**Data Analysis Strategy Document for Diabetes Risk Detection Project**

**1. Project Overview**

**1.1 Objective**

Develop a predictive analytics tool using machine learning to identify individuals at high risk of developing Type 2 diabetes based on a comprehensive set of risk factors.

**1.2 Stakeholders**

Project Team (Data Scientists, Analysts)

Healthcare Providers

Patients and Patient Advocacy Groups

Public Health Organizations

**2. Data Collection**

**2.1 Data Sources**

Public health databases (CDC, WHO)

Electronic Health Records (EHRs) from partnering healthcare providers

Clinical trials and research studies

Lifestyle and dietary surveys

**2.2 Data Requirements**

Demographic information (age, gender, ethnicity)

Medical history (previous diagnoses, comorbidities)

Lifestyle factors (diet, physical activity, smoking status)

Lab test results (blood glucose levels, cholesterol)

Genetic data (if available)

**3. Data Processing**

**3.1 Preprocessing Steps**

Cleaning: Address missing values and remove duplicates.

Transformation: Standardize lab test values and encode categorical variables.

Feature engineering: Develop new variables (e.g., BMI from height and weight) that may be predictive of diabetes risk.

**3.2 Data Splitting**

Split the dataset into training (80%) and testing (20%) sets to ensure model evaluation is performed on unseen data.

**4. Analysis Methodology**

**4.1 Machine Learning Models**

Initial Models: Logistic Regression (for baseline performance), Random Forest, and Gradient Boosting Machines (GBM).

Advanced Exploration: Neural Networks for deep learning analysis, conditional on data size and complexity.

**4.2 Feature Selection**

Utilize techniques like Recursive Feature Elimination (RFE) and model-based feature importance to identify key predictors of diabetes risk.

**4.3 Model Evaluation Metrics**

Primary Metrics: P-Value, Accuracy, Precision, Recall, F1 Score, and AUC-ROC.

Cross-validation: Employ k-fold cross-validation for assessing model stability and generalizability.

**5. Interpretation and Reporting**

**5.1 Key Findings**

Summarize the most predictive features of diabetes risk.

Report model performance across various metrics.

**5.2 Model Insights**

Provide insights into the relationship between risk factors and the likelihood of developing diabetes.

Highlight any unexpected findings or patterns in the data.

**5.3 Recommendations**

Suggest potential interventions or further studies based on model findings.

Propose how healthcare providers can integrate the predictive tool into patient care workflows.

**6. Ethical Considerations**

**6.1 Data Privacy**

Ensure all patient data is anonymized and securely stored.

Comply with all relevant data protection regulations (e.g., HIPAA, GDPR).

**6.2 Bias and Fairness**

Assess and mitigate potential biases in the data collection and model training processes.

Ensure the tool is fair and equitable across different patient demographics.

**7. Project Timeline**

Phase 1 - Data Collection and Processing: Weeks 1-4

Phase 2 - Model Development and Initial Testing: Weeks 5-8

Phase 3 - Model Refinement and Validation: Weeks 9-12

Phase 4 - Analysis and Reporting: Weeks 13-15

**8. Approval and Revision History**

**Document Created:** [02/02/2024]

**Reviewed By:** [Naija Thomas, Olayinka Ajao, Pamela Djan], [02/02/2024]

**Approved By:** [Naija Thomas, Olayinka Ajao, Pamela Djan], [02/02/2024]

**Revision History:** [List of revisions made to the document]