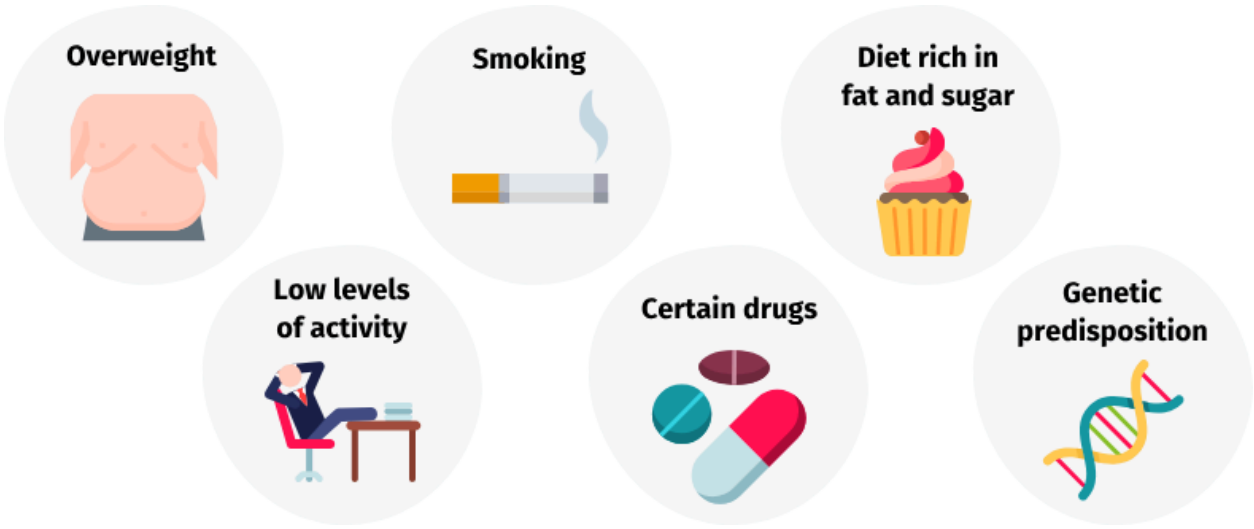
Februari t/m Juni 2024

A Retrospective Analysis of Type-2 Diabetes Risk 10 Years Later

Plan of Approach



ANCORA

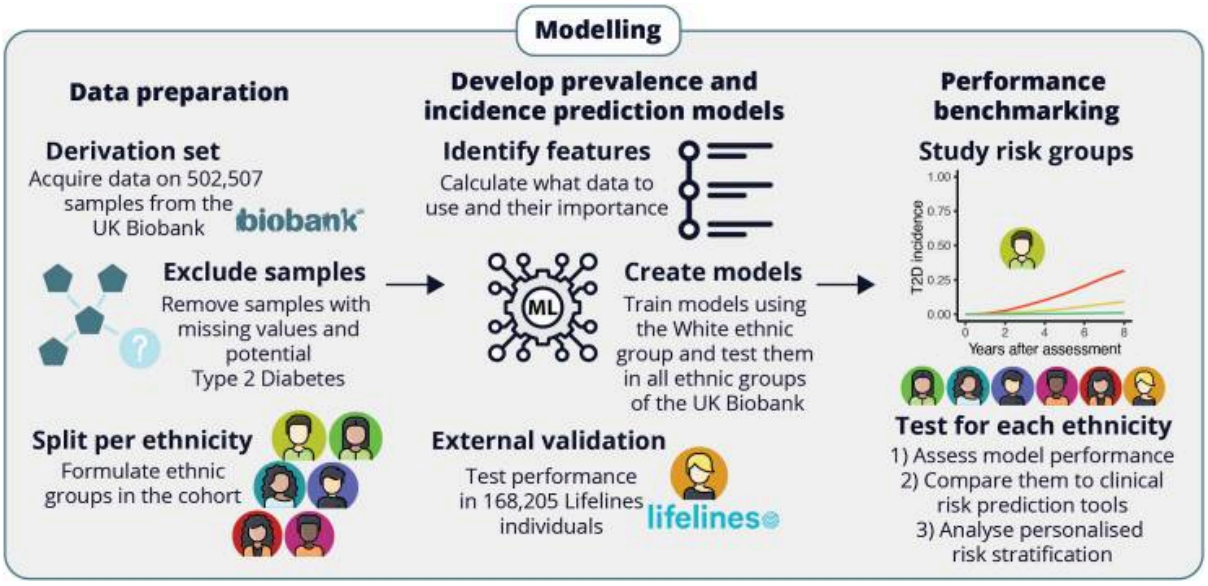
Hanzehogeschool Groningen

Student Jan Alfonso Busker j.a.busker@st.hanze.nl
Supervisor Sipko van Dam Sipkovandam@gmail.com
Supervisor Michiel Noback m.a.noback@pl.hanze.nl

Introduction

Background

Type-2 Diabetes (T2D) poses a significant global health challenge, characterized by its complex etiology and prolonged asymptomatic phase [1]. Early identification of individuals at risk is crucial for timely intervention and preventive measures. The landscape of T2D prediction has evolved, and predictive models have been developed to assess an individual's likelihood of developing the condition [1,2].



Workflow showing the steps taken to prepare the data and to create questionnaire-based prediction models for prevalent and incident type 2 diabetes.

The existing predictive models [3] have shown promise in discerning individuals with an elevated risk of T2D. However, to enhance the practical utility of these models, it is imperative to investigate their performance over an extended timeframe. T2D is known to manifest itself years after the initial risk assessment [1,3], emphasizing the need for a comprehensive evaluation of these predictive models' efficacy in anticipating long-term outcomes.

The main objective of this project is to assess the effectiveness of current T2D predictive models in identifying individuals who would develop the condition a decade later. This involves creating a user-friendly tool that incorporates a machine learning algorithm into a R Shiny application, allowing users to easily generate logistic regression models.

Research question: "How well does the existing Type-2 Diabetes risk prediction model identify individuals who develop T2D a decade after the initial assessment?"

This project aims to advance public health outcomes by enhancing T2D prediction accuracy, enabling early detection, and facilitating preventative interventions. The approach merges an

extensive dataset with cutting-edge machine learning strategies, setting a new standard for disease risk prediction.

Approach and Methods

The following approach and methods will be used:

Data Collection:

We will leverage an updated dataset with a larger representation of the Type-2 Diabetes (T2D) population for both model development and validation. The study data are sourced from the UK Biobank and Lifelines. Access to the data is available upon request and subject to permissions from the respective data providers.

Preprocessing and Model Development:

To ensure data quality and enhance model performance, we will implement a user-friendly preprocessing tool within an R Shiny application. This tool will facilitate the handling of missing values through imputation and provide options for data scaling. Subsequently, the R Shiny application will integrate a machine learning algorithm, enabling users to easily build logistic regression models for T2D prediction.

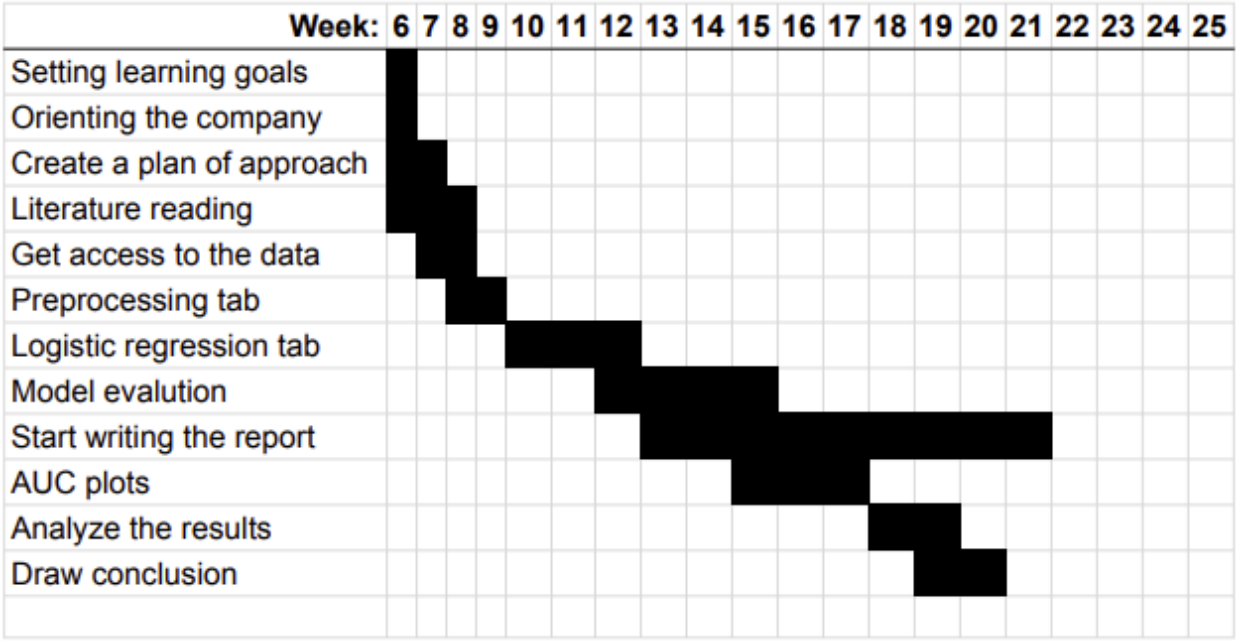
Model Evaluation:

We will evaluate the efficacy of existing predictive models in identifying individuals who develop T2D a decade after the initial assessment. To achieve this, we will employ advanced algorithms such as LASSO (Least Absolute Shrinkage and Selection Operator) to identify the most predictive variables. Additionally, techniques like Ridge regression will be utilized to mitigate overfitting and ensure the generalizability of the models. Model performance and stability will be assessed using k-fold cross-validation across different subsets of the dataset.

Calculation of AUC using pROC package:

After developing the logistic regression models, we will calculate the Area Under the ROC Curve (AUC) to quantify the models' discriminatory power. This metric provides a comprehensive assessment of the model's ability to distinguish between individuals who develop T2D and those who do not. We will utilize the pROC package in R to compute the AUC and generate ROC curves for visual representation.

Schedule


Week: 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Setting learning goals
Orienting the company
Create a plan of approach
Literature reading
Get access to the data
Preprocessing tab
Logistic regression tab
Model evalution
Start writing the report
AUC plots
Analyze the results
Draw conclusion

Organisation

This assignment is given by Ancora Health. Ancora Health is an organization dedicated to advancing public health outcomes through research, data analysis, and the development of predictive models for various health conditions. Founded with a mission to address global health challenges, Ancora Health collaborates with academic institutions, healthcare professionals, and research partners to drive innovation in disease prevention and management.
Internal guidance (from the Hanze) is provided by Dr. M.A. Noback. And external guidance is provided by Dr. S. van Dam.

References

[1] Reed J, Bain S, Kanamarlapudi V. A Review of Current Trends with Type 2 Diabetes Epidemiology, Aetiology, Pathogenesis, Treatments and Future Perspectives. Diabetes Metab Syndr Obes. 2021 Aug 10;14:3567-3602. doi: 10.2147/DMSO.S319895. PMID: 34413662; PMCID: PMC8369920.
[2] Singer ME, Dorrance KA, Oxenreiter MM, Yan KR, Close KL. The type 2 diabetes 'modern preventable pandemic' and replicable lessons from the COVID-19 crisis. Prev Med Rep. 2022 Feb;25:101636. doi: 10.1016/j.pmedr.2021.101636. Epub 2021 Nov 18. PMID: 34909369; PMCID: PMC8660571.
[3] Kokkorakis M, Folkertsma P, van Dam S, Sirotin N, Taheri S, Chagoury O, Idaghdour Y, Henning RH, Forte JC, Mantzoros CS, de Vries DH, Wolffenbuttel BHR. Effective questionnaire-based prediction models for type 2 diabetes across several ethnicities: a model development and validation study. EClinicalMedicine. 2023 Sep 25;64:102235. doi: 10.1016/j.eclinm.2023.102235. PMID: 37936659; PMCID: PMC10626169.