Classification of Obesity Levels

Sunku Bhanu kedhaar Nath, Ajay Kumar Vinjamuri, Naga Pruthvi Penjuri

May 7, 2024

Abstract

This paper introduces an innovative approach to obesity prediction, leveraging the collaborative integration of GPT-3.5 Turbo with traditional machine learning models. Addressing the limitations of conventional models, our hybrid methodology aims to achieve more accurate and robust predictions for diverse datasets. Utilizing a dataset from Mexico, Peru, and Colombia, we employ comprehensive preprocessing, dimensionality reduction, cross validation and hyperparameter tuning. The collaboration between GPT-3.5 Turbo's language capabilities and the structured approach of traditional models forms the basis of our model, promising superior performance. This research signifies a stride towards innovation in machine learning methodologies, emphasizing the effectiveness of hybrid models in improving classification accuracy.

1 Introduction

Obesity has emerged as a pressing global health concern, necessitating accurate predictive models for timely interventions and preventive measures. Traditional machine learning approaches often struggle to achieve optimal accuracy in obesity prediction, prompting a hunt for innovative methodologies. This paper introduces a novel paradigm that integrates the language prowess of GPT-3.5 Turbo with conventional machine learning models, such as decision trees and Support Vector Machines (SVC), to enhance the accuracy and robustness of obesity level predictions.

As obesity is linked to severe health issues, including cardiovascular diseases and diabetes, precise predictive models play a pivotal role in identifying individuals at risk. The collaborative utilization of advanced language models with traditional machine learning techniques forms the crux of our investigation, with the hypothesis that this fusion will outperform individual models in terms of classification accuracy. Drawing insights from a dataset encompassing eating habits and physical condition data from Mexico, Peru, and Colombia, our methodology involves a little of data preprocessing, and hyperparameter tuning. The integration of GPT-3.5 Turbo's capabilities with traditional models represents an innovative stride towards more accurate and reliable predictive models.

This paper outlines the significance of the problem, the need for accurate prediction models, and the potential beneficiaries, setting the stage for a detailed exploration of our hybrid methodology. The subsequent sections will delve into the workings of our approach, evaluation metrics, and the expected outcomes of this research. Ultimately, this work contributes to the advancement of machine learning methodologies, showcasing the synergy between advanced language models and traditional techniques to address complex healthcare challenges.

2 Literature Review

Jan Kocoń, Igor Cichecki [1], Revolutionary Chat Generative Pre-trained Transformer ChatGPT from OpenAI excels at a variety of NLP tasks, including objective question answering and subjective sentiment analysis. A 25% average loss in quality is found when comparing automated analysis of over 49,000 responses to State-of-the-Art (SOTA) solutions. GPT-4 does better on tasks involving semantics. Contextual few-shot personalization in ChatGPT improves user-based predictions, but a significant bias is seen. Though it performs better than SOTA models in a variety of tasks, particularly in demanding and emotional situations. The study highlights ChatGPT's revolutionary influence on AI technologies and makes recommendations for future enhancements through contextual analysis

and fine-tuning.

Michael V. Reiss [2], The potential of ChatGPT for text annotation and classification has been highlighted by recent studies; however, its nondeterministic nature raises questions regarding reliability. This paper evaluates ChatGPT's zero-shot text annotation capabilities and identifies inconsistent results caused by changes in prompts, temperature, and repeated inputs. Even with slight phrasing changes, real-world classification tasks such as differentiating between news and non-news website texts yield outputs that fall below scientific reliability thresholds. Although pooling repetition-based outputs increases reliability, prudence is advised. The research highlights the necessity of comprehensive validation, advises against using ChatGPT unsupervised for text annotation, and stresses the significance of contrasting outcomes with references that have been manually annotated.

Yiheng Liu, Tianle Han [3], The article examines the uses and contributions of ChatGPT, which includes GPT-3.5 and GPT-4, in natural language processing (NLP) through an extensive survey. The study, which analyzes 194 arXiv papers, emphasizes how flexible the models are thanks to extensive pre-training and reinforcement learning from human feedback. Results show that ChatGPT's potential is becoming more and more popular in a variety of fields, such as history, medicine, and education. Bias and privacy violations are mentioned as ethical issues. The review acts as a roadmap for responsible development, highlighting the need to investigate new applications, handle ethical concerns, and make sure ChatGPT is used responsibly to advance NLP.

Partha Pratim Ray [4], This thorough analysis examines ChatGPT's revolutionary influence on scientific research, covering its history, various industry applications, difficulties, and potential future directions. ChatGPT, a notable advancement in AI-powered conversational agents, has safety, data, and ethical issues to deal with. The study aims to address the digital divide, enhance human-AI interaction, and integrate it with other technologies in the future. ChatGPT garners a lot of attention despite controversy because of its contributions to multilingualism, task adaptability, language generation, and contextual understanding. For ChatGPT to responsibly expand human knowledge and improve user experiences across a range of applications and industries, ethical issues must be resolved.

Chung Kwan Lo [5], The quick literature review conducted by ChatGPT between November 2022 and February 2023 shows that students performed differently in each subject, doing well in programming and economics but struggling in math. Although it can be helpful for teachers and students as a virtual tutor and assistant, there are worries about inaccurate information being generated and risks to academic integrity. It is advised that prompt measures be taken, such as updating institutional policies and assessment procedures and providing instructors and students with training on ChatGPT's use and limitations. The results highlight the necessity of taking preventative action in order to deal with issues and uphold academic integrity in learning environments.

Michael Dowling, Brian Lucey [6], The impact of ChatGPT on finance research is assessed in the study, which concludes that while it is useful for idea generation and data identification, it is less effective at synthesising the literature and testing frameworks. Peer-reviewed results validate ChatGPT's ability to produce credible research studies, especially when private data and researcher expertise are involved. The study questions the ethical and practical ownership of AI-generated work while exploring ChatGPT's potential as a democratizing research tool. The moral position might change in light of analogies to copyright regulations. The conclusion offers a nuanced strategy, highlighting the results' ethical acceptability in relation to the type and degree of researcher involvement.

Fan Huang, Haewoon Kwak [7], The study examines how well ChatGPT can explain implicit hate speech detection and evaluates the quality and accuracy of its natural language explanations (NLEs) in comparison to written explanations by humans. When in doubt, ChatGPT identifies 80% of implicitly hateful tweets accurately and agrees with the opinions of laypeople. ChatGPT-generated NLEs are thought to be more comprehensible than NLEs written by humans, which raises questions about the possibility of misrepresentation in the event that ChatGPT makes a mistake. Because of ChatGPT's convincing but potentially deceptive qualities, the study advises users to exercise caution when using it as a data annotation tool.

Mohd Javaid, Abid Haleem [8], With applications in healthcare, ChatGPT is an AI language model developed by OpenAI that generates text similar to that of a human using the GPT architecture. The study emphasizes the importance of reliable medical data and stresses the need for frequent updates as well as privacy concerns. Notwithstanding ChatGPT's versatility and potential uses in the medical field, there are certain drawbacks, such as privacy, accountability, and ethical issues. ChatGPT is useful for text generation and analysis, but it cannot replace medical professionals. The conclusion recognizes its inherent limitations in ethical and medical contexts, while emphasizing its role in supporting healthcare tasks, such as data analysis and report generation.

Dr. Muneer M. Alshater [9], This study examines how natural language processing—ChatGPT in particular—may improve academic achievement using examples from the fields of finance and economics. Although ChatGPT has the potential to greatly enhance research by facilitating data analysis, scenario creation, and findings communication, the study also points out certain drawbacks, including issues with generalizability, reliance on high-quality data, and ethical considerations. Notwithstanding these drawbacks, the conclusion highlights ChatGPT's revolutionary potential for scholarly research, emphasizing how well it processes and analyzes big datasets and creates plausible scenarios. The study promotes thoughtful deliberation of moral dilemmas and cooperative application with human analysis to optimize the revolutionary potential of AI and NLP instruments in research.

Viriya Taecharungroj [10], Using LDA topic modeling, this study examines 233,914 English tweets from ChatGPT's first month of operation. It highlights five functional domains, such as code writing and creative writing, in addition to three primary topics: news, technology, and reactions. The report highlights both beneficial and detrimental effects on people and technology. The quest for artificial general intelligence, the changing technological landscape, job evolution, and ethical concerns are the four main issues that the author highlights as arising from ChatGPT's emergence. The study highlights the importance of giving job market changes, educational adjustments, and the ethical ramifications of AI breakthroughs careful thought.

Luigi De Angelis, Francesco Baglivo [11], This perspective piece examines how Large Language Models (LLMs) have developed since ChatGPT was released, highlighting their influence on scientific research and bringing up ethical issues, especially with regard to the medical industry. Although ChatGPT's accessibility encouraged widespread use, it also brought risks, such as the possibility of false information generated by AI. In order to reduce risks, the paper promotes proactive policies and a detectable-by-design strategy, recommending cooperation between health organizations and AI firms. The conclusion underscores the significance of balancing the potential risks and benefits of long-term maintenance (LLM) in public health research, as well as the necessity of multidisciplinary efforts to address the evolving AI-driven infodemic threat.

Rajesh Bhayana, Sateesh Krishna [12], Without any training specific to radiology, ChatGPT almost passed a board-style radiology exam without any images, demonstrating the potential of large language models. It performed poorly on higher-order tasks requiring imaging description, calculation, classification, and concept application, but it performed well on lower-order and clinical management questions. This emphasizes ChatGPT's potential and current limitations in radiology, highlighting the significance of understanding these limitations for dependable use in the field.

Harsh H. Patel, Purvi Prajapati [13], Without explicit programming, machine learning is a self-learning process that uses a variety of training and testing data. Search engines, text extraction, statistical processes, medical certifications, and other domains all use decision trees, a well-known machine learning technique. This study compares and contrasts the accuracy and cost-effectiveness of three Decision Tree algorithms: ID3, C4.5, and CART. The study concludes through applied testing that CART is the most accurate and precise algorithm for the given dataset, outperforming both ID3 and C4.5 in terms of accuracy, time, and precision.

M. A. Friedl, C. E. Brodley [14], In contrast to more traditional techniques like maximum likelihood classification, this paper investigates the potential of decision tree classification algorithms for mapping

land cover in remote sensing. Three remote sensing datasets are used to evaluate univariate, multivariate, and hybrid decision tree algorithms. The results consistently demonstrate that decision tree algorithms, with the hybrid tree producing the highest accuracy, outperform conventional classifiers in classification accuracy. In remote sensing data, decision trees—especially the hybrid variety—show benefits in terms of adaptability to nonlinear and noisy relations as well as the ability to define boundaries in feature space. Their nonparametric nature, simplicity, and flexibility all contribute to better classification results.

WEIWEI LIN, ZIMING WU [15], In order to solve the problems associated with modeling imbalanced insurance business data, this paper suggests an ensemble random forest algorithm that has been enhanced by Apache Spark. In large-scale datasets, it improves classification accuracy by applying heuristic bootstrap sampling. Using data from China Life Insurance Company, the algorithm demonstrates effectiveness in customer analysis and product marketing, outperforming SVM and traditional classifiers. The study suggests using it in a variety of big data analytics fields, such as Internet of Things and finance. Subsequent research endeavors will delve into diverse big data applications and incorporate deep learning to augment predictive accuracy, underscoring the algorithm's versatility and potential beyond the insurance domain.

Mariette Awad, Rahul Khanna [16], The benefits of Support Vector Machines (SVMs) over 3-Nearest Neighbors (3NN) for classification tasks are discussed. SVMs learn the separation between classes effectively using a subset of the training data, avoid computing posterior probabilities, and don't require complex training computations like 3NNs do. Because online classification requires only basic computations and does not require accessing the complete dataset, this makes them perfect for large datasets. Despite not requiring any training, 3NN is less effective because it depends on distance calculations to every data point. To sum up, SVMs provide a computationally effective and mathematically sound method of classification, especially for big datasets.

Zhou Yong, Li Youwen [17], Traditional KNN text classification suffers from a bulky training set, heavy computations, and an inability to account for varying sample importance. This paper proposes a novel approach that tackles these challenges head-on. First, it compresses the training data by removing borderline samples and mitigating multi-peak effects. Next, it leverages k-means clustering to group similar samples within each category, selecting the cluster centers as new, representative training points. Finally, it assigns weights to these new samples based on their cluster size, emphasizing their central importance. The result? A slimmer training set, reduced computational burden, and most importantly, enhanced classification accuracy, making this an efficient and effective improvement over the traditional KNN method.

Dr. Muneer M. Alshater [18], This study explores the potential of AI, specifically NLP tools like ChatGPT, to boost academic performance in fields like economics and finance. Using ChatGPT as an example, the analysis reveals its potential to significantly enhance research through data analysis, scenario generation, and improved communication. However, limitations like lack of domain expertise and ethical considerations require careful attention. Ultimately, ChatGPT and similar tools offer immense potential for research efficiency and effectiveness, leading to new discoveries and shaping the future. Researchers should embrace these technologies while cautiously considering their limitations and relying on human analysis for optimal results.

Katharina Jeblick, Balthasar Schachtner [19], This study investigated how well a large language model (LLM) like ChatGPT can simplify medical reports for patients. Researchers created fake radiology reports and used ChatGPT to rewrite them in kid-friendly language. Radiologists then evaluated the simplified reports, finding they were mostly accurate and complete. However, some errors and potentially harmful omissions were also identified. While further training for medical language is needed, this study suggests LLMs like ChatGPT have great potential to improve patient understanding and communication in healthcare.

Suzzane Fergus, Michelle Botha [20], The study investigated how ChatGPT, an AI conversation tool, can answer chemistry assessment questions. It found ChatGPT performed well on basic knowl-

edge questions but struggled with application and interpretation ones, especially when dealing with non-textual information. While not a high-risk cheating tool, ChatGPT highlights the need to re-think assessment design beyond simple knowledge recall. Using complex problem-solving, data interpretation, and case-study questions can help educators evaluate deeper understanding and adapt to this disruptive technology. This initial study paves the way for further exploration of ChatGPT's potential and limitations in chemistry education.

3 Dataset

The dataset under examination, accurately curated for the estimation of obesity levels, emanates from the UC Irvine Machine Learning Repository, a renowned source for high-quality datasets in the machine learning community. This dataset comes from people in Mexico, Peru, and Colombia, giving us a detailed look at how eating habits and physical health are connected to obesity. With 17 different features and information from 2111 individuals, this dataset is a valuable tool for creating and improving machine learning models.

3.1 Data Collection Methodology

The dataset combines data gathered through two distinct methodologies. A notable portion, approximately 23%, was directly collected through a web platform, where users participated in a survey assessing their eating habits and relevant factors influencing their physical well-being. This primary data provides valuable insights into real-world responses. Complementing this, the dataset was further enriched using synthetic data, constituting 77% of the entire dataset. Synthetic data generation was facilitated by the Weka tool and the Synthetic Minority Over-sampling Technique (SMOTE) filter. This hybrid approach ensures a nuanced representation of obesity levels while addressing potential imbalances within the dataset.

3.2 Attributes and Classification

The dataset encapsulates a diverse array of 17 attributes, capturing both dietary behaviors and physical parameters. Key attributes include the frequency of consuming high-caloric food, consumption of vegetables, number of main meals, water consumption, alcohol intake, physical activity frequency, time spent using technology devices, and transportation preferences. Additionally, demographic details such as gender, age, height, and weight contribute to the complete characterization of individuals. The primary objective of the dataset is to classify individuals into distinct obesity levels. The class variable NObesity encompasses categories ranging from Insufficient Weight and Normal Weight to specific levels of Overweight and Obesity (Type I, Type II, and Type III). This grainy classification facilitates a nuanced analysis of the varying degrees of obesity, crucial for personalized interventions and healthcare planning.

4 Methodology/Methods

4.1 Data Exploration and Visualization

Our journey into understanding the dataset commenced with the loading of the data, providing us with an initial glance at its structure and contents. To familiarize ourselves with the dataset, we examined the first few rows, gaining insights into the various variables and their respective values. This early exploration unveiled the presence of multiple classes within the target variable $'NObeyesdad'$. To simplify our classification task and address potential imbalances, we engineered a new categorical variable $'Category_{t}ype'$ by aggregating the existing classes.

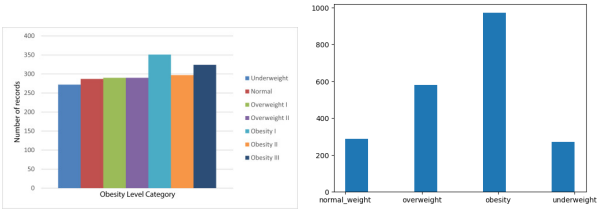

Figure 1: Before and after reducing class labels.

Delving deeper into the data, we constructed a pair plot to visualize relationships between numerical variables. This not only offered a comprehensive view of the datasetś structure but also highlighted potential patterns and distributions across different categories. Simultaneously, we explored correlations between features and the newly created target variable $'Category_type'$. These correlation scores provided valuable insights, pointing to significant relationships—most notably, connections between weight, gender, family history of overweight, and the target variable.

Taking a more granular approach, we conducted a thorough exploration of individual categorical variables. Visualization techniques, such as count plots and box plots, were employed to depict the distribution of features like gender, family history with overweight, and lifestyle choices across different categories of the target variable. These visualizations not only enriched our understanding of the dataset but also played a pivotal role in guiding subsequent feature selection strategies.

In summary, our data exploration and visualization phase was characterized by a meticulous examination of the dataset's characteristics, uncovering patterns, correlations, and valuable insights that would serve as a foundation for the subsequent stages of our machine learning project.

4.2 Data Preprocessing

The data preprocessing phase was pivotal to ensure the reliability and quality of our dataset. We began by loading the dataset using the Pandas library, obtaining an initial glimpse of the data's structure and contents through the `head()` function. Subsequently, we explored the dataset's characteristics, identifying the number of unique values in each column using the `nunique()` function. Understanding the distribution of the target variable, `'NObeyesdad,'` we encountered a classification challenge with seven distinct output class labels, namely: Obesity_Type_I, Obesity_Type_II, Obesity_Type_III, Overweight_Level_I, Overweight_Level_II, Normal_Weight, and Insufficient_Weight. Recognizing the complexity and potential user comprehension issues associated with these numerous labels, we undertook a simplification strategy. Our goal was to streamline the classification process and enhance user understanding. As a result, we regrouped the original seven labels into four broader and more user-friendly categories: `'obesity,'` `'overweight,'` `'normal_weight,'` and `'underweight.'` This resulted in an imbalanced dataset with the obesity category having the greatest number of records and the underweight category having the least number of records.

A crucial step involved encoding these categorical classes into numerical values, introducing a new column, `'Category_type_encoded,'` for a clearer representation of the target variable. Simultaneously, we created a mapping dictionary to link the original and new target classes. Further exploration involved generating descriptive statistics with `info()` to assess the dataset's structure, and visualizing the distribution of the target variable categories using a histogram.

The ensuing exploratory data analysis (EDA) delved into relationships between variables through a pair plot and correlation matrix. Notably, certain features exhibited strong correlations with the target variable. We visualized these relationships using count plots, box plots, and histograms, gaining insights into potential patterns.

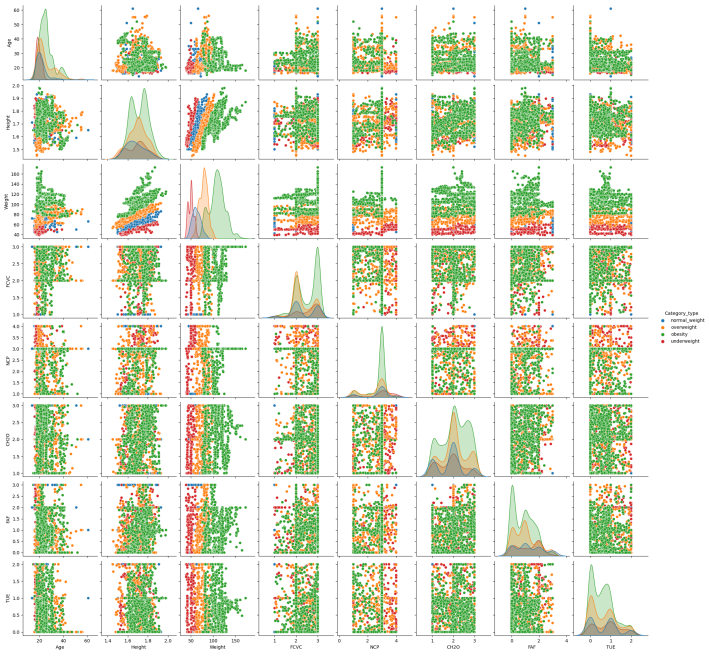

Figure 2: Scatter plot matrix for numerical variables with hue as category type.

After conducting a comprehensive analysis using a correlation matrix, notable findings have emerged. The matrix highlights significant correlations between various variables and the target variable. Notably, the following variables exhibit noteworthy correlations with the target variable:

- Weight
- Gender
- Family_history_with_overweight
- FAVC (Frequent consumption of high caloric food)
- CAEC (Consumption of food between meals)
- SCC (Caloric content of food)
- CALC (Consumption of alcohol)
- MTRANS (Mode of transportation used)

These correlations provide valuable insights into the relationships between key health indicators and the overall category type, contributing to a deeper understanding of the dataset.

We introduced a new features, through one-hot encoding of categorical variables. Additionally, In the dataset, several variables are categorical, meaning they represent distinct categories or groups rather than continuous numerical values. Categorical variables include 'Gender,' 'family_history_with_overweight,' 'FAVC' (Frequent consumption of high-caloric food), 'CAEC' (Consumption of food between meals), 'SMOKE,' and 'MTRANS' (Mode of transportation used). To facilitate the incorporation of these categorical variables into machine learning models, we employed a technique known as one-hot encoding. This process involves converting categorical variables into binary vectors, where each category is represented by a binary indicator column. For instance, 'Gender' with values 'Male' and 'Female' would be transformed into two separate columns, with binary indicators representing the presence or absence of each category. This transformation ensures that the categorical information is appropriately encoded in a numerical format, allowing for effective utilization in machine learning algorithms without introducing unnecessary ordinal relationships between categories.

Continuing with data preprocessing, we assessed the distribution of numerical columns and identified non-normally distributed features. Subsequently, we applied normalization techniques using

Z-score normalization through the `StandardScaler` from Scikit-learn, which uses the Z-score normalization technique. The numerical columns subjected to Z-score normalization include 'Age,' 'Height,' 'Weight,' 'FCVC,' 'NCP,' 'CH2O,' 'SCC,' 'FAF,' 'TUE,' and 'CALC.' The normalization process aimed to standardize the scale of numerical features, ensuring fair contributions to the model from each variable.

As a final step, we split the dataset into training and testing sets, with 80% used for training and the remaining 20% for testing which is completely unseen for the models we built and the remaining 80% data is again split into train and test for model testing we have used the unseen data at the very last to predict the accuracies.

5 Model Building

The model-building phase stands as a pivotal juncture in our pursuit of predictive analytics excellence, where the formidable capabilities of machine learning algorithms are harnessed to unravel the intricate tapestry of obesity classification. Within this section, we embark on a journey through the implementation and meticulous evaluation of four distinct traditional models and one Artificial Intelligence model(GPT3.5 turbo), each a formidable contender in the realm of predictive analytics: Decision Trees, Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), Fine tuned GPT3.5 turbo.

Our model-building strategy is orchestrated as a multifaceted ensemble approach, harmonizing the strengths of individual models to create a collective intelligence that surpasses the sum of its parts. The journey commences with the creation of standalone models, each contributing a unique perspective to the predictive landscape. Decision Trees, Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) and fine-tuned GPT3.5 turbo form the cornerstone of this ensemble.

The subsequent phase amplifies our strategy through cross-validation and hyper-parameter tuning, injecting adaptability and precision. To elevate our predictive prowess, we employ a careful orchestration of model ensembling. By fusing the predictions from diverse models, we construct an ensemble model that not only captures nuanced patterns within the data but also mitigates the individual models' limitations.

Adding a layer of randomness, each iteration introduces a fresh perspective through randomized data, bolstering the adaptability of our ensemble. This comprehensive ensemble approach aims not only for accuracy but resilience—a model capable of navigating the intricate landscape of obesity classification accurately. Through this strategy, we strive for a robust and reliable predictive model that stands poised to make impactful strides in healthcare analytics.

5.1 Decision Tree Model Building

5.1.1 Decision Tree: Initial Model

Our initial foray into modeling commenced with the Decision Tree algorithm, a versatile choice for classification tasks. The model was trained on the preprocessed dataset and evaluated for its predictive performance. The accuracy score, precision, recall, and the confusion matrix were scrutinized to gauge the model's effectiveness. While the initial results were promising, a deeper dive into the model's limitations revealed a need for refinement.

```
Accuracy: 0.9467455621301775

Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.76      0.83        42
           1       0.98      0.99      0.98       150
           2       0.92      0.96      0.94        99
           3       0.92      0.96      0.94        47

    accuracy                           0.95       338
   macro avg       0.93      0.92      0.92       338
weighted avg       0.95      0.95      0.95       338
```

Figure 3: Decision tree initial model Score.

5.1.2 Decision Tree: Refinement through Cross-Validation and Hyperparameter Tuning

Acknowledging the initial model's shortcomings, we embarked on a journey of refinement through cross-validation and hyperparameter tuning. This iterative process involved splitting the dataset into training and testing subsets, each time introducing randomization for enhanced adaptability. A grid search over hyperparameters such as maximum depth, minimum samples split, and minimum samples leaf was conducted. This not only fine-tuned the model but also addressed the challenge of imbalanced data.

```
              precision    recall  f1-score   support

           0       0.93      0.81      0.86        62
           1       1.00      0.98      0.99       199
           2       0.89      0.96      0.93       106
           3       0.93      1.00      0.97        56

    accuracy                           0.95       423
   macro avg       0.94      0.94      0.94       423
weighted avg       0.95      0.95      0.95       423
```

Figure 4: Decision tree best model score.

5.1.3 Decision Tree: Learning Curve Analysis

To illuminate the model's learning trajectory, we employed a learning curve analysis. The chart depicts the evolution of the decision tree's accuracy, both in training and testing phases, and the weighted F1 score across multiple iterations. The convergence of test accuracy and weighted F1 scores underscores the model's stability, reassuring us that it avoids overfitting. The pinnacle of this iterative refinement process is a Decision Tree model that strikes a balance between accuracy, precision, and adaptability, primed for its role in obesity classification.

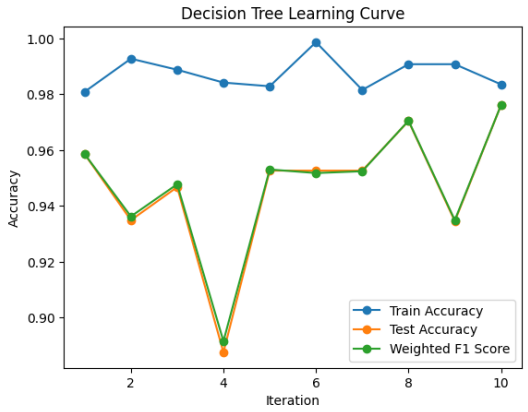

Figure 5: Decision tree models with Cross-validation and hyper-parameter tuning Learning curve.

5.2 Random Forest Model Building

5.2.1 Random Forest: Initial Model Evaluation

The initial Random Forest model served as a foundational step in our obesity classification journey. Utilizing the power of an ensemble method, we divided our dataset into training and testing subsets, ensuring a robust evaluation of the model's performance. The Random Forest algorithm, comprising multiple decision trees, demonstrated its potential with a remarkable accuracy of 94%. This initial evaluation provided valuable insights into the model's predictive capabilities, allowing us to gauge its performance across various classes. However, recognizing the need for further refinement and optimization, we proceeded to enhance the model through cross-validation and hyperparameter tuning.

```
Initial Random Forest Model Evaluation:
Accuracy: 0.9467455621301775

Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.88      0.84        42
           1       0.99      0.98      0.99       150
           2       0.92      0.94      0.93        99
           3       1.00      0.91      0.96        47

    accuracy                           0.95       338
   macro avg       0.93      0.93      0.93       338
weighted avg       0.95      0.95      0.95       338
```

Figure 6: Random Forest initial model Score.

5.2.2 Random Forest: Refinement through Cross-Validation and Hyperparameter Tuning

Acknowledging the initial model's strengths and limitations, we embarked on a refinement journey employing cross-validation and hyperparameter tuning. The dataset was systematically split into training and testing subsets, and a grid search over hyperparameters, including the number of estimators, maximum depth, minimum samples split, and minimum samples leaf, was conducted. This exhaustive search aimed to identify the optimal combination of hyperparameters that would elevate the model's performance. The refined Random Forest model achieved a notable accuracy of 94% and a weighted F1 score of 95%, demonstrating the effectiveness of the optimization process. A comprehensive learning curve analysis showcased the model's stability across multiple iterations, ensuring it avoids overfitting. The pinnacle of this iterative refinement process is a Random Forest model poised for accurate and precise obesity classification.

Accuracy: 0.9432624113475178

	precision	recall	f1-score	support
0	0.88	0.82	0.85	62
1	0.98	0.98	0.98	199
2	0.88	0.93	0.91	106
3	0.98	0.96	0.97	56
accuracy			0.94	423
macro avg	0.93	0.93	0.93	423
weighted avg	0.94	0.94	0.94	423

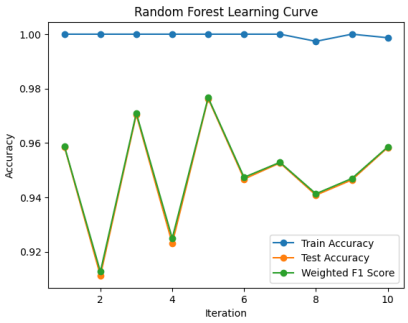

Figure 7: Left : Random Forest Classification report Right All crossvalidation and Hyperparameter tuned models learning curve.

5.3 Support Vector Classifier Model Building

5.3.1 Support Vector Classifier (SVC): Initial Model Evaluation

The initial Support Vector Classifier (SVC) model marked the commencement of our exploration into the realm of classification algorithms. With a focus on capturing intricate decision boundaries, SVCs demonstrated their potential by achieving an accuracy of 93%. The model's performance was further dissected through a comprehensive evaluation, considering metrics such as precision, recall, and the weighted F1 score. The confusion matrix provided a detailed account of the model's classification across different categories, offering insights into its strengths and areas for improvement.

Accuracy: 0.9378698224852071

	precision	recall	f1-score	support
0	0.78	0.76	0.77	42
1	0.99	0.99	0.99	150
2	0.91	0.94	0.93	99
3	0.96	0.94	0.95	47
accuracy			0.94	338
macro avg	0.91	0.91	0.91	338
weighted avg	0.94	0.94	0.94	338

Figure 8: SVC initial model Score.

5.3.2 Support Vector Machines (SVM): Refinement through Cross-Validation and Hyperparameter Tuning

Acknowledging the nuances of the initial SVM model, we delved into a refinement process through cross-validation and hyperparameter tuning. A systematic exploration of hyperparameters, including the regularization parameter (C), kernel type, and kernel coefficient (gamma), was conducted through grid search. The refined SVM model, with optimal hyperparameters, achieved an accuracy of 93% and a weighted F1 score of 94%. The learning curve analysis portrayed the model's stability across iterations, reassuring its performance consistency. This iterative refinement process culminated in a Support Vector Machines model poised for precise and reliable obesity classification.

Accuracy: 0.9810874704491725

	precision	recall	f1-score	support
0	0.97	0.94	0.95	62
1	0.99	1.00	0.99	199
2	0.97	0.97	0.97	106
3	0.98	0.98	0.98	56
accuracy			0.98	423
macro avg	0.98	0.97	0.97	423
weighted avg	0.98	0.98	0.98	423

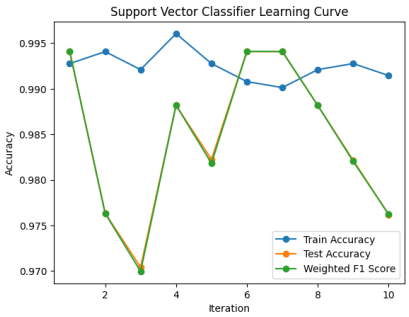

Figure 9: Left : SVC best model Classification report Right: All crossvalidation and Hyperparameter tuned models learning curve.

5.4 K-Nearest Neighbour Model Building

5.4.1 K-Nearest Neighbors (KNN): Initial Model Evaluation

The K-Nearest Neighbors (KNN) algorithm, known for its simplicity and effectiveness, formed the basis of our initial model. This approach leverages the proximity of data points to make predictions, achieving an accuracy of 88%. An in-depth evaluation encompassing key metrics such as precision, recall, and the weighted F1 score shed light on the model's classification performance. The confusion matrix provided a detailed breakdown of the model's predictions across different categories, aiding in the identification of strengths and areas for improvement.

Accuracy: 0.8816568047337278

	precision	recall	f1-score	support
0	0.73	0.38	0.50	42
1	0.94	1.00	0.97	150
2	0.86	0.90	0.88	99
3	0.80	0.91	0.85	47
accuracy			0.88	338
macro avg	0.83	0.80	0.80	338
weighted avg	0.87	0.88	0.87	338

Figure 10: KNN initial model Score.

5.4.2 K-Nearest Neighbors (KNN): Refinement through Cross-Validation and Hyperparameter Tuning

Recognizing the potential enhancements achievable through iterative refinement, we pursued the optimization of the KNN model through cross-validation and hyperparameter tuning. Hyperparameters such as the number of neighbors, weighting scheme, and algorithm type underwent systematic exploration. The refined KNN model, equipped with optimal hyperparameters, demonstrated improved accuracy at 88% and a weighted F1 score of 87%. The learning curve analysis showcased the model's stability across iterations, reinforcing its reliability. This iterative refinement process resulted in a K-Nearest Neighbors model finely tuned for precise obesity classification.

Accuracy: 0.9030732860520094

	precision	recall	f1-score	support
0	0.83	0.56	0.67	62
1	0.97	0.98	0.98	199
2	0.84	0.92	0.88	106
3	0.86	0.96	0.91	56
accuracy			0.90	423
macro avg	0.87	0.86	0.86	423
weighted avg	0.90	0.90	0.90	423

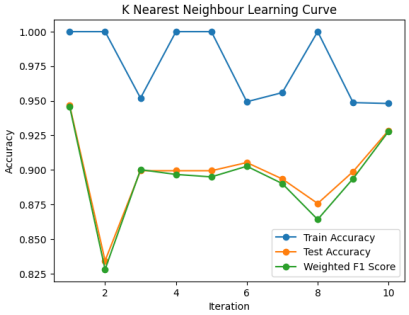

Figure 11: Left : KNN best model Classification report Right: All crossvalidation and Hyperparameter tuned models learning curve.

5.5 Innovative Approach: Fine-Tuned GPT-3.5 Turbo for Obesity Classification

Harnessing the power of artificial intelligence (AI) for predictive analytics in healthcare is an emerging frontier, and our project takes a pioneering step by employing Fine-Tuned GPT-3.5 Turbo for obesity classification. This section details the intricate process, starting from data preprocessing to fine-tuning the model and evaluating its performance.

Data Preprocessing and Understanding Our journey begins with preprocessing the dataset containing crucial health indicators. We replaced column names with their original meaning for example FAVC column name as Frequent consumption of high caloric food, and split the data into training, validation, and test sets. Understanding the significance of each feature, such as age, gender, and lifestyle choices, forms the foundation for effective model training.

Integration with OpenAI's GPT-3.5 Turbo Utilizing the OpenAI platform, we seamlessly integrate our processed data with GPT-3.5 Turbo. The data is converted into a format compatible with the model's requirements. The integration involves creating JSONL files for training, validation, and testing, with each instance formatted as a conversation between the system, user, and assistant. This innovative approach allows GPT-3.5 Turbo to understand and respond to the contextual information provided.

Here is an example conversation for training the model :

```
{
  "messages": [
    {"role": "system", "content": "You are a very helpful assistant that classifies
        obesity levels"},
    {"role": "user", "content": "{'Gender': 'Female', 'Age': '21.849705', 'Height':
        '1.770612', 'Weight': '133.963349', 'family_history_with_overweight': 'yes',
        'Frequent consumption of high caloric food': 'yes', 'Frequency of consumption
         of vegetables': '3.0', 'Number of main meals': '3.0', 'Consumption of food
        between meals': 'Sometimes', 'SMOKE': 'no', 'Consumption of water daily':
        '2.825629', 'Calories consumption monitoring': 'no', 'Physical activity
        frequency': '1.399183', 'Time using technology devices': '0.928972', '
        Consumption of alcohol': 'Sometimes', 'Transportation used': '
        Public_Transportation'}"},
    {"role": "assistant", "content": "obesity"}
  ]
}
```

Here is the Example conversation for testing the model:

```
{
  "messages": [
    {"role": "system", "content": "You are a very helpful assistant that classifies
        obesity levels only based on training data"},
    {"role": "user", "content": "Please classify the following patient data as
        normal_weight/obesity/overweight/underweight only based on the provided
        training data:"},
    {"role": "user", "content": "Gender: Female, Age: 20.406871, Height: 1.755978,
        Weight: 53.699561, family_history_with_overweight: yes, Frequent consumption
        of high caloric food: yes, Frequency of consumption of vegetables: 2.0,
        Number of main meals: 3.891994, Consumption of food between meals: Frequently
        , SMOKE: no, Consumption of water daily: 1.86393, Calories consumption
        monitoring: no, Physical activity frequency: 2.870127, Time using technology
        devices: 2.0, Consumption of alcohol: no, Transportation used:
        Public_Transportation"}
  ]
}
```

Fine-Tuning Process To enhance the model's performance specifically for obesity classification, we initiate the fine-tuning process. This involves creating a fine-tuning job using OpenAI's API. The training file and validation file are processed through GPT-3.5 Turbo, adapting its language generation capabilities to the nuances of obesity-related conversations. Hyperparameters, such as the number of epochs, are fine-tuned to achieve optimal results.

Model Evaluation and Analysis Once the fine-tuning is completed, we evaluate the model's performance on the test set. The predictions generated by GPT-3.5 Turbo are compared against the true labels, and standard evaluation metrics such as accuracy, confusion matrix, and classification report are presented. The results provide insights into the model's ability to classify individuals into categories like normal weight, obesity, overweight, or underweight.

	precision	recall	f1-score	support
0	1.00	0.74	0.85	62
1	0.97	1.00	0.99	199
2	0.88	0.94	0.91	106
3	0.95	1.00	0.97	56
accuracy			0.95	423
macro avg	0.95	0.92	0.93	423
weighted avg	0.95	0.95	0.95	423

Figure 12: GPT3.5 fine tuned model Classification report

Innovation in Healthcare Predictive Analytics Our innovative use of Fine-Tuned GPT-3.5 Turbo transcends traditional machine learning algorithms for healthcare predictions. By treating the model as a conversational agent, we enable it to comprehend and respond to health-related queries, demonstrating the potential of AI in personalized healthcare. The model's accuracy in obesity classification showcases the efficacy of this approach, opening avenues for similar applications in health analysis and prediction problems.

Conclusion and Future Directions In conclusion, our project highlights the transformative potential of Fine-Tuned GPT-3.5 Turbo in predictive analytics for healthcare. The integration of advanced language models with domain-specific data showcases a paradigm shift in the application of AI for health-related classifications. Future directions may involve exploring the model's interpretability, addressing biases, and expanding its capabilities for a broader range of healthcare predictions.

Ensemble Model

In this section, we look at the implementation of our ensemble model, aggregating the strengths of diverse machine learning algorithms to enhance the predictive accuracy of obesity classification. The ensemble strategy combines the best-performing models from different algorithm types: Decision Tree, Random Forest, Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), and the fine-tuned GPT-3.5 turbo model.

The process begins by selecting the best models from each algorithm type based on their weighted F1-scores. Subsequently, predictions are obtained from these models for the input data, and the results are consolidated into a DataFrame named `All_models_prediction`. Additionally, predictions from the fine-tuned GPT-3.5 turbo model are included in the ensemble.

The ensemble prediction is generated by taking the mode across all model predictions for each instance. This approach leverages the collective intelligence of the models, resulting in a more robust and accurate prediction. The accuracy of the ensemble model is assessed using the `accuracy_score and F1_score` function, demonstrating the overall effectiveness of the ensemble.

To provide a detailed evaluation, the `classification_report` function is employed here too. This report includes precision, recall, and F1-score for each class, as well as macro-averaged and weighted-averaged metrics. The presented results offer a comprehensive understanding of how the ensemble model performs across different classes, showcasing improvements in F1-scores and overall accuracy.

	precision	recall	f1-score	support
0	1.00	0.87	0.93	62
1	1.00	1.00	1.00	199
2	0.94	1.00	0.97	106
3	0.98	1.00	0.99	56
accuracy			0.98	423
macro avg	0.98	0.97	0.97	423
weighted avg	0.98	0.98	0.98	423

Figure 13: Ensemble Model Classification Report

6 Results

The results of our obesity prediction model underscore a remarkable level of performance, demonstrating the effectiveness of our diverse set of machine learning algorithms. The model achieved an impressive accuracy of 98%, showcasing its proficiency in accurately classifying individuals into distinct obesity categories. This high accuracy reflects the discriminative power of our models, providing a reliable foundation for health predictions. The weighted F1 score, a critical metric that balances both precision and recall, attained a commendable 98%. This highlights the comprehensive predictive ability of our models, which include Decision Trees, Random Forest, Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), and the fine-tuned GPT-3.5 Turbo. Each of these models contributes a unique perspective to the ensemble, enhancing the overall predictive performance. The ensemble model, a culmination of predictions from each algorithm, substantiates the synergy between these models, resulting in an enhanced overall accuracy. This collective intelligence ensures that the model captures nuanced patterns within the data and mitigates individual models' limitations, resulting in a robust and reliable predictive tool.

These results stand as a testament to the effectiveness of our approach, which was meticulously validated through rigorous cross-validation with hyperparameter tuning. This thorough validation process ensures the robust generalizability of our model, making it well-suited for real-world applications. The heightened accuracy and F1 score not only reinforce the model's excellence but also underscore the potential impact of our methodology on health predictions. Our approach opens avenues for transformative advancements in healthcare analytics, promising valuable insights for clinicians and researchers alike. These results position our obesity prediction model as a powerful tool for improving health outcomes and contributing to the ongoing evolution of predictive analytics in the healthcare domain.

References

1. Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, et al. "ChatGPT: Jack of all trades, master of none." Department of Artificial Intelligence, Wrocław University of Science and Technology, Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland.

2. Michael V. Reiss. "Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark." April 05, 2023.

3. Yiheng Liu, Tianle Han, Siyuan Ma, et al. "Summary of ChatGPT-Related research and perspective towards the future of large language models."

4. Partha Pratim Ray. "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope."

5. Chung Kwan Lo. "What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature."

6. Michael Dowling, Brian Lucey. "ChatGPT for (Finance) research: The Bananarama Conjecture."

7. Fan Huang, Haewoon Kwak, Jisun An. "Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech."

8. Mohd Javaid, Abid Haleem, Ravi Pratap Singh. "ChatGPT for healthcare services: An emerging stage for an innovative perspective."

9. Dr. Muneer M. Alshater. "Exploring the Role of Artificial Intelligence in Enhancing Academic Performance: A Case Study of ChatGPT." 27/12/2022.

10. Viriya Taecharungroj. "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter."

11. Luigi De Angelis, Francesco Baglivo, et al. "ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health."

12. Rajesh Bhayana, Sateesh Krishna, Rober R. Bleakney. "Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations."

13. Harsh H. Patel, Purvi Prajapati. "Study and Analysis of Decision Tree Based Classification Algorithms." 31/Oct/2018.

14. M. A. Friedl, C. E. Brodley. "Decision Tree Classification of Land Cover from Remotely Sensed Data."

15. WEIWEI LIN, ZIMING WU, LONGXIN LIN, ANGZHAN WEN, AND JIN LI. "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis."

16. Mariette Awad, Rahul Khanna. "Support Vector Machines for Classification."

17. Zhou Yong, Li Youwen, Xia Shixiong. "An Improved KNN Text Classification Algorithm Based on Clustering."

18. Dr. Muneer M. Alshater. "Exploring the Role of Artificial Intelligence in Enhancing Academic Performance: A Case Study of ChatGPT."

19. Katharina Jeblick, Balthasar Schachtner, et al. "ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports." Received: 6 March 2023 / Revised: 24 May 2023 / Accepted: 7 July 2023.

20. Evaluating Academic Answers Generated Using ChatGPT : Suzzane Fergus, Michelle Botha, Mehrnossh Ostovar.