Analyzing Child Growth Trajectories Final Report

Inzish Khan
Shirmeen Amir
Khadija Javed

May 12, 2024

Abstract

The healthy growth and development of infants and young children is critical for the realization of their full physical and mental potential. Recognized globally as a key indicator of physical well-being in children, child growth holds significant importance in achieving international goals, including targets set by the World Health Assembly for 2025 concerning stunting, wasting, and overweight among children under 5 years. The ramifications of inadequate child growth extend to mortality, morbidity, and compromised cognitive development, with enduring effects into adulthood, influencing body size, work performance, reproductive health, and susceptibility to chronic diseases. Although assessing child growth is not inherently complex, it requires adherence to fundamental principles and meticulous attention to detail to ensure accurate evaluation and timely intervention. This paper delves into the classification of child growth patterns, namely undergrowth, normal, and overgrowth, employing machine learning techniques to enhance predictive capabilities and contribute to improved pediatric health monitoring and intervention strategies.

1 Introduction

The future of human societies depends on children being able to achieve optimal growth and development.Child growth and development are fundamental aspects of pediatric health, with far-reaching implications for individual well-being and social progress. Understanding the dynamics of child growth involves examining a variety of factors, from genetic predispositions to environmental influences. By analyzing large data sets on child growth, researchers can uncover patterns and trends that offer valuable insight into pediatric health outcomes. This project aims to leverage machine learning techniques to classify child growth patterns and identify potential health concerns early on.

1.1 Motivation

To deepen our understanding of child growth and its determinants, it's crucial to focus on specific datasets that provide concrete examples. In this project, we used a comprehensive data set that includes various attributes associated with child development. This data set offers a wealth of information on various aspects of child growth, including physical measurements and developmental milestones. By analyzing these data, we aim to gain practical insights into the factors influencing child growth and develop tools for better monitoring and intervention strategies.

1.2 Dataset and Description

Our dataset encompasses a comprehensive set of attributes crucial for understanding child growth and development. Addresses the challenges inherent in inconsistent monitoring and aims to provide insight into developmental trajectories. Each entry includes a diverse range of attributes that span the physical, cognitive, social, emotional, and communication domains. Key attributes comprise weight, height, age, motor skills (both gross and fine), communication abilities, and various cognitive capabilities.

An illustration of the structure of the dataset is shown below. Each entry in the data set represents an individual child.

Table 1: Description of Child Growth Dataset

Weight	Height	Age	Gross Motor	...	Growth
16	113	1	poor	...	normal
7	54	3	excellent	...	normal

The data set seen above is biased because all the target labels are almost equally distributed. The targets (growth) in the 25th percentile is 0 [over growth], 50th percentile is 1 [under growth] and 75th percentile is 2 [normal. The total dataset comprises of 80000 rows comprising of [overgrowth] having 26771 , [normal] having 26616 and [undergrowth] having 26613 instances in contribution to the dataset.Throughout the paper, we will utilize the **"Growth"** column as the target class and consider all other attributes as independent variables. Notably, we have excluded the **"Fine Motor"**, **"Gross Motor"**, and **"Perception of Directions"** columns from our analysis. These attributes were deemed to lack a direct influence on child growth, and their removal has shown improvement in model accuracy.One possible task that can be done on this dataset is conducting multi-class classification to predict a child's growth status across three categories based on their developmental attributes. Moreover, the availability of ground truth labels within the dataset enables the utilization of supervised learning algorithms for model training and evaluation.

1.3 Multi-Class Classification

Multi-class classification is a machine learning task where the goal is to assign each instance to one of multiple classes or categories. Following this, we will focus on the multi-class classification task throughout this paper. Each instance in the dataset corresponds to a child, and the objective is to classify them into one of several growth categories based on their developmental attributes. The accuracy of the classification will be measured against the ground truth labels available in the dataset. This task is essential for identifying developmental concerns and monitoring a child's progress over time.

2 MULTI-CLASS CLASSIFICATION

We will delve into the multi-class Classification problem on our dataset. Since we have access to ground truth labels for each child, we can apply supervised learning approaches for classification. Specifically, we will implement, test, and compare various methods suitable for multi-class classification on our child growth dataset. These methods include traditional machine learning algorithms such as Decision Trees, Random Forests, and Support Vector Machines (SVM), as well as more advanced techniques such as Neural Networks and Gradient Boosting Machines.

Given that our dataset includes both categorical and continuous columns, we will work our way through converting them into appropriate data types suitable for the machine learning model applied.

2.1 Preprocessing

Preprocessing the child growth dataset involves several essential steps to ensure data quality and prepare it for subsequent analysis and modeling.

Initially, missing values are addressed through appropriate techniques such as removal to ensure completeness and integrity of the data. Exploratory data analysis (EDA) techniques are then employed to visualize the distribution of the target variable (child growth categories) and explore relationships between features.

Feature engineering is conducted to extract relevant features from the dataset that are likely to have significant predictive power in determining child growth patterns which included dropping 3 Features from the dataset consisting of **Fine Motor**, **Gross Motor** and **perception of directions**.

Categorical features are encoded using method like one hot encoding to convert them into numerical representations, ensuring compatibility with machine learning algorithms.

Duplicate records are identified and removed to ensure data consistency which were seen as None in the dataset. Outliers are detected and handled using appropriate techniques to mitigate their impact on model performance. Visualizing the class Distribution as shown below in the Figure 1.

Figure 1: Visualizing the class Distribution

Hence we can clearly see that there is no class imbalance problem as all the target labels are almost equally distributed. Finally, the dataset is split into training and testing sets to facilitate model evaluation, with the training set used for model training and the testing set kept separate for unbiased evaluation of model performance.We kept the size of train/test split as 0.2. Which means that 80 percent of data used for training and rest 20 percent used for testing.

We tested algorithms like KNN, Random Forest, MultiLayer Perceptron , Decision Trees,Naive Bayes,Gaussian Naive Bayes, Logistic Regression, SVM and gradient boosting(XG Boosting) for classification purposes.

1. **Classification Algorithms:** The first algorithm which we applied for classification purposes is gradient Boosting. After applying the model we got was 0.32 having precision of 0.31. The second Model we applied was MLP(MultiLayer Perceptron).The result of this model included of an accuracy of 0.33 and precision of 0.33.Further in this we applied Decision Trees which gave us an accuracy of 0.33 and precision of 0.33. Moving on we applied more models on our dataset that is shown in the table below

Table 2: TABLE SHOWING ACCURACY AND PRECISION OF DIFFERENT MODELS

Model	Accuracy	Precision
Gradient Boosting	0.32	0.31
MLP (Multilayer Perceptron)	0.31	0.31
KNN (K-Nearest Neighbors)	0.33	0.34
Naive Bayes	0.33	0.33
Gaussian Naive Bayes	0.323	0.32
Logistic Regression	0.31	0.31
Decision Trees	0.33	0.33
Random Forest	0.32	0.32
SVM	0.325	0.32

Looking only at the accuracy column of Table 2 shown above , we can interpret a lot about our model. Accuracy reflects the proportion of correctly classified instances out of the total instances, while precision measures the proportion of true positive predictions out of all positive predictions made by the model. From the results, it is apparent that the models exhibit a range of accuracy scores. Notably, KNN achieved the highest accuracy score of 0.33, followed closely by Decision Trees with an accuracy of 0.33. However, the overall accuracy of the models, including Gradient Boosting, MLP, KNN, Naive Bayes, Gaussian Naive Bayes, Random Forest, and SVM, remains relatively low, ranging from 0.317 to 0.33. These results suggest that while certain models perform moderately well in accurately classifying child growth patterns, there is room for improvement in achieving higher accuracy across the board.

Hence only taking in account the accuracy might not be the right approach as in some cases the accuracy might not be providing pattern measures within our models applied.In accordance with this especially in medical related fields we put more emphasis on the precision of the model being applied.From the table, it is evident that the models exhibit varying degrees of performance. Notably, logistic regression achieved the highest precision of 0.35, indicating its ability to accurately identify positive instances of child growth.

Visualizing the models Testing Accuracies for understanding. The figure below shows the accuracies of the models applied on our dataset.

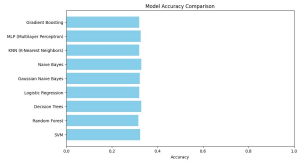

Figure 2: Test accuracy comparison of different models

In the pursuit of maximizing the predictive performance of our classification models for child growth trajectories, hyperparameter tuning emerged as a crucial step in our methodology. Delving into more detail about the Hyperparameter Tuning

2.2 Hyperparameter Tuning:

Recognizing the significance of hyperparameters in influencing model complexity and generalization, we embarked on a systematic exploration of hyperparameter spaces to optimize our models' performance. Leveraging techniques such as GridSearchCV, RandomizedSearchCV, and cross-validation scores, we meticulously tuned hyperparameters to strike a balance between bias and variance, ultimately aiming to improve model accuracy and robustness. By systematically varying hyperparameters such as learning rate, regularization strength, number of layers, and activation functions, we aimed to identify the configurations that yield the most optimal performance for our classification task. Through this iterative process, we anticipated not only achieving higher accuracy but also enhancing the models' ability to generalize well to unseen data, thus bolstering the reliability and effectiveness of our predictive models for child growth classification.

In the pursuit of optimizing the Decision Tree classifier for accurately classifying child growth trajectories, our focus centered on refining key hyperparameters, namely 'max depth', 'min samples split', and 'min samples leaf'. We conducted an exhaustive exploration of parameter values within predefined ranges, with 'max depth' spanning from None to 100 in increments of 10, 'min samples split' ranging from 2 to 20, and 'min samples leaf' varying from 1 to 10. Employing RandomizedSearchCV, we systematically evaluated the model's performance across 100 iterations, utilizing 5-fold cross-validation to assess accuracy. Through this iterative process, nested cross-validation allowed for the determination of the mean accuracy and standard deviation, providing insights into model robustness. Subsequently, the optimal parameters identified through hyperparameter tuning were incorporated into the model, resulting in enhanced accuracy in the classification of child growth trajectories.Similarly, applying GridsearchCV on Random Fores,MLP and Logistic Regression etc. The Table 3 below shows that accuracy and Precision of the models after Hyperparameter Tuning.

Table 3: TABLE SHOWING ACCURACY AND PRECISION OF DIFFERENT MODELS AFTER HYPERPARAMETER TUNINING

Model	Accuracy	Precision
Gradient Boosting	0.32	0.32
MLP (Multilayer Perceptron)	0.32	0.31
KNN (K-Nearest Neighbors)	0.32	0.33
Naive Bayes	0.33	0.33
Gaussian Naive Bayes	0.33	0.32
Logistic Regression	0.31	0.35
Decision Trees	0.86	0.87
Random Forest	0.33	0.33

After hyperparameter tuning, the performance of various classification models for predicting child growth trajectories was assessed based on accuracy and precision metrics. Notably, Decision Trees exhibited a significant improvement in accuracy, achieving the highest score of 0.86, accompanied by a precision of 0.87. This highlights the efficacy of hyperparameter optimization in enhancing the predictive capabilities of Decision Trees for this specific task. Conversely, models such as Gradient Boosting, MLP, KNN, Naive Bayes, and Gaussian Naive Bayes demonstrated modest accuracy scores ranging from 0.31 to 0.33, with precision values varying slightly. While Random Forest and Logistic Regression maintained a similar accuracy level post-tuning, it exhibited consistent precision values across iterations. Overall, the results underscore the importance of hyperparameter tuning in fine-tuning model performance and highlight **Decision Trees** as a promising candidate for accurately predicting child growth trajectories.

Visualizing the Accuracies of all the models in the Figure 3 shown below

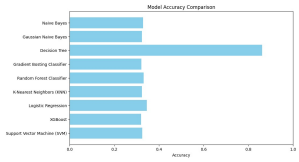

Figure 3: Test accuracy comparison of different models

Overall Comparing the results of before and after Hyperparameter Tuning we can clearly see that the after Hyperparameter Tuning the Decision Trees model is giving the Highest Accuracy of 0.86 and a precision of 0.87 which is considered as better performing then the previous models and before hyperparameter Tuning.

Confusion Matrix for the Decision Tree is shown below showcasing the performance measures of the model in graph form:

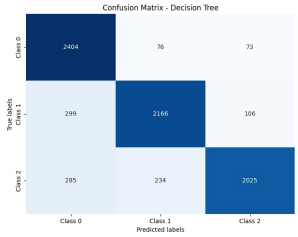

Figure 4: Confusion Matrix of Decision Tree

Furthermore, Hyperparameter Tuning the models resulted in achieving a higher Accuracy and Precision of the Model applied on the dataset which in turn is turning in the benefit of the improved predictive performance and reliability for classifying child growth trajectories.

2.3 Isolated Forest:

Isolated Forest, also known as Isolation Forest, is an anomaly detection algorithm based on the principle of isolating anomalies in the data. Unlike traditional methods that seek to model normal behavior explicitly, Isolated Forest focuses on isolating anomalies by randomly selecting features and partitioning data points into subsets. This process continues recursively until anomalies are isolated into small partitions, making them easier to detect.

We applied the Isolated Forest algorithm to our classification task of child growth due to its effectiveness in identifying outliers or anomalies, which could be indicative of abnormal growth patterns or developmental issues in children.An illustration of the model performance is shown in the table 4 shown below.

Table 4: TABLE SHOWING ACCURACY AND PRECISION OF ISOLATED FOREST

Model	Accuracy	Precision
Isolated Forest	0.89	0.09008

Given the importance of precision in medical applications, Isolated Forest's ability to accurately identify anomalies makes it particularly suitable for our dataset, where the detection of potential growth concerns is paramount. Despite previous attempts with models such as logistic regression, naive Bayes, and multi-layer perceptron (MLP), achieving a maximum accuracy of 0.86 with decision trees, the Isolated Forest algorithm yielded promising results with an accuracy of 0.89 and a precision of 0.90. This performance underscores the potential of Isolated Forest in accurately classifying child growth patterns and detecting anomalies that may signify underlying health issues, thus facilitating timely interventions and healthcare decisions.

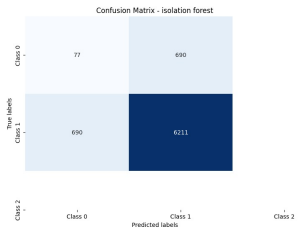

Figure 5: Confusion Matrix of Isolated Forest

As seen from the above confusion matrix we can identify that the isolation forest model is effective at identifying true positive cases for this particular subject of study as indicated by a high number of true positives. However, the presence of false positives and false negatives suggests that while the model is precise, it may still misclassified in some instances, which is an important consideration for medical diagnosis accuracy as in our case which can be seen in the above image, the precision being high enough as to cater the misclassifications.

3 Conclusion and Future Direction :

In this study, we explored various machine learning models to classify child growth trajectories, focusing on achieving accurate predictions to facilitate early detection of developmental concerns. Among the models tested, Isolated Forest emerged as the most promising, attaining an accuracy of 0.89 and a precision of 0.90, surpassing the performance of other classifiers such as Gradient Boosting, MLP, KNN, Naive Bayes, Logistic Regression, Decision Trees, and Random Forest. Despite initial attempts yielding a maximum accuracy of 0.86 with Decision Trees after hyperparameter tuning, Isolated Forest demonstrated superior performance, underscoring its potential in accurately classifying child growth patterns.

The Future direction for this project could venture into several promising avenues. Firstly, longitudinal data analysis could be integrated to monitor individual growth trajectories over time, allowing for personalized interventions and early identification of developmental delays. Secondly, incorporating additional data sources such as genetic profiles, environmental factors, and socioeconomic indicators could offer a more comprehensive understanding of the multifaceted influences on child growth. Furthermore, exploring advanced machine learning techniques, including deep learning models and ensemble methods, may enhance predictive accuracy and unveil subtle patterns in child growth data that were previously undetected. Lastly, forging collaborations with healthcare professionals and policymakers could facilitate the development of targeted intervention programs and policies aimed at fostering optimal child development and mitigating health inequalities.

4 Indivisual Contributions :

Shirmeen Amir studied the differnet EDA techniques and applied them on the dataset. Khadija Javed along with Shirmeen divided the models among themselves and applied them on the dataset. Inzish Khan worked with Khadija Javed in hyperparamter tuning in order to improve the accuracy of the already applied models.Inzish and Shirmeen both contributed in writing the report and putting the results all in one place.