Secure Diet Recommendation System using Federated Learning

Arunima Barik [1], Gaurav Golchha [1]

Abstract

Modern diets, primarily focusing on processed, high-calorie, low-nutrient foods, have significantly impacted human health. A balanced diet is essential for proper functioning of the body. Calories stored in food are crucial for various bodily functions. On an average, a person needs 2000 calories per day, but intake depends on factors like weight, height, age, and gender. Food choices impact health today, tomorrow and in the future. A tailored diet plan based on individual characteristics, such as height, weight, age, gender, physical activity levels, BMI and BMR values, can provide the necessary nutrients for proper function. Such specialized diet plans can help reduce obesity rates, cardiovascular diseases and cancer risk. Federated Learning, a decentralized, privacy-preserving framework provides an effective means for addressing the challenges of data privacy and security when data is abundant but sensitive. Through the integration of federated learning, the privacy risks can be mitigated by allowing individuals to keep their dietary data locally while still benefiting from collective intelligence. This paper demonstrates how this integration enables the model to learn and adapt from distributed data sources, offering highly personalized dietary recommendations that are not only effective but also inherently respectful of individual privacy.

1 Introduction

The quest for a healthy lifestyle has never been more prevalent in today's society. Amidst the myriad of factors influencing health and well-being, diet plays a pivotal role. However, one plan does not fit all when it comes to nutrition. Individuals are unique, with varying physiological characteristics, activity levels , and nutritional needs. Thus, a personalized approach to diet planning becomes essential for achieving optimal health outcomes.

This research endeavors to bridge the gap between generic dietary recommendations and the specific requirements of individuals. We propose an innovative method that harnesses the power of data analysis and nutrition science to provide tailored diet plans. The crux of our approach lies in the consideration of individual attributes, such as height, weight, age, gender, and physical activity levels, to calculate the Body Mass Index (BMI) and Basal Metabolic Rate (BMR). These fundamental metrics form the foundation of our personalized diet recommendations, ensuring that each individual's unique nutritional needs are met.

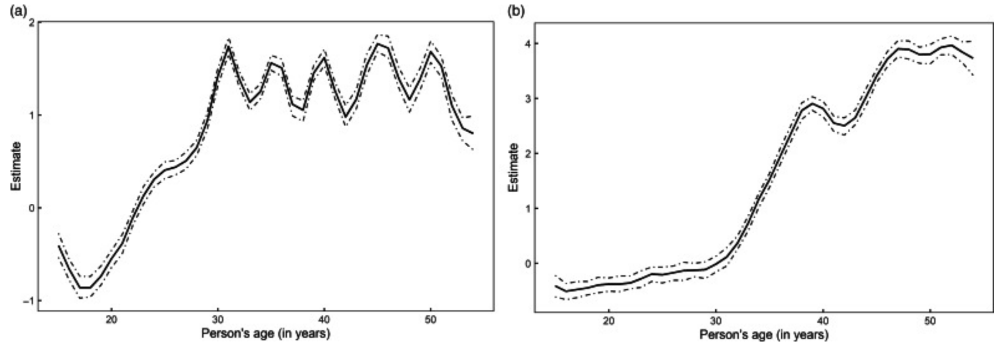


Figure 1: The impact of men's age on (a) being underweight and (b) being overweight has nonlinear effects. [7]

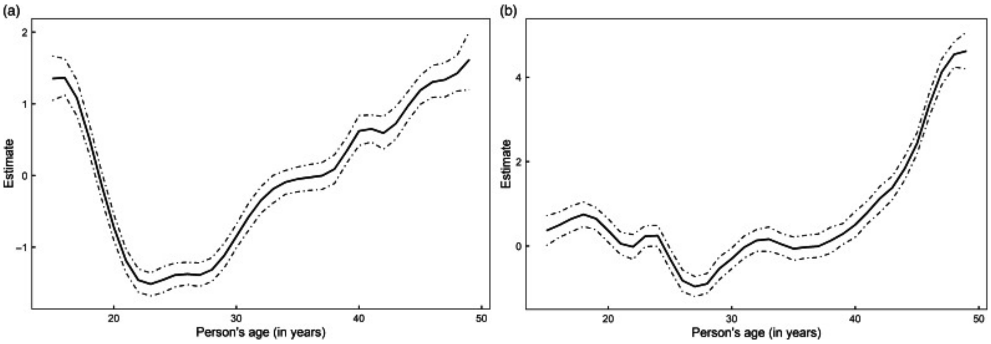


Figure 2: The impact of women's age on (a) being underweight and (b) being overweight has non-linear effects. [7]

As we delve deeper into the complexities of diet planning, it becomes evident that a one-size-fits-all approach is inadequate. To address this issue, our research introduces a sophisticated algorithm that takes into account not only the quantitative data but also qualitative preferences and dietary restrictions. We aim to provide individuals with a user-friendly tool that not only improves their overall health but also encourages lasting lifestyle changes.

The realm of health and nutrition has been increasingly shaped by the data-driven insights derived from diet recommendation systems. These systems have proven invaluable in aiding individuals on their journey toward healthier lifestyles, as well as supporting healthcare professionals in designing tailored dietary plans. However, the efficacy of such systems hinges on access to vast and diverse datasets, which, in the traditional centralized approach, often raise significant privacy and security concerns. As we venture into the digital age, the need to reconcile the advantages of data-driven diet recommendations with stringent data protection regulations has grown ever more pressing. It is within this context that federated learning emerges as a transformative approach, revolutionizing the integration of privacy-preserving machine learning in diet recommendation systems.

Federated Learning, a recent and pioneering development in machine learning, enables the collaborative training of models across decentralized data sources while safeguarding the privacy of each contributing entity. Traditional centralized models often require participants to share their dietary habits and health-related information directly, which, while beneficial for recommendation accuracy, can be a significant deterrent due to privacy concerns. The introduction of federated learning offers a promising resolution to this dichotomy. By allowing individuals to retain their dietary data on their own devices, we strike a balance between recommendation system performance and data security. This paradigm shift fosters the development of highly personalized dietary recommendations while preserving the privacy of individual users. In essence, the integration of federated learning transforms diet recommendation systems into privacy-conscious, data-driven tools, empowering individuals to embark on their nutritional journey with confidence, while safeguarding their sensitive health information.

1.1 Literature Review

Federated Learning is becoming a crucial approach for safeguarding data privacy and security, especially in healthcare environments where the protection of sensitive patient information is of utmost importance. The paper "Polymorphic Cipher Functional Encryption" by [10] presents a strategy that guarantees both anonymity and end-to-end data protection in healthcare. This scheme allows for the training of several machine learning models for different security circumstances. Another method in privacy-preserving federated learning for healthcare, as highlighted in the same context, prioritizes openness, repeatability, and the implementation of standardized frameworks [5]. This system integrates advanced techniques such as improved Differential Privacy, secure multiparty computing, and homomorphic encryption to ensure safe model updates.

Efficiency and resource optimization are vital aspects of healthcare federated learning. A framework has been developed using edge computing [12] that encompasses cloud, edge and application modules, offering personalized healthcare insights while preserving privacy. It also addresses challenges such as resource management and user acceptance of technology. In line with efficiency, another approach introduces the FedCare framework for federated learning in resource-constrained Internet of Medical Things (IoMT) devices [11]. It utilizes split learning to distribute the deep learning model across multiple institutions, ensuring the privacy and security of patient data, while also optimizing resource efficiency.

The limited resources of Internet of Medical Things (IoMT) devices make security a crucial challenge in federated learning. Network security may be improved by decreasing the size of the network using parameter quantization approaches [16]. The findings demonstrate that this method attains superior precision and reduced quantization error, all the while guaranteeing the confidentiality and protection of data throughout the transmission of the model. The Federated Edge Aggregator architecture integrates edge intelligence and differential privacy to maintain the privacy of individual clients and data in smart healthcare systems [6]. The efficacy of the framework is evidenced by empirical findings, emphasizing its capacity to attain a high level of testing precision.

Decentralized healthcare systems are increasingly relying on federated learning for data analysis while preserving user privacy. Deep Federated Learning (DFL) framework integrates IoT and federated learning technologies to build a skin-disease detection model [9]. This framework represents a significant leap in ensuring the privacy of healthcare data. Furthermore, another approach uses the Web of Objects (WoO) ontology to assess the quality of client contextual data in healthcare applications [14]. By incorporating deep learning models and a weighting factor estimation algorithm, the scheme enhances the quality of the global model, promoting efficiency and intelligence in healthcare applications.

Active involvement of data owners is essential for the successful implementation of federated learning, and incentives play a pivotal role in this process. The DPFL framework employs a multidimensional contract approach that provides a just and equitable incentive mechanism for data owners [15]. This approach takes into account the unequal access to information between the owner of the model and the owners of the data and offers incentives depending on different sorts of costs. The F-RCCE approach employs reinforcement learning techniques to assess the contributions of federated participants [17]. This method effectively and precisely evaluates the contributions of clients while maintaining their privacy.

Collaborative data analytics is essential in federated learning, especially in secure multi-party setups. EaSTFLy, a privacy-preserving framework for federated learning that enables secure collaboration among multiple parties [8]. It utilizes ternary federated learning and employs protocols to protect against adversaries while optimizing computation and communication overheads.

The PEFL approach specifically targets privacy problems in the context of federated learning for industrial artificial intelligence[13]. This system guarantees confidentiality during the training process, even in cases when adversaries collaborate with several entities, offering an effective and privacy-enhanced solution for industrial AI.

Table 1 provides a comprehensive overview of recent advancements in federated learning for healthcare applications. Various studies leverage techniques such as federated learning with privacy-preserving mechanisms, novel incentive structures, and frameworks integrating edge intelligence. Authors propose innovative approaches like split learning, minimum square quantization error, and reinforcement learning. The table summarizes key highlights, datasets used, results, and pros and cons of each study, offering valuable insights into the evolving landscape of federated learning in healthcare.

1.2 Motivation

This research is driven by a combination of critical factors, rooted in the compelling necessity to address the increasing global health challenges associated with diet-related issues, including obesity, malnutrition, and chronic diet-related diseases. Conventional dietary guidelines, while undoubtedly valuable, often fail to account for the substantial individual variations that significantly affect nutritional requirements.

First and foremost, our primary motivation is to provide a solution that caters to the imperative of personalized diet planning systems. This endeavor seeks to empower individuals to take control of their health and nutrition. The objective is to furnish a practical and accessible tool

Table 1: Literature review table

Author	Year	Key Highlights	Dataset Used	Result	Pros	Cons
Gobinath et al. [10]	2023	Utilizing Federated Learning (FL) in conjunction with Differential Privacy and Functional Encryption can improve model accuracy, protect privacy, and mitigate collusion concerns.	Diabetes data	Avg. encryption time - 25.2 to 37.5 secs Avg. decryption time - 52.8 to 63.7 secs Avg. computation time is between 2.225 and 3.2 seconds.	Enhanced data security during transfers and more accurate predictions	
Jeon et al. [14]	2023	Proposed a data quality index method based on the WoO ontology for assessing healthcare data quality	Data from wearable and non-wearable sensors	Data quality of dataset in accordance with application purpose increases from 87.1% to 88.5%	Promotes quality of the global model, leading to higher efficiency and intelligence of application capabilities	
Gupta et al. [11]	2023	Proposed a novel approach in federated learning called split learning, which enables stragglers to collectively train by utilizing the nearest edge node.	Camera based IOMT devices	Decrease of 3.6 hours in video training duration The server achieved a global accuracy rate of 90.32	Optimizes computational resources by involving resource-constrained devices in split learning, reduces training time and scalability	Communication overhead and challenges related to device heterogeneity, potentially impacting system performance and reliability
Abaoud et al. [5]	2023	The process entails combining Differential Privacy methods, specifically Local and Global DP, with sophisticated Secure Multi-Party Computation and Homomorphic Encryption for updating the model.	Medical Information Mart for Intensive Care III and the Synthetic Health Dataset	Average accuracy rate of 97.69%	Exhibits superior computational efficiency, privacy preservation and utility-privacy trade-off.	Challenges include ensuring data security, addressing communication overhead, and striking a balance between collaboration and privacy.
Xu et al. [16]	2023	Created a novel framework, denoted as FED_{MSQE} (Federated Learning with Minimum Square Quantization Error), to minimize the quantization error for each client in the Federated Learning (FL) scenario.	Used MNIST and CIFAR-10 dataset for experiment purposes.	The quantization error of FED_{MSQE} is approximately 42% lower than $FED_{UNIFORM}$ and 72% lower than FED_{APOT}.	Safeguards data privacy by enhancing data transmission security through the quantization of neural network parameters.	Large-scale neural networks pose challenges due to power and memory limitations, and privacy leakage risks when models are intercepted.
Akter et al.[6]	2022	Integrates edge intelligence and differential privacy to protect individual client privacy and ensure data privacy in smart healthcare systems	Study carried out on MNIST, CIFAR10, STL10, COVID19 Chest X-RAY dataset	90% testing accuracy at the 20th and 75th communication rounds for the FEA model, 62% and 78% testing accuracy for the same communication rounds in terms of non-IID distribution for NbAFL model	Safeguards individual client privacy and data privacy in smart healthcare systems by integrating edge intelligence and differential privacy	Additional computational resources, Communication overhead
Elayan et al.[9]	2021	In the Proposed model, the local DL model is trained on specific data, the parameters are shared with the central architecture, the global model is trained, and the old local model is replaced by the new global model.	Atlas Dermatology dataset (skin disease)	Area under the Receiver Operating Characteristic Curve increased to 97.4% for federated rounds	Decentralized data analysis, preserves user privacy, and reduces the need for extensive feature engineering.	AI reliability and compatibility issues with IoT devices, limited availability of healthcare datasets, and potential impact on quality of service due to increased conversion time
Wu et al. [15]	2021	Developed a three-dimensional contract-based incentive system to determine the most advantageous payment and local training data size for various categories of data owners in the presence of information asymmetry.	MNIST dataset	As the allocation of resources for privacy declines, the accuracy of the tests also lowers. A smaller data budget, with a fixed dataset size, leads to more noise and a larger convergence error. Training in the absence of incentives results in increased training loss and decreased test accuracy.	Incorporates differential privacy methods and contribution and costs of data owners	Discusses only about incentives and thus has limited real world exposure
Zhao et al. [17]	2021	F-RCCE uses reinforcement learning techniques to train a data value estimator and determines the contribution of each client. Simulated with 50 to 500 clients and about 1000 epochs	SMS Spam Dataset	Accurately evaluates the contribution of gradients provided by each client. Time increases by 6% when the number of clients increases from 100 to 500.	Time cost remains almost constant with increased number of clients. Suitable for applications with large B2C models.	Model yet to effectively distribute benefits to all participants using the clients' contribution

Hakak et al. [12]	2020	Proposed a framework for edge-assisted data analytics that use Federated Learning to update local machine learning models with user-generated data from wearable devices. This framework aims to assess mobility levels and behaviors for illness diagnosis and prediction.	Public datasets	Proposed framework can be used for disease management, mental health tracking, real time health monitoring etc.	Improved efficiency in healthcare data analytics, personalized healthcare insights, and privacy preservation.	Resource management, optimization of local storage and computation, and user acceptance of emerging technologies
Dong et al.[8]	2020	Introduced EaSTFLy - ternary federated learning with secret sharing and homomorphic encryption, optimized computation and communication overheads	MNIST and SVHN	EaSTFLy performs way better than floating point privacy preserving protocols. EaSTFLy uses protocols similar to TSS and PHE for privacy of ternary federated learning.	Decreased computation and communication overheads on ternary federated learning	Scalability of the models and tradeoff between privacy preservation and model accuracy not discussed
Hao et al.[13]	2020	PEFL scheme for industrial artificial intelligence. Combines differential privacy, encryption using BGV scheme and secure aggregation using A-LWE ciphertext	MNIST	For differential privacy budget = 0.5, the neural network achieved 90.8% and 91.5% accuracy for = 105 and = 103	Incorporates differential privacy, encryption, secure aggregation techniques, providing post-quantum security and protection against collusion attacks.	The absence of a thorough comparison between the proposed approach and current privacy-preserving federated learning schemes is evident.

that accommodates individuals from diverse socio-economic backgrounds and lifestyles. Our aim is to motivate individuals to embrace healthier dietary practices by delivering tailored dietary recommendations that align with their unique requirements, thereby contributing to a reduction in the prevalence of diet-related health issues.

Secondly, our motivation includes the domain of Federated Learning. In recognition of the paramount importance of data privacy and security, especially within the context of health and nutrition, we seek to leverage the potential of this pioneering approach. Federated Learning allows for the provision of personalized dietary guidance without necessitating the sharing of sensitive health information, thereby ensuring data privacy while delivering valuable recommendations.

1.3 Contribution

The contributions of this paper are as follows -

1.3.1 Identifies parameters affecting the diet

Examines age, height, weight and activity level as key factors influencing diet, laying the groundwork for personalized recommendations

1.3.2 Deduces healthy calorie intake using ML

Utilizes advanced machine learning to predict optimal daily calorie intake based on individual characteristics, promoting personalized and effective nutrition.

1.3.3 Uses federated learning for privacy-preserving training

Adopts federated learning to train the model across decentralized devices, ensuring privacy while deriving collective insights from distributed data.

1.3.4 Suggests 3-course meal based on healthy calorie intake

Provides practical dietary guidance with personalized 3-course meal suggestions, enhancing the application of research for individuals seeking balanced nutrition.

Table 2: Health Care Parameters

Parameter	Description
Body Mass Index (BMI)	BMI is equated using weight and height of the body. The range is divided into 4 categories. BMI <18.5 (Underweight range), 18.5 <BMI <24.9 (Healthy weight range), 25.0 <BMI <29.9 (Overweight range), BMI >30.0 (Obese range)
Percentage Body Fat (%BF)	Body Fat Percentage is equated by dividing the total mass of fat divided by total body mass multiplied by 100. %BP provides an overview about essential body weight and storage fat
Blood Serum Cholesterol (BSC)	BSC is the amount of cholesterol in the bloodstream. Unit of measurement is milligrams per deciliter (mg/dL) or millimoles per liter (mmol/L)
Systolic and Diastolic Blood Pressure	Systolic BP represents the pressure in your arteries when your heart contracts (beats) and pumps blood into the arteries. Diastolic BP represents the pressure in your arteries when your heart is at rest between beats, or during the relaxation phase of the cardiac cycle

1.4 Organization

The rest of the paper is organized as follows. Section II talks about the problem statement in detail. Section III consists of the methodology used to build a secure diet recommendation system. Section IV discusses the proposed solution and results are shown in Section V.

2 Problem Statement

The development and deployment of diet recommendation systems present a critical need in the context of modern healthcare and lifestyle management. These systems, powered by machine learning algorithms, have the potential to provide individuals with personalized dietary guidance, thereby significantly contributing to the prevention and management of chronic diseases and the promotion of overall health and well-being.

However, a substantial barrier to the widespread adoption of diet recommendation systems is the inherent tension between data-driven recommendation accuracy and the preservation of individual privacy. In conventional centralized models, users are often required to share sensitive dietary and health data, raising valid concerns about data security and the protection of personal information. The critical problem at hand is how to effectively harness the power of data while preserving individual privacy, ensuring that users can benefit from tailored dietary guidance without compromising their confidential health information. Federated Learning offers a promising solution by allowing the collaborative training of recommendation models without direct data sharing, addressing this conundrum. This research seeks to explore the implementation of Federated Learning as a groundbreaking solution to the privacy dilemma in diet recommendation systems and offer a thorough comprehension of the consequences and possibilities it brings in the domain of health and nutrition.

Table 2 outlines essential health parameters with their respective descriptions. It includes Body Mass Index (BMI), categorized into underweight, healthy weight, overweight, and obese ranges based on weight and height. Percentage Body Fat (%BF) is calculated by dividing total fat mass by total body mass and multiplying by 100, offering insights into body weight and fat distribution. Blood Serum Cholesterol (BSC) is measured in milligrams per deciliter (mg/dL) or millimoles per liter (mmol/L), indicating cholesterol levels in the bloodstream. Systolic and Diastolic Blood Pressure measurements provide insights into arterial pressure during heart contraction and relaxation phases, respectively.

3 Methodology

This section outlines the comprehensive methodology employed to develop and implement our innovative diet plan recommendation system.

3.1 Diet Plan Recommendation Software

This paper proposes a Diet Plan Recommendation System for users to get suggestions for a healthy diet. The software requires inputs from the user about their physical status. The inputs include: age, height, weight, gender and activity level.

The software internally computes the amount of calories the person should intake depending on his/her physical status. This computation is done using a machine learning model and requires data of other people about their healthy calorie intake. This data needs to be obtained from health data sources. The primary sources of health statistics are surveys, administrative and medical records, health care claims data, vital records, surveillance, illness registries, grey literature, and peer-reviewed publications [2].

This information about calories can be used to suggest recipes that match the computed calorie intake level. The software predicts the total calorie intake for a day which can be split into a 3 or 4 course meal. The implementation here is for a 3-course meal where weightage of breakfast, lunch and dinner has been taken as 35%, 40% and 25% respectively.

3.2 Security breaches in healthcare data

Security breaches in medical healthcare data have profound and far-reaching effects, compromising both individual well-being and organizational integrity. These breaches, ranging from cyber attacks, insider threats or inadvertent exposures, compromise with patient privacy by revealing sensitive medical details such as diagnoses and treatment plans. The consequences include identity theft, insurance fraud and medical identity theft, leading to financial strain for both individuals and healthcare providers. Beyond the immediate financial implications, breaches erode trust between patients and healthcare organizations, potentially deterring patients from sharing crucial information.

The critical nature of patients' health data is paramount for accurate medical predictions and estimations. However, if patients withhold their data due to concerns about potential data breaches, it creates a significant impediment to the advancement of healthcare analytics and personalized medicine. Predictive models heavily rely on diverse and extensive datasets to generate meaningful insights. Without access to comprehensive patient information, healthcare professionals face limitations in their ability to make accurate predictions, hindering the development of treatment plans and potentially impacting overall healthcare outcomes.

3.3 Distributed processing techniques: Federated Learning

To tackle the growing apprehensions regarding data breaches and the confidentiality of patient information, the adoption of distributed processing in managing healthcare data emerges as a crucial and innovative solution. One prominent technique in this realm is Federated Learning, which plays a pivotal role in ensuring that patient data remains decentralized, thereby minimizing the vulnerability to centralized breaches. Through the implementation of distributed processing methods like federated learning, the emphasis is on keeping patient data localized, thereby reducing the potential risks associated with centralized breaches.

Federated Learning, as a cutting-edge approach, facilitates collaborative model training across a network of decentralized devices. This approach enables predictive algorithms to glean insights and improve their accuracy by learning from a diverse range of datasets, all without exposing the raw and sensitive patient information. The key advantage of this strategy is its ability to strike a delicate balance between harnessing the power of healthcare data for advancements while safeguarding individual privacy.

Not only does federated learning preserve the privacy of individual patients, but it also contributes significantly to fortifying the overall security of healthcare analytics. By leveraging the capabilities of distributed processing, the healthcare industry can successfully navigate the terrain of data-driven progress while upholding the paramount importance of patient confidentiality. This not only instills trust in healthcare systems but also lays the groundwork for more effective and reliable medical predictions, ultimately benefiting both patients and healthcare providers. The integration of distributed processing techniques in healthcare not only addresses current concerns but also paves the way for a more secure and trustworthy future in medical data management.

3.4 Federated Learning in Healthcare Systems

Within the healthcare sector, Federated Learning is ushering a paradigm shift in both data privacy and model development. Its impact extends across a spectrum of applications, notably in collaborative disease prediction. Here, the amalgamation of insights derived from a diverse array of patient data contributes to markedly more accurate prognoses, all while steadfastly safeguarding individual privacy. This innovative approach not only enhances the precision of disease predictions but also sets a precedent for prioritizing patient confidentiality amid the continuous evolution of data-driven medical practices.

Federated Learning's prowess is further evident in its application to personalized treatment plans. By allowing models to glean knowledge from decentralized datasets, it facilitates the development of tailored interventions. This tailored approach ensures that medical interventions are finely tuned to the unique characteristics of individual patients, optimizing treatment outcomes without compromising the sensitive nature of personal health data.

Examining a specific example, consider a diet plan recommendation system. Federated Learning plays a pivotal role in ensuring the secure prediction of calorie needs by leveraging user data without centralizing it. This not only guarantees the privacy of users but also showcases the versatility of federated learning in addressing diverse healthcare scenarios. The system's ability to extract meaningful insights from decentralized datasets while maintaining data security redefines the landscape of personalized healthcare recommendations.

As federated learning continues to reshape predictive analytics in healthcare, it establishes itself as a transformative force, creating a new standard in the industry. This standard places patient privacy at the forefront, acknowledging its paramount importance in the era of data-driven medical advancements. The ongoing integration of federated learning in healthcare not only refines existing practices but also opens avenues for further innovation, ensuring a future where medical progress aligns seamlessly with the imperative of preserving individual privacy.

4 Proposed Solution

This section talks about the proposed diet recommendation system. The approach integrates Linear Regression for calorie prediction and Nearest Neighbors for meal recommendations. The entire system operates on a federated learning framework, allowing users to retain their data while contributing to the collective knowledge.

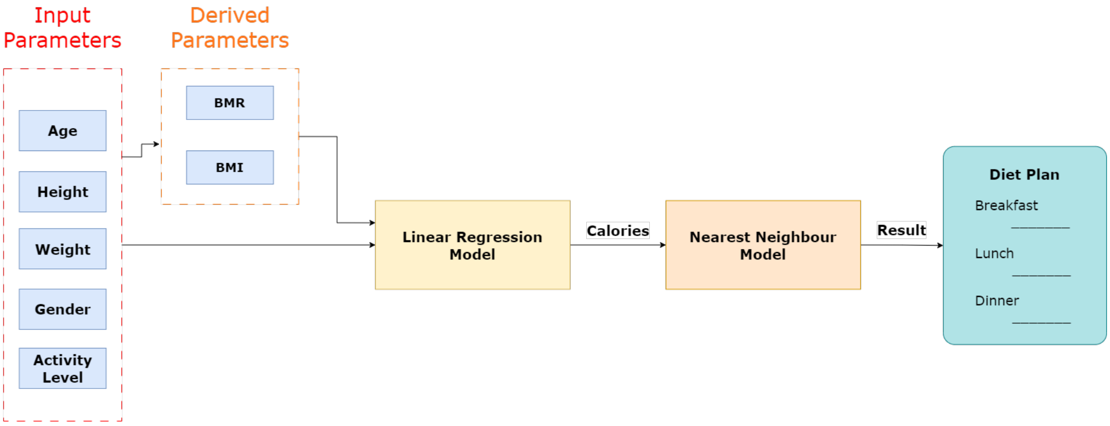


Figure 3: Proposed Architecture

4.1 Calorie Prediction

The first step in diet recommendation system is Calorie Prediction.

4.1.1 Input parameters from user

The user needs to provide the following as input to the diet recommendation system:

- Age (in yrs)
- Height(in m)
- Weight (in kg)
- Gender (Male or Female)
- Activity Level (between 0 to 2 with 0 - No Exercise and 2 - Heavy Exercise)

4.1.2 Derived parameters

The following parameters are calculated from the user's input:

- **Basal Metabolic Ratio (BMR)**

  BMR, or Basal Metabolic Rate, denotes the amount of calories required by your body while at rest to sustain fundamental physiological processes, such as respiration and circulation. It is impacted by variables such as age, gender, weight, height, and body composition. BMR, or Basal Metabolic Rate, is a crucial factor in determining the number of calories needed on a daily basis. This information is valuable for managing weight and planning nutritional intake.

  It is calculated using Mifflin-St Jeor equation

  For male:
  $BMR = 10 * weight(kg) + 6.25 * height(m) * 100 - 5 * age(yrs) + 5$

  For female:
  $BMR = 10 * weight(kg) + 6.25 * height(m) * 100 - 5 * age(yrs) - 161$

- **Body Mass Index (BMI)**

  BMI is a metric that quantifies body fat by considering an individual's weight and height. It is widely used to classify individuals into distinct weight status groups, including underweight, normal weight, overweight, and obesity. Nevertheless, it does not provide a direct assessment of body fat percentage or distribution, nor does it consider variables such as muscle mass.

  $BMI = \frac{weight(kg)}{height^2(m)}$

The input parameters from user and derived parameters are collectively fed to the model to predict healthy calorie intake for the user.

4.1.3 Dataset Used

A dataset from Kaggle [4] is used to train the machine learning model to predict calories.

The dataset contains 10 columns and 10,726 rows. Of these, 7 columns (age, height, weight, gender, activity level, BMI and BMR) are features and calories is the target column.

Before training, all categorical values have been encoded using Label Encoder and scaled using Standard Scaler. 80% of the data is used as train data and 20% as test data.

4.1.4 Machine Learning models

Multiple regression machine learning models were used to predict calories:

- **Linear Regression**

  Linear Regression models the relationship between the dependent variable Y and independent variables X with a linear equation:

  $$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_n X_n$$

  Here, b_0 is the intercept, b_i are coefficients, and X_i are feature values. Parameters to optimize include the coefficients and intercept.

- **Decision Tree Regressor**

  Decision Tree Regressor partitions data into segments based on feature splits, predicting the average of the target variable within each segment. The algorithm selects the most informative features at each node to maximize the reduction in variance. Key parameters include the maximum depth of the tree and the minimum number of samples required to split a node.

- **Random Forest Regressor**

  Random Forest Regressor is an ensemble of Decision Trees, providing robust predictions by averaging or voting on the outputs of individual trees. It introduces randomness by training each tree on a subset of data. Parameters include the number of trees, the criteria for splitting, and the maximum depth of each tree.

- **SVR**

  Support Vector Regressor (SVR) constructs a hyperplane to predict continuous values. In a linear SVM, the hyperplane equation is $w \cdot X - b = 0$, where w is the weight vector, X is the input vector, and b is the bias term. The objective is to minimize the loss function while satisfying margin constraints. Parameters involve the choice of kernel function, regularization term, and kernel-specific parameters.

- **SGD Regressor**

  Stochastic Gradient Descent (SGD) Regressor optimizes a linear model using stochastic gradient descent. The model iteratively updates parameters to minimize a specified loss function. The objective is to find the optimal coefficients for the linear equation. Parameters include the learning rate, regularization term, and the number of iterations.

4.1.5 Evaluation Parameters

- **Mean Squared Error (MSE)**

  The Mean Squared Error (MSE) is a vital measure used to assess the performance of regression models. The metric calculates the mean squared deviation between expected and actual values, with a focus on greater errors.

  The formula for MSE is given by:

  $$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

  Here, n is the number of data points, Y_i is the actual value, and $\hat{Y}_i$ is the predicted value. Squaring the differences ensures that both overestimations and underestimations contribute positively, providing a comprehensive measure of model accuracy. A lower MSE indicates better predictive performance, with zero denoting a perfect match between predicted and actual values.

- **Mean Absolute Error (MAE)**

  Mean Absolute Error (MAE) is a useful statistic for evaluating regression models, providing a more understandable estimate of the average error in predictions. MAE, in contrast to MSE, takes into account the absolute discrepancies between projected and actual values, resulting in reduced sensitivity to outliers.

  The formula for MAE is given by:

  $$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

  MAE provides a direct representation of the average magnitude of errors, making it suitable for scenarios where the impact of individual errors is crucial. Similar to MSE, a lower MAE signifies superior model performance, with zero indicating perfect predictions.

4.1.6 Federated Learning

Extensive comparison of machine learning models draws an opinion that Linear Regression may prove to be a good fit for Federated Learning due to its simplicity in updating parameters and accurate results.

As discussed earlier, federated learning has been used in this solution to address privacy concerns associated with user data. This model enables training on local devices while keeping raw data secure, as opposed to centralized training that poses privacy risks.

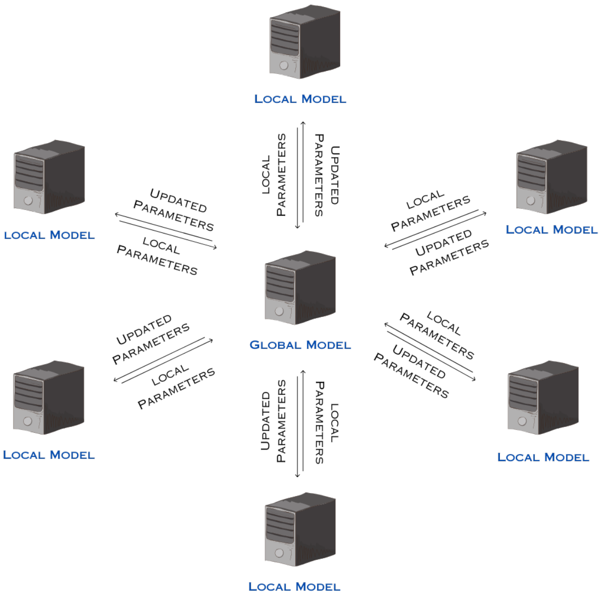


Figure 4: Federated Learning framework

Federated Linear Regression combines the strengths of both concepts. It leverages the simplicity and efficiency of Linear Regression for accurate calorie predictions while preserving user privacy through the decentralized nature of Federated Learning. This hybrid approach minimizes the risk of data breaches and enhances the ethical considerations associated with predictive modeling.

- **Flower: Federated Learning framework**

  The Flower Federated Learning Framework [3] is a cutting-edge platform designed to facilitate decentralized model training across multiple devices while prioritizing privacy and security. Flower abstracts the complexities of federated learning, providing a user-friendly interface for researchers and developers. This framework empowers the creation of collaborative machine learning models without compromising individual data privacy. Flower supports various machine learning frameworks and architectures, enabling seamless integration with popular tools. It employs a client-server architecture, allowing model updates to occur on local devices, and aggregates these updates securely to improve the global model.

- **Server Client Architecture**

  The client-server architecture of Federated Linear Regression involves a collaborative model training process that maintains data privacy. In this setup, two clients and one server participate in the training procedure. The primary focus is on optimizing the linear regression model's parameters, namely the coefficients and intercept.

  At the initiation of each training round, the server distributes the current global model parameters, including the coefficients and intercept, to the two clients. Each client independently computes local updates based on its own data, adjusting the coefficients and intercept to better align with its specific dataset.

  After the local computations, the clients send their respective updated model parameters back to the server. The server then aggregates these updates to adjust the global model parameters, ensuring that the collaborative knowledge from both clients contributes to the improvement of the overall linear regression model.

  This client-server architecture maintains the privacy of individual datasets by keeping raw data localized on each client device. The model parameters, specifically the coefficients and intercept, are the only information exchanged between the clients and the server, minimizing the risk of data exposure. Through this collaborative and privacy-preserving approach, Federated Linear Regression enables accurate predictions while respecting the confidentiality of individual data.

- **Dataset partition**

  50% of the train data has been used by each client for the training process.

  The client datasets can reside on different devices and still communicate with the server by using a standardized communication protocol. This abstraction of complexity ensures compatibility across diverse platforms, allowing seamless collaboration between devices.

- **Federated Averaging (FedAvg)**

  Federated Averaging, commonly known as FedAvg, is a decentralized machine learning approach that aggregates model updates from local devices to construct a global model. In FedAvg, participating devices, such as smartphones or edge devices, train a model locally using their data and share only the model weights instead of raw data. The global model is then updated by averaging these local parameters. This collaborative process allows for privacy preservation and reduced communication overhead.

  Alternative federated learning methods include Federated Learning with Differential Privacy (FedDP), which introduces noise to individual updates for enhanced privacy and Federated Averaging with Non-IID Data (FedAvg-N), designed to address the challenges posed by non-independent and identically distributed data across devices. These alternatives showcase the adaptability of federated learning to various privacy and data distribution challenges.

Algorithm 1 outlines the process of Federated Linear Regression, a decentralized machine learning approach that enables model training across multiple clients while preserving data privacy. The algorithm involves a central server and multiple clients. The global model parameters, including coefficients w_{global} and intercept b_{global}, are initialized on the server. The training process spans several rounds.

In each round, the server communicates with each client, exchanging global model parameters and receiving local updates. The clients independently train their local models using their private data and the current global parameters. The server aggregates these local updates using the Federated Averaging (FedAvg) method, computing the weighted average of the local model parameters across all clients. The updated global parameters are then broadcasted to all clients for the next round.

Algorithm 1 Federated Linear Regression

```
Server:
Initialize global model parameters: $w_{\text{global}}$ (coefficients), $b_{\text{global}}$ (intercept)
Set number of clients: $C$
for round = 1 to 5 do
    for each client $i$ do
        Send current global model parameters to client $i$
        Receive local model parameters: $w^{(i)}_{\text{local}}$, $b^{(i)}_{\text{local}}$ from client $i$
    end for
    Update global model parameters using FedAvg:
        $w_{\text{global}} = \frac{1}{C} \sum_{1}^{C} w^{(i)}_{\text{local}}, \quad b_{\text{global}} = \frac{1}{C} \sum_{1}^{C} b^{(i)}_{\text{local}}$
end for
End Server
Client $i$ (for each client):
for round = 1 to 5 do
    Receive global model parameters: $w_{\text{global}}$, $b_{\text{global}}$ from the server
    Train local model with client's data using $w_{\text{global}}$, $b_{\text{global}}$
    Compute local model parameters: $w^{(i)}_{\text{local}}$, $b^{(i)}_{\text{local}}$
    Send $w^{(i)}_{\text{local}}$, $b^{(i)}_{\text{local}}$ to the server
end for
End Client
```

Mathematically, the FedAvg update is expressed as:

$$w_{\text{global}} = \frac{1}{C} \sum_{1}^{C} w^{(i)}_{\text{local}}, \quad b_{\text{global}} = \frac{1}{C} \sum_{1}^{C} b^{(i)}_{\text{local}}$$

This process iterates for a predefined number of rounds, facilitating collaborative model training while keeping client data decentralized and secure. The algorithm strikes a balance between global model coordination and local data privacy in the context of linear regression tasks.

4.2 Meal Recommendations

The second step in the diet recommendation system is recommending meals using predicted calories.

4.2.1 Input Parameters

Calorie: The calories predicted from the previous step is used as the input parameter. Calories are used to quantify the energy content of food and the energy expended by the body through physical activity.

The calories from the previous step refer to the total daily intake of an individual. The total daily calories (calories) are distributed into breakfast (breakfast_calories), lunch (lunch_calories) and dinner (dinner_calories) based on the given percentages:

$$\begin{aligned} \text{breakfast\_calories} &= 0.35 \times \text{calories} \\ \text{lunch\_calories} &= 0.40 \times \text{calories} \\ \text{dinner\_calories} &= 0.25 \times \text{calories} \end{aligned}$$

This division is for a 3-course meal. The same can be applied to a 4-course meal as well.

4.2.2 Dataset Used

A recipes dataset from Kaggle [1] is the list of recipes used for recommending a meal. The user can customize his/her meals by changing the recipes list - adding vegetarian and non-vegetarian foods, adding vegan products only etc.

The dataset contains 28 columns and 5,22,517 rows. Of these, 1 column (calories) is the feature and recipe_names is the target column.

4.2.3 Nearest Neighbors Model

Nearest Neighbors model offers robust functionality for unsupervised learning, forming the foundation for various methods such as manifold learning and spectral clustering. In unsupervised nearest neighbors, the algorithm seeks to identify a predefined number of training samples closest in distance to a new data point, influencing predictions based on this proximity. The user can define a constant (k) or vary the number based on local density (radius-based neighbor learning). The distance metric is typically Euclidean, although other metrics can be employed.

The unsupervised nearest neighbors algorithm seeks to predict the label ($\hat{y}$) for a new point (x_{new}) based on a predefined number of training samples (x_i) closest in distance:

$$\hat{y} = \arg\min_{y} \left(\sum_{i} \text{dist}(x_{\text{new}}, x_i) \right)$$

Here, dist represents the distance metric, which is often the Euclidean distance. The number of neighbors (k) or the distance threshold (radius) can be user-defined, influencing the algorithm's predictions.

Here, Nearest Neighbors model is used to predict 5 recipes per meal depending on the amount of calorie intake per meal.

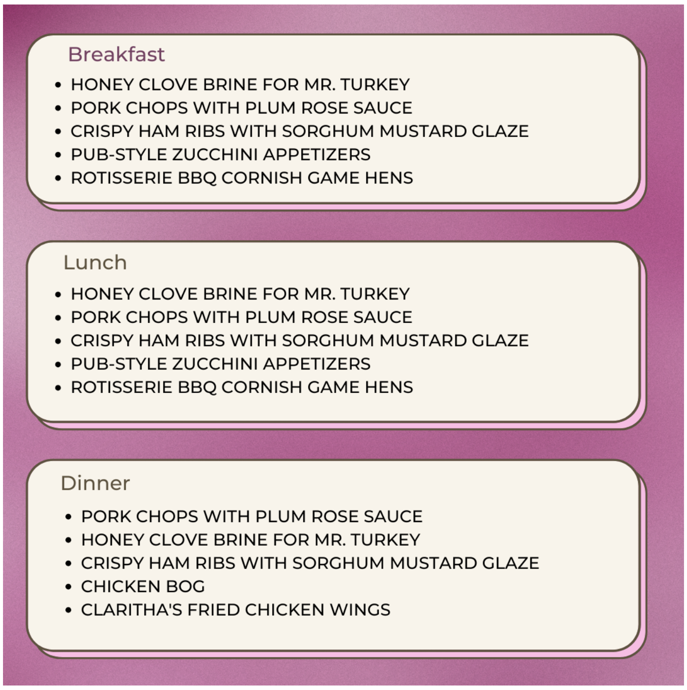
4.2.4 Diet Prediction

Consider the following example:

```
age = 26
height = 1.63
weight = 66
gender = 'F'
activity level = 1.9

Since the gender is female (F),
    BMR = 10 * 66 + 6.25 * 1.63 * 100 - 5 * 26 - 161 = 1387.75

BMI = 66 / (1.63 * 1.63)  = 24.84

Total calories as predicted by Linear Regression Model = 1345.4822

Weightage of calories in Breakfast = 35% of Total calories = 470.91

Weightage of calories in Lunch = 40% of Total calories = 538.193

Weightage of calories in Dinner = 25% of Total calories = 336.37
```

Applying the nearest neighbors algorithm with the number of items in each meal to be 5, we get the diet plan in Figure 5.

The Diet Recommendation Algorithm (Algorithm 2) computes personalized dietary suggestions based on user information. It begins by calculating the Body Mass Index (BMI) from the user's weight and height. Subsequently, it determines the Basal Metabolic Rate (BMR), incorporating Mifflin-St Jeor equation. Using these metrics, a Linear Regression model predicts the daily calorie



Breakfast

- HONEY CLOVE BRINE FOR MR. TURKEY
- PORK CHOPS WITH PLUM ROSE SAUCE
- CRISPY HAM RIBS WITH SORGHUM MUSTARD GLAZE
- PUB-STYLE ZUCCHINI APPETIZERS
- ROTISSERIE BBQ CORNISH GAME HENS

Lunch

- HONEY CLOVE BRINE FOR MR. TURKEY
- PORK CHOPS WITH PLUM ROSE SAUCE
- CRISPY HAM RIBS WITH SORGHUM MUSTARD GLAZE
- PUB-STYLE ZUCCHINI APPETIZERS
- ROTISSERIE BBQ CORNISH GAME HENS

Dinner

- PORK CHOPS WITH PLUM ROSE SAUCE
- HONEY CLOVE BRINE FOR MR. TURKEY
- CRISPY HAM RIBS WITH SORGHUM MUSTARD GLAZE
- CHICKEN BOG
- CLARITHA'S FRIED CHICKEN WINGS

Figure 5: Recommended diet plan

Algorithm 2 Diet Recommendation Algorithm

```
Input:
    $age$ (yrs), $height$ (m), $weight$ (kg), $gender$ (male or female), $activity\_level$ (between 0 and 2)
Calculate BMI:
$bmi \leftarrow \frac{weight}{height^2}$

if $gender$ is male then
    $bmr \leftarrow 10 \times weight + 6.25 \times height \times 100 - 5 \times age + 5$

else
    $bmr \leftarrow 10 \times weight + 6.25 \times height \times 100 - 5 \times age - 161$
end if
Predict Calories using Linear Regression:
    $calories \leftarrow \text{LinearRegressionPredict}(age, height, weight, gender, activity\_level, bmi, bmr)$
Divide Calories for Meals:
    $breakfast\_calories \leftarrow 0.35 \times calories$
    $lunch\_calories \leftarrow 0.40 \times calories$
    $dinner\_calories \leftarrow 0.25 \times calories$
Predict Recipes using Nearest Neighbors:
    $breakfast\_recipes \leftarrow \text{NearestNeighborsPredict}(breakfast\_calories)$
    $lunch\_recipes \leftarrow \text{NearestNeighborsPredict}(lunch\_calories)$
    $dinner\_recipes \leftarrow \text{NearestNeighborsPredict}(dinner\_calories)$
Output:
    $calories$, $breakfast\_recipes$, $lunch\_recipes$, $dinner\_recipes$
```

intake. The total calories are then partitioned for breakfast, lunch, and dinner. Nearest Neighbors prediction is employed to suggest recipes aligned with the allocated calorie percentages. The algorithm outputs the total predicted calories and recommended recipes for each meal, offering tailored nutritional guidance for the user's specific characteristics and dietary preferences.

The entire project implementation can be found on: `https://github.com/arunimabarik75/Secure-Diet-Recommendation-using-Federated-Learning`

5 Results

This section enlists the outcomes of the conducted experiments and research.

5.1 Machine Learning models

Figures 6 and 7 shows a comparison of the different machine learning models in terms of MAE and MSE. The graphs conclude that Random Forest performs the best followed by SVR model.

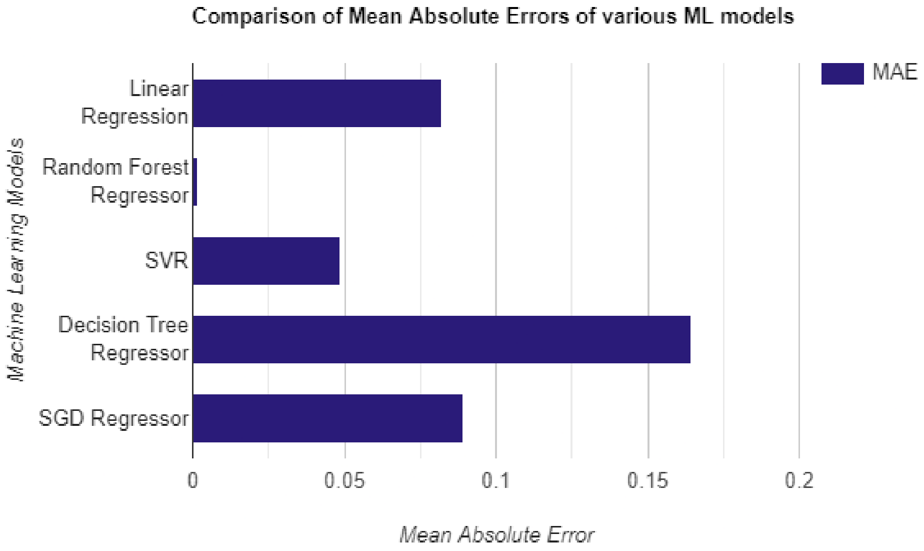


Figure 6: Comparison of MAE of various ML models

5.2 Federated Learning model

The federated linear regression model has been trained using 2 client nodes and 5 epochs. Figure 8 shows the variation of MSE and MAE for each epoch.

After completion of model training the Linear Regression equation is found to be -

$$\text{calorie} = 0.068252 \times \text{age} - 0.910724 \times \text{weight} + 0.385860 \times \text{height} - 0.049887 \times \text{gender} + 0.561099 \times \text{bmi} \\ +0.999884 \times \text{bmr} + 0.872894 \times \text{activity\_level} - 7.644943 \times 10^{-16}$$

The MAE of the model is noted as 0.082734 and MSE as 0.013596 after 5 rounds of training.

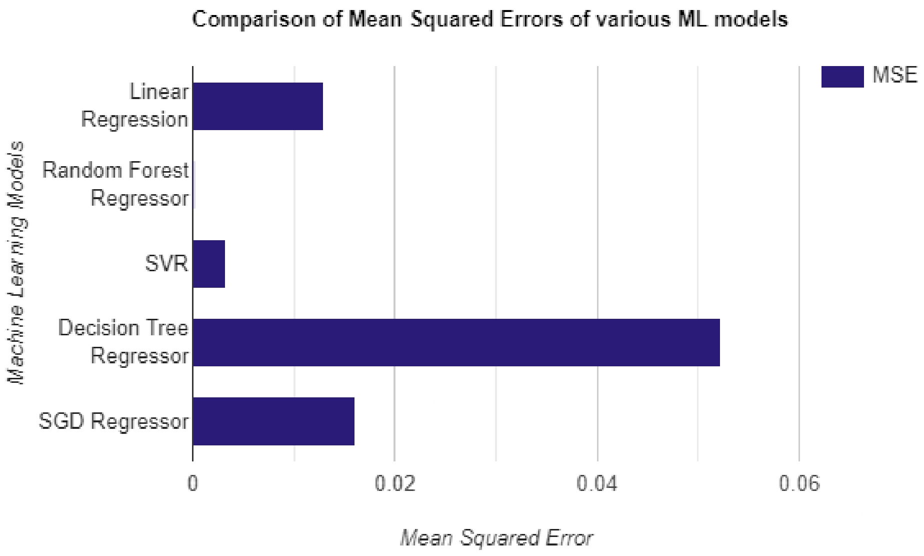


Figure 7: Comparison of MSE of various ML models

5.3 Nearest Neighbors model

The nearest neighbor model, utilizes cosine distance and brute-force algorithm and suggests 5 recipes per meal based on similarities in features or characteristics.

$$\text{model} = \text{NearestNeighbors}(n\_neighbors = 5, \text{algorithm='brute'}, \text{metric='cosine'})$$

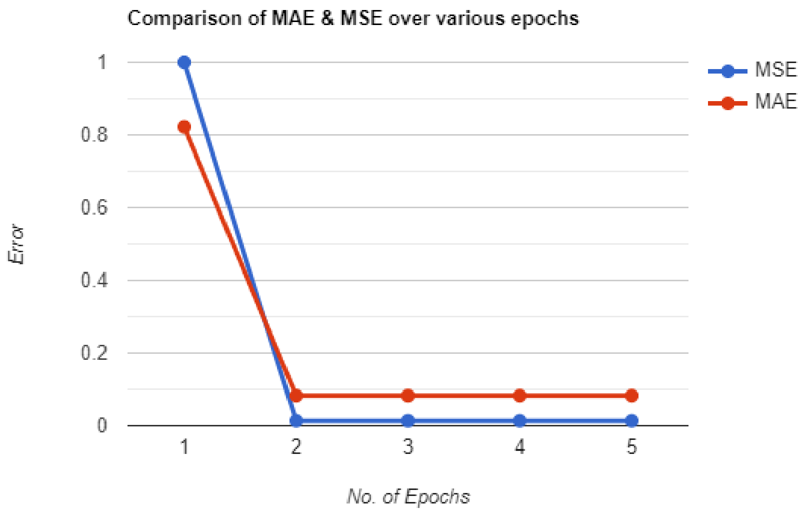


Figure 8: Comparison of MAE & MSE over various epochs

References

[1] Alvin (2020) food.com - recipes and reviews, kaggle. `https://www.kaggle.com/datasets/irkaal/foodcom-recipes-and-reviews?select=recipes.csv`.

[2] Finding and using health statistics. `https://www.nlm.nih.gov/oet/ed/stats/03-000.html`.

[3] A friendly federated learning framework, flower. `https://flower.dev/`.

[4] Kishoreharshankumar (2023) diet-plan-recommendation, kaggle. `https://www.kaggle.com/code/kishoreharshankumar/diet-plan-recommendation/input?select=Dataset.csv`.

[5] Mohammed AbaOud, Muqrin Almuqrin, and Mohammad Faisal Khan. Advancing federated learning through novel mechanism for privacy preservation in healthcare applications. *IEEE Access*, 2023.

[6] Mahmuda Akter, Nour Moustafa, Timothy Lynar, and Imran Razzak. Edge intelligence: Federated learning-based privacy protection framework for smart healthcare systems. *IEEE Journal of Biomedical and Health Informatics*, 26(12):5805–5816, 2022.

[7] Pravat Bhandari, Ezra Gayawan, and Suryakant Yadav. Double burden of underweight and overweight among indian adults: spatial patterns and social determinants. *Public Health Nutrition*, 24(10):2808–2822, 2021.

[8] Ye Dong, Xiaojun Chen, Liyan Shen, and Dakui Wang. Eastfly: Efficient and secure ternary federated learning. *Computers & Security*, 94:101824, 2020.

[9] Haya Elayan, Moayad Aloqaily, and Mohsen Guizani. Deep federated learning for iot-based decentralized healthcare systems. In *2021 International Wireless Communications and Mobile Computing (IWCMC)*, pages 105–109. IEEE, 2021.

[10] B Gobinath and TK Shanmugam. Healthcare data protection using federated learning technology. In *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 436–442. IEEE, 2023.

[11] Anshita Gupta, Sudip Misra, Nidhi Pathak, and Debanjan Das. Fedcare: Federated learning for resource-constrained healthcare devices in iomt system. *IEEE Transactions on Computational Social Systems*, 2023.

[12] Saqib Hakak, Suprio Ray, Wazir Zada Khan, and Erik Scheme. A framework for edge-assisted healthcare data analytics using federated learning. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3423–3427. IEEE, 2020.

[13] Meng Hao, Hongwei Li, Xizhao Luo, Guowen Xu, Haomiao Yang, and Sen Liu. Efficient and privacy-enhanced federated learning for industrial artificial intelligence. *IEEE Transactions on Industrial Informatics*, 16(10):6532–6542, 2019.

[14] Kyung-Chan Jeon, Geun-Seok Han, Chae-Yun Han, and Ilyoung Chong. Federated learning model for contextual sensitive data quality applications: Healthcare use case. In *2023 31st Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE, 2023.

[15] Maoqiang Wu, Dongdong Ye, Jiahao Ding, Yuanxiong Guo, Rong Yu, and Miao Pan. Incentivizing differentially private federated learning: A multidimensional contract approach. *IEEE Internet of Things Journal*, 8(13):10639–10651, 2021.

[16] Zhiang Xu, Yijia Guo, Chinmay Chakraborty, Qiaozhi Hua, Shengbo Chen, and Keping Yu. A simple federated learning-based scheme for security enhancement over internet of medical things. *IEEE Journal of Biomedical and Health Informatics*, 27(2):652–663, 2022.

[17] Jie Zhao, Xinghua Zhu, Jianzong Wang, and Jing Xiao. Efficient client contribution evaluation for horizontal federated learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3060–3064. IEEE, 2021.