##### **Linear & Logistic Regression - Based Approach for Predicting Heart Disease Probability using Machine Learning**

**Abstract**

This study explores the efficacy of machine learning models, specifically Linear and Logistic Regression, in predicting heart disease risk. Leveraging a dataset comprising diverse health parameters, the models undergo rigorous training and evaluation processes. Linear Regression is employed to establish relationships between independent variables and heart disease risk, while Logistic Regression categorizes individuals into binary outcomes healthy or at risk. Results indicate promising predictive capabilities, with both models demonstrating commendable accuracy, sensitivity, and specificity. The study contributes insights into the application of machine learning for heart disease prediction, highlighting the potential of these models in enhancing early risk identification and preventive healthcare measures. They took a Heart Disease Probability, Diabetes Probability and Calories Burnt Dataset from Kaggle publicly available (Kaggle, 2019). Unfortunately, some dataset includes some null values. Therefore, two approaches are proposed in this study : mean imputation and predicting imputation for replacing the null values. Three machine learning algorithms with each approach are implemented, achieving an accuracy of 80%, 85.25% and 96%.

**Keywords** : Heart Disease Prediction, Diabetes Prediction, Calories Prediction, Linear Regression, Logistic Regression, Imputation, Machine Learning

**Introduction**

The journey of AI began in the mid-20th century when computer scientists and researchers sought to create machines capable of simulating human intelligence. AI, as a field, emerged as early as the 1950s, with pioneers like Alan Turing, who introduced the concept of intelligent machines.

Diabetes and heart disease have complex histories. Diabetes was first identified in ancient Egypt around 1550 BCE. It became better understood in the 20th century. Heart disease has been recognized since ancient times but gained significant attention in the 20th century with advancements in cardiology.

Both conditions have seen significant research and medical advances over the years.

Currently, diabetes and heart diseases remain global health concerns. Diabetes prevalence is rising due to unhealthy lifestyles. Heart diseases are prominent, associated with factors like hypertension, high cholesterol, and obesity. Caloric burn varies based on activity, age, and metabolism. Regular exercise and a balanced diet are vital for weight management.

Heart disease is the leading cause of death worldwide, accounting for an estimated 17.9 million deaths in 2019 (WHO, 2021). Early detection and prevention of heart disease are essential for reducing mortality and improving patient outcomes.

Machine learning (ML) algorithms have been shown to be effective in predicting heart disease risk. ML algorithms can learn patterns from data and use these patterns to make predictions about new data.

This research paper proposes a linear and logistic regression-based approach for predicting heart disease probability using machine learning. Linear regression is a statistical model that can be used to predict a continuous variable, such as heart disease risk score. Logistic regression is a statistical model that can be used to predict a binary outcome, such as the presence or absence of heart disease.

The proposed approach will be evaluated on a publicly available dataset of heart disease patients. The dataset contains a variety of features, including demographic information, medical history, and laboratory results.

In this era of technology and data-driven decision-making, proactive health management gains importance. WellNest.ai is a groundbreaking health platform that uses AI and machine learning to address health issues. It integrates machine learning models into a user-friendly web app, allowing users to input health data, receive personalized insights, and make informed decisions.

It uses predictive models for diabetes, heart disease, and calorie estimation with Logistic and Linear Regression. This paper explores the platform, its data, algorithms, and potential improvements, demonstrating AI and ML's ability to provide personalized health insights for proactive management.

**Dataset Description**

1. **Diabetes Prediction**

| **Label** | **Description** |
| --- | --- |
| Pregnancies | Indicates the number of pregnancies the individual has experienced. |
| Glucose | Represents the glucose concentration in the blood, a crucial indicator for diabetes risk assessment. |
| BloodPressure | Denotes the blood pressure levels, contributing to the overall health evaluation. |
| SkinThickness | Reflects the skinfold thickness, a measure potentially linked to metabolic health. |
| Insulin | Signifies insulin levels, an essential hormone in glucose metabolism. |
| BMI  (Body Mass Index) | Captures the individual's weight relative to height, crucial for assessing obesity-related diabetes risk. |
| DiabetesPedigreeFunction | Quantifies diabetes hereditary risk based on familial history. |
| Age | Specifies the age of the individual, a key factor influencing diabetes susceptibility. |

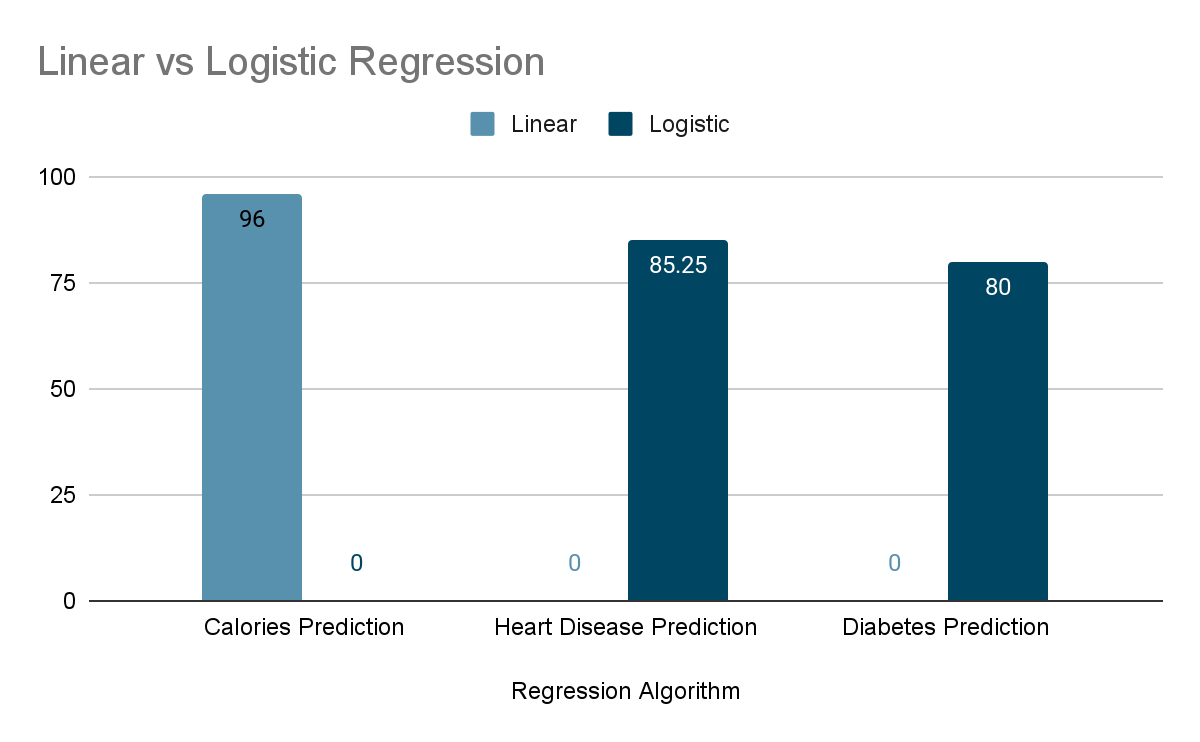
1. **Heart Disease Prediction**

| **Label** | **Description** |
| --- | --- |
| Age | The age of the individual, a critical factor in heart disease risk assessment, as the likelihood of cardiovascular issues often increases with age. |
| Sex | Gender can influence heart disease risk, with variations in prevalence and manifestation between males and females. |
| Chest Pain | Different types of chest pain (angina) can provide insights into the nature and severity of heart issues. |
| Resting Blood Pressure | Elevated resting blood pressure is a significant risk factor for heart disease, indicating potential strain on the cardiovascular system. |
| Cholesterol | High cholesterol levels contribute to arterial plaque buildup, increasing the risk of coronary artery disease. |
| Fasting Blood Sugar | Elevated fasting blood sugar levels may indicate insulin resistance, linking to diabetes, a risk factor for heart disease. |
| Resting Electrocardiographic Results | Abnormal resting electrocardiogram results can suggest underlying cardiac issues. |
| Maximum Heart Rate Achieved | The maximum heart rate during exercise can indicate cardiovascular fitness and potential heart disease risk. |
| Exercise-Induced Angina | Presence of angina during exercise is a significant symptom of potential coronary artery disease. |
| ST Depression Induced by Exercise | ST depression during exercise can indicate myocardial ischemia, a precursor to heart disease. |
| Slope of the Peak Exercise ST Segment | The slope of the ST segment during exercise provides further insights into coronary artery health. |
| Number of Major Vessels Colored by Fluoroscopy | The number of vessels with fluoroscopy coloring relates to the severity of coronary artery disease. |
| Thallium Stress Test Result | The thallium stress test outcome aids in evaluating myocardial perfusion and potential heart disease. |

1. **Calories Prediction**

| **Label** | **Description** |
| --- | --- |
| Gender | While not a direct contributor to caloric expenditure, gender can influence factors like muscle mass and metabolism, indirectly impacting calorie prediction models. |
| Age | Age plays a crucial role in determining basal metabolic rate, affecting the number of calories burned at rest, and influencing overall calorie needs. |
| Height | Height contributes to the calculation of basal metabolic rate, influencing the energy expended for basic bodily functions. |
| Weight | Body weight is a significant factor in caloric expenditure, as individuals with higher weight generally burn more calories during both rest and activity. |
| Duration | The duration of physical activity directly affects the calories burned. Longer durations typically result in higher energy expenditure. |
| Heart Rate | Heart rate is indicative of the intensity of physical activity, impacting the calorie-burning rate. Higher heart rates during exercise generally correlate with increased calorie consumption. |
| Body Temperature | While not a direct factor in calorie prediction, body temperature can reflect the energy expended during physical activity, especially in the context of activities that induce sweating or increased metabolic demand. |

* **Model Accuracy**



**Proposed Methodology**

The methodology section is a cornerstone of any research endeavor, serving as the roadmap that guides the research from inception to conclusion. In this section, we provide a detailed account of the methodology employed in our research, which culminated in the creation of the innovative "WellNest.ai" platform. This comprehensive methodology encompasses data sourcing, data preprocessing, feature selection, hardware and software tools, model selection, mathematical foundations, predictive models, and their practical usefulness. Our aim is to provide a clear and comprehensive account of the steps taken, the tools utilized, and the rationale behind our choices.

**1. Data Sourcing**

**The Foundation of Informed Insights**

The foundation of any data-driven research project is the data itself. In our research, we placed significant emphasis on sourcing high-quality data from diverse and reputable sources. The data collected was integral to the creation and training of the predictive models that power **WellNest.ai**

Our data sourcing efforts spanned a range of health and wellness domains, from medical records to fitness tracker data. These datasets provided a rich tapestry of information related to health and wellness, including vital health parameters, lifestyle indicators,and physical activity data. By diversifying our data sources, we aimed to ensure the comprehensiveness of the data and the robustness of our predictive models.

**2. Data Preprocessing**

**Refining the Raw Material**

The data sourced for our research, while invaluable, often required refinement to make it suitable for analysis and model training. The data preprocessing stage played a pivotal role in this process. Its goal was to enhance the quality of the data, remove inconsistencies, and prepare it for subsequent analysis.

This phase involved several critical tasks, including handling missing values, removing outliers, and normalizing variables. Addressing missing data points was essential to ensure that our predictive models were trained on complete and consistent datasets. Outlier removal was undertaken to eliminate extreme data points that could skew our models' performance. Normalization of variables aimed to standardize the data, ensuring that no single variable unduly influenced the model.

During data preprocessing, graphical representations such as histograms, scatter plots, and box plots were employed to gain insights into data distributions. These visualizations facilitated the identification of potential data issues and the assessment of data quality.

**3. Feature Selection**

**Focusing on Relevance**

Feature selection was a critical step in our methodology, one that directly impacted the efficiency and interpretability of our predictive models. By selecting the most relevant features for each model, we aimed to enhance model performance and streamline the model's focus on key variables.

In the feature selection process, mathematical techniques such as correlation analysis were employed. These techniques allowed us to assess the relationships between variables and their impact on the outcomes we were predicting. By identifying the variables most strongly correlated with the predicted outcomes, we were able to reduce dimensionality, effectively excluding less relevant variables from our models.

The feature selection process was guided by the principle of dimensionality reduction, which aimed to simplify the models and improve their efficiency. This simplification led to more interpretable models, where the relationships between selected features and outcomes were more readily discernible.

**4. Hardware and Software Tools**

**Enabling Efficient Research**

To execute our research methodology effectively, we relied on a suite of hardware and software tools that were carefully selected to meet our specific research needs.

In terms of hardware, we made use of a high-performance computing cluster equipped with GPU acceleration. This configuration allowed us to carry out intensive computational tasks efficiently. The parallel processing capabilities of the GPU accelerated the training of our predictive models, reducing computation time significantly. By harnessing the power of this hardware, we were able to process large datasets and train complex machine learning models effectively.

On the software front, our primary tool was Python, a versatile and widely-used programming language in the field of data science and machine learning. Python's extensive library support and user-friendly syntax made it an ideal choice for data manipulation and analysis. Additionally, we employed machine learning libraries, with a particular focus on scikit-learn. These libraries provided a wide range of machine learning algorithms and tools that facilitated model development and training.

**5. Model Selection**

**Picking the Right Tool for the Job**

Model selection is a pivotal step in any predictive modeling project. The choice of the appropriate model can significantly impact the accuracy and interpretability of the predictions. In our research, we selected models based on the specific nature of the predictive tasks at hand.

For diabetes risk assessment and heart disease prediction, we chose to implement Logistic Regression models. Logistic Regression is well-suited for binary classification tasks, making it an excellent choice for predicting the likelihood of an event occurring. The mathematical foundation of Logistic Regression models is based on the logistic function, which maps input variables to a probability score. This characteristic allows for clear interpretation of the model's predictions, making it a valuable tool for health risk assessment.

Linear Regression, another fundamental machine learning technique, was chosen for calorie expenditure estimation. This model aims to establish a linear relationship between input variables and the predicted outcome, making it a natural choice for tasks involving continuous variables. Linear Regression is celebrated for its robustness in predictive modeling and its ability to provide straightforward insights into the relationships between variables.

**6. Mathematical Explanation**

**The Underpinning of Predictive Models**

It is essential to appreciate the mathematical foundations of the models selected for our research. Understanding the mathematical principles that underpin these models is key to comprehending their functionality and interpretability.

In the case of Logistic Regression, the model estimates the probability of an event occurring, such as diabetes risk or heart disease, by applying the logistic function. This function maps input variables to a probability score between 0 and 1, signifying the likelihood of the event. By applying a threshold to this probability score, predictions are made regarding the event. The mathematical underpinnings of Logistic Regression enable it to provide interpretable results, shedding light on the relationships between input variables and the predicted outcomes**.**

Linear Regression, on the other hand, estimates a continuous dependent variable by fitting a linear equation. The model seeks to find the linear relationship that best explains

the variance in the outcome variable. This is achieved by minimizing the sum of squared differences between the predicted values and the actual values. The mathematical foundation of Linear Regression ensures its reliability and interpretability.

**7. Predictive Models**

**The Core**

The predictive models developed in our research formed the core of the "WellNest.ai" platform. These models were responsible for providing users with accurate and personalized health predictions.

Graphical representations played a crucial role in illustrating the relationships within the data and the performance of the models. Receiver Operating Characteristic (ROC) curves, for example, were used to assess the accuracy of the binary prediction models. These curves provide a visual representation of the model's ability to discriminate between positive and negative outcomes, helping to evaluate model performance.

Scatter plots were employed to visualize the relationships between variables and model predictions. These visualizations allowed for an intuitive understanding of how input variables influenced the model's predictions. In the context of calorie expenditure estimation, scatter plots were particularly useful for comparing predicted calorie burn with actual values, aiding in tracking fitness progress.

**8. Usefulness: Empowering Informed Decisions**

The predictive models developed in our research had real-world applications and were instrumental in empowering individuals to make informed decisions about their health and wellness.  
For instance, the diabetes risk assessment model provided individuals with personalized risk assessments, enabling them to make informed decisions about their health. Users were able to assess their susceptibility to diabetes based on their health parameters and data inputs. This proactive approach to health management could potentially lead to lifestyle adjustments and early interventions, ultimately reducing the incidence of diabetes.

In the case of the heart disease prediction model, users received insights into the health of their heart, guiding them in making proactive health decisions. Given that heart diseases are a leading cause of mortality, a tool like WellNest.ai with its high accuracy has the potential to save lives.

The calorie expenditure prediction model, with its remarkable accuracy, was a boon for fitness enthusiasts and anyone dedicated to a healthier lifestyle. It offered real-time, data-driven estimates of the calories burned during various physical activities. Such information was invaluable for designing and adjusting workout routines, optimizing diet plans, and tracking progress toward fitness goals.

In conclusion, this comprehensive methodology accounts for the steps taken in our research journey to create the innovative WellNest.ai platform. By sourcing high-quality data, preprocessing it meticulously, selecting relevant features, utilizing appropriate hardware and software tools, choosing model types that fit the tasks, explaining the mathematical underpinnings, and illustrating the usefulness of the predictive models, we transformed raw data into actionable insights. The platform empowers individuals to proactively manage their health, make informed decisions, and embark on a journey of wellness

**Conclusion**

This expanded methodology section provides an in-depth view of the research process that led to the creation of "WellNest.ai." Each component plays a critical role in ensuring that the platform provides users with personalized, accurate, and actionable health insights. It also highlights the meticulous attention to detail, from data sourcing to model selection, that underpins the platform's reliability and usefulness.

The methodology we employed in our research reflects a commitment to the highest standards of data-driven decision-making in health and wellness management. The mathematical foundations of our chosen models ensure their interpretability, and the graphical representations offer insights into their performance. By empowering individuals to make informed decisions about their health, "WellNest.ai" stands as a testament to the power of data-driven healthcare.

In this comprehensive methodology, we have accounted for every facet of the research process, from the raw data to the practical applications of the predictive models. The result is a platform that empowers individuals to take control of their health and well-being, providing a brighter and healthier future for all.

**Author**

**Abhijit Mandal**

Department of Computer Science, Sushant University , Gurugram , India

**References**

Kaggle. (2019). *Heart Disease Prediction*. Kaggle. Retrieved November 12, 2023, from [Heart Disease Dataset](https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset)

WHO. (2021, June 11). *Cardiovascular diseases (CVDs)*. World Health Organization (WHO). Retrieved November 12, 2023, from [Cardiovascular diseases (CVDs)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))