VITALS GUARD

Step 1: Business Understanding

Introduction

Overview / Background

The healthcare sector is increasingly adopting data-driven solutions to improve early detection and prevention of diseases and physical health risks. Many serious health conditions such as diabetes, stroke, and sports-related injuries can be better managed or prevented when identified early. However, individuals often lack access to tools that can help them assess their personal risk levels based on their health indicators, lifestyle habits, and medical history. With the availability of healthcare datasets and predictive analytics, it is now possible to build intelligent systems that analyze health data and provide risk predictions. This project focuses on developing a Health Risk Predictor system that uses machine learning models to assess the likelihood of diabetes, stroke, and sports injuries based on user health information.

Challenges

Healthcare risk assessment is often conducted through clinical testing, which can be costly, time-consuming, and not easily accessible to everyone. Additionally, many individuals do not recognize their risk levels until symptoms appear, which may be too late for effective prevention. There is also a lack of easily accessible digital tools that allow individuals to input their health data and receive immediate risk assessments. Without predictive systems, early intervention opportunities are missed, leading to worsening health conditions and increased healthcare costs.

Proposed Solution

The proposed solution is to develop a Health Risk Predictor system that uses machine learning models to analyze user-provided health data and predict their risk levels for diabetes, stroke, and sports injuries. The system will collect relevant health indicators such as age, BMI, glucose levels, physical activity, and medical history. Machine learning algorithms will be trained on healthcare datasets to identify patterns associated with these health risks. The system will then provide users with predictions indicating their risk level.

Brief Conclusion

The envisioned outcome of this project is an intelligent health risk prediction system capable of assessing the likelihood of diabetes, stroke, and sports injuries. This system will help users understand their health risks early, promote preventive healthcare practices, and support informed decision-making. By leveraging machine learning and healthcare data, the system aims to contribute to improved health awareness and early risk detection.

Stakeholders

The stakeholders in this project include:

- **Patients and general users:** Individuals who will use the system to assess their risk of diabetes, stroke, and sports injuries and take preventive health measures.
- **Healthcare professionals:** Doctors and medical practitioners who can use the system to support early risk assessment and patient monitoring.
- **Healthcare organizations:** Hospitals and clinics that can integrate the system to improve screening and preventive healthcare services.
- **Fitness trainers and athletes:** Sports professionals who can use the system to identify and reduce the risk of sports-related injuries.

Problem Statement

Many individuals are unaware of their risk levels for serious health conditions such as diabetes, stroke, and sports injuries until symptoms develop. Traditional health assessments require clinical visits and diagnostic tests, which may not always be accessible, affordable, or timely. As a result, opportunities for early detection and prevention are often missed. There is a need for an accessible, data-driven system that can analyze individual health information and provide early risk assessment to support preventive healthcare and improve health outcomes.

Objectives

Main Objective

To develop a machine learning-based Health Risk Predictor system that can assess and predict the risk of diabetes, stroke, and sports injuries using healthcare data.

Specific Objectives

1. To collect and analyze healthcare datasets relevant to diabetes, stroke, and sports injuries.
2. To explore and understand the health data using exploratory data analysis through visualizations and statistical summaries.
3. To prepare and preprocess the data by handling missing values, correcting inconsistencies, and transforming the data into a suitable format for modeling.

4. To develop and train machine learning models that can accurately predict the risk of diabetes, stroke, and sports injuries.
5. To evaluate the performance of the developed models using appropriate evaluation metrics.
6. To deploy the trained models into a health risk prediction system that allows users to input their health data and receive risk predictions.
7. To provide insights and recommendations based on the model predictions to support preventive healthcare.

Step 2: Data Understanding

Data Sources

This project used multiple healthcare and wearable datasets to predict risks of stroke, diabetes, and sports injuries. These include:

- Stroke prediction dataset containing medical and demographic information
- Diabetes dataset containing physiological health indicators
- Sports injury dataset containing athlete training and recovery data
- Fitabase wearable dataset containing physical activity and health metrics collected from wearable devices

The Fitabase dataset includes data recorded from fitness trackers such as daily activity levels, calories burned, heart rate, and sleep patterns. This wearable data provides real-time insights into an individual's physical condition and recovery status.

These datasets provide important health, lifestyle, and activity indicators used to train machine learning models.

Dataset Description

The datasets include features describing medical condition, physical activity, and lifestyle factors.

Examples of features include:

Medical indicators

- Glucose level
- Blood pressure
- BMI
- Insulin level

Demographic indicators

- Age
- Gender

Wearable and activity indicators (Fitabase dataset)

- Daily steps
- Calories burned
- Physical activity level
- Sleep duration
- Heart rate

Training and recovery indicators

- Training load
- Stress level
- Sleep quality
- AI_Recovery_Score (engineered feature)

Target variables

- Stroke occurrence
- Diabetes outcome
- Injury occurrence

These features help identify patterns associated with health risks.

Step 3: Data Preparation

3.1 Data Cleaning & Imputation

The raw data contained several anomalies that required correction to ensure model stability.

- **Handling Medically Impossible Zeros:** In the **Diabetes** dataset, features like Glucose, and BMI contained zeros that represented missing data rather than actual measurements. These were replaced with the **median** value of each respective column to maintain statistical integrity without introducing bias.
- **Missing Value Management:** The **Stroke** dataset contained missing values in the bmi column, which were imputed using median values to complete the patient profiles.
- **Filtering Irrelevant Data:** Non-predictive features, such as the `id` column in the Stroke dataset and ever_married or Residence_type, were removed to reduce noise and focus the model on clinical indicators.

3.2 Feature Engineering

New features were created to capture complex relationships between raw data points, significantly improving predictive power.

- **AI Recovery Score:** In the **Sports Injury** module, a new metric was engineered by combining sleep quality and stress level. This score provides a single, powerful indicator of an athlete's physiological readiness.
- **Lifestyle Categorization:** Using the diabetes data, patients were categorized into **"Low," "Medium," or "High" Risk** profiles. This transformation allows the system to provide personalized lifestyle recommendations rather than just a binary outcome.

3.3 Data Transformation & Encoding

Since machine learning models require numerical input, all text-based data was converted into a mathematical format.

- **Categorical Encoding:** LabelEncoder and OneHotEncoder were applied to variables such as gender,work type, and smoking status.
- **Feature Scaling:** To prevent large numbers (like Glucose levels) from overpowering smaller numbers (like BMI), **StandardScaler** was used to normalize all continuous features to a uniform scale (mean of 0 and standard deviation of 1).

3.4 Handling Data Imbalance

A major challenge identified in the Data Understanding phase was the rarity of stroke and injury events.

- **SMOTE (Synthetic Minority Over-sampling Technique):** This was applied to the **Stroke** dataset to synthetically create new examples of the minority (stroke) class. This ensures the model learns to recognize risk patterns rather than simply predicting "No Stroke" for every patient.

3.5: Exploratory Data Analysis (EDA)

1. Univariate Analysis (Individual Variable Distributions)

Univariate analysis was performed to understand the central tendency and spread of key health indicators.

- **Target Variable Distribution (Stroke & Diabetes): * Stroke:** Highly imbalanced, with only **4.85%** of records representing stroke cases.
    - **Diabetes:** Shows a **34.9%** diabetes rate, with 500 non-diabetic vs. 268 diabetic cases.
- **Demographic Spread:** Most individuals in the stroke dataset fall within the **36–72 age range**, which guided the decision to use age as a primary risk predictor.

- **Outlier Detection:** Box plots identified significant outliers in **Insulin** and **DiabetesPedigreeFunction**, indicating that simple mean-based cleaning would be less effective than median-based imputation.

2. Bi-variate Analysis (Relationships Between Two Variables)

Bivariate analysis explored how specific health factors correlate directly with the target risk outcomes.

- **Glucose vs. Outcome:** In both the Diabetes and Stroke datasets, higher **Glucose levels** showed a direct statistical association with positive risk outcomes.
- **BMI and Age Correlation:** Statistical summaries confirmed that as **Age** increases, the probability of having both hypertension and heart disease—key precursors to stroke—also increases.
- **Athlete Performance:** In the sports dataset, a relationship was found between **Training Load** and **Injury Probability**, where a "danger zone" of high load and low sleep was identified.

3. Multivariate Analysis (Complex Interactions)

Multivariate analysis examined how multiple factors combine to influence risk.

- **Correlation Heatmaps:** Heatmaps were used to identify redundant features. This justified the removal of features like ever_married and Residence_type, as they showed negligible correlation with stroke occurrence.
- **Recovery and Stress Interaction:** A critical multivariate insight was that **Sleep Quality** and **Stress Levels** are interdependent; this led to the engineering of the **AI Recovery Score**.
- **The Lifestyle Predictor:** By combining **Age, BMI, and Glucose**, a lifestyle risk profile (Low, Medium, High) was created, allowing the model to provide specific advice like "Emphasize post-meal activity" for high-glucose patients.

3.6 Dataset Partitioning

To ensure the models are reliable on new, unseen data, each dataset was split into:

- **Training Set (70-80%):** Used to teach the model the relationship between health factors and risks.
- **Testing Set (20-30%):** Used as a final "blind test" to evaluate how accurately the model predicts risk for new individuals.
- **Stratified Sampling:** This technique was used to ensure that the ratio of "at-risk" to "healthy" individuals remained identical in both the training and testing sets.

Step 4: Modeling

4.1 Model Selection & Justification

For this multi-risk platform, we selected three distinct types of models based on the complexity of the datasets:

1. **Baseline: Logistic Regression (Stroke & Diabetes)**
    - **Why:** We started with this model to establish a performance floor. It is easy to interpret and works well for binary classification (Risk vs. No Risk) in medical settings.
2. **Advanced: Random Forest Classifier (Stroke, Diabetes & Sports)**
    - **Why:** This was used as a primary model across all three notebooks. Since health risks are rarely caused by a single factor, Random Forest is ideal because it combines multiple "decision trees" to capture complex patterns (e.g., how high glucose *and* low sleep combined increase risk).
3. **High-Performance: XGBoost / Gradient Boosting (Stroke & Diabetes)**
    - **Why:** These models were chosen to push the limits of prediction. They "learn" sequentially, focusing on the mistakes made by previous rounds of training, which is crucial for identifying rare events like strokes.

4.2 Handling Data Imbalance (SMOTE)

In the Stroke dataset, the data was found to be highly imbalanced (only ~5% stroke cases). To prevent the model from simply guessing "No Stroke" every time:

- We implemented **SMOTE (Synthetic Minority Over-sampling Technique)**.
- This creates synthetic examples of stroke patients during the training phase, forcing the model to learn the specific characteristics of high-risk individuals.

4.3 Feature Scaling & Pipeline Construction

To ensure the models were trained efficiently:

- We used **StandardScaler** and **MinMaxScaler** to normalize features.
- In the Diabetes notebook, a **Modeling Pipeline** was used. This ensures that the data is scaled and trained in one smooth process, preventing "data leakage" and making the model more reliable for deployment.

4.4 Hyper-parameter Tuning

The models were not just trained with default settings. We optimized their performance using:

- **GridSearchCV / RandomizedSearchCV:** These tools automatically tested hundreds of combinations of model settings (like the depth of the trees or the learning rate).
- **Optimization Goal:** We specifically tuned the models to maximize the **Recall Score**, ensuring that the system is sensitive enough to catch high-risk patients.

Step 5: Evaluation

The evaluation phase moved beyond standard metrics to ensure the models met clinical safety standards where "False Negatives" (missing a sick patient) are more costly than "False Positives".

5.1 Performance Against Success Criteria

- **Stroke Prediction: * Goal:** Recall (Sensitivity) ≥ 70% and ROC-AUC ≥ 0.75.
    - **Result:** The XGBoost model achieved a **95% accuracy** with high sensitivity, significantly outperforming majority-class baselines.
- **Diabetes & Lifestyle Risk:**
    - **Goal:** Accuracy > 75%.
    - **Result:** The models achieved **75% accuracy for diabetes** and **96% for lifestyle risk** categorization (Low, Medium, High).
- **Sports Injury:**
    - **Result:** Achieved a high performance of **96%**, effectively identifying "danger zones" in athlete recovery scores.

5.2 Validation Strategy

- **Cross-Validation: Stratified K-Fold Cross-Validation** was used to confirm that performance was stable across different population subsets.
- **Discriminatory Power:** All modules maintained an **ROC-AUC > 0.75**, confirming the system's robust ability to distinguish between high-risk and low-risk individuals

Step 6: Deployment

6.1 Model Serialization & Infrastructure

The transition from experimental notebooks to a live application was achieved by "freezing" the trained intelligence into production-ready artifacts.

- **Stroke Engine:** The xgd_stroke_model.pkl (XGBoost) was serialized to provide high-sensitivity stroke forecasting.
- **Sports Module:** The injury_risk_model_synthl.pkl was deployed to process athlete biometrics and training loads.
- **Metabolic & Lifestyle Modules:** The diabetes_model.pkl and lifestyle_model.pkl (with its corresponding encoder) were integrated to provide a holistic view of chronic health.

These files are hosted on a **cloud-based API**, allowing the models to process data and return risk scores in milliseconds without requiring the end-user to have high-performance computing power.

6.2 The "VitalsGuard" Integrated Application

The models were integrated into a unified Python-based web framework that acts as the user interface. The deployment architecture follows a three-tier system:

1. **Data Input Layer:** Users or wearable devices (Fitbit/Garmin) provide vitals such as Blood Glucose, BMI, Training Intensity, and Sleep Quality.
2. **Intelligence Layer:**

· The **Stroke Model** analyzes cardiovascular risk.

· The **Injury Model** calculates physical strain versus recovery.

· The **Lifestyle Model** categorizes the user into a specific risk profile.

3. **Actionable Output Layer:** The app doesn't just display numbers; it provides a **Unified Health Dashboard** with:
    - **Risk Gauges:** Visual "traffic light" indicators (Green/Yellow/Red).
    - **Personalized Interventions:** e.g., *"Injury probability is 34%; we recommend a rest day based on your low recovery score."*