Diabetes: Prediction as a means of prevention

Anthony Forde
Dublin City University, Glasnevin, Dublin 9, Ireland
Email: anthony.forde23@dcu.mail.ie

***Abstract*—Diabetes is a chronic disease which poses a serious impact to a sufferer's health and lifestyle. It effects the body's ability to regulate insulin in breaking down sugar [11]. Uncontrolled, diabetes can lead to cardiovascular complications, loss of vision, lower limb amputation and death. While there is no cure for diabetes, it is no longer fatal in most treated cases. Simple changes to diet and lifestyle can help prevent the onset of diabetes, while advances in medicine and technology have enabled people with diabetes to lead relatively normal lives by monitoring and controlling their insulin levels. Adversely, this may also be a contributory factor to complacency as suggested by the alarming increase in diabetes in modern society. It is estimated there are 266,000 people living with diabetes Ireland – approximately 5% of the Irish population – and over half a billion people worldwide [12]. This is a huge health issue and financial burden, which is exacerbated by bad lifestyle choices and lack of education. As with any illness, early detection and diagnosis are key to successful treatment and precaution can mean prevention. There are a number of health and socioeconomic factors which can contribute to likelihood of developing diabetes. Creating awareness and promoting health education are considered key to tackling the issue.**

Keywords—Diabetes, Prediction, Machine Learning

I. INTRODUCTION

We live in an age where information has never been more abundantly accessible. Combined with remarkable advances in medicine, technology and the explosion of social media, we are in a unique position where information is power. In parallel with information retrieval, the tools and technologies to process that information have developed beyond the expectations of previous decades, allowing data mining and machine learning algorithms to easily draw automated conclusions which previously would have seemed untenable by manual human endeavour. Information in itself has become a valuable commodity and with it comes power. This power can be used for profit or for the benefit of humankind – or, indeed, both in the medical industry.

Early detection of many illnesses can be crucial in determining the success of any prescribed treatment. The expanding role of AI in healthcare innovation is more and more apparent. For example, AI image recognition can now be used to identify malignant melanomas, often with more accuracy than human diagnoses. We see Natural Language Processing (NLP) applications help diagnose cognitive and psychological disorders. AI retina scaning is also proving successful in multiple diagnoses.

Diabetes is no exception. The rapid advancement in diabetes treatment and ease of use and accuracy of personal glucose monitors has exponentially increased the quality of life of diabetics since Frederick Banting, a physician in Ontario, Canada, first used insulin to successfully treat a patient in 1922 [13]. Prior to this, diabetes was fatal. Now diabetic patients can rely on implanted sensors for automatic monitoring of their glucose levels, forgoing the finger-prick test (itself a marvel of innovation in its time).

Taking it a step further, what if AI could be used to predict, thus prevent, rather than just detect and monitor illnesses? This is one goal of data mining in the medical field.

Diabetes manifests itself in different forms – referred to as type 1, type 2 and gestational, the latter developing in women during pregnancy and usually temporary. Prediabetes is a further condition where blood sugar is higher than normal but not high enough to be considered full diabetes [14]. While type 1 can be genetic, the exact cause is unknown. Type 2 diabetes, however, can be acquired and the risks are heavily influenced by lifestyle factors such as bad diet, lack of physical exercise and general poor health. These inputs understandably also contribute to other conditions such as cardiovascular disease, high cholesterol, some cancers, to list but a few. However, less obvious factors can also play a part: education, income, urbanisation and mental health.

It is evident that diabetes type 2 is on the rise in western societies. It is difficult to rule out a similar trend in developing countries due to lack of data. What is clear is the financial cost involved in patient care. Given diabetes is more prevalent in lower socioeconomic groups, cost is often a blocker to diagnosis and treatment, creating unfair segregation from a treatment point of view, particularly in countries such as the US where free medical care does not exist. Many who fall into this category go undiagnosed until diabetic symptoms arise, when it is too late to prevent, despite screening being quick and non-invasive. Education and predictive analysis of diabetes could help level the playing field here.

In this paper, we take a look at a dataset of diabetes, health and lifestyle information and try to ascertain:

- Which are the most predictive factors of diabetes?
- What are the right questions to ask?
- Is inequality a factor?

To achieve this, we follow a data mining methodology to analyse and process the data. We explore different methods of statistical testing, as appropriate. We then trial several machine learning techniques for diabetes prediction and score and rate them to determine which model performs best. These include: Decision Tree Classification, Naïve Bayes, K-Nearest Neighbours (KNN), Random Forest.

The rest of this paper is organised as follows. In the next section, we review some related research which influenced this

work. We then examine the dataset. Following this, we take an in-depth look at the implemented solution, including the data mining methodology selected for the project, and illustrate how we utilised this in formulating the solution. Next we evaluate the results and implications of our research. We draw our conclusions from the work and try to convey impartial assessment of what worked well, what could be improved upon, any limitations and any unexpected findings. Before closing out, we will consider, based on the study, further worthwhile research in the same field. The final sections include details and locations of all code and data used for the research as well as a video presentation – access is shared public.

II. RELATED WORK

D. Dutta, D. Paul and P. Ghosh wrote a paper on the analysis of feature importance in diabetes prediction. This provided a good insight into the success of different methods such as Logistic Regression, Support Vector Machines (SVM) and Random Forest – the latter which performed best and inspired me to use it in my own research. Another key take away was the negative impact of using an imbalanced dataset which, again, contributed to my own decision on the data for this research. There were good correlational analyses between age, glucose and diabetes. Unfortunately, glucose level was not included in the dataset which I used, however, this would not have been possible as the data was not easily available for such a vast population size [1].

T.Joshi and P.Chawan validated the use of SVN, Logistic Regression, and Artificial Neural Networks (ANN) for implementing diabetes prediction systems. However, the work was largely theoretical – no results provided [2].

Q. Zou, K. Qu1, Y. Luo, D. Yin, Y. Ju and H. Tang illustrated the successful use of Decision Trees, Random Forest, ANN and Principal Component Analysis (PCA) in machine learning predictive models for diabetes. What I found particularly interesting was the impact PCA made in the results. However, my own research data did not contain sufficient continuous data to warrant PCA, being mostly categorical. The results showed that prediction with Random Forest worked best, which concurred with paper above and justified my own use of this classification. The importance of glucose data is also stressed here, as previous paper. [3].

P. Sonar and K. JayaMalini discussed the use of Decision Tree SVN, Naïve Bayes classifier and ANN. They endorse the use of Decision Trees – which I used – for ease of use and robustness. I also used Naïve Bayes in my research. They advise caution using ANN on large datasets due to huge processing requirements – advice which I also took on board. [4]

A. Lynam, J. Dennis, K. Owen et al. promoted the use of classic Logistic Regression in diabetes classification where there are a small number strongly predictive variables, claiming it performed as well as more advanced machine algorithms. I decided to put this to the test [5].

A. Anand and D. Shakti provided good insight into best use of categorical data in predictive modelling and feature importance in diabetes, including interpretation of Chi-Square test statistics. They also provided good analysis of health and lifestyle factors and advice influencing diabetes. As will be illustrated later, my own test results complemented theirs when it came to high blood pressure [6].

T. LaVeist, R.Thorpe, J. Galarraga, K. Bower, T. Gary-Webb wrote a fascinating paper on the influence of socioeconomic versus environmental factors across the race divide in determining likelihood of developing diabetes. According to the authors, when living in the same communities and social conditions, it is economic environment more than cultural or ethnic background which is more influential. This refutes the argument that some racial groups are more prone to diabetes than others under the same conditions. However, while the results were summarised, they gave no illustrations to the methods used [7]

A. Tol, G. Sharifirad, D. Shojaezadeh, E. Tavasoli L. Azadbakht argue that there needs to be more focus on socioeconomic factors as a preventative measure in type 2 diabetes. They also suggest supportive strategies and educational resources. This illustrated the importance of education in health, made good background sociological reading rather than predictive modelling [8].

P. Gomez-Galvez, C. S. Mejías and L. Fernandez-Luque discuss the increasing popularity in use of social media, mobile applications and “wearables” which are transforming the lives of people with diabetes. On a precautionary measure, they highlight the risks of over reliance on technology in relation to data mining, privacy issues and isolation among the older population who may be less inclined to utilise it [9].

O. Geman, R. Toderean, M. Lungu, I. Chiuchisan and M. Covasa emphasise the importance of inclusivity in the rise in therapeutic support applications and tools. They insist that in order for technological therapy to be successful and data analytics to be valid, they must be customisable for all individual requirements, preferences, and cultural traditions. These are important points to consider before selecting any dataset for research, the source and method of collation cannot be underestimated in determining its validity [10].

III. DATA

The data, sourced from Kaggle.com, originated from the Centers for Disease Control and Prevention (CDC) in the US. It comprised results from the Behavioral Risk Factor Surveillance System (BRFSS) from 2015. The BRFSS is a health and lifestyle phone survey conducted annually by the CDC. In this case, the dataset related to diabetes and potential risk factors,

listing the answers to 22 questions including whether or not the participant had diabetes [15].

The original BRFSS raw data for the year contained more than 250 thousand unprocessed records of interviewees' answers, one row per individual interviewed, for the health and lifestyle questionnaire [16]. However, we opted to use an available subset of approximately 72 thousand records which constituted a balanced dataset of equal samples where the outcomes were diabetes and no diabetes, i.e. an equal number of individuals responded "yes" and "no" to having diabetes. The dataset, therefore, was broken into 21 feature variables and one target binary outcome (diabetes or no diabetes). The features, listed and described below, comprised different types of data – categorical, continuous and discreet.

TABLE I. Independent variables

Feature Name	Description	Type
HighBP	High blood pressure	Y / N
HighChol	High cholesterol	Y / N
CholCheck	Cholesterol checked in last five years	Y / N
BMI	Body mass index	Numeric
Smoker	Ever been a smoker	Y / N
Stroke	Stroke	Y / N
HeartDiseaseorAttack	Heart disease or heart attack	Y / N
PhysActivity	Physical activity	Y / N
Fruits	Consumes fruit regularly	Y / N
Veggies	Consumes vegetables regularly	Y / N
HvyAlcoholConsump	Heavy drinker	Y / N
AnyHealthcare	Has health insurance	Y / N
NoDocbcCost	Unable to afford doctor cost	Y / N
GenHlth	State of general health	Rate 1-5
MentHlth	State of mental health	Rate 1-30
PhysHlth	State of physical health	Rate 1-30
DiffWalk	Difficulty walking	Y / N
Sex	Gender	F / M
Age	Age category	Category 1-13
Education	Level of education	Category 1-6
Income	Income category	Category 1-8

IV. SOLUTION DESIGN

A. *Data Mining Methodology*

A CRISP-DM style iterative process model was employed for the analytics solution. For the analysis, planning and implementation process, the following workflow was applied:

1. Understand and analyse data
2. Prepare the data (repeat if necessary)
3. Create model
4. Validate the model (revise and reiterate as required)
5. Deploy

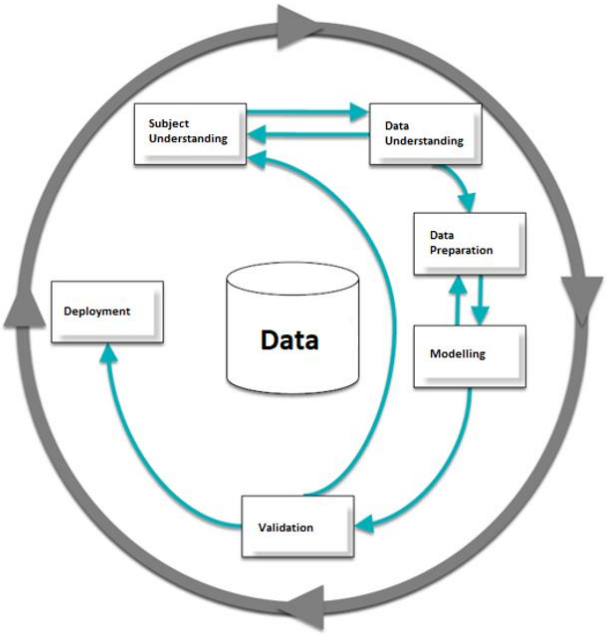

Fig. 1. CRISP-DM [17]

B. *Tools and Technologies*

The following tools and technologies were employed to implement the solution:

- MS Excel: To read source data and perform initial analysis
- Python: Programming / scripting language to:
  - Manipulate, interpret and derive statistics from data
  - Create visualisations
  - Formulate and evaluate hypotheses tests
  - Create machine learning models
- Jupyter Notebook: Computing notebook to edit and run Python code.
- Google Colab: To store and run notebooks
- GitHub: To store, back up and share all data and code
- Loom: Video presentation

C. *Solution Process*

The following steps were followed in the solution process:

1. Bring in the data
2. Perform initial analysis
3. Preprocess the data
4. Check for correlations
5. Check for multicollinearity
6. Split the data
7. Run tests for categorical data
8. Run tests for numeric data

9. Evaluate machine learning models

D. Data analysis and preprocessing

Relating to steps 1-5 in *D. Solution Process*, above, this is where we imported the data, checked for any missing data or inconsistencies and cleaned the data if required. Initial analysis was performed using pandas_profiling in python which generated a detailed description of the data, including correlation matrix, output to readable report.

E. Multicollinearity

The next step was to check our data for any multicollinearity in among the featured independent variables. Highly correlated independent variables can adversely effect the reliability of our model. We used Variance Inflation Factor (VIF) to detect multicollinearity in our data which led to our decision to remove three independent variables (in order to bring highest VIF to below 10):

TABLE II. Multicollinearity in independent variables

Feature Name	**Description**
CholCheck	Cholesterol checked in last five years
AnyHealthcare	Has health insurance
BMI	Body mass index

F. Splitting the data

At this stage, we had our finalized features and we split the base data into multiple subset dataframes to prepare for testing. The first split was by data type, i.e. split into categorical, discreet or continuous data, to determine what type of tests we could run.

We also split the base data into logical categories, grouping the features under common themes or headings where possible, namely: *health*, *lifestyle*, *medical* and *socioeconomic*.

G. Categorica data– Chi-Square Test

The dependent variable under test, diabetes_binary, was a categorical binary field. In order to test with other categorical data, this necessitated Chi-Square testing. We used the Scipy library in Python to test which of our independent categorical features had more influence over diabetes outcome.

TABLE III. Categorical variables

Feature Name	**Description**	**Type**
HighBP	High blood pressure	Y / N
HighChol	High cholesterol	Y / N
Smoker	Ever been a smoker	Y / N
Stroke	Stroke	Y / N
HeartDiseaseorAttack	Heart disease or heart attack	Y / N
PhysActivity	Physical activity	Y / N
Fruits	Consumes fruit regularly	Y / N
Veggies	Consumes vegetables regularly	Y / N
HvyAlcoholConsump	Heavy drinker	Y / N
NoDocbcCost	Unable to afford doctor cost	Y / N
GenHlth	State of general health	Rate 1-5
DiffWalk	Difficulty walking	Y / N
Sex	Gender	F / M
Education	Level of education	Category 1-6

H. Numeric data – Logistic Regression

To compare our numeric features against the categorical target field, diabetes_binary, we performed Logistic Regression testing. The tests evaluated the effectiveness of the numeric independent variables as predictors of the target diabetes outcome. This was implemented in Python using the Sklearn, Statsmodels and Seaborn libraries. The numeric fields were of type: discrete, interval and ratio:

TABLE IV. Numeric variables

Feature Name	**Description**	**Type**
MentHlth	State of mental health	Rate 1-30
PhysHlth	State of physical health	Rate 1-30
Age	Age category	Category 1-13
Income	Income category	Category 1-8

I. Machine Learning Prediction Models

Using Python library Sklearn, we tried a number of different machine learning models for diabetes prediction and evaluated their accuracy and performance accordingly:

- Decision Tree Classification
- Naïve Bayes
- KNN
- Random Forest

V. EVALUATION/RESULTS

A. Chi-Square Tests

First we, ran tests to check that the p-values for our independent categorical features were significant, i.e. less than 0.05, to justify using them in our analysis. All features proved significant in hypothesis testing, thus were included. We then compared the Chi-Square statistic for each feature to see which

features were more important as diabetes risk factors. General health status and high blood pressure were clearly the most influential factors, followed by high cholesterol. See visualisation in Fig 2.

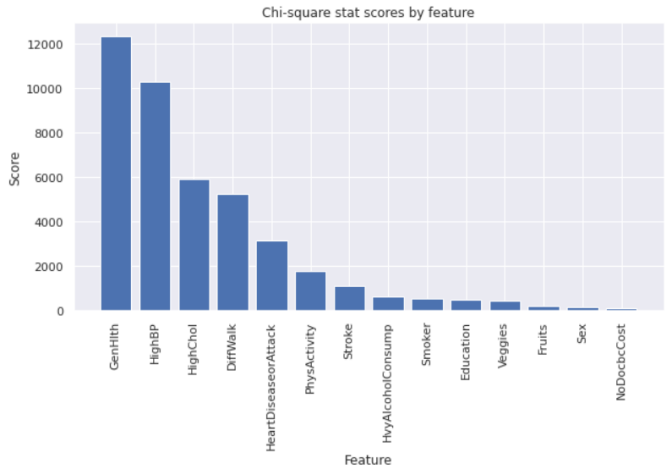

Fig. 2. Comparison of Chi-Square scores

We will take a look at some individual features here.

Fig. 3. illustrates the breakdown of general health categories and their Chi-Square scores. We can see significant correlation between lower status of health (fair and poor) and having diabetes.

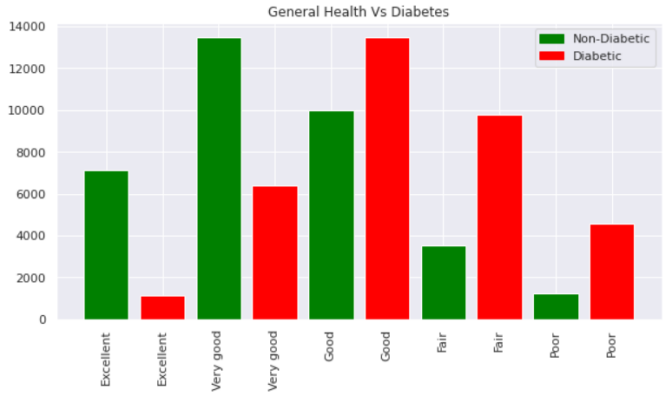

Fig. 3. Scores by general health status

The results for high blood pressure, Fig. 4, were as expected, similar to poor health – a significant factor in diabetes.

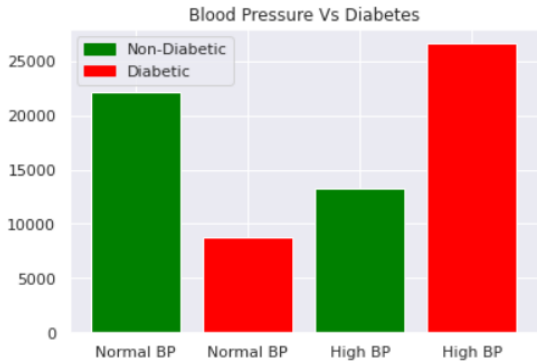

Fig 4. High blood pressure scores

By contrast, gender (Fig. 5), although significant enough by p-value to include in hypothesis testing, was not as significant as the aforementioned health factors.

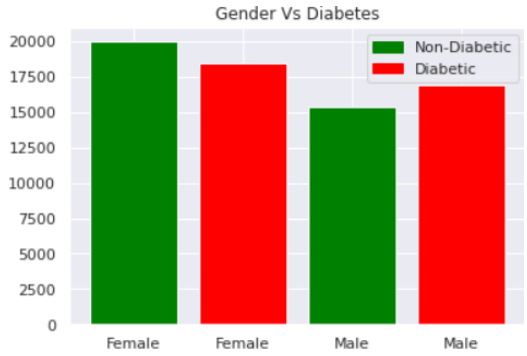

Fig 5. Gender scores

Education was broken down into the following categories:

1 - Never attended school or only kindergarten

2 - Grades 1 through 8 (Elementary)

3 - Grades 9 through 11 (Some high school)

4 - Grade 12 or GED (High school graduate)

5 - College 1 year to 3 years (Some college or technical school)

6 - College 4 years or more (College graduate)

We can see that there is significantly less diabetes among the higher educated in level 6. While the same is true for levels 1-3, we must take into account these (outliers excepted) relate to children where diabetes may be less prevalent.

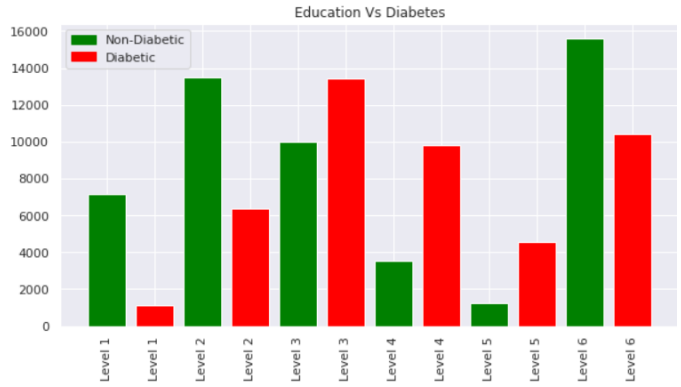

Fig. 6. Scores by education level where: 1 is lowest, 6 is highest

B. *Logistic Regression Tests*

Using Logistic Regression, we tested the reliability of using the numeric features, below, in predicting diabetes.

1. Mental health: Number of bad days in last month from 1 (best) to 30 (worst)
2. Physical health: Number of bad days in last month from 1 (best) to 30 (worst)
3. Age: Categories ranging from 1 (age 18-24) through 13 (80 or older)
4. Income: Categories ranging from 1 (0 to US$9,999) through 8 ($75,000 or more)

Reviewing the results, Fig. 7, we found that LLR p-value was significant (less than 0.05) which told us that our Logistic Regression performed better than the null model. However, it did not generate a great pseudo r-squared value of 0.0957 (this should ideally be between 0.2 and 0.4 for a good result). In saying that, the individual p-values were significant (0) for physical health, age and income but proved insignificant (greater than 0.05) for mental health.

Logit Regression Results

Dep. Variable:	Diabetes_binary	No. Observations:	70692
Model:	Logit	Df Residuals:	70688
Method:	MLE	Df Model:	3
Date:	Sun, 03 Apr 2022	Pseudo R-squ.:	0.09571
Time:	09:43:27	Log-Likelihood:	-44310.
converged:	True	LL-Null:	-49000.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
MentHlth	-0.0003	0.001	-0.261	0.794	-0.002	0.002
PhysHlth	0.0289	0.001	31.352	0.000	0.027	0.031
Age	0.1375	0.002	67.245	0.000	0.134	0.142
Income	-0.2262	0.003	-78.265	0.000	-0.232	-0.221

Fig. 7. Logistic regression results summary

We generated a confusion matrix, Fig. 7, for more information. The results, again, were not high performing.

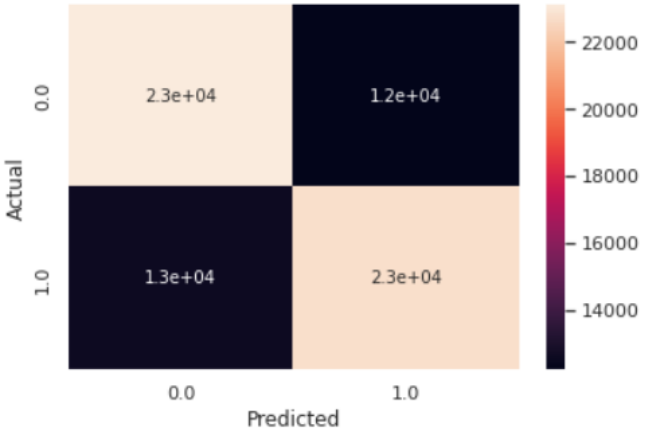

Fig. 8. Confusion matrix

The approximate percentage breakdown of above was:

- TP (true positive): 32.5
- TN (true negative): 32.5
- FP (false positive): 18
- FN (false negative): 18

Giving us an unimpressive precision of 65%:

	precision	recall	f1-score	support
0.0	0.65	0.65	0.65	35346
1.0	0.65	0.64	0.65	35346
accuracy			0.65	70692
macro avg	0.65	0.65	0.65	70692
weighted avg	0.65	0.65	0.65	70692

Fig. 9. Scoring metrics

While age and income were highly correlated, their VIF scores were within acceptable range. However, we tested the model iteratively excluding one feature at a time to be certain: mental health (which had an insignificant p-value), age (highly correlated with income) and income (highly correlated with age). The results did not fare any better. The overall conclusion was Logistic Regression did not perform as well as expected in predicting diabetes with the independent feature variables available.

C. Machine Learning Prediction Models

The F1 score results of the four prediction models we tested, using an 80/20 training/testing split, are displayed in the Table V below. The training sets scored better than testing, as expected, with Decision Tree Classifier and Random Forest joint best. In testing, Random Forest topped the results with an F1 score of 0.72, followed closely by Naïve Bayes at 0.69. Neither of these results are convincing; we would be aiming for at least 0.9 accuracy for a more reliable model.

TABLE V. Model F1 scores

Model	**F1 Score: Training**	**F1 Score: Testing**
Decision Tree	0.97	0.65
Naïve Bayes	0.69	0.70
KNN	0.79	0.69
Random Forest	0.97	0.72

VI. CONCLUSIONS

When I set out on this research project, I had a reasonable layperson's knowledge of diabetes and risk factors related to the disease. As with many illnesses, particularly acquired conditions, I assumed the obvious health and lifestyle factors would be key indicators: diet, physical exercise, smoking, alcohol consumption, etc. Not surprisingly, this proved to be true. However, my goal was not to prove what we already know: that not looking after your health could increase your chances of developing diabetes. Rather, from the extensive list of risk factors, I wanted to find out if we could narrow this down and isolate the more important contributors to diabetes. Even from a medical point of view, if a doctor could advise a patient to curtail or improve three or four aspects of their lifestyle, instead of giving up 10 (often enjoyable) habits or pastimes, the patient would surely be more likely to adhere to their doctor's advice? Perhaps the same could be true for raising health awareness among the general public.

It was no surprise that general health, high blood pressure and high cholesterol made up the top three most important factors of the categorical dataset. But, equally, it would not have surprised me had the top three been gender, fruit consumption and smoking. That we cannot discount correlation among the independent features is a given. But, coming back to the individual's perspective, it does not discount the usefulness of narrowing down the risk list to help avoid diabetes.

While all features included in the dataset were statistically significant, that is influential enough on diabetes to reject the null hypothesis, it was interesting to see some of the lower

scoring individual features such as smoking, fruit and vegetable intake. Again, it must be pointed out that these features may contribute to the table toppers of general health, high blood pressure and cholesterol. Gender, on the other hand, although significant enough to rate in statistical testing, there was not a huge difference in diabetes risk between male a female.

In addition to health and lifestyle, I wanted to explore the idea that socioeconomic factors such as income and education might influence the likelihood of developing diabetes. There was definitely evidence to suggest that, among the adult population, those with a higher level of education are less susceptible. Income, included in numeric logistic regression testing, although not proving as influential as one would have expected, was significant enough to overrule the null hypothesis. Which brings us to the test results.

On the whole, the Chi-Square tests produced interesting but not overwhelmingly surprising results for our categorical variables. The Logistic Regression testing for our numeric variables did not perform as well as I expected. Admittedly, they were a mixture of discreet, interval and ratio variables, as opposed to more suitable continuous data, which perhaps may explain this. In saying that, they still performed better than the null model.

As with Logistic Regression, the results for our predictive models, while reasonably good, were not exceptional. The best model, Random Forest classification, scored 97% on training data but only 72% on test data. I would also have expected Decision Tree classification to fare better in this binary prediction, but it came in at 97% training / 65% testing.

To conclude, common sense and moderation prevail, as one would expect, when it comes to health and diabetes. More research into socioeconomic factors would be well merited. While modern, affluent, western societies appear to be more prone to the rise in diabetes [18], it is the poorer sections of those same societies which may be worst impacted by diabetes. In tackling the problem, education may prove more effective than insulin.

VII. CODE AND DATA REPOSITORY

A. *Github*

All data and code for the project are stored in the following shared Github repository:

https://github.com/anthonyforde/CA683I_Assignment

Here, you will find:

1. **Diabetes_Prediction.pdf:** A copy of this document.
2. **Diabetes_Prediction.ipynb:** Jupyter notebook containing all Python code, saved with latest output, for this research project.
3. **diabetes_binary_5050split_health_indicators_BRFSS2015.csv:** Raw dataset used for the research.
4. **ProfileReport.html:** Python generated profile report of raw dataset used for initial analysis.
5. **Correlations.xlsx:** Correlation matrix of features in the dataset, used for initial analysis.

VIII. VIDEO PRESENTATION

A summary video presentation of the research can be found (shared public) on Loom:

https://www.loom.com/share/b4ae4b53432e4eaca5858254db620023

IX. ACKNOWLEDGEMENTS

The dataset was sourced from Kaggle: https://www.kaggle.com/code/alexteboul/diabetes-health-indicators-dataset-notebook/data.

REFERENCES

[1] D. Dutta, D. Paul, P. Ghosh, "Analysing Feature Importances for Diabetes Prediction using Machine Learning," November 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8614871

[2] T. Joshi, P. Chawan, "Diabetes Prediction Using Machine Learning Techniques," January 2018. [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/56913852/C0801020913-with-cover-page-v2.pdf?Expires=1649168850&Signature=bv6-fIJcuWn~5I2ThOQDfK6Z9MSotvY7IQ9DuDzpOaj1kXSfyG6ntRwLv-3-fkAA3Ng2lSiuf9kgAG15jDqPcNNETkFQa31Y46gGqitxr0XvzLeOiVrA4YnHYVwlQdG3zQ1cyjIUdoTAJEDQNx0LEWNgDIrk8Eh1Pk2u~V-swkvuZv-UTZxllZHHyDKE0~eL1t8QjR4dGR75vpdF8nd2fVvNYp4ppMkRJUX4pEGmKaJ-9vtsapzdQPYsxdugilK8it3CIn1CM~2LBNHsniGTsa85KZLT6JVwt0~-SziMMqAaYpFs3icktM8pEjwiCiYGExaiEVaRNuurzRV1SpCh~g__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

[3] Q. Zou, K. Qu1, Y. Luo, D. Yin, Y. Ju and H. Tang "Predicting Diabetes Mellitus With Machine Learning Techniques," November 2018. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full

[4] P. Sonar, K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," March 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8819841/authors#authors

[5] A. Lynam, J. Dennis, K. Owen, R. Oram, A. Jones, B. Shields, L. Ferrat "Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults," June 2020. [Online]. Available: https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-020-00075-2

[6] A. Anand, D. Shakti "Prediction of diabetes based on personal lifestyle indicators," September 2015. [Online]. Available: https://ieeexplore.ieee.org/document/7375206

[7] T. LaVeist, R.Thorpe, J. Galarraga, K. Bower, T. Gary-Webb, "Environmental and Socio-Economic Factors as Contributors to Racial Disparities in Diabetes Prevalence," August 2009, [Online]. Available: https://link.springer.com/article/10.1007/s11606-009-1085-7

[8] A. Tol, G. Sharifirad, D. Shojaezadeh, E. Tavasoli L. Azadbakht, "Socio-economic factors and diabetes consequences among patients with type 2 diabetes ," February 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3778578/

[9] P. Gomez-Galvez, C. S. Mejías and L. Fernandez-Luque, "Social media for empowering people with diabetes: Current status and future trends," August 2015. [Online]. Available: [Online]. Available: https://ieeexplore.ieee.org/document/7318811

[10] O. Geman, R. Toderean, M. M. Lungu, I. Chiuchisan, M. Covasa, "Challenges in nutrition education using smart sensors and personalized tools for prevention and control of type 2 diabetes," October 2017. [Online]. Available: https://ieeexplore.ieee.org/document/8259943

[11] "What is Diabetes?," December 2021. [Online]. Available: https://www.cdc.gov/diabetes/basics/diabetes.html#:~:text=Diabetes%20

is%20a%20chronic%20(long,your%20pancreas%20to%20release%20insulin

[12] “DIABETES PREVALENCE IN IRELAND,” January 2022. [Online]. Available: https://www.diabetes.ie/about-us/diabetes-in-ireland/

[13] K.McCoy, “The History of Diabetes,” November 2009. [Online]. Available: https://www.everydayhealth.com/diabetes/understanding/diabetes-mellitus-through-time.aspx#:~:text=The%20first%20known%20mention%20of,people%20who%20had%20this%20disease

[14] J. Larson, “How many types of diabetes are there?,” November 2021. [Online]. Available: https://www.singlecare.com/blog/types-of-diabetes/

[15] “2015 Behavioral Risk Factor Surveillance System Questionnaire,” 2015. [Online]. Available: https://www.cdc.gov/brfss/questionnaires/pdf-ques/2015-brfss-questionnaire-12-29-14.pdf

[16] “2015 BRFSS Survey Data and Documentation,” 2015. [Online]. Available: https://www.cdc.gov/brfss/annual_data/annual_2015.html

[17] A. D’Arcy, “CRISP-DM V’s Agile – The Face Off,” 2022. [Online]. Available: https://krisolis.ie/crisp-dm-vs-agile-the-face-off/

[18] “U.S. Leads Developed Nations in Diabetes Prevalence,” December 2015. [Online]. Available: https://endocrinenews.endocrine.org/u-s-leads-developed-nations-in-diabetes-prevalence/