Diabetes risk predictor

Data Science Bootcamp

Sprint 1

Created By:
Diego Villanueva

Diabetes has become one of the biggest epidemics in human history

It is estimated that 422 million people are living with diabetes in the world...

422 M cases

1 in 11 of the world’s adult population.

4.8 M cases

90 % are type 2

... and almost half of them have not been diagnosed

46% of global cases are undiagnosed

In the UK, up to 1M people have diabetes that is yet to be diagnosed.

Sources: https://www.who.int/health-topics/diabetes#tab=tab_1
https://www.diabetes.co.uk/diabetes-prevalence.html#:~:text=World%20diabetes%20prevalence,with%20diabetes%20worldwide%20by%202040.

A Data Science solution

A **machine learning solution** can predict the risk of someone developing diabetes based on readily available health indicators

Proactive approach

Prediction tool for medical professionals

People use their data to understand **risk factors**

… can bring both health and financial benefits

Increase life expectancy

- 5th most common reason for death
- The life expectancy decreases 10 years

Financial benefits

- Globally: 825B USD in 2016 -> $2.5T by 2030
- UK: 10% of NHS budget = £9B /year

Sources: https://www.who.int/health-topics/diabetes#tab=tab_1
https://www.diabetes.co.uk/diabetes-prevalence.html#:~:text=World%20diabetes%20prevalence,with%20diabetes%20worldwide%20by%202040
https://www.imperial.ac.uk/news/171619/cost-diabetes-hits-825-billion-dollars

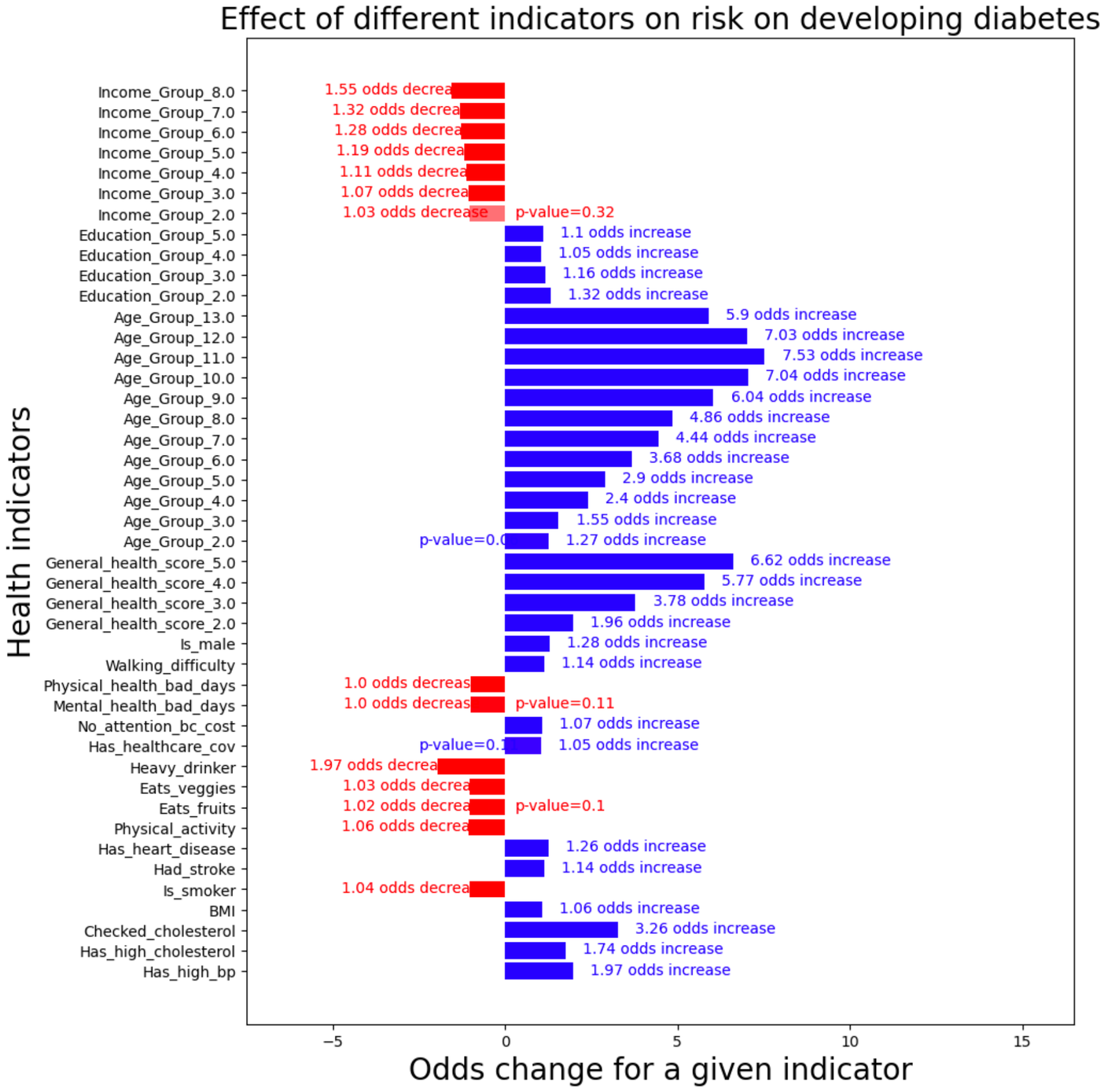
The project uses a dataset of 20 health indicators

While the dataset is relatively clean and can be trusted...

Trusted source	• Behavioral Risk Factor Surveillance System (BRFSS) • By Center for Disease Control and Prevention
Target Variable	• Does the participant have diabetes?
Features include:	• Lifestyle • Medical history • Physical characteristics

... data cleaning and preprocessing was needed before modelling

- Initial observations
- Duplicates
- Missing data
- Univariate and bivariate analysis
- Correlation analysis
- Data types
- One-hot encoding


Effect of different indicators on risk on developing diabetes
Health indicators
Odds change for a given indicator
Income_Group_8.0
Income_Group_7.0
Income_Group_6.0
Income_Group_5.0
Income_Group_4.0
Income_Group_3.0
Income_Group_2.0
Education_Group_5.0
Education_Group_4.0
Education_Group_3.0
Education_Group_2.0
Age_Group_13.0
Age_Group_12.0
Age_Group_11.0
Age_Group_10.0
Age_Group_9.0
Age_Group_8.0
Age_Group_7.0
Age_Group_6.0
Age_Group_5.0
Age_Group_4.0
Age_Group_3.0
Age_Group_2.0
General_health_score_5.0
General_health_score_4.0
General_health_score_3.0
General_health_score_2.0
Is_male
Walking_difficulty
Physical_health_bad_days
Mental_health_bad_days
No_attention_bc_cost
Has_healthcare_cov
Heavy_drinker
Eats_veggies
Eats_fruits
Physical_activity
Has_heart_disease
Had_stroke
Is_smoker
BMI
Checked_cholesterol
Has_high_cholesterol
Has_high_bp
1.55 odds decrea
1.32 odds decrea
1.28 odds decrea
1.19 odds decrea
1.11 odds decrea
1.07 odds decrea
1.03 odds decrease
p-value=0.32
1.1 odds increase
1.05 odds increase
1.16 odds increase
1.32 odds increase
5.9 odds increase
7.03 odds increase
7.53 odds increase
7.04 odds increase
6.04 odds increase
4.86 odds increase
4.44 odds increase
3.68 odds increase
2.9 odds increase
2.4 odds increase
1.55 odds increase
p-value=0.0
1.27 odds increase
6.62 odds increase
5.77 odds increase
3.78 odds increase
1.96 odds increase
1.28 odds increase
1.14 odds increase
1.0 odds decreas
1.0 odds decreas
p-value=0.11
1.07 odds increase
p-value=0.1
1.05 odds increase
1.97 odds decrea
1.03 odds decrea
1.02 odds decrea
p-value=0.1
1.06 odds decrea
1.26 odds increase
1.14 odds increase
1.04 odds decrea
1.06 odds increase
3.26 odds increase
1.74 odds increase
1.97 odds increase
−5
0
5
10
15

Next steps

- Further analysis on indicators with un-expected results (e.g., heavy drinker)
- Examine specific groups of features and iterate using different combinations. Focus on changes people can actively implement
- Try different models to achieve higher accuracy