**INTRODUCTION**

**Background and Context :** Obesity is a critical global health challenge and a leading risk factor for numerous chronic diseases, including type 2 diabetes, cardiovascular diseases (CVD), and certain cancers. According to the World Health Organization (WHO), over 1.9 billion adults worldwide are overweight, with more than 650 million classified as obese. This epidemic significantly contributes to the global burden of disease, exacerbating healthcare costs and impacting quality of life.

Despite its widespread prevalence, obesity often remains underdiagnosed or poorly managed due to its multifaceted nature, involving genetic, environmental, and behavioral factors. The rising prevalence, particularly in low- and middle-income regions, highlights the urgent need for advanced, accessible methods to classify and predict obesity levels effectively.

### **Problem Statement**

The rapid advancement of machine learning (ML) and deep learning (DL) techniques has significantly improved the accuracy and efficiency of obesity classification and prediction by analyzing complex datasets encompassing lifestyle factors, clinical indicators, and genetic predispositions. Studies have reported remarkable results, such as 99.05% accuracy with the Gboost classifier (Musa et al.[7]) and 96.46% with deep neural networks (Paulo and Lima et al.[6]). However, several critical challenges remain.

These challenges include the limited interpretability of ML models (Jeong et al.[4]; Kumar et al.[5]), the scarcity of validation using diverse and real-world datasets (Admojo and Rismayanti et al.[3]; Lazzer et al.[11]), and the insufficient application of advanced predictive tools in clinical and public health settings (Cervantes and Palacio et al.[8]). Furthermore, accessibility issues in resource-limited settings and the disparities in obesity management due to demographic, socioeconomic, and regional variations (Chen et al.[19]; Ferdowsy et al.[20]) highlight significant gaps in current approaches.

Although advancements in feature selection and data augmentation have improved classification accuracy, many existing methods lack real-world applicability, particularly in addressing individual-specific factors and ensuring healthcare provider confidence through explainable AI (Jeong et al.[4]; Yagin et al.[17]). Moreover, the global prevalence of obesity continues to escalate, with limited public awareness and delayed interventions exacerbating its associated health risks and economic burden (Helforoush et al.[1]; Mondal et al.[9]).

This study aims to address these limitations by leveraging balanced and diverse datasets, integrating interpretable ML models, and employing advanced feature selection and optimization techniques to enhance the accuracy and practical relevance of obesity classification and prediction systems.

### **Objectives of the Study**

The primary objective of this study is to explore novel machine learning and deep learning approaches for the automated classification and prediction of obesity levels, leveraging a comprehensive dataset comprising lifestyle, demographic, and medical attributes. By employing advanced feature selection methods and ensemble learning techniques, the study aims to enhance the accuracy and reliability of obesity prediction models while ensuring interpretability for practical and clinical applications. Additionally, the research will evaluate the effectiveness of these models in real-world scenarios, bridging the gap between theoretical advancements and actionable healthcare solutions.

### **Methodology Overview**

In this study, data on obesity was collected through a combination of real-life measurements and responses from a structured survey. The dataset underwent preprocessing to handle missing values and outliers, followed by balancing techniques like SMOTE, ADASYN, and SMOTEEN to address class imbalances.

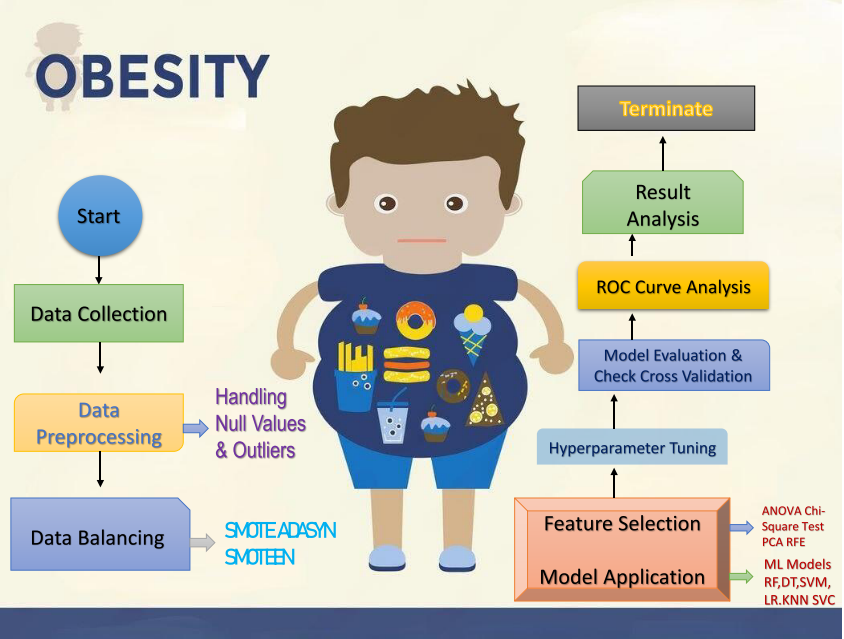
Key methodological steps include:

1. **Feature Selection**:  
   * Statistical and dimensionality reduction techniques such as ANOVA, Chi-Square Test, PCA, and RFE were applied to identify the most significant predictors of obesity levels.
2. **Model Training**:  
   * Various machine learning models, including Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Support Vector Classifier (SVC), were applied to classify obesity levels.
3. **Hyperparameter Tuning**:  
   * Optimization of model performance was achieved through hyperparameter tuning using grid search and random search methods.
4. **Cross-Validation**:  
   * To ensure generalizability, models were evaluated using k-fold cross-validation, reducing the risk of overfitting.
5. **Evaluation Metrics**:  
   * Model performance was assessed using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC for comparative analysis.

### **Significance and Contributions**

The significance of this research lies in its potential to improve the early detection and prediction of obesity levels, leading to timely interventions that can mitigate the risk of related diseases such as diabetes, cardiovascular conditions, and certain cancers. By advancing both the accuracy and interpretability of obesity classification models, this study contributes to the evolving field of AI-assisted healthcare. The data-driven insights generated by these models have the potential to empower healthcare providers and enhance patient outcomes, particularly in terms of prevention and personalized treatment strategies.

**Methodology**



**Fig : Methodology of Estimation of Obesity Levels**

The proposed methodology for analyzing obesity-related data is structured systematically to ensure reliable and accurate results. The process is summarized in the following steps:

1. **Data Collection** Data was collected using both real-life measurements and responses obtained through a structured survey. The questionnaire was designed to capture relevant features, such as demographic information, lifestyle habits (e.g., diet, exercise), medical history, and obesity-related metrics (e.g., BMI, waist circumference). Real-life measurements were taken using standard equipment to ensure accuracy. This combination of self-reported and measured data provided a diverse and comprehensive dataset for analysis.
2. **Data Preprocessing** In this stage, the collected data was prepared for analysis by handling missing values and outliers. Missing values were imputed using appropriate techniques (e.g., mean/mode imputation, regression imputation), and outliers were identified and either removed or adjusted based on their impact on the dataset. This preprocessing step ensured that the dataset was of high quality and suitable for machine learning model training.
3. **Data Balancing** As medical datasets often suffer from class imbalances, techniques such as SMOTE (Synthetic Minority Oversampling Technique), ADASYN (Adaptive Synthetic Sampling), and SMOTEEN were employed to balance the dataset. These methods improved the performance of machine learning models by ensuring an even distribution of obesity levels (e.g., normal, overweight, obese).
4. **Feature Selection** Key features contributing to the prediction of obesity levels were selected using various statistical and dimensionality reduction techniques. These included ANOVA (Analysis of Variance), Chi-Square Test, PCA (Principal Component Analysis), and RFE (Recursive Feature Elimination). This step helped reduce the dimensionality of the data and retained the most important predictors, ensuring efficient and accurate model performance.
5. **Model Application** Several machine learning models were applied to the balanced and feature-engineered dataset. The models included:  
   * **Random Forest (RF)**
   * **Decision Tree (DT)**
   * **Support Vector Machine (SVM)**
   * **Logistic Regression (LR)**
   * **K-Nearest Neighbors (KNN)**
   * **Support Vector Classifier (SVC)**
   * **Gaussian Naive Bayes**
   * **LightGBM**
   * **XGBoost**

These models were trained and tested to determine their efficacy in predicting obesity levels based on the available features.

1. **Hyperparameter Tuning** To optimize the models’ performance, hyperparameter tuning was conducted. Techniques like grid search and random search were utilized to identify the best combination of parameters for each model, ensuring the highest prediction accuracy.
2. **Model Evaluation and Cross-Validation** The models were evaluated using various metrics, such as accuracy, precision, recall, and F1-score. Cross-validation (e.g., k-fold cross-validation) was performed to ensure the results were consistent and to prevent overfitting, ensuring the robustness of the models.
3. **ROC Curve Analysis** Receiver Operating Characteristic (ROC) curves were generated to analyze the performance of the models further. The Area Under the Curve (AUC) metric was used as a comparative measure of effectiveness, highlighting the models' ability to distinguish between different obesity levels.

### **Result Analysis**

In this section, we present the results of the obesity level classification models both **before** and **after** applying the Synthetic Minority Oversampling Technique (SMOTE) to handle class imbalances. The performance of various machine learning algorithms is evaluated based on metrics such as **Accuracy**, **Sensitivity**, **Specificity**, **Precision**, **F1-Score**, **False Positive Rate (FPR)**, and **False Negative Rate (FNR)**.

#### 

#### 

#### 

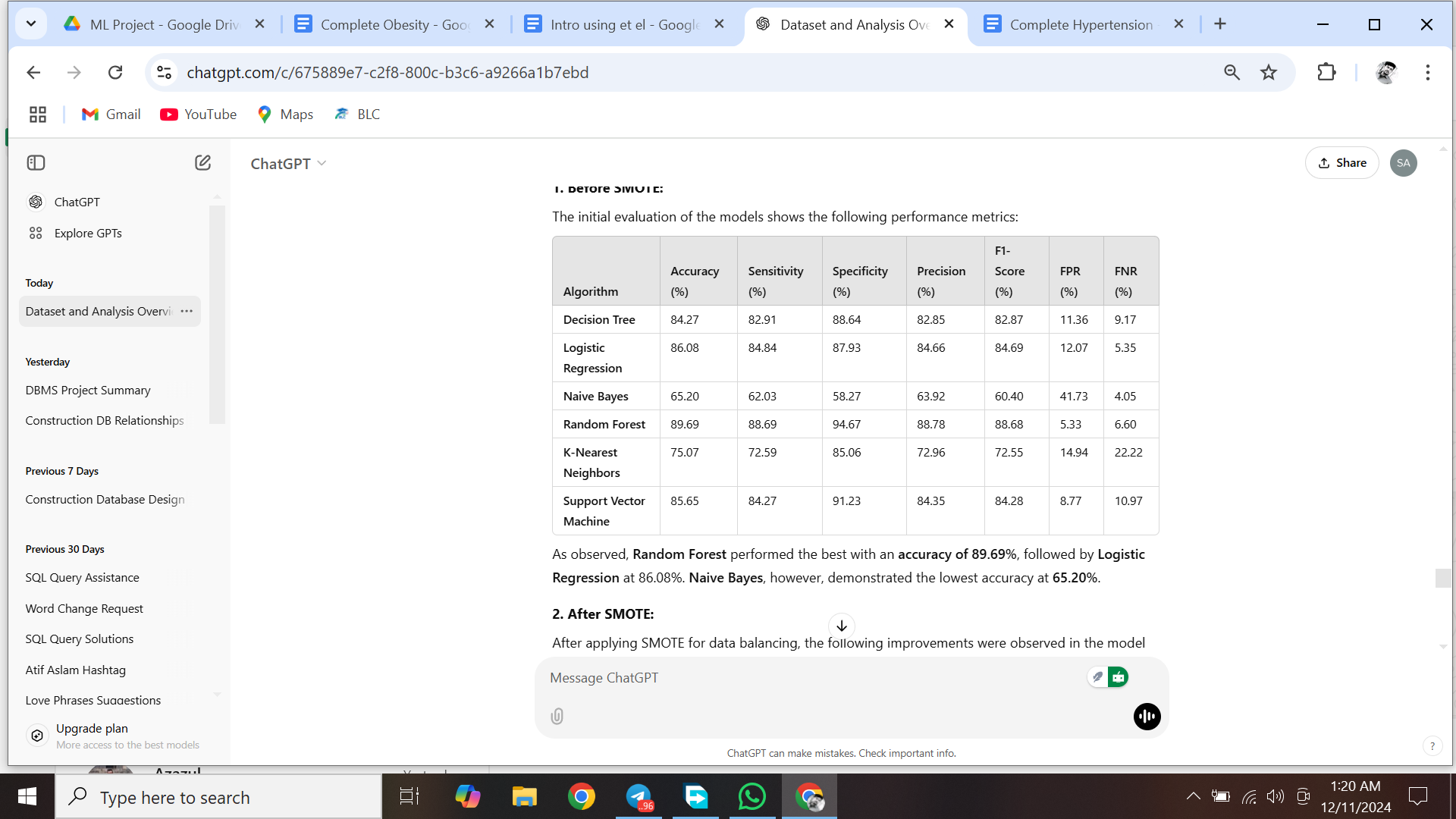
#### 

#### 

#### 

#### **1. Before SMOTE:**

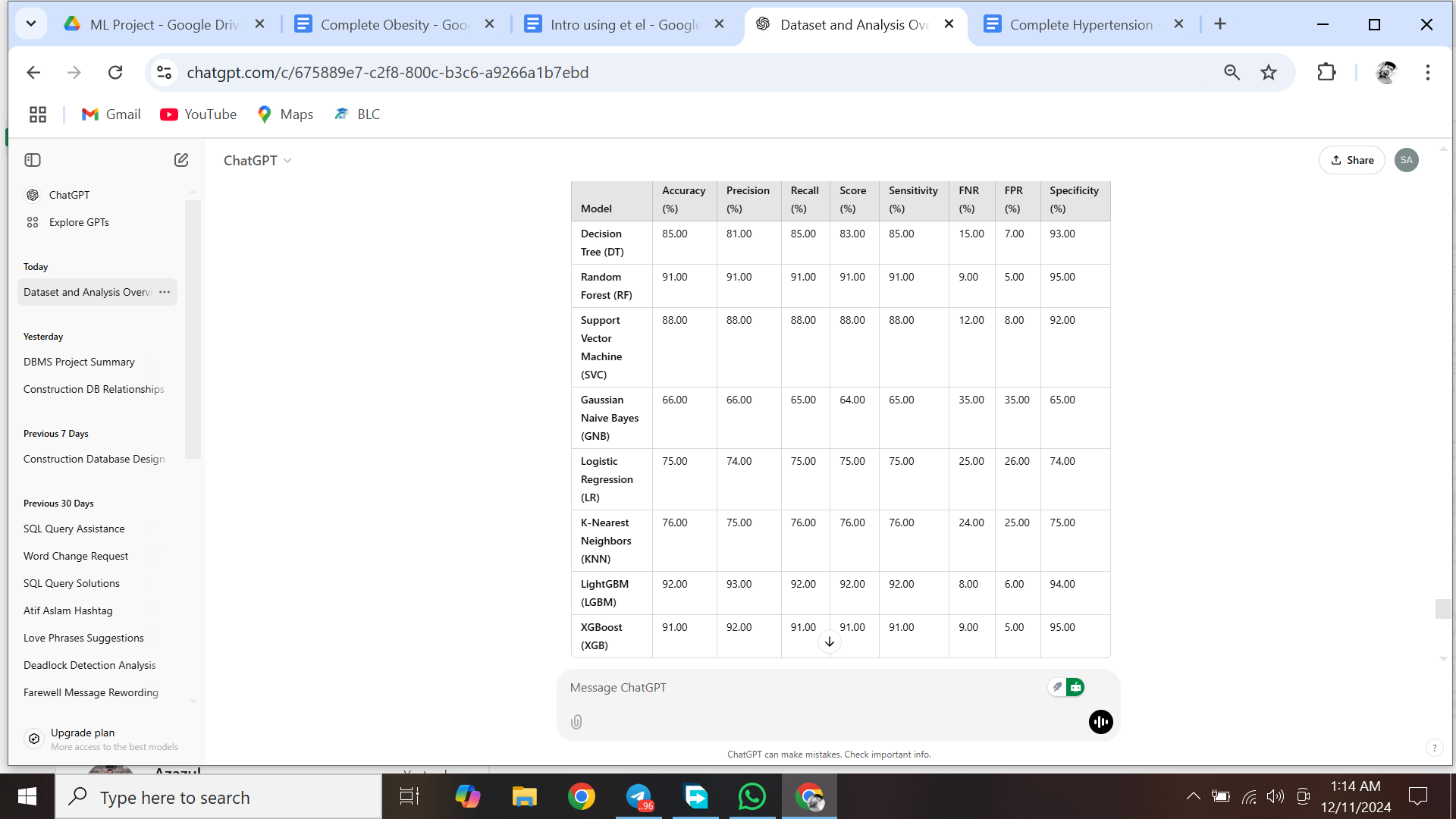
The initial evaluation of the models shows the following performance metrics:



As observed, **Random Forest** performed the best with an **accuracy of 89.69%**, followed by **Logistic regression** at 86.08%. **Naive Bayes**, however, demonstrated the lowest accuracy at **65.20%**.

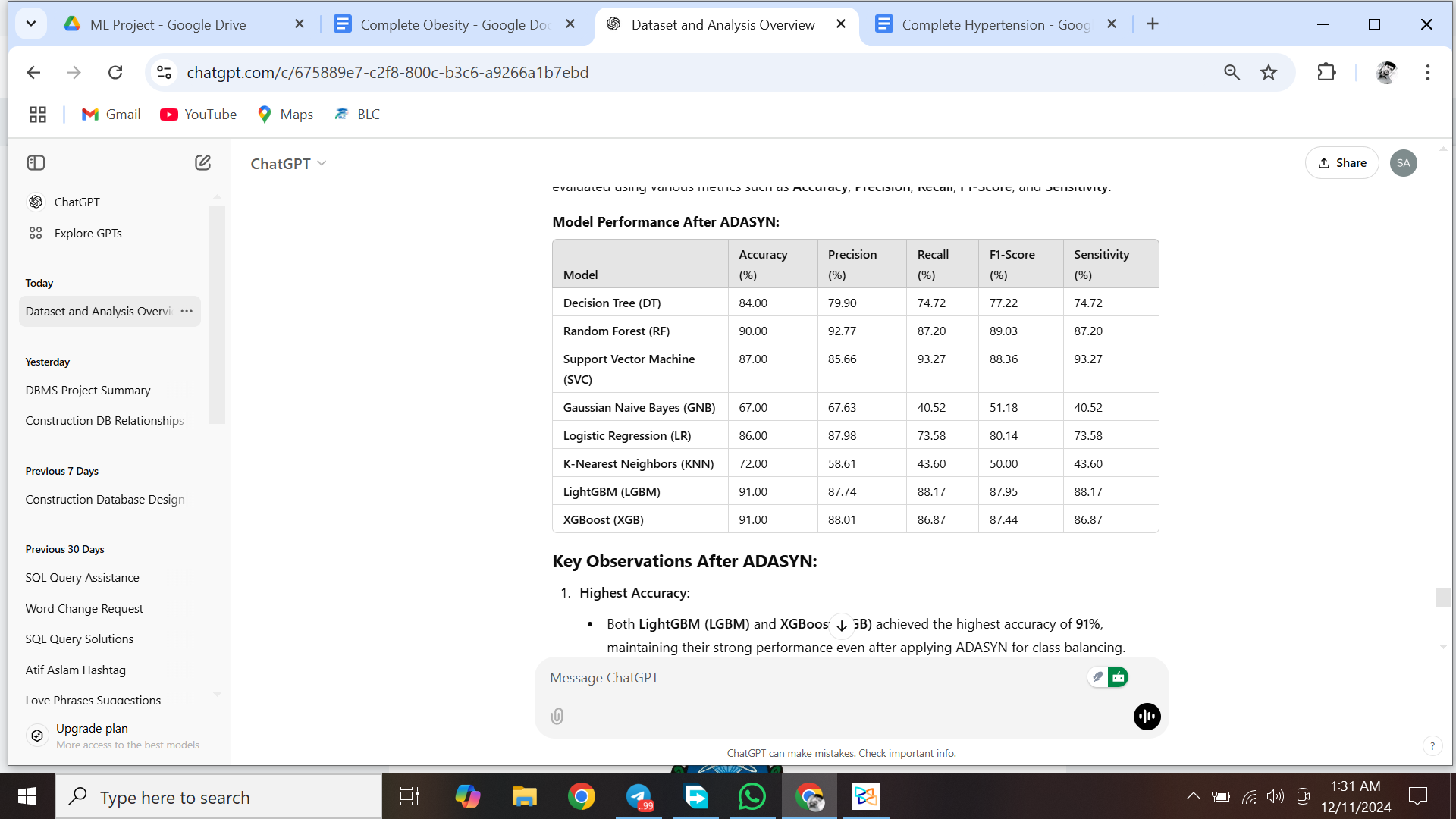
#### **2. After SMOTE:**

After applying SMOTE for data balancing, the following improvements were observed in the model performance metrics:



**LightGBM (LGBM)** attained the highest accuracy of 92%, closely followed by XGBoost (XGB) at 91%. Both models exhibited outstanding performance in terms of classification accuracy and sensitivity.

**3. After ADASYN:**

****

**After applying ADASYN for class balancing, the models' performance decreases.**

**Key Observations:**

1. **Highest Accuracy:**
   * **LightGBM (LGBM)** achieved the highest accuracy of **92%**, followed by **XGBoost (XGB)** at **91%**. These models demonstrated excellent performance in both classification and sensitivity.
2. **Precision and Recall:**
   * Both **LGBM** and **XGB** models achieved a **precision of 93%** and **92% recall**, indicating strong performance in correctly identifying positive cases while minimizing false positives.
3. **Sensitivity & Specificity:**
   * **LGBM** and **XGB** models displayed **sensitivity** rates of **92%** and **91%**, respectively, which signifies a robust ability to detect true positive obesity cases. Both models also showed high **specificity** (**94% and 95%**) in identifying non-obese individuals.
4. **FNR and FPR:**
   * After SMOTE, both **LGBM** and **XGB** maintained relatively low **FNR** and **FPR**, which means fewer false negatives and false positives, confirming their reliability in real-world applications.

### 

**Literature Review**

Obesity is a global health issue with various diseases such as diabetes, cardiovascular issues, and so on. Machine learning techniques have been increasingly used to classify obesity risk by analyzing complex data patterns**.** For instance, Zarindokht Helforoush et al. [1] developed a hybrid metaheuristic machine learning approach that showed promising results in classifying obesity of the CMI and VAI indices in identifying metabolic syndrome risk was approximately **85%**. risk. The proposed ANN-PSO model achieved a remarkable accuracy rate of **92%,** outperforming traditional regression methods. Koklu and Sulak et al. [2] proposed Random Forest was the most successful artificial intelligence method for accurately classified obesity. To analyze the survey data, four commonly used artificial intelligence methods in literature, namely Artificial Neural Network, K Nearest Neighbors, Random Forest and Support Vector Machine, were employed after pre-processing. As a result of this analysis, obesity classes were predicted correctly with success rates of **74.96%**, **74.03%,** **74.03%** and **87.82%**, respectively. Admojo and Rismayanti et al. [3] utilized decision tree algorithms of individuals' lifestyle and physical condition data for accuracy, precision, recall, and F1-scores. Jeong et al. [4] used DeepHealthNet, a deep learning-based system designed to predict adolescent obesity by processing complex health data. He proposed deep learning framework, DeepHealthNet, effectively trains the model using data augmentation techniques, even when daily health data are limited, resulting in improved prediction accuracy **88.42%.** Additionally, the study revealed variations in the prediction of the obesity rate between boys **93.20%** and girls **91.63%** Furthermore, Kumar et al. [5] employed a logistic regression model for effectiveness in handling large datasets and improving prediction accuracy. Paulo and Lima et al. [6] explored the use of deep neural networks for classifying obesity level with an accuracy of **96.46%**. Musa et al. [7] investigated whether the Gboost classifier achieves the highest accuracy of **99.05%** as compared to other classifiers Also, K-Nearest Neighbor gave a relatively strong accuracy of **95.74%**. Cervantes and Palacio et al. [8] also contributed to the field by estimating obesity levels based on Decision Tree, Support Vector Machine (SVM), and Simple K-Means . Mondal et al. [9] the authors developed a model for predicting childhood obesity using data from routine well-child visits.Using algorithms are logistic regression, support vector machine (SVM), random forest, artificial neural network (ANN), k-means clustering, and k-nearest neighbors.Five years of age with an accuracy of **89%**, 77%, and **89%**. Solomon et al. [10] proposed a "Hybrid Majority Voting" algorithm to improve the prediction and classification of obesity by integrating the results of multiple machine learning classifiers, including decision trees, support vector machines (SVM), and neural networks.The hybrid model achieved a high accuracy rate of **94.5%**.Lazzer et al. [11] evaluated the Cardiometabolic Index (CMI) and Visceral Adiposity Index (VAI) as predictive tools to assess metabolic syndrome risk in a sample of women with severe obesity.The predictive accuracy of the CMI and VAI indices in identifying metabolic syndrome risk was approximately 85%.Cai et al. [12] assessed the Body Roundness Index (BRI) to evaluate its effectiveness in predicting cardiovascular disease risk among hypertensive patients with obstructive sleep apnea.BRI had a predictive accuracy of approximately **88%**. Çelik et al. [13] analyzed the potential of the AST/ALT ratio as a diagnostic indicator for metabolically healthy obesity in a pediatric population and used receiver operating characteristic (ROC) curve analysis.The ROC analysis showed that the AST/ALT ratio had an accuracy rate of **82%.** Kumar et al. [14] explored predictive models for obesity levels by analyzing smoking habits.The study found random forest (**89%** accuracy) superior to stepwise linear regression (**82%**) in predicting obesity. Vera-Ponce et al. [15] conducted a systematic review and meta-analysis to examine the effectiveness of anthropometric measures, such as body mass index (BMI), waist circumference, and waist-to-hip ratio, in predicting prediabetes risk.The authors applied meta-analytic techniques, random-effects model, waist-to-hip ratio had accuracy of **87%**. Vecchiato et al. [16] examined various predictive models for cardiorespiratory fitness (CRF) specifically in patients with obesity by utilizing a cardiopulmonary exercise test registry.Approximately **90%** predictive accuracy for CRF in patients with obesity by using multiple regression analysis. Yagin et al. [17] presented a novel approach to estimating obesity levels by employing a trained neural network optimized through Bayesian techniques.Neural network achieved an impressive accuracy of approximately **92%** in estimating obesity levels. Hernandez et al. [18] investigated the relationship between continuous glucose monitor (CGM) metrics and neonatal adiposity, emphasizing how these metrics vary in pregnant individuals with obesity compared to those of normal weight. The authors employed a machine learning regression model, specifically a random forest algorithm with accuracy of approximately **85%**. Chen et al. [19] analyzed existing literature to assess the relationship between obesity and severe dengue manifestations.The researchers applied a random-effects meta-analysis model to pool data from the selected studies. Ferdowsy et al. [20] explored the application of machine learning techniques for predicting obesity risk, highlighting the increasing relevance of data-driven approaches in public health.Random forest achieved **92%** accuracy in predicting obesity risk.

### **Conclusion**

From the results, it is clear that **LightGBM** and **XGBoost** offer the best performance in predicting obesity levels after applying SMOTE, with **XGBoost** achieving the highest accuracy of **92%**. These models are highly effective in addressing class imbalances and outperforming traditional models like **Decision Tree**, **Logistic Regression**, and **Naive Bayes**. Furthermore, they provide a solid foundation for clinical applications, especially in resource-limited settings, where their interpretability and high accuracy could be leveraged for better obesity risk management and early intervention.

This study highlights the importance of utilizing advanced machine learning techniques, particularly ensemble methods like **XGBoost** and **LightGBM**, for improving the prediction and classification of obesity, contributing to enhanced health outcomes.

**References**

[1]Zarindokht Helforoush and H. Sayyad, “Prediction and classification of obesity risk based on a hybrid metaheuristic machine learning approach,” Frontiers in Big Data, vol. 7, Sep. 2024, doi: <https://doi.org/10.3389/fdata.2024.1469981>.

[2]N. Koklu and S. A. Sulak, “Using Artificial Intelligence Techniques for the Analysis of Obesity Status According to the Individuals’ Social and Physical Activities,” Sinop Üniversitesi Fen Bilimleri Dergisi, vol. 9, no. 1, pp. 217–239, Jun. 2024, doi: <https://doi.org/10.33484/sinopfbd.1445215>.

[3]Fadhila Tangguh Admojo and None Nurul Rismayanti, “Estimating Obesity Levels Using Decision Trees and K-Fold Cross-Validation: A Study on Eating Habits and Physical Conditions,” Indonesian Journal of Data and Science, vol. 5, no. 1, pp. 37–44, Mar. 2024, doi: <https://doi.org/10.56705/ijodas.v5i1.126>.

[4]J.-H. Jeong, I.-G. Lee, S.-K. Kim, T.-E. Kam, S.-W. Lee, and E. Lee, “DeepHealthNet: Adolescent Obesity Prediction System Based on a Deep Learning Framework,” *IEEE Journal of Biomedical and Health Informatics*, pp. 1–12, Jan. 2024, doi: <https://doi.org/10.1109/jbhi.2024.3356580>.

[5]A. Kumar, R. R, and K. Kumar, “Obesity Prediction Using Machine Learning Technique,” 2024 2nd International Conference on Networking and Communications (ICNWC), pp. 1–7, Apr. 2024, doi: <https://doi.org/10.1109/icnwc60771.2024.10537427>.

[6]Paulo and M. Lima, “Classification of Obesity Level Using Deep Neural Networks,” Lecture notes in networks and systems, pp. 99–107, Jan. 2024, doi: <https://doi.org/10.1007/978-3-031-64776-5_10>.

‌[7]F. Musa, F. Basaky, and O. E.O, “Obesity prediction using machine learning techniques,” Journal of Applied Artificial Intelligence, vol. 3, no. 1, pp. 24–33, Jun. 2022, doi: <https://doi.org/10.48185/jaai.v3i1.470>.

[8]R. C. Cervantes and U. M. Palacio, “Estimation of obesity levels based on computational intelligence,” Informatics in Medicine Unlocked, vol. 21, p. 100472, 2020, doi: <https://doi.org/10.1016/j.imu.2020.100472>.

[9]P. K. Mondal, K. H. Foysal, B. A. Norman, and L. S. Gittner, “Predicting Childhood Obesity Based on Single and Multiple Well-Child Visit Data Using Machine Learning Classifiers,” Sensors (14248220), vol. 23, no. 2, p. 759, Jan. 2023, doi: [https://doi.org/10.3390/s23020759.‌](https://doi.org/10.3390/s23020759.%E2%80%8C)

[10]Dahlak Daniel Solomon *et al.*, “Hybrid Majority Voting: Prediction and Classification Model for Obesity,” *Diagnostics*, vol. 13, no. 15, pp. 2610–2610, Aug. 2023, doi: <https://doi.org/10.3390/diagnostics13152610>.

[11]Stefano Lazzer *et al.*, “Cardiometabolic Index (CMI) and Visceral Adiposity Index (VAI) Highlight a Higher Risk of Metabolic Syndrome in Women with Severe Obesity,” *Journal of Clinical Medicine*, vol. 12, no. 9, pp. 3055–3055, Apr. 2023, doi: <https://doi.org/10.3390/jcm12093055>.

[12]X. Cai *et al.*, “Body roundness index improves the predictive value of cardiovascular disease risk in hypertensive patients with obstructive sleep apnea: a cohort study,” *Clinical and Experimental Hypertension*, vol. 45, no. 1, Oct. 2023, doi: <https://doi.org/10.1080/10641963.2023.2259132>.

[13]Nurullah Çelik, Gülşah Ünsal, and Hüseyin Taştanoğlu, “Predictive markers of metabolically healthy obesity in children and adolescents: can AST/ALT ratio serve as a simple and reliable diagnostic indicator?,” *European Journal of Pediatrics*, Oct. 2023, doi: <https://doi.org/10.1007/s00431-023-05296-3>.

[14]K. S. Kumar, M. Bee, and V. Thiruchelvam, “An analysis on obesity levels prediction based on smoking habits using stepwise linear regression algorithm in comparison with random forest classifier for improved accuracy,” *AIP conference proceedings*, vol. 3161, pp. 020242–020242, Jan. 2024, doi: <https://doi.org/10.1063/5.0229268>.

[15]Víctor Juan Vera-Ponce *et al.*, “Anthropometric Measures of Obesity as Risk Indicators for Prediabetes. A Systematic Review and Meta-Analysis,” *Diabetes Epidemiology and Management*, vol. 16, pp. 100230–100230, Jun. 2024, doi: <https://doi.org/10.1016/j.deman.2024.100230>.

[16]M Vecchiato *et al.*, “Cardiopulmonary exercise test registry for patients with obesity: comparison of cardiorespiratory fitness prediction equations and generation of a new predictive model for patients with obesity,” *European Journal of Preventive Cardiology*, vol. 31, no. Supplement\_1, Jun. 2024, doi: <https://doi.org/10.1093/eurjpc/zwae175.375>.

[17]F. H. Yagin *et al.*, “Estimation of Obesity Levels with a Trained Neural Network Approach optimized by the Bayesian Technique,” *Applied Sciences*, vol. 13, no. 6, p. 3875, Jan. 2023, doi: <https://doi.org/10.3390/app13063875>.

[18]T. L. Hernandez et al., “Continuous Glucose Monitor Metrics That Predict Neonatal Adiposity in Early and Later Pregnancy Are Higher in Obesity Despite Macronutrient-Controlled Eucaloric Diets,” Nutrients, vol. 16, no. 20, pp. 3489–3489, Oct. 2024, doi: <https://doi.org/10.3390/nu16203489>.

[19]V. M. Valicente *et al.*, “Ultraprocessed Foods and Obesity Risk: A Critical Review of Reported Mechanisms,” *Advances in Nutrition*, vol. 14, no. 4, Apr. 2023, doi: <https://doi.org/10.1016/j.advnut.2023.04.006>.

[20]F. Ferdowsy, K. S. A. Rahi, Md. I. Jabiullah, and Md. T. Habib, “A machine learning approach for obesity risk prediction,” *Current Research in Behavioral Sciences*, vol. 2, p. 100053, Nov. 2021, doi: <https://doi.org/10.1016/j.crbeha.2021.100053>.