understand the distribution of age better, boxplot is constructed. The boxplot in Figure 4(b) shows that the median age for parents is around 32-35 and consists of outliers.

Figure 4

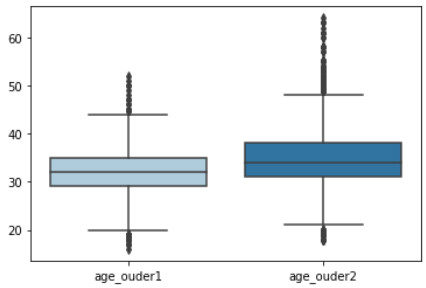
Plots for continuous features: (a) Distribution plot of continuous features; (b) Boxplot for age of parents

(a)





(b)



5.6 Experiments

As the dataset is resampled, it is essential to test the performance of the proposed models on hold-out set. The dataset after preprocessing is split into 80-20 ratio for train and hold-out sets. This 80% train set is resampled and further split into train and test sets for training of the four different models. All the experiments are implemented using Python language. List of packages used in the experiments can be found in Appendix A.

5.6.1 Experiment 1: Feature Selection

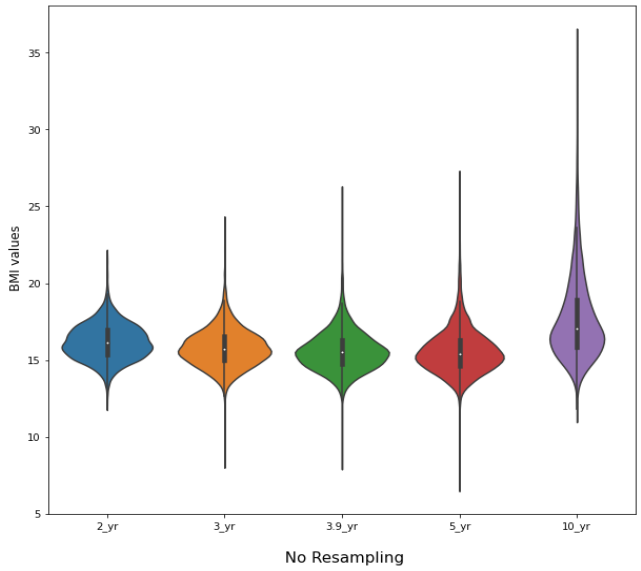
SelectKBest feature selection technique was implemented to select the best contributory features apart from longitudinal BMI features. As the dataset for this study does not consist of very high number of features, model performance with all the features included is considered i.e., all the models are trained and tested on the same features.

5.6.2 Experiment 2: Applying Resampling Technique To Balance The Dataset

As mentioned in the earlier section, the dataset after preprocessing is highly unbalanced with only 17% of records with BMI greater than 20. The Figure 5 violin plot shows distribution of BMI values in each age-group, especially in the target column. As seen from the width of the violin plot for the target feature, 'normal' BMI values occurs more frequently than 'extreme' values. The median value of the target feature is 17.03. In order to balance the dataset two resampling techniques are applied, random under sampling (RU) and SMOTE for regression (SMOTEr). As RU works only with classification, the target feature is classified based on BMI cut-off values for each gender. This added a new feature - 'category' to the dataset. RU was then applied on this newly added feature. This column was used for resampling and exploratory data analysis only and was removed later. SMOTER applies combination of under sampling majority class and over sampling minority class and returns under sampled and synthetic over sampled observations (Torgo et al., 2013). By using two resampling techniques, two new datasets (RU and SMOTEr) were created on which the models were trained.

Figure 5

Violin plot of unbalanced dataset



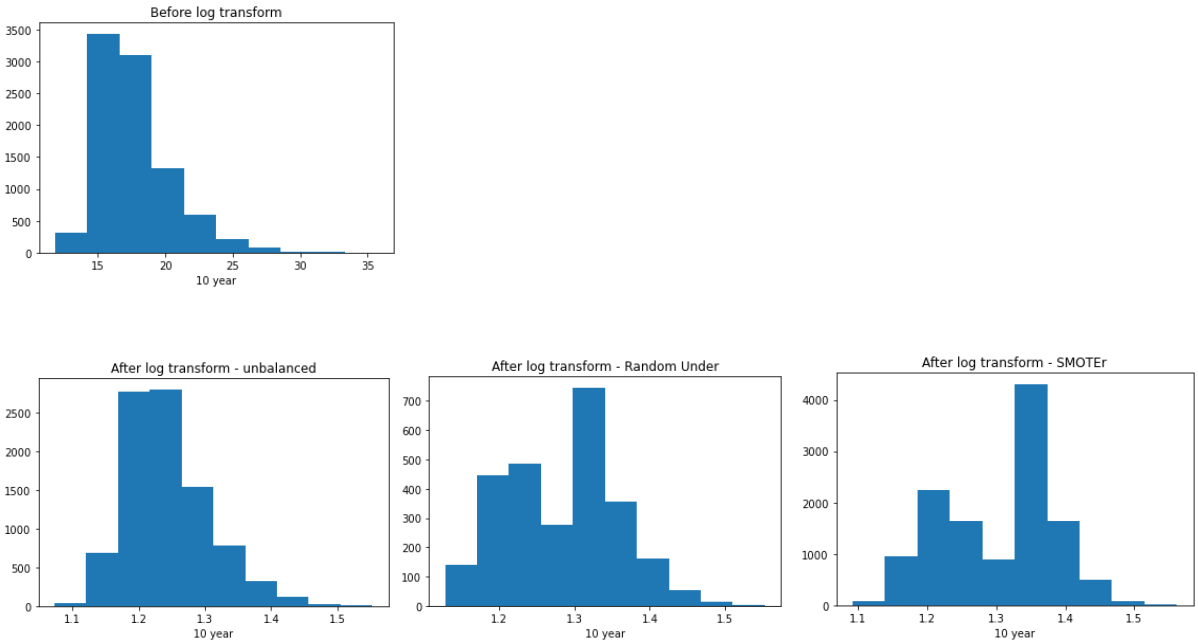
5.6.3 Experiment 3: Model Comparison For BMI Prediction At Age 10

As seen in the literature review, both statistical models as well as machine learning models have been implemented for predicting obesity in children. As prediction of BMI values is a regression problem, a Linear Regression (LR) model as a baseline is implemented. The first model, ARIMA model is built using the pmdarima package in python. The function, auto.arima selects the optimal parameters, without the need for the user to specify it explicitly. Based on the user defined range for p, d and q it searches on its multiple combinations stepwise and chooses the model with lowest AIC (Akaike Information Criterion). The second model, XGBoost is quite popular in medical domain because of its ability to handle missing data as electronic health records have missing information due to various reasons. Past research shows the many advantages of using neural networks as predictive algorithms due to their capability to handle the non-linear relationships in longitudinal data (Gamboa, 2017).

For optimal performance of XGBoost, RNN and LSTM models, hyperparameter tuning is performed for each balanced and unbalanced dataset using Bayesian optimization. The first step in Bayesian optimization is to provide a range of values for the parameters to be optimized. This optimization technique creates models with different hyperparameter values but it uses information from previous model to select the hyperparameter value for the new model. The parameters which are optimized for this study is listed in Appendix E. To evaluate model generalizability and prevent overfitting, a 5-fold cross validation is implemented with the best hyperparameters from Bayesian optimization on the training set and evaluated on the test set. This was done for all three datasets, namely – unbalanced, RU and SMOTEr. Numeric features present in each data set were log transformed. Real world data is not always normally distributed. Log transformation is one of the methods to change a skewed distribution to normal. Before and after log transformation for the target feature is shown in Figure 6. As the dataset consist of both categorical and numeric features, embedding layers were implemented for the categorical features for the neural network models. The study consists of four time-steps and predicting on the fifth step, hence cross validation cannot be applied for ARIMA as the range of time is too small (Hyndman, R 2010).

Figure 6

Before and After Log transformation of target feature



To improve the predictions, custom loss functions are implemented. Customizing loss function in XGBoost model requires defining a function that takes actuals and their predictions and return two arrays of the gradient and hessian of each observation. The gradient and hessian are the first and second order derivative of the objective function – in this case squared loss. The loss function was customized such that it gives 5 times more penalty when the predictions are less than actual values. Similarly, the custom loss function for RNN and LSTM models were implemented such that they would penalize the error by a factor of 5 if residual is high. Performance of the models with custom loss function and with standard loss (MAE) is compared.

To sum up, using the three datasets and best hyperparameters, models were trained and the final evaluation and comparison was on the hold-out set. To check if there is a difference in gender specific models and general models, the three datasets are split gender wise and new models are trained using these gender specific datasets.

5.7 Evaluation

In order to evaluate performance of the models, different evaluation metrics are used. Generally, for regression, mean absolute error (MAE), mean squared error (MSE), R^2 (r-squared) and mean absolute percentage error (MAPE) are important metrics to understand performance of a model. MAE is the mean of the absolute value of the errors and is quite robust to outliers. MAE is unambiguous and better suited for model comparisons (Willmott et al., 2005). MSE is the average of the square of the prediction error. Lower value of MSE means the predicted values is close to actual. R^2 shows how much of the variance in the dependent variable can explain the variance in the independent variable. Higher the R-squared, the more variation is explained by the dependent variables and better is the model. MAPE is a very popular metric to compare how well the time series model fit. Smaller values indicate better fit of the model.

Since the dataset is resampled, the final evaluation of the models is on the hold-out set.

6.Results

In order to answer the sub-research question and research question, experiments were setup and conducted. The experiments were conducted after applying the pre-processing steps (as explained in section 5.2) on the original dataset.

6.1 Experiment 1: Feature Selection

Longitudinal BMI values are by default important contributory features; hence they were not added in this feature selection step. Using SelectKBest feature algorithm, a list of other best features was obtained. The graph showing importance of features other than BMI values is in Appendix B. The description of all the features is in Appendix C. There are two important observations which can be made from the graph. First, the feature 'CM_4-12years' which the algorithm deems as an important one denotes the number of times a child has a 'contact moment - CM' or number of visits to the pediatric doctor in the age group 4-12. Since the target prediction age is 10, this feature adds value. However, selection of this feature is debatable which is discussed in the next section. Second, the

features ‘parent1_Nederland’ and ‘parent2_Nederland’ emerge as another set of important features. As shown in the EDA section above, the dataset does consist of more than 80% of parents born in the Netherlands.

As mentioned before, all features were included irrespective of their magnitude of contribution to the target feature in all the models.

6.2 Experiment 2: Applying Resampling Technique To Balance The Dataset

The first step in this experiment was the creation of balanced datasets by applying resampling techniques as described in section 3.5. Since random under sampling (RU) balances the minority class, the size of the dataset is reduced considerably. Before applying RU, the dataset had normal, overweight and underweight categories. The size of these classes can be represented in increasing size as underweight < overweight <<< normal. Since the goal of performing resampling is to have more(balanced) overweight class, records for underweight class were discarded. Hence the final RU dataset has same number of normal and overweight records. As SMOTEr generates synthetic data,

Figure 7

Violin plot of balanced datasets

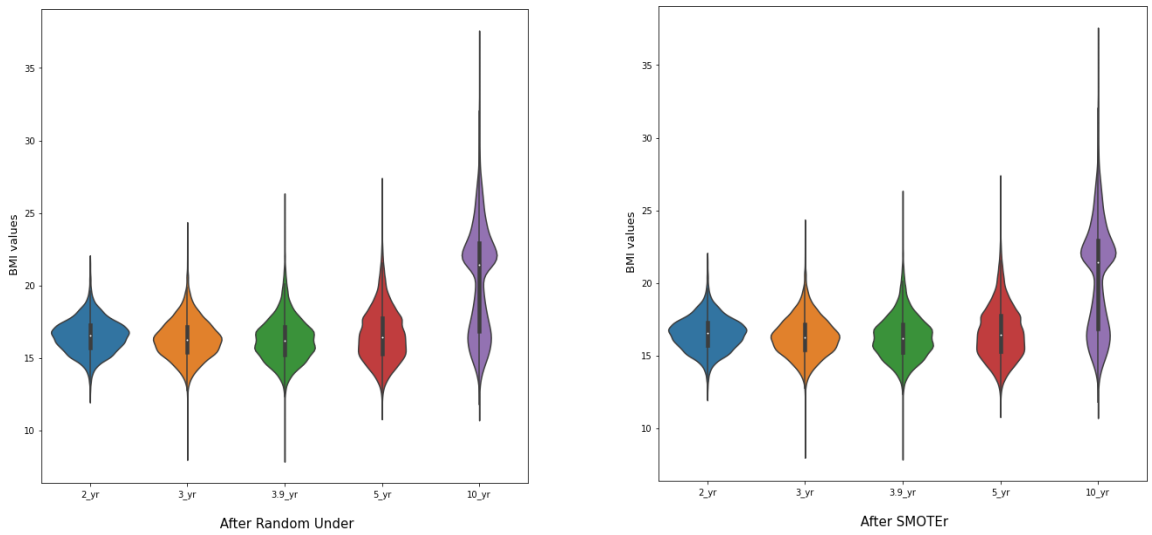


Table 2

Descriptive Statistics on balanced datasets

	BMI	mean	std	min	max
Normal Count=1342	2_yr	16.03	1.03	12.70	20.72
	3_yr	15.58	1.02	11.72	19.38
	3.9_yr	15.35	1.00	12.34	18.69
	5_yr	15.22	1.04	11.96	18.59
	10_yr	16.65	1.43	13.43	19.83
Overweight Count=1342	2_yr	16.88	1.25	13.34	21.80
	3_yr	16.78	1.39	12.42	23.95
	3.9_yr	16.82	1.55	11.66	25.89
	5_yr	17.30	1.79	6.92	26.85
	10_yr	22.27	2.28	19.84	35.7

(a) Random Under

	BMI	mean	std	min	max
Normal Count=5282	2_yr	16.03	1.07	12.11	21.72
	3_yr	15.58	1.06	8.37	20.44
	3.9_yr	15.35	1.04	8.29	19.42
	5_yr	15.23	1.05	11.75	19.99
	10_yr	16.71	1.47	13.43	19.85
Overweight Count=7028	2_yr	16.97	1.10	13.34	21.80
	3_yr	16.96	1.28	12.42	23.95
	3.9_yr	17.1	1.44	11.66	25.89
	5_yr	17.71	1.65	6.92	26.85
	10_yr	23.25	2.03	19.84	36.45

(b) SMOTEr

the size of the dataset increases and the minority class is better represented in the target feature. Figure 7 shows violin plots of dataset after applying resampling techniques with a median value of 19.8 and 21.5 for random under and SMOTEr resampling respectively. The target feature is now a balanced representation of the BMI values as compared to unbalanced feature shown in Figure 5. The size of datasets after applying resampling techniques is in Appendix D. Table 2 shows descriptive statistics for the BMI features split by BMI category after resampling. After resampling, the distribution of the target feature for both the resampled datasets are similar in each BMI category.

6.3 Experiment 3: Model Comparison For BMI Prediction At Age 10

The first step in this experiment is selecting hyperparameters using Bayesian Optimization (BO) on the three datasets which will give optimum performance for the XGBoost, RNN and LSTM models. The

final values for the hyperparameters for each model are listed in Appendix E. For the ARIMA model, function auto.arima was implemented which selects the best parameters automatically. For all the three datasets, auto.arima selected 0 for parameters p, d and q. Using these parameters from auto.arima three separate models were built for the three datasets. The final evaluation was on the hold-out set. Similarly, for XGBoost, RNN and LSTM models, using best hyperparameters from BO, three models were built for each dataset. Five-fold cross validation was implemented on the training set and the final performance evaluation was on the hold-out set. Table 3 shows the cross-validation results for XGBoost, RNN and LSTM models. Mean absolute error (MAE) was obtained for each fold and which was then averaged to get the mean error. Each model when trained with SMOTEr dataset has lower error as compared to when trained with other datasets.

Table 3

Cross-Validation results for general model-with custom loss

	Dataset	Mean MAE
XGBoost	Unbalanced	1.50
	Random Under sampled	1.91
	SMOTEr	1.41
RNN	Unbalanced	1.71
	Random Under sampled	1.83
	SMOTEr	1.57
LSTM	Unbalanced	1.28
	Random Under sampled	1.73
	SMOTEr	1.55

The performance of all the models on the hold-out set with custom and standard loss is shown in Table 4. ARIMA does not have a loss function, hence same results are repeated. The performance of models trained with custom loss function is slightly better than those trained on standard loss. However, the performance difference is not very significant. In-depth analysis of the model performance with custom loss showed that none of the models could outperform the baseline. Rather the baseline and the four models gave almost similar performance results on all the evaluation metrics. On comparing the MAE metric, surprisingly all the models performed better with an

Table 4

(A) Performance on hold-out set for the general model- with custom loss

	Dataset	MAPE	MAE	MSE	R^2
Baseline - LR	Unbalanced	0.072	1.321	3.176	0.567
	Random Under sampled	0.087	1.528	3.611	0.508
	SMOTEr	0.090	1.580	3.923	0.466
ARIMA	Unbalanced	0.073	1.337	3.204	0.563
	Random Under sampled	0.091	1.589	3.826	0.479
	SMOTEr	0.096	1.677	4.257	0.420
XGBoost	Unbalanced	0.070	1.308	3.202	0.564
	Random Under sampled	0.087	1.542	3.814	0.480
	SMOTEr	0.085	1.526	3.948	0.462
RNN	Unbalanced	0.071	1.328	3.311	0.549
	Random Under sampled	0.075	1.369	3.441	0.531
	SMOTEr	0.098	1.731	5.166	0.296
LSTM	Unbalanced	0.070	1.301	3.235	0.559
	Random Under sampled	0.073	1.339	3.192	0.565
	SMOTEr	0.081	1.464	3.745	0.490

(B) Performance on hold-out set for the general model- with standard loss

	Dataset	MAPE	MAE	MSE	R^2
Baseline - LR	Unbalanced	0.072	1.321	3.176	0.567
	Random Under sampled	0.087	1.528	3.611	0.508
	SMOTEr	0.090	1.580	3.923	0.466
ARIMA	Unbalanced	0.073	1.337	3.204	0.563
	Random Under sampled	0.091	1.589	3.826	0.479
	SMOTEr	0.096	1.677	4.257	0.420
XGBoost	Unbalanced	0.072	1.307	3.089	0.549
	Random Under sampled	0.089	1.540	3.655	0.466
	SMOTEr	0.089	1.562	4.078	0.405
RNN	Unbalanced	0.072	1.368	5.975	0.128
	Random Under sampled	0.085	1.499	3.907	0.430
	SMOTEr	0.080	1.467	4.181	0.390
LSTM	Unbalanced	0.072	1.288	2.880	0.579
	Random Under sampled	0.074	1.309	2.966	0.567
	SMOTEr	0.079	1.410	3.517	0.487

unbalanced dataset as compared to balanced sets. A similar interpretation can be drawn for the R-squared metric. As observed from Table 4(A), the R-squared values are higher for unbalanced set, which means that the model fits the unbalanced data better. However, since the main goal of the study is to predict better for extreme BMI values it is essential to understand how the model

performed on extreme values. An extreme BMI value for a 10-year-old is defined as BMI greater than 19.86 for males and BMI greater than 19.84 for females (GROEIDIAGRAMMEN IN PDF-FORMAAT). Using this definition, the actual and predicted values were classified into underweight, normal and overweight categories which was used to further evaluate model performance with the means of confusion matrix. Table 5 shows mean absolute error (MAE) and mean absolute percentage error (MAPE) metric for the extreme BMI values for models trained with custom and standard loss. It can be observed from Table 5 that with a custom loss there is an improvement in prediction of extreme BMI values, but it is not a significant improvement which is a surprising outcome. Both the error metrics have a reduced value as the size of dataset increases. This answers the second research question that a balanced dataset does improve the predictions. Among the two resampling techniques, lower error is achieved with SMOTEr dataset. This is also confirmed in the confusion matrix of all models shown in Appendix F.

To answer the sub research question of whether gender specific models predict better than the general model, the three datasets are split gender wise and same procedure as general model is

Table 5

MAE and MAPE for Extreme BMI values - with custom loss

Dataset	**ARIMA**		**XGBoost**		**RNN**		**LSTM**	
	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE
Unbalanced	2.46	0.10	2.64	0.11	2.77	0.123	2.65	0.11
Random Under sampled	2.00	0.08	1.85	0.09	2.49	0.11	2.35	0.10
SMOTEr	1.92	0.08	1.88	0.08	2.02	0.09	2.00	0.08

MAE and MAPE for Extreme BMI values – with standard loss

Dataset	**ARIMA**		**XGBoost**		**RNN**		**LSTM**	
	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE
Unbalanced	2.46	0.10	2.63	0.11	3.49	0.15	3.15	0.12
Random Under sampled	2.00	0.08	2.03	0.08	2.25	0.10	2.50	0.10
SMOTEr	1.92	0.08	2.00	0.08	2.52	0.11	2.20	0.09

followed to train new gender specific models. The hyperparameters for gender specific models were same as the general model. The 5-fold cross-validation results for gender specific models is shown in Table 6. The performance of gender specific models with custom and standard loss on hold-out set is shown in Table 7. Similar to the general model, there is an improvement when models are trained with custom loss function. A closer inspection of Table 7(A) shows that performance wise there is no substantial difference between gender specific models and the general model. The R-squared for the 'female' gender shows that ARIMA and LSTM fit much better as compared to other two models. Similar to the result from the general model, the MAE is lowest for the unbalanced set for all the models for both genders.

Table 6

Cross-Validation results for Gender specific models

	Dataset	**Male**	**Female**
		Mean MAE	Mean MAE
XGBoost	Unbalanced	1.50	1.51
	Random Under sampled	1.92	1.91
	SMOTEr	1.45	1.35
RNN	Unbalanced	1.53	1.35
	Random Under sampled	1.85	1.78
	SMOTEr	1.58	1.57
LSTM	Unbalanced	1.31	1.38
	Random Under sampled	1.77	1.76
	SMOTEr	1.67	1.65

Table 7

(A)Performance on hold-out set for gender specific models - with custom loss

(a)For Gender='Male' model

		MAPE	**MAE**	**MSE**	R^2
Baseline - LR	Unbalanced	0.072	1.336	3.375	0.489
	Random Under sampled	0.072	1.336	4.234	0.489
	SMOTEr	0.088	1.572	4.073	0.384
ARIMA	Unbalanced	0.074	1.360	3.499	0.471
	Random Under sampled	0.091	1.604	3.99	0.397
	SMOTEr	0.094	1.656	4.399	0.335
XGBoost	Unbalanced	0.072	1.334	3.333	0.496
	Random Under sampled	0.086	1.525	3.784	0.428
	SMOTEr	0.087	1.563	4.088	0.382

	Unbalanced	0.082	1.472	3.681	0.499
RNN	Random Under sampled	0.087	1.615	4.644	0.367
	SMOTEr	0.087	1.558	4.159	0.433
	Unbalanced	0.072	1.329	3.295	0.551
LSTM	Random Under sampled	0.075	1.372	3.435	0.532
	SMOTEr	0.086	1.527	3.926	0.465

(b)For Gender='Female' model

		MAPE	**MAE**	**MSE**	R^2
	Unbalanced	0.071	1.303	2.971	0.631
Baseline - LR	Random Under sampled	0.071	1.303	2.971	0.631
	SMOTEr	0.090	1.577	3.783	0.531
	Unbalanced	0.072	1.309	2.956	0.633
ARIMA	Random Under sampled	0.091	1.578	3.668	0.545
	SMOTEr	0.098	1.688	4.176	0.482
	Unbalanced	0.071	1.315	3.283	0.593
XGBoost	Random Under sampled	0.092	1.626	4.205	0.478
	SMOTEr	0.088	1.578	4.183	0.481
	Unbalanced	0.072	1.362	3.669	0.50
RNN	Random Under sampled	0.085	1.511	4.428	0.397
	SMOTEr	0.092	1.658	4.969	0.323
	Unbalanced	0.072	1.332	3.202	0.564
LSTM	Random Under sampled	0.084	1.495	3.519	0.520
	SMOTEr	0.09	1.590	4.156	0.434

(B)Performance on hold-out set for gender specific models - with standard loss

(a)For Gender='Male' model

		MAPE	**MAE**	**MSE**	R^2
	Unbalanced	0.072	1.336	3.375	0.489
Baseline - LR	Random Under sampled	0.072	1.336	4.234	0.489
	SMOTEr	0.088	1.572	4.073	0.384
	Unbalanced	0.074	1.360	3.499	0.471
ARIMA	Random Under sampled	0.091	1.604	3.99	0.397
	SMOTEr	0.094	1.656	4.399	0.335
	Unbalanced	0.073	1.347	3.410	0.514
XGBoost	Random Under sampled	0.088	1.555	3.928	0.440
	SMOTEr	0.089	1.573	4.179	0.404
	Unbalanced	0.070	1.286	3.371	0.508
RNN	Random Under sampled	0.072	1.902	3.251	0.512
	SMOTEr	0.092	1.889	3.265	0.524
	Unbalanced	0.072	1.296	3.000	0.562
LSTM	Random Under sampled	0.095	1.648	4.171	0.391
	SMOTEr	0.081	1.441	3.516	0.487

(b)For Gender='Female' model

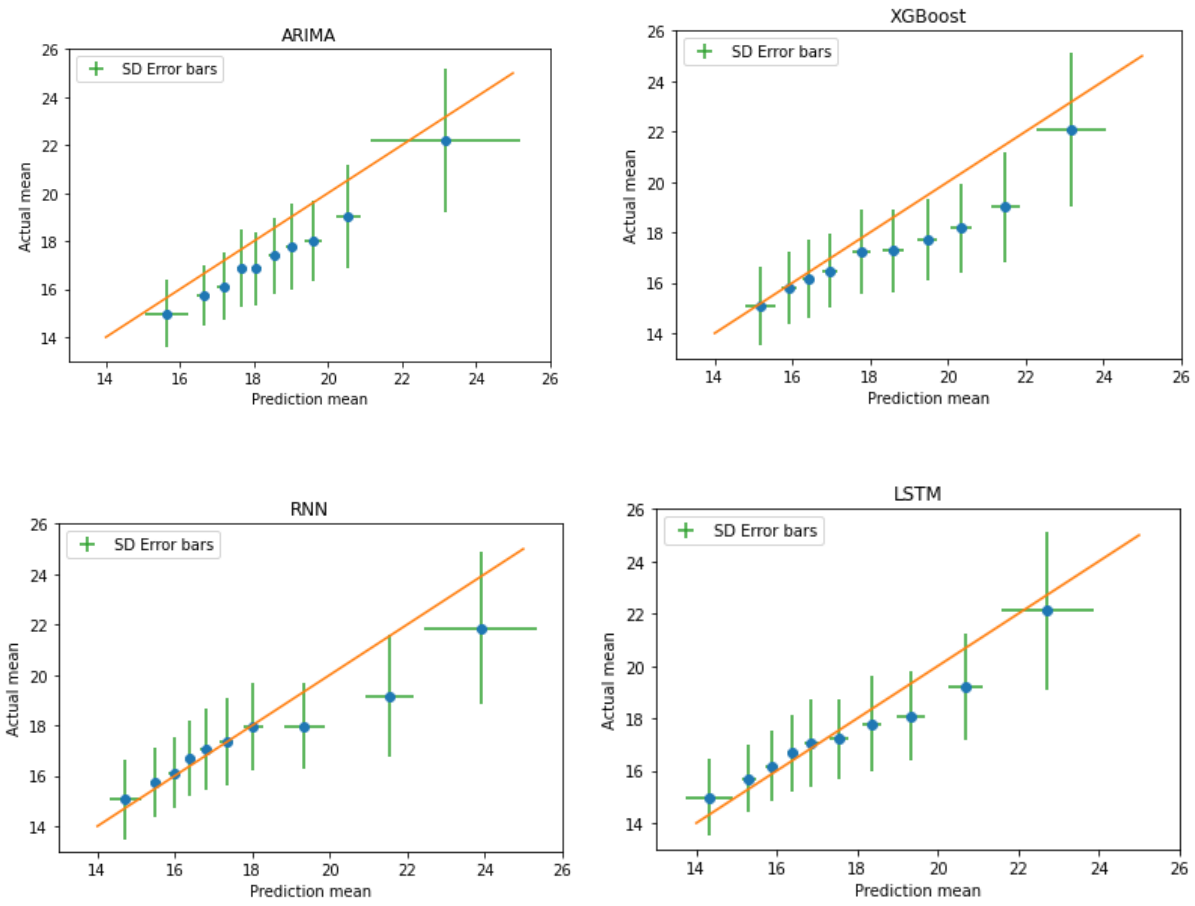
		MAPE	MAE	MSE	R^2
Baseline - LR	Unbalanced	0.071	1.303	2.971	0.631
	Random Under sampled	0.071	1.303	2.971	0.631
	SMOTEr	0.090	1.577	3.783	0.531
ARIMA	Unbalanced	0.072	1.309	2.956	0.633
	Random Under sampled	0.091	1.578	3.668	0.545
	SMOTEr	0.098	1.688	4.176	0.482
XGBoost	Unbalanced	0.073	1.312	3.049	0.544
	Random Under sampled	0.095	1.614	4.054	0.394
	SMOTEr	0.092	1.605	4.209	0.371
RNN	Unbalanced	0.084	1.470	4.470	0.348
	Random Under sampled	0.093	1.621	4.511	0.342
	SMOTEr	0.102	2.092	4.310	0.452
LSTM	Unbalanced	0.074	1.314	2.904	0.576
	Random Under sampled	0.077	1.361	3.113	0.545
	SMOTEr	0.085	1.508	3.757	0.452

To understand how well did the model fit, binned calibration plots for mean predicted vs actual values were plotted. As SMOTEr dataset gave the best performance, this fit was analyzed on predictions given by models trained on SMOTEr dataset. The predictions were sorted and binned into 10 bins, with 10% data in each bin. The mean predicted and mean actual value were plotted along with standard deviation (SD) for each bin. As seen from Figure 8, LSTM performed slightly better than others with the mean values for each bin lying on or near the identity line. However, it shows large SD similar to the other models especially for extreme BMI values.

To sum up, the answer to the first research question and sub question - none of the models stands out as the best performing model when taking into account all the evaluation criteria and there is not much significant performance difference between the general and the gender specific models. The answer to the second research question – a balanced dataset certainly does improve the model performance.

Figure 8

Binned Calibration Plot



7. Discussion

With the rise of obesity rates world over, it is essential to predict its development early on. In order to achieve this, it is important that the health trajectory is predicted accurately. The main goal of this study is to determine the optimal predictive model which can accurately predict BMI for children aged 10, with more emphasis on overweight children. This was done by comparing the performances of different models trained on longitudinal BMI data and demographic features. Since the data used for this study consisted of a time-series component, four models are selected which are considered the best in time-series analysis. The four models are - ARIMA, XGBoost, RNN and LSTM.

The first step was feature selection. Using SelectKBest as the feature selection method, a list of features contributing most to the target variable was generated. Due to small number of features and the performance of models with only the best features and all the features not yielding any significant performance difference, all the features were included in the final models. However, selection of the feature 'CM_4-12 years' warrants some attention. As mentioned previously this feature denotes the number of visits to the doctor in the specified age group. In the period of 4-12 years, there is no way to distinguish the number of visits at a particular age, plus it cannot be determined if the visits were related to obesity. For example, if a child visiting the doctor at age 4 due to some illness will cause the count in the CM variable to rise. Another child visiting the doctor for obesity intervention will also cause the count to increase. For now, there is no way to incorporate this intervention information in the model and out of scope of this study. For purpose of simplicity, this feature was added to all the models and further research is required.

The second step was to balance the dataset as there was imbalance of data with the majority data consisting of normal BMI values. For the model to learn on better predicting the extreme BMI values, the training data must be a good representation of all the classes. Hence data was balanced using random under sampling and SMOTEr techniques. Performance is also compared on balanced and unbalanced datasets. Next, hyperparameter tuning is performed by using Bayesian Optimization to find the best model. The next step was identifying which machine learning model is better at predicting the BMI for 10-year-old, with more weight on better predicting the extreme BMI values or outliers. This was done by building custom loss functions. Models were trained using the best set of hyperparameters and their performance was evaluated on the hold-out set. As expected, all the models performed better with SMOTEr dataset as this dataset had the largest size and more balanced samples. Custom loss functions added slight improvement to the predictions. XGBoost gave least MAE for predicting extreme BMI values. Upon examining the fit of the models, LSTM performed better than the others. On comparing MAE, LSTM and XGBoost performed slightly better for 'male' and 'female'

gender respectively. From R-squared perspective, ARIMA and LSTM performed better for 'male' and 'female' respectively. Unfortunately, not one model performed the best on all the evaluation criteria.

This study contributes to existing literature in two ways. First, as shown in study by Cavadas et al., 2015, training a model with a balanced dataset improves the model performance. As studies implementing data balancing in a regression study are rare, this study can be a reference to show how resampling improves the performance of machine learning models. Second, as majority of the existing literature is about classification of obesity (Gupta et al., 2019, Singh at al., 2020, Pang et al., 2021) this study adds a new dimension to BMI prediction by predicting BMI values. BMI values help to determine the distribution of BMI in the general population. Since 1858, there has been an increase in height and weight among the population of Dutch children due to many contributory factors like improved nutrition, better hygiene and overall child health (Fredriks et al., 2000). Change in height and weight suggest that BMI categories should also be adjusted based on changes in the population.

As with any research, this study also has some limitations. First, since the source of data, the JGZ was digitized in 2011, there is quite a substantial amount of missing data from the initial stages. In order to have a complete record, data was considered from age 2 onwards. This is a limitation as studies have shown that BMI information from early life stages is quite important in predicting future BMI (Smego et al., 2017, Roy et al., 2015). Second, as per different literatures, maternal and paternal BMI and income features are also a good predictor of obesity/BMI in a child (Morandi et al., 2012, Redsell et al., 2016, Graversen et al., 2015). Since this information is not available it was not included in this study. Third, as mentioned in the introduction, a child has regular visits till the age of 5 and none till the age of 10. The time between 5-10 years is important as it is during this time a child develops his/her lifestyle choices – playing sports, eating habits for instance. With no regular visits from age 5-10, this information is missed. Lastly, since the dataset only had data from the western part of the Netherlands, the findings from this study cannot be generalized.

Lastly, GDPR regulations and ethics come into effect when implementing any machine learning model into healthcare. Under GDPR it is important to protect privacy of an individual. With healthcare information being sensitive data, it becomes more important to protect the identity of the subject. For this study, the data was anonymized hence privacy of the child is protected. Ethics come into play when a predictive model is implemented in healthcare. Will it be ethical to pinpoint a child saying he/she will be obese in future based on the model outcome – is a debatable question. On one hand it will be beneficial for the child as he/she can get additional support and education for his/her well-being. On the other hand, this information can also lead to negative consequences like stigma associated with obesity. Another aspect is incorrect prediction. If the model predicts that a child will not be obese but he/she develops obesity at age 10 – such children (and their families) will be less likely to pay much attention to their overall well-being because of the prediction. Generally speaking, as predictive models are never 100% accurate, it will be a challenge to find a balance between these two ethical aspects.

8.Conclusion

Obesity in adults and in children is a chronic disease affecting millions globally. Once the disease sets in it affects the life of an individual in a negative manner. Studies have shown that obesity can be predicted in an early age and hence it is important to get accurate prediction sooner to get proper medical intervention and support. With accurate predictions, proper interventions can be planned for those identified as ‘at risk’. This will help not only the individual but also better allocation of medical resources and personnel. In the past many studies have employed models for classification of children as obese vs non-obese. In this study, four machine learning models were compared and evaluated to understand which algorithm is the best in predicting BMI values. On comparison, the performance of all the models were similar to each other. None of the models could outperform the baseline linear regression. The results obtained do not look very satisfactory when looking at the performance of each individual model due to the limitations mentioned in previous section. However, further research

using an ensemble of these or similar models and a larger training set might give better results. From an ethical point of view a clustering approach might be beneficial so as to have concentrated effort on the high-risk cluster instead of individual children.

References

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., & Cox, D. D. (2015). Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, *8*(1), 014008.

Branco, P., Torgo, L., & Ribeiro, R. P. (2017, October). SMOGN: a pre-processing approach for imbalanced regression. In *First international workshop on learning with imbalanced domains: Theory and applications* (pp. 36-50). PMLR.

Branco, P., Torgo, L., & Ribeiro, R. P. (2019). Pre-processing approaches for imbalanced distributions in regression. Neurocomputing, 343, 76-99.

Brownlee J. (2021, March 22). *A Gentle Introduction to XGBoost Loss Functions*. https://machinelearningmastery.com/xgboost-loss-functions/

Cavadas, B., Branco, P., & Pereira, S. (2015, September). Crime prediction using regression and resources optimization. In Portuguese Conference on Artificial Intelligence (pp. 513-524). Springer, Cham.

Chauhan, A., & Singh, A. (2017). An ARIMA model for the forecasting of healthcare waste generation in the Garhwal region of Uttarakhand, India. *International Journal of Services Operations and Informatics*, *8*(4), 352-366.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Choudhury, A., & Urena, E. (2020). Forecasting hourly emergency department arrival using time series analysis. *British Journal of Healthcare Management*, *26*(1), 34-43.

Chung, A.; Backholer, K.; Wong, E.; Palermo, C.; Keating, C.; Peeters, A. Trends in child and adolescent obesity prevalence in economically advanced countries according to socioeconomic position: A systematic review. *Obes. Rev.* 2016, *17*, 276–295.

Colmenarejo, G. (2020). Machine Learning Models to Predict Childhood and Adolescent Obesity: A Review. Nutrients, 12(8),2466. MDPI AG. Retrieved from http://dx.doi.org/10.3390/nu12082466

Davis, J., Juarez, D., & Hodges, K. (2013). Relationship of ethnicity and body mass index with the development of hypertension and hyperlipidemia. Ethnicity & disease, 23(1), 65–70.

Fredriks, A. M., Van Buuren, S., Burgmeijer, R. J., Meulmeester, J. F., Beuker, R. J., Brugman, E., ... & Wit, J. M. (2000). Continuing positive secular growth change in The Netherlands 1955–1997. Pediatric research, 47(3), 316-323.

Gamboa, J. C. B. (2017). Deep learning for time-series analysis. *arXiv preprint arXiv:1701.01887*.

GDPR EU Recital 26. (2016, May 4). *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL*. http://data.europa.eu/eli/reg/2016/679/oj

Gupta, M., Phan, T. L. T., Bunnell, T., & Beheshti, R. (2019). Obesity Prediction with EHR Data: A deep learning approach with interpretable elements. *arXiv preprint arXiv:1912.02655*.

Graversen, L., Sørensen, T. I., Petersen, L., Sovio, U., Kaakinen, M., Sandbaek, A., ... & Obel, C. (2014). Preschool weight and body mass index in relation to central obesity and metabolic syndrome in adulthood. *PloS one*, *9*(3), e89986.

GROEIDIAGRAMMEN IN PDF-FORMAAT. https://www.tno.nl/nl/aandachtsgebieden/gezond-leven/roadmaps/youth/groeidiagrammen-in-pdf-formaat/

Hammond R, Athanasiadou R, Curado S, Aphinyanaphongs Y, Abrams C, Messito MJ, Gross R, Katzow M, Jay M, Razavian N, Elbel B. Predicting childhood obesity using electronic health records and publicly available data. PLoS One. 2019 Apr 22;14(4):e0215571. doi: 10.1371/journal.pone.0215571. Erratum in: PLoS One. 2019 Oct 7;14(10):e0223796. PMID: 31009509; PMCID: PMC6476510.

Heymsfield, S. B., Peterson, C. M., Thomas, D. M., Heo, M., & Schuna, J. M., Jr (2016). Why are

there race/ethnic differences in adult body mass index-adiposity relationships? A quantitative critical review. Obesity reviews : an official journal of the International Association for the Study of Obesity, 17(3), 262–275. https://doi.org/10.1111/obr.12358

Hinton, G. E. (1992). How Neural Networks Learn from Experience. *Scientific American*, *267*(3), 144–151. http://www.jstor.org/stable/24939221

Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on November 12, 2021

Hyndman, R. (2010) Why every statistician should know about cross-validation. https://robjhyndman.com/hyndsight/crossvalidation/

Kaushik, S., Choudhury, A., Sheron, P. K., Dasgupta, N., Natarajan, S., Pickett, L. A., & Dutt, V. (2020). AI in healthcare: time-series forecasting using statistical, neural, and ensemble architectures. *Frontiers in big data*, *3*, 4.

Koehrsen Will. (2018). *A Conceptual Explanation of Bayesian Hyperparameter Optimization for Machine Learning.* https://towardsdatascience.com/a-conceptual-explanation-of-bayesian-model-based-hyperparameter-optimization-for-machine-learning-b8172278050f

Moniz, N., Branco, P., & Torgo, L. (2016, October). Resampling strategies for imbalanced time series. In 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 282-291). IEEE.

Monteiro, P. O. A., & Victora, C. G. (2005). Rapid growth in infancy and childhood and obesity in later life–a systematic review. *Obesity reviews*, *6*(2), 143-154.

Morandi A, Meyre D, Lobbens S, Kleinman K, Kaakinen M, Rifas-Shiman SL, et al. Estimation of newborn risk for child or adolescent obesity: lessons from longitudinal birth cohorts. PLoS One. 2012;7(11):e49919. Epub 2012/12/05. pmid:23209618; PubMed Central PMCID: PMCPMC3509134.

Morid, M. A., Sheng, O. R. L., Kawamoto, K., Ault, T., Dorius, J., & Abdelrahman, S. (2019). Healthcare cost prediction: Leveraging fine-grain temporal patterns. *Journal of biomedical informatics*, *91*, 103113.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *The New England journal of medicine*, *375*(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

Pang, X.; Forrest, C.B.; Le-Scherban, F.; Masino, A.J. Understanding early childhood obesity via interpretation of machine learning model predictions. In 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA); IEEE: Boca Raton, FL, USA, 2019; pp. 1438–1443.

Pang X, Forrest CB, Lê-Scherban F, Masino AJ. Prediction of early childhood obesity with machine learning and electronic health record data. Int J Med Inform. 2021 Jun;150:104454. doi: 10.1016/j.ijmedinf.2021.104454. Epub 2021 Apr 9. PMID: 33866231.

Monteiro P. O. A and C. G. Victora, "Rapid growth in infancy and childhood and obesity in later life—a systematic review," (in eng), Obesity Reviews: An Official Journal of the International Association for the Study of Obesity, vol. 6, no. 2, pp. 143-154, 2005/05// 2005, doi: 10.1111/j.1467-789X.2005.00183.x.

Prabhakaran S. (2021, August 22). *ARIMA Model – Complete Guide to Time Series Forecasting in Python.* https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting- python/

Putatunda, S., & Rama, K. (2018, November). A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost. In *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning* (pp. 6-10).

Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M, Sundberg P, Yee H, Zhang K, Zhang Y, Flores G, Duggan GE, Irvine J, Le Q, Litsch K, Mossin A, Tansuwan J, Wang D, Wexler J, Wilson J, Ludwig D, Volchenboum SL, Chou K, Pearson M, Madabushi S, Shah NH, Butte AJ, Howell MD, Cui C, Corrado GS, Dean J. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018 May 8;1:18. doi: 10.1038/s41746-018-0029-1. PMID: 31304302; PMCID: PMC6550175.

S. A. Redsell, S. Weng, J. A. Swift, D. Nathan, C. Glazebrook, Validation, optimal threshold determination, and clinical utility of the infant risk of overweight checklist for early prevention of child overweight, Childhood Obesity 12 (3) (2016) 202–209. doi:10.1089/chi.2015.0246.

Ritchie, H & Roser, M (2017) - "*Obesity*". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/obesity'

Roy, S. M., Chesi, A., Mentch, F., Xiao, R., Chiavacci, R., Mitchell, J. A., ... & McCormack, S. E. (2015). Body mass index (BMI) trajectories in infancy differ by population ancestry and may presage disparities in early childhood obesity. The Journal of Clinical Endocrinology & Metabolism, 100(4), 1551-1560.

Siddiqui, H., Rattani, A., Kisku, D. R., & Dean, T. (2020, December). AI-based BMI Inference from Facial Images: An Application to Weight Monitoring. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1101-1105). IEEE.

Singh, A. S., Mulder, C., Twisk, J. W., Van Mechelen, W., & Chinapaw, M. J. (2008). Tracking of childhood overweight into adulthood: a systematic review of the literature. Obesity reviews, 9(5), 474-488.

Singh, B., & Tawfik, H. (2020). Machine Learning Approach for the Early Prediction of the Risk of Overweight and Obesity in Young People. *Computational Science – ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part IV*, *12140*, 523–535. https://doi.org/10.1007/978-3-030-50423-6_39

Smego, A., Woo, J. G., Klein, J., Suh, C., Bansal, D., Bliss, S., ... & Crimmins, N. A. (2017). High body mass index in infancy may predict severe obesity in early childhood. The Journal of pediatrics, 183, 87-93.

The Netherlands in numbers. https://opendata.cbs.nl/statline/#/CBS/nl/

Torgo L, Ribeiro R. P, Pfahringer B, and Branco P. Smote for regression. In Progress in Artificial Intelligence, pages 378–389. Springer, 2013.

WHO. (2021, June 9). *Obesity and Overweight.* **https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight**

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82. http://www.jstor.org/stable/24869236

Weng SF, Redsell SA, Nathan D, Swift JA, Yang M, Glazebrook C. Estimating overweight risk in childhood from predictors during infancy. Pediatrics. 2013;132(2):e414–21. Epub 2013/07/17. pmid:23858427.

Zheng Z and Ruggiero K, "Using machine learning to predict obesity in high school students," *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2017, pp. 2132-2138, doi: 10.1109/BIBM.2017.8217988.

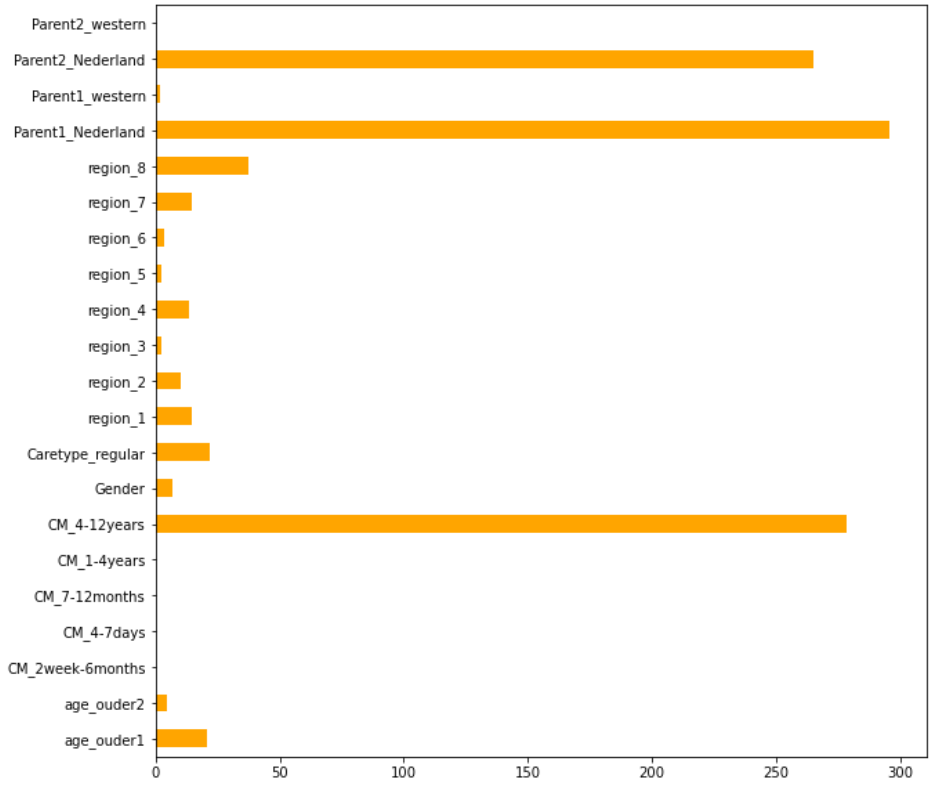
Zhou, L., Zhao, P., Wu, D., Cheng, C., & Huang, H. (2018). Time series model for forecasting the number of new admission inpatients. *BMC medical informatics and decision making*, *18*(1), 1-11.

Appendices

Appendix A: List Of Python Packages

Python Packages
numpy
pandas
matplotlib
seaborn
sklearn
keras
pmdarima
tensorflow
hyperopt
smogn
Imblearn
NaNImputer
pickle

Appendix B: Feature Importance



Appendix C: Feature Description

Feature Name	Description
2_yr	BMI at age 2
3_yr	BMI at age 3
3.9_yr	BMI at age 3.9
5-6_yr	BMI at age 5 or 6
10-11_yr	BMI at age 10 or 11 – Target feature
age_ouder1	Age of parent 1
age_ouder2	Age of parent 2
CM_2week-6months	Number of Contact Moment at age 2week-6months
CM_4-7days	Number of Contact Moment at age 4-7 days
CM_7-12months	Number of Contact Moment at age 7-12 months
CM_1-4years	Number of Contact Moment at age 1-4 years
CM_4-12years	Number of Contact Moment at age 4-12 years
Gender	Gender of the child
Caretype_regular	Whether child is under special care. Will have a value of 1 if under regular care
All columns ending with '_region'	Current region of stay. Will have value of 1 based on region of stay.
Parent1_Nederland	Parent 1 birth country is Netherlands
Parent1_western	Parent 1 birth country is a Western country
Parent2_Nederland	Parent 2 birth country is Netherlands
Parent2_western	Parent 2 birth country is a Western country

Appendix D: Size Of Datasets After Resampling

Resampling Method	Size after resampling
No Resampling	7256
Random Under Sampling	2602
SMOTEr	12335

Appendix E: Hyperparameter Configuration For The Models

	colsample_bytree	**gamma**	**learning_rate**	**max_depth**	**min_child_weight**	**n_estimators**	**reg_alpha**	**reg_lambda**	**subsample**
No Resampling	0.844	5.40	0.27	6	4	125	169	0.015	1.0
Random Under Sampling	0.662	4.51	0.12	17	6	1242	50	0.454	0.9
SMOTEr	0.61	2.62	0.05	16	9	271	177	0.812	0.85

XGBoost HyperParameter settings.

	learning_rate	**epochs**	**neurons-1st layer**	**neurons-2nd layer**	**neurons-3rd layer**	**activation functions**
No Resampling	0.0001	50	544	544	544	relu
Random Under Sampling	0.0001	50	64	64	64	relu
SMOTEr	0.0001	50	416	416	416	relu

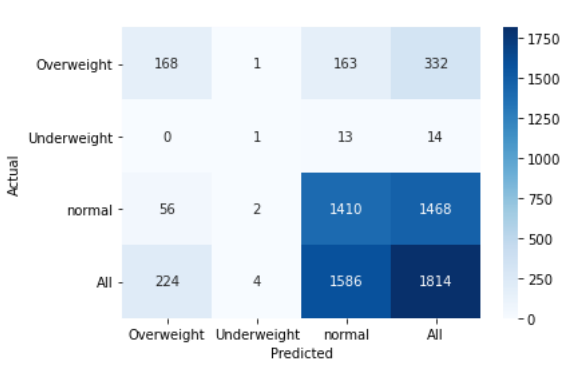
RNN HyperParameter settings

	learning_rate	**epochs**	**neurons-1st layer**	**neurons-2nd layer**	**neurons-3rd layer**	**activation functions**
No Resampling	0.0001	50	96	96	96	linear
Random Under Sampling	0.0001	50	320	320	320	relu
SMOTEr	0.0001	50	288	288	288	linear

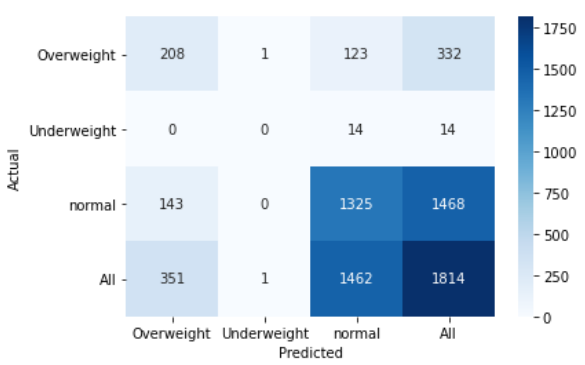
LSTM HyperParameter settings

Appendix F: Confusion Matrix For All Models

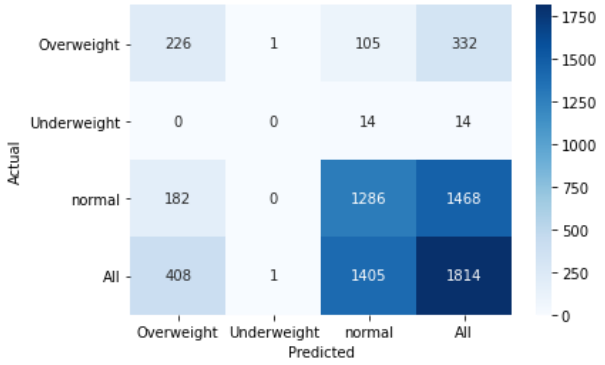
ARIMA



a) Unbalanced dataset

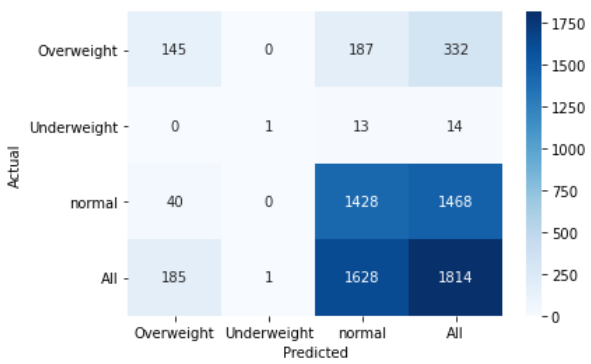


b) Random Under sampled

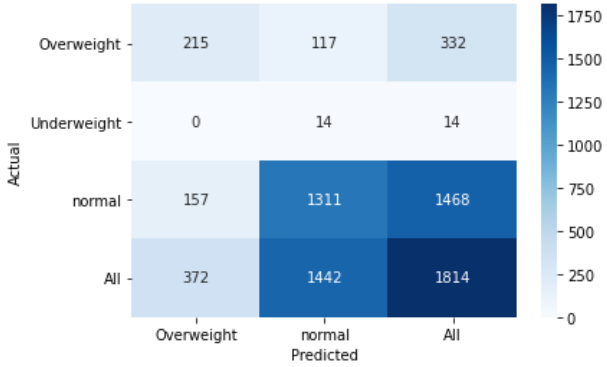


c) SMOTEr dataset

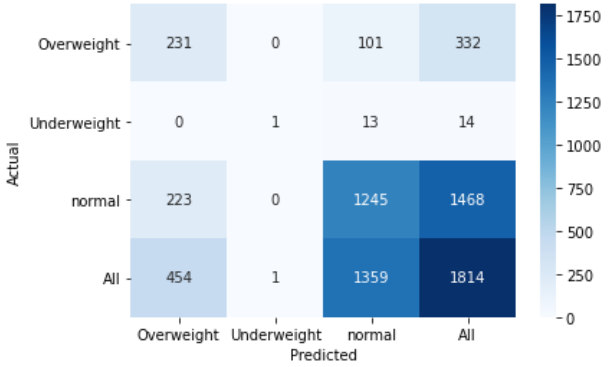
XGBoost



a) Unbalanced dataset

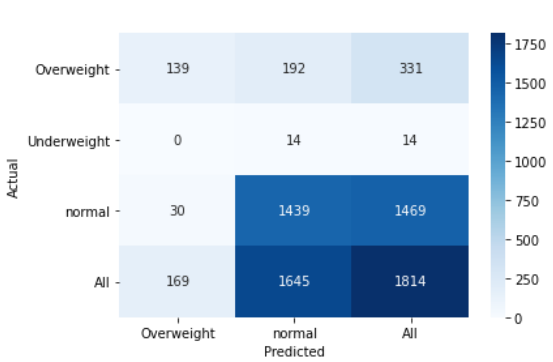


b) Random Under Sampled

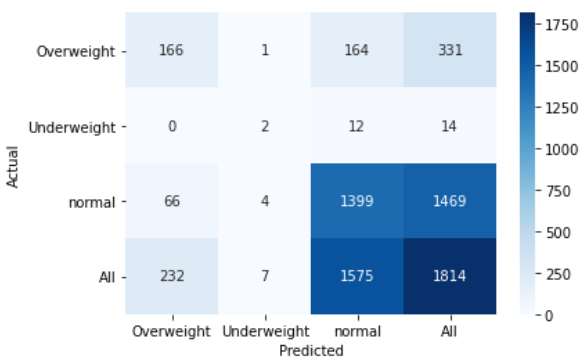


c) SMOTEr dataset

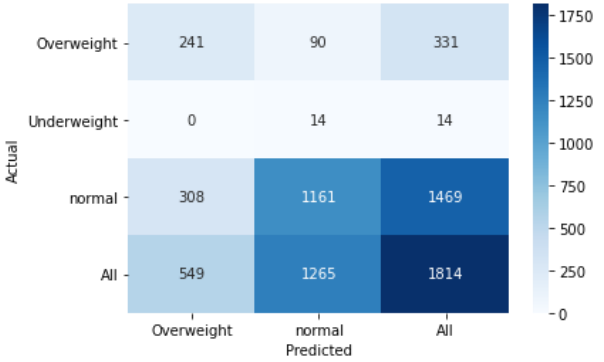
RNN



a) Unbalanced dataset

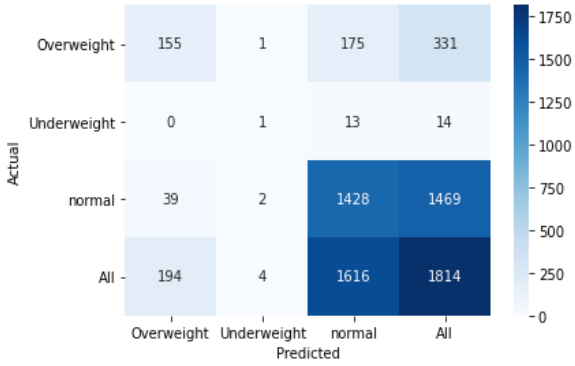


b) Random Under Sampled

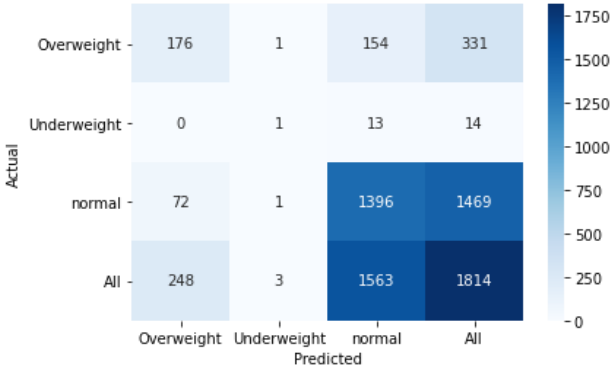


c) SMOTEr dataset

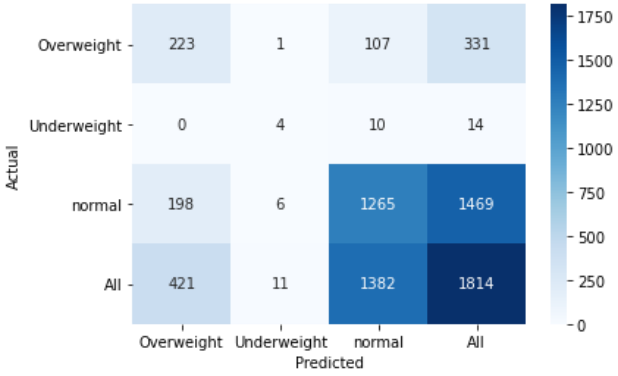
LSTM



a) Unbalanced dataset



b) Random Under Sampled



c) SMOTEr dataset