Comparing Diabetes Risk Assessment Scores

Question: Can we create a better predictive system for T2DM risk via machine learning or weighting factors as compared to the classical pen and paper method?

Background:

Diabetes Mellitus Type 2 (T2DM(is a chronic metabolic disease involving resistance to the hormone insulin that afflicts nearly 3.8 million in the UK and is one of the most significant comorbidities for a range of diseases including cardiovascular disease, strokes, and obesity. Although it currently has no cure, several factors including weight loss are known to prevent or even lead to remission of the disease, implying that if we can identify target patients, there is potential to focus efforts on prophylactic treatments. T2DM currently costs the NHS billions to treat, meaning there is a great need for these sorts of efforts.

The NHS currently offers a diabetes risk calculator created by Diabetesuk, the University of Leicester and University Hospitals of Leicester NHS Trust that allows patients to input their parameters such as their ages, genders, ethnicity, etc. This score will then output a patient's percentile risk factor for T2DM. There are also pen and paper methods that involve simply adding points for each risk factor a patient presents with that are used in clinical settings.

There are multiple techniques to apply machine learning to data, including neural networks and random forests. Machine learning is currently being applied to several medical fields and represents outstanding potential to improve patient care and clinical outcomes. The random forest technique from machine learning would be helpful in this study; it is widely used to process continuous or discrete data to find the best-fit pattern of a data set. 'Scikit-learn' is a library that can be used to perform the random forest method.

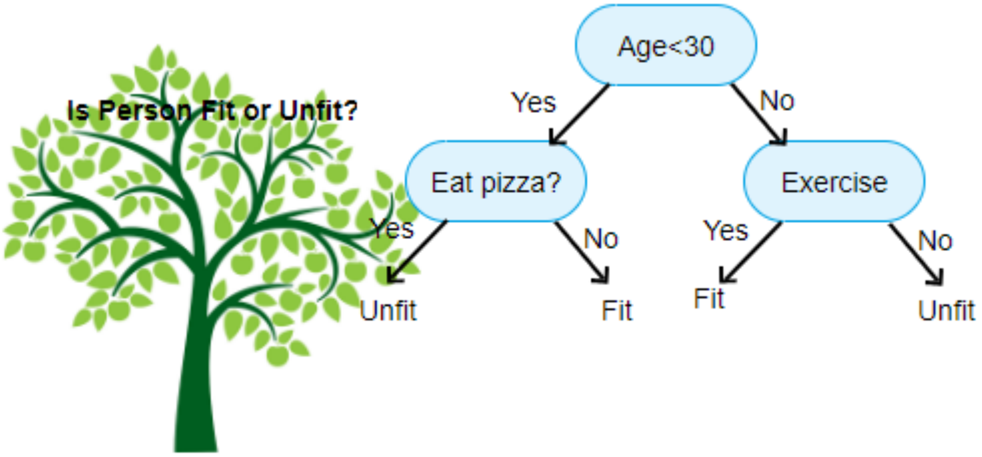
We can also create a 1-layer factor-weighting model where we can solve for each factor's weighting using pseudoinverse matrices. Here, we will attempt to solve for x where Ax=B, with A being our input parameters and B our risk assessment. Moore-Penrose pseudoinverses allow us to generate the pseudoinverse of matrix A even if it's not a square matrix (which it won't be due to our data points). We can find Moore-Penrose pseudoinverses classically with $(A^TA)^{-1}$, but also via Singular Value Decomposition. Here, we found it easier to use some Python functions from the numpy library, but these values could have been relatively easily calculated via Gauss-Jordan Elimination. This is effectively fitting the patients parameters linearly to a curve.

We will determine whether such methods create a better predictive model than the NHS's current pencil and paper method, which we will apply to the data via Python. We found patient data collected in India on several hundred T2DM patients that we then used to compare methods.

While we hope to obtain a better AI model, we also recognize the utility of easy pen and paper models in clinical settings. Thus, we will also try to develop a system requiring only simple math but offering more predictive power than current methods guided by our factor weighting models.

Disciplinary Aspects: We will combine knowledge from mathematics and computing and apply it towards predictive medical models. While the current NHS model is sufficient for most clinical needs, we hope to see how it fares against modern computational methods. Here, we combine aspects of mathematics such as basic linear algebra to solve for weighting factors as well as machine learning through random forest techniques. We will also utilize team members' medical backgrounds by applying these methods to medicine as well as researching the current status quo. Finally, we will use current medical research to guide the factors we use to determine weightings to help alleviate overfitting. This research might improve current predictive models for T2DM using commonly collected patient characteristics; thus, we would effectively obtain more "bang for our buck."

Breakdown to Individual Steps:
The first step is to examine our patient data sample, which includes 768 patients, and determine whether we will need to discard any statistical outliers/clean data. In our preliminary research, the data appears complete and ready for analysis.
Machine Learning:
We will do all further data modifications using scikit-learn. We plan to use random forest techniques, where 10% of the data will be kept as a validation set and another 10% for a testing set. Decision trees are the basis of random forest techniques. Similar to the figure below, a decision tree is a series of binary decisions to classify input data. After the decision tree goes over all the input variables, it will determine the best way to group the elements by aiming to split data in two using a single classification factor with the most accurate differentiation of outcomes (i.e., the most predictive factor). The computer will then calculate the decrease in the entropy of each tree and decide which tree is the best model. This is the general idea behind a decision tree, and the random forest is simply a forest of decision trees with less stratification. For example, if there are seven arguments to classify, there would be seven levels in a classical decision tree; meanwhile, each tree in the random forest only chooses two arguments from the seven, forming decision trees with two levels. The program would generate thousands of trees based on the input argument, calculate the weighting of each tree based on its accuracy, and finally present a model. In other words, random forest techniques use multiple uncorrelated (randomly selected factor choices) decision trees that seek to optimize classification of random samples with replacement of elements and seeing which factors are best across multiple trees.



Factor Weighting:
We wish to use classical factor weighting methods to develop a model for T2DM risk. We do not anticipate any difficulties with this as many team members have experience with this method. We will either do it entirely with software or create a matrix of data points, find the pseudoinverse, and use this to solve for the weightings. Where n is the number of patients and p is the number of patient parameters, we will generate Ax=b, with A being rows of patient parameters (nxp), x being a columnar matrix of weights (nx1), and b being our outputs, most likely glucose levels (nx1). Solving for x via A's pseudoinverse, we will then process the glucose levels to fit diabetes risk stratification levels. We might also test whether we can work the system with binary outputs. To make this more interesting, we might generate non-linear factors from the given patient parameters, ie. multiplying two factors or squaring one to account for correlations we'd expect to see from medical literature.

Classical Methods:
Members will examine the literature to find out how the NHS determines risk for T2DM. We will then apply this method to our data set.

(Intro & Search Method)
The NHS determines T2DM risk in two main ways. One is in a clinical setting where labs and blood results are available while the other is done by the patient themselves. Although the former (clinical) provides greater precision and accuracy, it does not fit our project as patients may not have the necessary equipment. Instead, we looked at

several major diabetes scoring sheets that can be performed by patients with little/no equipment to give a numerical score that informs patients of their risk of developing T2DM. This included sheets from the American Diabetic Association, Australian Type 2 Diabetes Risk Assessment Tool, Finnish Diabetes Risk Score, and the Cambridge Diabetes score.

(Example Score Sheet - American Diabetes Association, ADA, Risk Calculator)

Risk Factors	***Point Distribution***			
Age	<40 (0p)	40-49 (1p)	50-50 (2p)	>60 (3p)
BMI	<25 (0p)	25-30 (1p)	30-40 (2p)	>40 (3p)
Gender	Female (0p)	Male (1p)		
1st Degree Relative w/ DM	No (0p)	Yes (1p)		
Hypertension	No (0p)	Yes (1p)		
Physically Active	No (0p)	Yes (1p)		
Total Points	More than 5 points = High risk			

- Simply select the category which you belong to and add up the total points for each risk factor that you fit into
- If the total score is higher than 5, you are at risk of developing T2DM

(Designing Score Sheet)
Amongst the various scoring tools we looked at, several common risk factors are accounted for in most of the tools. These include a variety of risk factors, see table below. Although it would be ideal to include all of these risk factors into our calculator, we do not have the information for all of these risk factors in our sample data. Thus, we will be mainly looking at six main risk factors that we do have access to (Age, Gender/Sex, Family History, Hypertension, BMI, Blood Glucose).

Risk Factors	Most Commonly Assessed	Have access to in our data	Risk Factors we will be analysing
Age	✓	✓	✓
Gender/Sex	✓	**? all female?**	✓
Family History / 1st degree relative with diabetes	✓	✓	✓
Hypertension or antihypertensive meds	✓	✓	✓
Physical Activity	✓		
BMI/waist	✓	✓	✓
Glucose Levels	✓	✓	✓
Pregnancy		✓	
Skin Thickness		✓	
Insulin		✓	✓

(Score Weighting)
Based on the scoring sheets we have looked at, we generated approximate weightings that each risk factor considered should hold compared to each other. (Eg. Old Age is 2x the risk of Family History and 3x the risk of Gender). This was done by examining the scores given to each risk factor by the score sheets chosen. Certain risk factors, such as age, were given a greater maximum score compared to other risk factors. These maximum scores were ratio-ed to give approximate relative weightings, which will be used to seek non-linear factor-weightings in our models.

Risk Score (Score to be considered high risk)	ADA (5+)	AUSRISK (20+)	FINDRISC (20+)	Cambridge (???)	Mean weight
Age	3/9	8/30	4/19	✓	27%
Gender/Sex (Male)	1/9	3/30	??	✓	10%
1st degree relative with diabetes	1/9	3/30	5/19	✓	15%
Hypertension or antihypertensive meds	1/9	2/30	2/19	✓	9%
BMI/waist	3/9	8/30	3/19	✓	25%
Glucose Levels		6/30	5/19		23%

x/y
X - how many points its worth in diabetes calculator
Y - how many points considered in total (sum of accounted x's)

(Python and Interface)
Our application will replicate what many tools have done by assigning numerical values to the categorical and continuous data of risk factors we have obtained. Patients enter their personal data for each risk factor and Python assigns a predetermined numerical score. These scores will be summed up to give an approximate risk of developing T2DM of a patient back to the patient.

Extensions:
Further extensions could be using neural networks to predict when patients have low blood sugar based on their past history. Furthermore, we can also implement AI via Dialogueflow into the chatbot to give a better patient experience.

Required Tools: We will be using the Scikit software to generate our random forest models. We do not anticipate any other tools being necessary.

Required datasets: We will be using publicly available datasets available from https://www.kaggle.com/uciml/pima-indians-diabetes-database.

The dataset contains 768 female Pima Indian patients with T2DM and their patient parameters: pregnancies, blood pressure, skin thickness, BMI, glucose, insulin, age, outcome, diabetes pedigree, outcome of diabetes diagnosis.

Agreed Contributions from each member:
Ka: Will work on the user interface, where patients will know what data to give. He will also research pen and paper methods and design a python program that uses these to calculate a score for patients. He will work with Joshua to advise on which factors are useful.
Joshua: Will do much of the writing. Will also work on the factor weighting methods. Joshua will also advise on which factors to work on.
Hanyuan: Will learn and perform the machine learning techniques on the given data and try to improve the accuracy of the results. Because he and Zhihao are living together, they would collaborate on most tasks.
Zhihao: Will help on performing the factor method via Python. As mentioned, he and Hanyuan will collaborate on carrying out the machine learning method. Will also look into the technique of machine learning and try to develop our own program based on the existing package.
Everyone: We will use statistical techniques to compare the models.

Timeline:

1) Complete proposal draft by Feb. 19
2) Determine important parameters by Feb. 23
3) Develop a random forest algorithm to classify patients on our data set by March 1.
4) By March 1, complete the factor weighting algorithm.
5) By March 4, produce an algorithm that follows the pencil and paper method.
6) By March 5, everyone has audited the other’s code.
7) March 8-ish, Ka and Joshua work on a better pen and paper method based on their analysis.
8) By March 10, determine a statistical ranking system for each algorithm. At the current moment, we plan to use our models to create a predicted blood glucose level and use clinical guidelines to determine a risk level from this. We can also use accuracy of morbidity predictions. Methods will be ranked based on their deviations from observed values.
9) Edit and finalize until the deadline.

Communication:
We will meet twice a week to discuss progress.
Github: https://github.com/tian-xuan/Interdisplinary-computing-projrct

References:

Sklearn: https://scikit-learn.org/stable/
Diabetes data set:https://www.kaggle.com/uciml/pima-indians-diabetes-database
Factor Weighting Methods: https://www.johndcook.com/blog/2018/05/05/svd/

QDiabetes - https://qdiabetes.org/
ADA - https://www.diabetes.org/risk-test
AUDRISK - https://www.health.gov.au/resources/apps-and-tools/the-australian-type-2-diabetes-risk-assessment-tool-ausdrisk#:~:text=The%20Australian%20type%202%20diabetes%20risk%20assessment%20tool%20(AUSDRISK)%20is,over%20the%20next%205%20years.
FINDRISC - https://www.mdcalc.com/findrisc-finnish-diabetes-risk-score
Cambridge - https://www.ncbi.nlm.nih.gov/books/NBK260920/