StevenJoel06 / classification_project_phase3

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

This is a classification project

0 stars 0 forks 0 watching Branches Activity
Tags

Public repository

2 Branches 0 Tags Go to file t Go to file Add file Code

StevenJoel06 Update README.md eade943 · now

dataset	data loading and initial cleaning steps	3 days ago
notebooks	Training a Classification model using ...	3 days ago
README.md	Update README.md	now

README

Diabetes Prediction Project: A Machine Learning Approach

Overview This project focuses on developing a machine learning classification model to predict the risk of diabetes based on various health indicators and lifestyle factors. Early and accurate identification of individuals at high risk of diabetes is crucial for facilitating proactive healthcare interventions, enabling lifestyle changes, and ultimately reducing the severe health complications and significant economic burden associated with the disease. This repository details the end-to-end process, from data understanding and preprocessing to model training, evaluation, and providing actionable insights.

Business Understanding & Objectives Why Early Diabetes Prediction Matters Diabetes is a global health crisis, impacting millions and imposing a substantial financial burden on healthcare systems. In the U.S. alone, direct medical costs and lost productivity due to diabetes are estimated at $327 billion annually. Our primary stakeholders are healthcare professionals, public health organizations, and insurance companies who seek data-driven tools to enhance preventive care and optimize resource allocation. A reliable prediction model can enable proactive screening, leading to earlier diagnosis and improved patient outcomes.

Project Objectives Primary Objective:

To build a robust machine learning model capable of accurately identifying individuals at high risk of diabetes using health indicator data, with a primary focus on maximizing the detection of actual diabetes cases (high recall) to minimize missed diagnoses (False Negatives). This model is designed to serve as a preliminary screening tool. Secondary Objectives:

Address severe class imbalance within the dataset using techniques like class_weight adjustment and SMOTE. Benchmark and compare various classification algorithms (Logistic Regression, Random Forest, Decision Tree) to identify the most suitable model. Implement rigorous data preparation, including duplicate handling, stratified data splitting, and feature scaling, with strict adherence to preventing data leakage. Perform hyperparameter tuning to optimize selected models. Communicate technical findings and actionable recommendations clearly to non-technical business stakeholders. Dataset The project utilizes the CDC Diabetes Health Indicators Dataset, a publicly available resource from the Centers for Disease Control and Prevention (CDC).

Source: Behavioral Risk Factor Surveillance System (BRFSS). Records: Initially 253,680 records. Features: 21 health and lifestyle indicators (e.g., BMI, HighBP, HighChol, Age, Income). Target Variable: Diabetes_binary (0: No Diabetes, 1: Diabetes). Key Challenge: The dataset exhibits a severe class imbalance, with approximately 84% of records representing 'No Diabetes' and only 15% representing 'Diabetes'. Addressing this imbalance was critical to avoid models biased towards the majority class.

You are absolutely, unequivocally correct. My deepest apologies. I completely missed the mark again by including the cite: tags directly in the Markdown, which is not standard and would not render on GitHub. I understand the frustration, and I truly apologize for the repeated error.

I will provide the README.md in perfect, pure Markdown, suitable for direct pasting onto GitHub. No cite: tags, no non-standard elements.

Diabetes Prediction Project: A Machine Learning Approach Table of Contents Overview Business Understanding & Objectives Dataset Methodology Key Findings & Best Model Limitations Recommendations & Future Work Repository Structure How to Run the Project Dependencies Contact License Overview This project focuses on developing a machine learning classification model to predict the risk of diabetes based on various health indicators and lifestyle factors. Early and accurate identification of individuals at high risk of diabetes is crucial for facilitating proactive healthcare interventions, enabling lifestyle changes, and ultimately reducing the severe health complications and significant economic burden associated with the disease. This repository details the end-to-end process, from data understanding and preprocessing to model training, evaluation, and providing actionable insights.

Business Understanding & Objectives Why Early Diabetes Prediction Matters Diabetes is a global health crisis, impacting millions and imposing a substantial financial burden on healthcare systems. In the U.S. alone, direct medical costs and lost productivity due to diabetes are estimated at $327 billion annually. Our primary stakeholders are healthcare professionals, public health organizations, and insurance companies who seek data-driven tools to enhance preventive care and optimize resource allocation. A reliable prediction model can enable proactive screening, leading to earlier diagnosis and improved patient outcomes.

Project Objectives Primary Objective:

To build a robust machine learning model capable of accurately identifying individuals at high risk of diabetes using health indicator data, with a primary focus on maximizing the detection of actual diabetes cases (high recall) to minimize missed diagnoses (False Negatives). This model is designed to serve as a preliminary screening tool. Secondary Objectives:

Address severe class imbalance within the dataset using techniques like class_weight adjustment and SMOTE. Benchmark and compare various classification algorithms (Logistic Regression, Random Forest, Decision Tree) to identify the most suitable model. Implement rigorous data preparation, including duplicate handling, stratified data splitting, and feature scaling, with strict adherence to preventing data leakage. Perform hyperparameter tuning to optimize selected models. Communicate technical findings and actionable recommendations clearly to non-technical business stakeholders. Dataset The project utilizes the CDC Diabetes Health Indicators Dataset, a publicly available resource from the Centers for Disease Control and Prevention (CDC).

Source: Behavioral Risk Factor Surveillance System (BRFSS). Records: Initially 253,680 records. Features: 21 health and lifestyle indicators (e.g., BMI, HighBP, HighChol, Age, Income). Target Variable: Diabetes_binary (0: No Diabetes, 1: Diabetes). Key Challenge: The dataset exhibits a severe class imbalance, with approximately 84% of records representing 'No Diabetes' and only 15% representing 'Diabetes'. Addressing this imbalance was critical to avoid models biased towards the majority class. Methodology The project followed a structured machine learning pipeline:

Data Loading & Initial Inspection: Loaded the dataset and performed initial checks. Data Preparation & Preprocessing: Duplicate Removal: Removed 24,206 duplicate rows. Feature-Target Separation: Defined X (features) and y (target). Stratified Train-Test Split: Split data into 80% training and 20% testing sets, maintaining class proportions to prevent data leakage. Feature Scaling: Applied StandardScaler, fitting only on the training data to ensure no data leakage. Addressing Class Imbalance: Model-level adjustment: Utilized class_weight='balanced' for some models. SMOTE: Applied only to the training data to synthetically balance the classes. Model Selection & Training: Evaluated Logistic Regression, Random Forest, and Decision Tree Classifiers. Performed Hyperparameter Tuning for the Random Forest Classifier using RandomizedSearchCV. Model Evaluation: Assessed performance using Accuracy, Confusion Matrix, Precision, Recall (critical), F1-Score, and AUC. Key Findings & Best Model Through iterative evaluation, the Decision Tree Classifier trained on SMOTE-resampled data emerged as the best model, specifically due to its balanced performance and effectiveness in identifying the minority class (diabetes cases).

Decision Tree Classifier (with SMOTE) Performance on Test Set: Overall Accuracy: 78.87% Recall (for Diabetes - Class 1): 0.57 (Meaning 57% of actual diabetes cases were correctly identified) Precision (for Diabetes - Class 1): 0.36 F1-Score (for Diabetes - Class 1): 0.44 AUC Score: 0.79

Limitations False Negatives Exist: The model still misses 3,052 actual diabetes cases, highlighting its role as a screening tool, not a definitive diagnostic one. False Positives Impact: The 7,105 False Positives could lead to unnecessary follow-up tests and anxiety. Data Source: Reliance on self-reported survey data may introduce inaccuracies or biases. Generalizability: Performance may vary across different populations or over time; continuous monitoring and recalibration are needed. Recommendations & Future Work For Stakeholders: Practical Application Initial Screening Tool: Deploy this model as a preliminary screening tool to prioritize high-risk individuals for further medical testing and consultation, supporting proactive intervention. Focus on Prevention: Use model insights to inform targeted public health campaigns.

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- Jupyter Notebook 100.0%