TEAM JOHNSON-SIRLEAF PROJECT PROPOSAL

PROPOSED PROJECT TOPIC: MACHINE LEARNING APPROACH TO PREDICTING DIABETES RISKS.

INTRODUCTION AND PROBLEM DESCRIPTION

Diabetes Mellitus, fondly called scientifically, is one of the major health concerns in the public health space. This disease is one of the four primary diseases being tracked by the World Health Organization (WHO), including cardiovascular diseases, cancer, and chronic respiratory diseases. Diabetes occurs either when there is a deficiency in the production of insulin by the pancreas, a gland located behind the stomach, or when there is a mismanagement of the insulin produced by the body. Insulin is a critical hormone that regulates the body's energy storage and blood sugar levels, it helps keep the blood sugar in a healthy range by facilitating the uptake of glucose into the cells of the body, promoting glucose storage, and preventing excessive sugar buildup in the bloodstream. Some of the factors that contribute to the development of diabetes include age, diet, obesity, physical inactivity, high blood pressure, and genetics. It can also lead to severe complications such as cardiovascular issues, kidney disease, and amputations. For example, according to the World Health Organization (WHO), in 2019 diabetes and kidney disease due to diseases led to approximately 2 million deaths and there has been a 3% increase in diabetes-related deaths between 2000 and 2019 (World Health Organization 2021).

The detection of this disease at an infancy level has been a persistent problem for public health professionals. Knowing that even as noxious as diabetes is, its harmful effect can be obviated if detected at an early stage and treated consequently medically. In the field of machine learning,

efficient algorithms can be developed and employed to detect the potential of an individual developing diabetes (Di Angelantonio & Lancer 2010). This would significantly help reduce the cases of death cases relating to diabetes and also reduce the workload on medical professionals. This project aims to make diabetes-related data generated in the medical field actionable by employing the power of data and technology to develop algorithms that can help predict and detect the development of diabetes in the body at an early stage.

The Objectives

The objective of this machine learning approach to predicting diabetes risks are:

- To accurately identify individuals at high risk of developing diabetes based on a combination of their clinical, genetic, environmental, and lifestyle factors, thereby enabling timely interventions and personalized care.

- Predicting diabetes risks using a machine learning approach has the potential to revolutionize healthcare by providing timely and accurate risk assessments. Here are five primary objectives of such an approach:

- Early Detection and Prevention: One of the main objectives is to identify individuals at high risk of developing diabetes in the future. Early detection allows for timely interventions, which can prevent or delay the onset of the disease. Preventative measures can include lifestyle modifications, such as diet and exercise changes, or medication for those at high risk.

- Personalized Risk Assessments: Different individuals may have varying risk factors for diabetes based on their genetics, lifestyle, and other health conditions. A machine learning approach can analyze complex datasets to provide personalized risk assessments based on a myriad of factors. This allows for recommendations to be tailored to the individual rather than applying a one-size-fits-all approach.

- Efficient Resource Allocation: In healthcare systems with limited resources, it's essential to prioritize individuals who are at the highest risk. By accurately predicting diabetes risks, healthcare professionals can allocate resources and interventions to those who need them most, ensuring that preventive efforts are as effective and efficient as possible.

- Continuous Learning and Improvement: A significant advantage of machine learning models is their ability to continuously learn and improve. As more data is collected, the model can be retrained to improve its predictions. This ensures that the model remains up-to-date with the latest trends and insights, making predictions even more accurate over time.

- Comprehensive Analysis of Risk Factors: Traditional approaches to predicting diabetes might focus on a few well-known risk factors. Machine learning can analyze a vast number of potential predictors simultaneously, from genetic markers to lifestyle factors, and even combine information from different types of data (e.g., imaging data, electronic health records, wearable device data). This comprehensive analysis can reveal previously

unknown risk factors or highlight interactions between factors that are significant in predicting diabetes risk.

EXISTING SOLUTION

- Clinical Risk Assessment Tools: Healthcare providers often use clinical risk assessment tools and questionnaires to estimate an individual's risk of developing diabetes. These tools take into account factors like age, family history, BMI, and lifestyle habits to assess risk. Examples include the Diabetes Risk Score (FINDRISC) and the American Diabetes Association's Risk Test.

- Machine Learning-Based Predictive Models: Many research studies and healthcare institutions have developed predictive models for diabetes risk using machine learning. These models utilize features like age, gender, BMI, and various health parameters to

make predictions. Some of these models have been integrated into electronic health records to assist healthcare professionals in identifying high-risk patients.

- mHealth Apps and Wearables: Mobile health (mHealth) applications and wearable devices are increasingly being used for diabetes risk assessment and management. These apps can track health parameters, provide risk assessments, and offer lifestyle recommendations. Some can even predict hypoglycemic events for individuals with diabetes.

- Telemedicine and Remote Monitoring: Telemedicine services and remote monitoring solutions allow healthcare providers to remotely monitor patients with diabetes or those at risk. These platforms often use real-time data from connected devices and sensors to assess and manage diabetes risk and progress.

- Population Health Management Systems: Healthcare systems and insurance companies use population health management systems to identify individuals at risk of diabetes. These systems aggregate data from various sources to assess population health and target interventions for at-risk groups.

- Artificial Intelligence for Diabetic Retinopathy Detection: AI is also being used to detect diabetic retinopathy, a complication of diabetes that affects the eyes. AI-based retinal screening systems can identify early signs of retinopathy, helping in the early diagnosis and treatment of this condition.

- Research Initiatives and Public Health Campaigns: Public health organizations and research institutions often conduct studies and campaigns to identify and educate individuals at risk of diabetes. These initiatives may include community screenings, educational materials, and awareness programs.

- Medication and Treatment Advances: While not predictive in nature, advances in diabetes medications and treatments have significantly improved the management of the condition. Medications, insulin delivery systems, and glucose monitoring technologies have become more effective and user-friendly.

It's important to note that the effectiveness of these solutions may vary, and they often complement each other. Combining clinical assessments, machine learning models, mHealth tools, and population health management systems can provide a comprehensive approach to predicting and managing diabetes risk. Additionally, the integration of electronic health records and health information exchange systems allows healthcare providers to share information and coordinate care for individuals at risk or those already diagnosed with diabetes.

PROPOSED METHOD

1. Data Collection and Preprocessing

1.1 Data Sources

The dataset to be used was obtained from Kaggle. The dataset can be accessed with this link. This dataset contains demographic data from patients as well as their diabetes status. This dataset contains 10,000 records with each record having eight features as well as the diabetes state of the patient. The features are explained below:

1. **Gender (‘gender)**: This is the gender of the patient. 59% of the patients are females, 41% are males and 0% are others.
2. **Age (’age’)**: This is the age of the patient within the range of 0 to 80.
3. **Hypertension (‘hypertension’)**: This records if the patient has hypertension or not. 1 indicates that the patient has hypertension while 0 indicates that the patient does not have hypertension.
4. **Heart Disease (‘heart_disease’)**: This records if the patient has heart disease or not. 1 indicates that the patient has heart disease while 0 indicates that the patient does not have heart disease.
5. **Smoking History (‘smoking_history’)**: This record shows the smoking characteristics of the patient. The values in this field are 'never', 'No Info', 'current', 'former', 'ever', and 'not current'.
6. **BMI (‘bmi’)**: This record shows the Body Mass Index of the patient.
7. **Hemoglobin A1c Level (‘HbA1c_level’)**: This record shows the Hemoglobin A1c level of the patient. This is the average blood sugar level of the patient over the past 2-3 months.
8. **Blood Glucose Level (‘blood_glucose_level’)**: This record shows the blood glucose level of the patient at the time of taking the record.
9. **Diabetes (‘diabetes’)**: This is the target label. The presence of diabetes is indicated with 1 while the absence of diabetes is indicated with 0.

The first five records in this dataset are shown below:

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0

1.2 Data Preprocessing

- Data Cleaning: The dataset will be inspected for missing values, duplicates, and outliers, and appropriate actions will be taken to address these issues.
- Feature Selection: We will analyze the relevance of each feature in the dataset and select the most informative variables for the predictive models.
- Encoding: Categorical variables, such as gender and smoking history, will be one-hot encoded for compatibility with machine learning algorithms.

2. Exploratory Data Analysis (EDA)

2.1 Descriptive Statistics

We will compute descriptive statistics to gain insights into the distributions of key variables, providing an initial understanding of the data.

2.2 Data Visualization

Data visualization techniques, such as histograms, box plots, and correlation matrices, will be employed to visualize the relationships between variables and identify patterns within the data.

3. Feature Engineering

Feature engineering will involve the creation of new variables and transformations to enhance the model's predictive power. This may include scaling, standardization, and the creation of interaction terms or composite variables.

4. Model Selection

4.1 Machine Learning Models

Various machine learning algorithms will be considered for predicting diabetes risk. These may include:

- Logistic Regression
- Decision Trees
- Random Forest
- Support Vector Machines
- Neural Networks (Deep Learning)
- Gradient Boosting

4.2 Model Evaluation

To evaluate the models' performance, we will employ metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Cross-validation techniques will be used to assess generalization performance, and hyperparameter tuning will be performed to optimize model parameters. The machine learning models can also be trained with base hyperparameters and hyperparameter tuning can be performed on the best model.

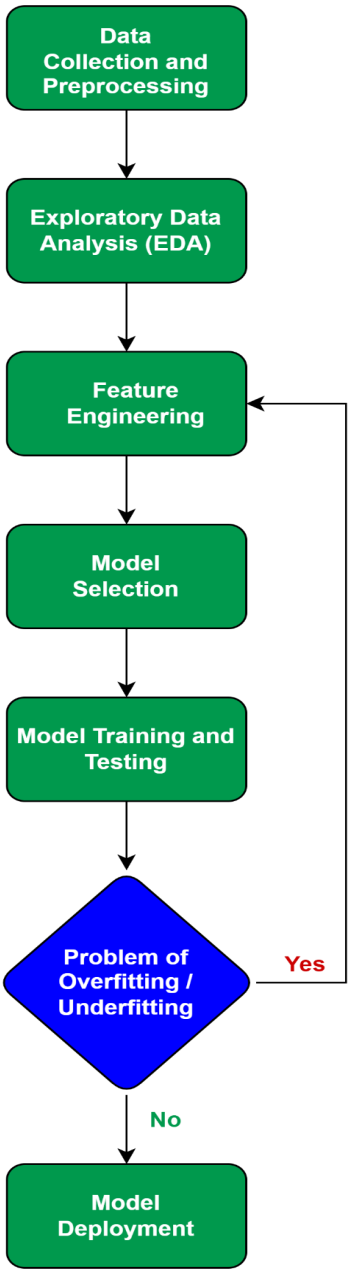
5. Model Training and Testing

The dataset will be split into training and testing sets to train and validate the models. A portion of the data will be set aside for testing to ensure the model's predictive performance on unseen data.

6. Model Deployment

Upon selecting the best-performing model, we will deploy it to predict diabetes risk in real-world scenarios. This could involve the development of a user-friendly application or integration into a healthcare system.

Project Flowchart



PROPOSED SPLIT

GLORY - DATA SOURCING & PREP

FEMI - DATA CLEANING

USMAN - MODEL

DOLAMU - MODEL

GREY - DEPLOYMENT

PROPOSED TIMELINE

Data collection and Research on Data.
16-Oct - 22-Oct (1week)

Data Cleaning and Data Transformation
23-Oct - 29-Oct (1 week)

Data Visualization and insight extraction
30-Oct - 5-Nov (1 week)

Model building
6-Nov - 12-Nov (1 week)

Model Tuning, Review, Comparison and Selection
13-Nov - 19-Nov (1 week)

Model Deployment
20-Nov - 26-Nov (1 week)

Feedback Generation, Testing and Maintenance
27-Nov - 3-Dec (1 week)

CONCLUSION

In conclusion, this is the proposed outline for the Team Johnson-Sirleaf. The project was introduced, following the aim and objectives of this project. The methodology outlines the steps to be undertaken for predicting diabetes risk using a machine learning approach, leveraging the two datasets provided. The analysis and model development will lead to a tool for early diabetes risk prediction, which can be valuable in healthcare settings for proactive patient management. Also, the timeline needed to execute this project is clearly stated and a proposed split of the work between the team members is also included.