University of Bucharest

Faculty of Mathematics and Informatics

Diabetes prediction

-project for Exploratory Data Analysis course-

Students:
Andrei Nicolae, group 407
Nandor Birta, group 407
Gabriela Giurea, group 407

Introduction

Diabetes is one of the most common chronic diseases in the world. In vague terms, it affects how an individual's body can transform food into energy. Usually, food is broken down into sugars, and released into the bloodstream. These sugars are controlled by a substance called insulin, that acts as a gatekeeper, without which cells cannot get access to the energy provided in the form of sugars. Diabetes is the condition, under which the body either cannot create enough insulin or it can not utilise its insulin production. Being aware of such a condition is essential, since the long-term complications can include heart diseases, vision loss and various kidney-related diseases (CDC, 2021).

Based on calculations by the Center for Disease Control and Prevention (CDC) an estimated 34.2 million American citizens are affected by diabetes and another 88 million show the signs of a condition called prediabetes. Moreover, it is estimated that 1 in 5 people are unaware of their condition in the case of diabetes. Similarly, 8 in 10 pre-diabetics are oblivious in terms of potential risks or complications (*Kaggle*, n.d.).

As of today, there is no cure for diabetes, however there are ways to manage it. There are a number of medical treatments that can reduce potential complications. Additionally, lifestyle related factors, such as a healthy diet plan, an active way of life and weight loss can contribute to reducing the risk of developing diabetes and can also help in managing the disease itself. It is essential to diagnose it early–be it the disease itself or the condition known as pre-diabetes–as it can impact the life of an individual quite severely.

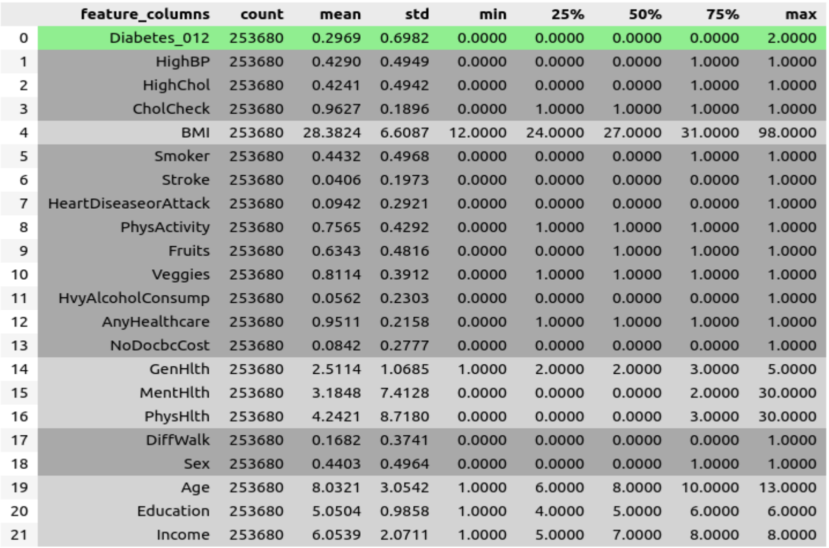
Of course, diabetes is an umbrella term, it encompasses a number of variants. The most common one is type II diabetes, which affects a vast segment of the population. Prediabetes denotes the condition when an individual's blood sugar is higher than normal, while it is below the threshold to qualify as type II diabetes (*NIDDK*, n.d.). These diseases are prevalent in various age groups, and their progression can be dependent on various socioeconomic factors, such as income, education and location. Furthermore, they can also pose a serious financial burden and thus impediment on the less fortunate. From an economic perspective, the costs are also considerable, diagnosed diabetes–including diagnosis, treatment and various other stages–is estimated to account for over $300 billion dollars in expenditure and considering the prediabetic condition as well, costs can reach up to $400 billion dollars on a yearly basis (*Kaggle*, n.d.).

And this is where predictive modeling comes into the picture. The dataset at hand was collected by the Behavioral Risk Factor Surveillance System (BRFSS), a public health-related telephone survey that is conducted by the CDC year by year. The sampling pool includes over 400.000 U.S residents and the questionnaire focuses on topics such as health-related risk behavior, chronic health conditions and services associated with prevention and treatment (*Kaggle*, n.d.).

The approach of the team

After some preliminary data analysis and cleaning, each member of our team of three, picked various model classes for experimentation. Since these were different models, we believed that it would make more sense to find the best combination of input features for our particular model. But even before that, we have separated 20% of the data for testing purposes, to be used only to quantify the performance of our models, making these somewhat comparable and to reduce the risk of potential information leaks.

Feature presentation



	feature_columns	count	mean	std	min	25%	50%	75%	max
0	Diabetes_012	253680	0.2969	0.6982	0.0000	0.0000	0.0000	0.0000	2.0000
1	HighBP	253680	0.4290	0.4949	0.0000	0.0000	0.0000	1.0000	1.0000
2	HighChol	253680	0.4241	0.4942	0.0000	0.0000	0.0000	1.0000	1.0000
3	CholCheck	253680	0.9627	0.1896	0.0000	1.0000	1.0000	1.0000	1.0000
4	BMI	253680	28.3824	6.6087	12.0000	24.0000	27.0000	31.0000	98.0000
5	Smoker	253680	0.4432	0.4968	0.0000	0.0000	0.0000	1.0000	1.0000
6	Stroke	253680	0.0406	0.1973	0.0000	0.0000	0.0000	0.0000	1.0000
7	HeartDiseaseorAttack	253680	0.0942	0.2921	0.0000	0.0000	0.0000	0.0000	1.0000
8	PhysActivity	253680	0.7565	0.4292	0.0000	1.0000	1.0000	1.0000	1.0000
9	Fruits	253680	0.6343	0.4816	0.0000	0.0000	1.0000	1.0000	1.0000
10	Veggies	253680	0.8114	0.3912	0.0000	1.0000	1.0000	1.0000	1.0000
11	HvyAlcoholConsump	253680	0.0562	0.2303	0.0000	0.0000	0.0000	0.0000	1.0000
12	AnyHealthcare	253680	0.9511	0.2158	0.0000	1.0000	1.0000	1.0000	1.0000
13	NoDocbcCost	253680	0.0842	0.2777	0.0000	0.0000	0.0000	0.0000	1.0000
14	GenHlth	253680	2.5114	1.0685	1.0000	2.0000	2.0000	3.0000	5.0000
15	MentHlth	253680	3.1848	7.4128	0.0000	0.0000	0.0000	2.0000	30.0000
16	PhysHlth	253680	4.2421	8.7180	0.0000	0.0000	0.0000	3.0000	30.0000
17	DiffWalk	253680	0.1682	0.3741	0.0000	0.0000	0.0000	0.0000	1.0000
18	Sex	253680	0.4403	0.4964	0.0000	0.0000	0.0000	1.0000	1.0000
19	Age	253680	8.0321	3.0542	1.0000	6.0000	8.0000	10.0000	13.0000
20	Education	253680	5.0504	0.9858	1.0000	4.0000	5.0000	6.0000	6.0000
21	Income	253680	6.0539	2.0711	1.0000	5.0000	7.0000	8.0000	8.0000

Fig.1 Descriptive statistics for each individual feature

The dataset includes approximately 250.000 entries and 20 features. The table (position) shows various descriptive statistics; highlighted in green is the target feature, 'Diabetes_012' which includes three classes. Zero represents no diabetes, class 1 denotes the prediabetic condition and class 2 is diabetes. As it can be seen in the plot showing the distribution of the target variable, the classes are quite unbalanced. Almost 85% of the entries belong to class 0, less than 2% belong to class 1 and around 13% to class 2. In table (number) dark gray highlights all the binary features (14 in total), while light gray implies either continuous or categorical features. Among others, the binary features represent medical related aspects (high blood pressure, cholesterine, previous heart attacks, strokes, having difficulty at walking), lifestyle related indicators (physical activity, smoking habits, eating habits) pointers in terms the individual's relationship with the healthcare system (having any healthcare, benefiting from medical services, being able to finance visits to the doctor). As it can be observed CholCheck, Stroke, HeartDiseaseorAttack, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost and DiffWalk are all highly unbalanced.

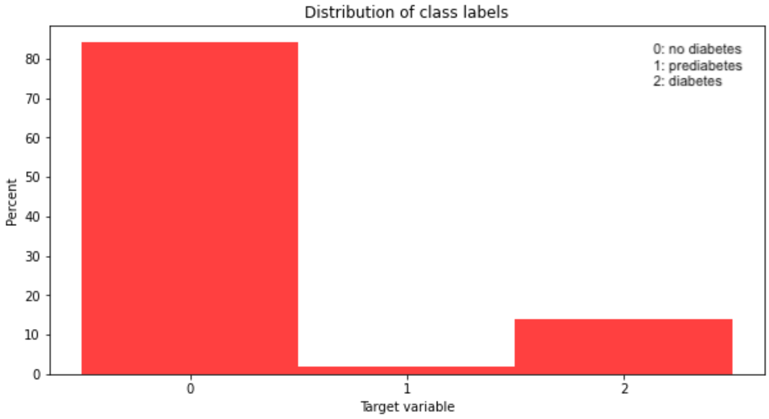


Fig.2 Distribution of target labels

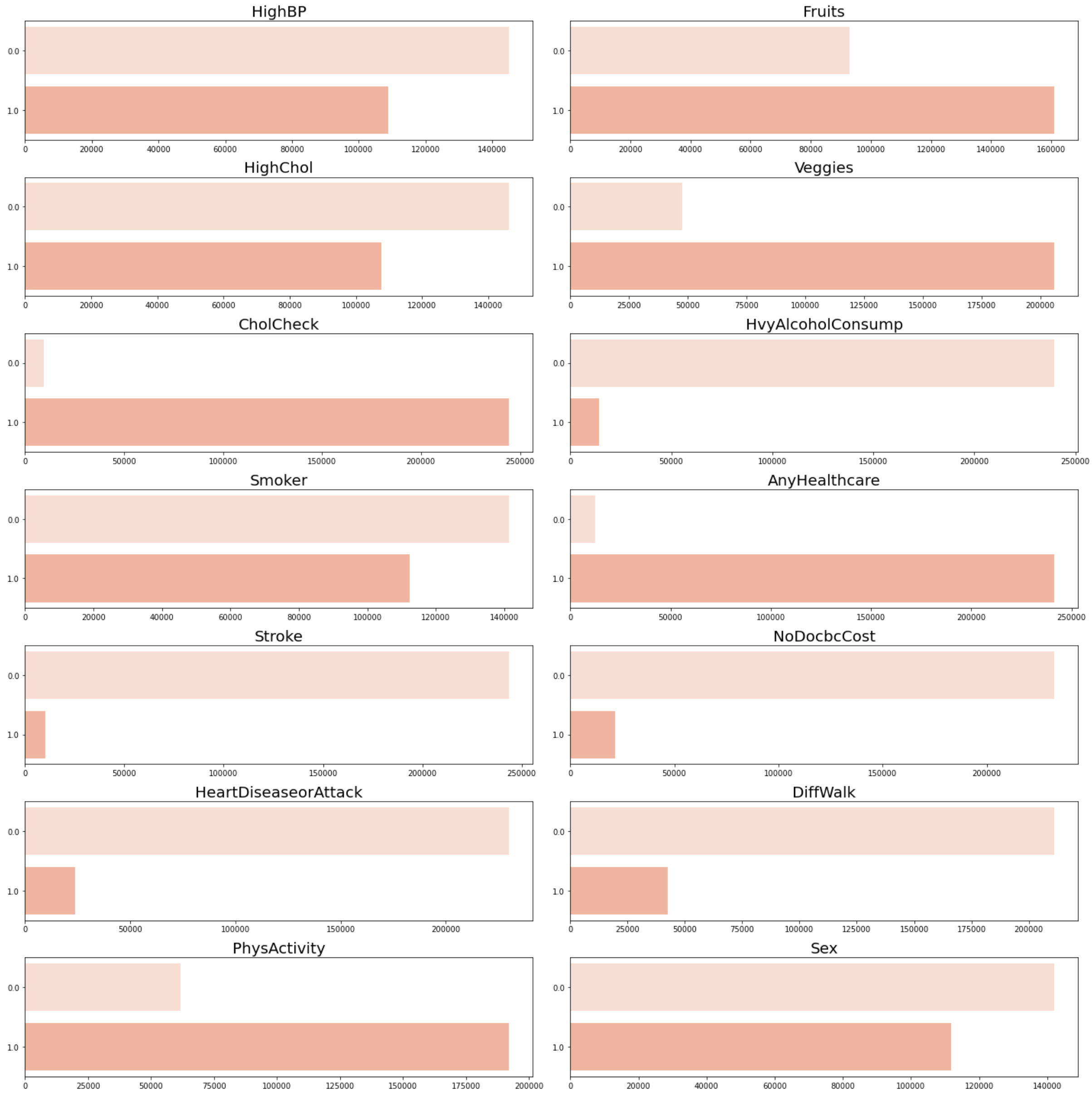


Fig.3 Distribution of binary features

The remaining six features, and their distributions are plotted and discussed at the feature engineering section. Body Mass Index (BMI) is a continuous value that is derived from the height and weight of a person. General health (GenHlth), mental health (MentlHlth) and physical health (PhysHlth) are all placed on a scale; age is represented through binned values, basically representing various age-groups, although they cannot be directly identified. Lastly, there are education and income, both features being categorical. One thing to note, is the fact that based on education and income, it seems that the higher socioeconomic classes are overrepresented in this survey.

It can be observed from the correlation matrix that GenHlth has the highest correlation with the target, at 0.3. A poor general Health has the highest values in the set of values for GenHlth and the positive correlation with the target indicates that in the context of a degraded health status, diabetes is more likely to develop.

Income has the highest absolute value of the negative correlations with the target, at -0.17. This indicates that a wealthy person has lower chances of developing diabetes than one with a poor financial status.

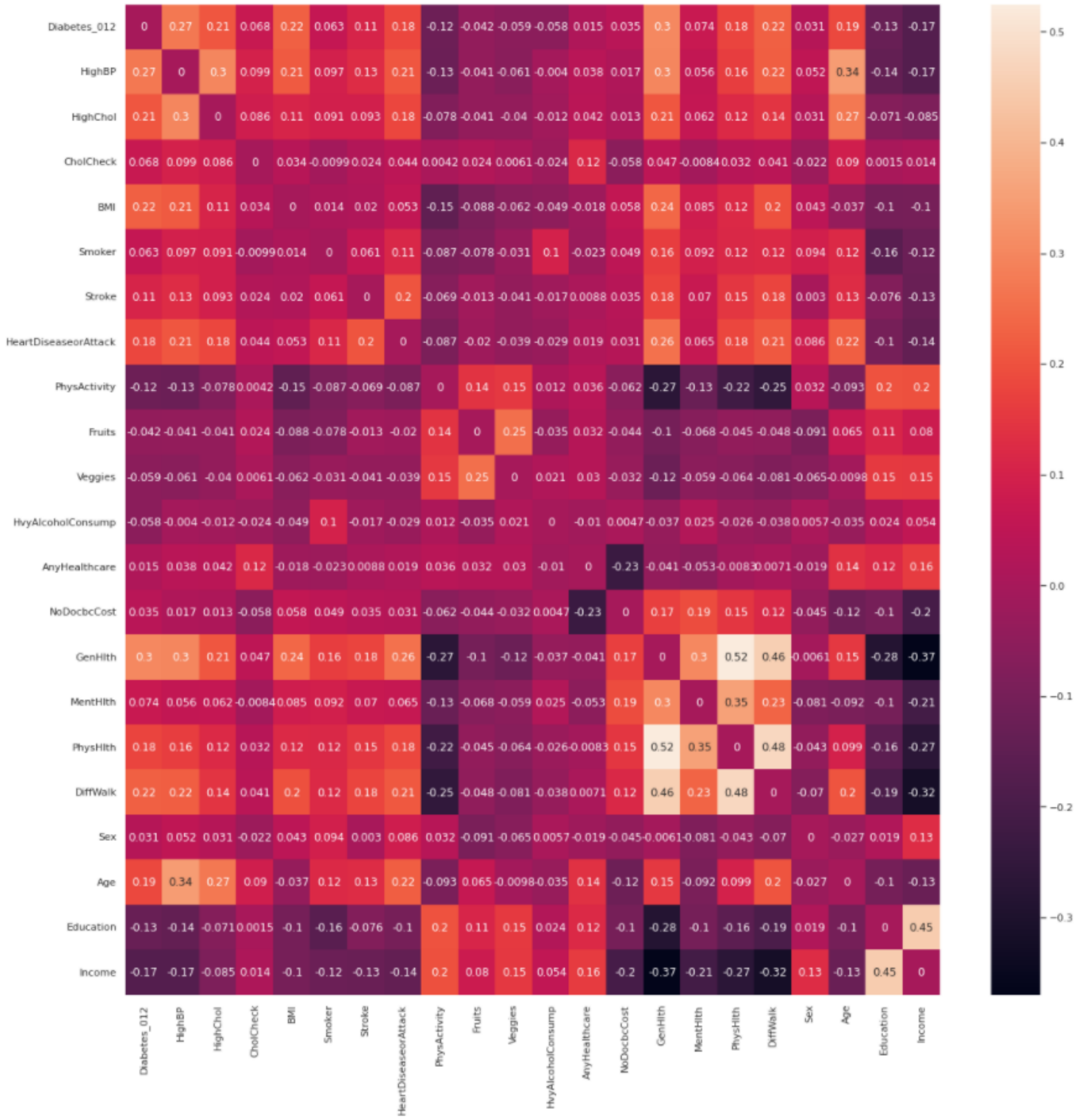


Fig. 4 Correlation matrix for all the features and the target variable

Feature engineering

Skewness is the degree of asymmetry of a distribution. If the frequency distribution has a longer “tail” to the right of the central maximum than to the left, the distribution is said to be skewed to the right (or to have a positive skewness). If the reverse is true, it is said to be skewed to the left (or to have a negative skewness). For a skewed distribution, the mean tends to lie on the same side of the mode as the longer tail.

The histograms of features whose skewness significantly changed also contain the distribution before the transformation.
For right skewed data, every value was transformed by applying log(1+x).
MentHlth skewness changed from 2.72 to 1.52.
PhysHlth skewness changed from 2.21 to 1.24.
BMI skewness changed from 2.10 to 0.71.

For left skewed data, every value was transformed by applying x^3.
Education skewness changed from -0.77 to -0.17.
Income skewness changed from -0.89 to -0.11.

Out of all the features, the set of BMI possible values has the highest cardinality, being equal to 84. Its distribution is closer to a Gaussian Distribution and the right skew adjustment can be observed in the histogram below.

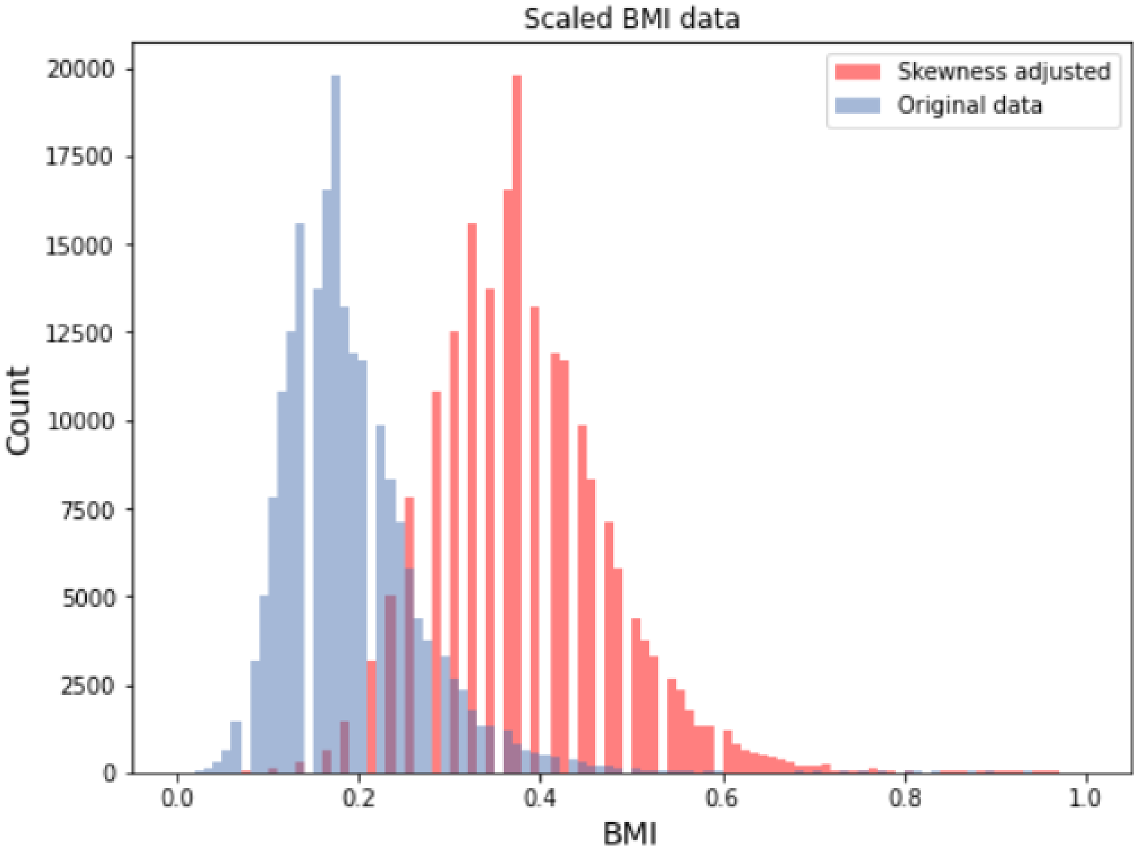


Fig. 5 BMI before and after right skewness adjustment

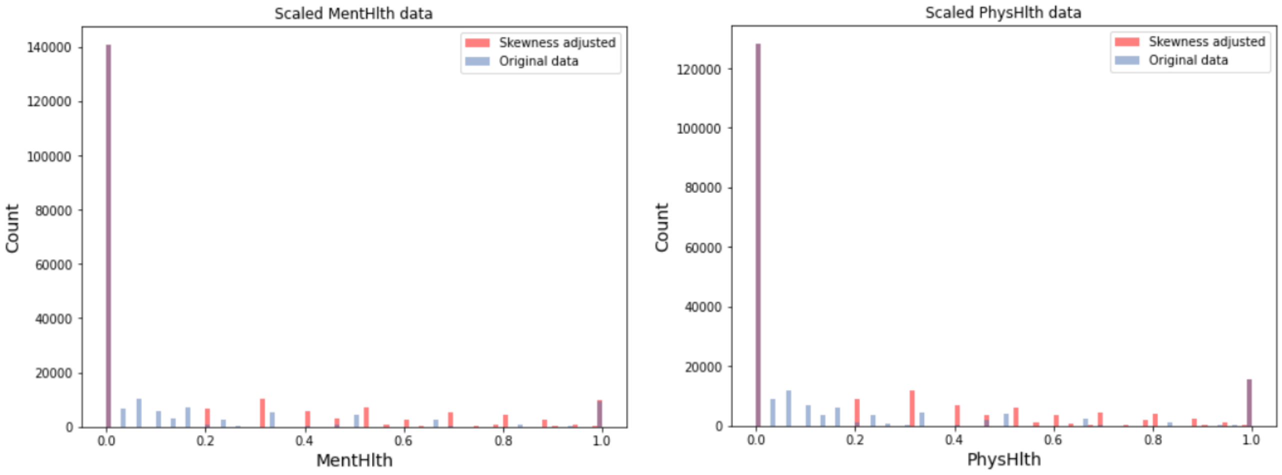


Fig. 6 MentHlth, PhysHlth before and after right skewness adjustment

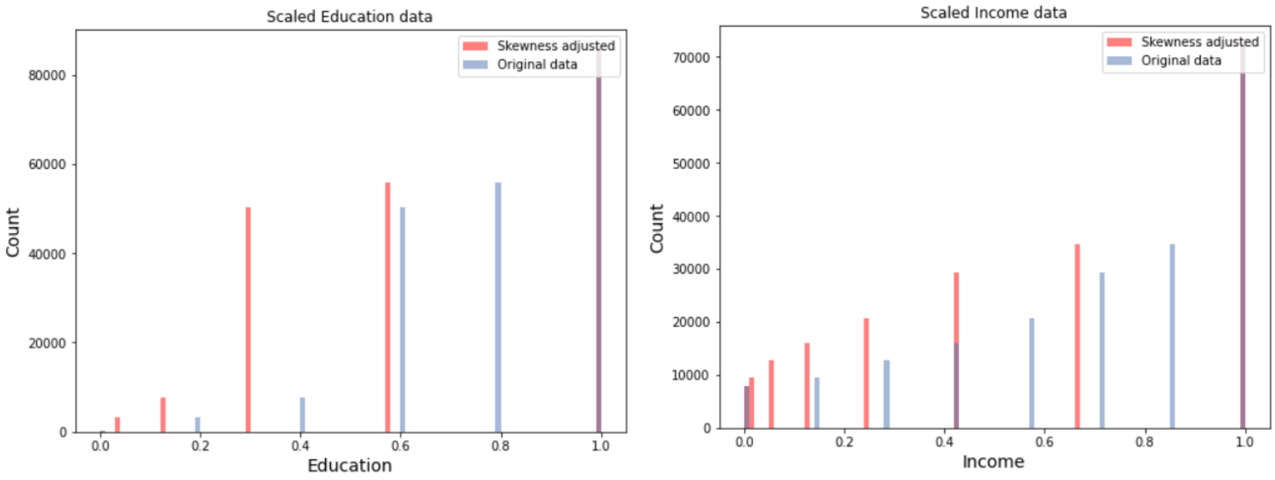
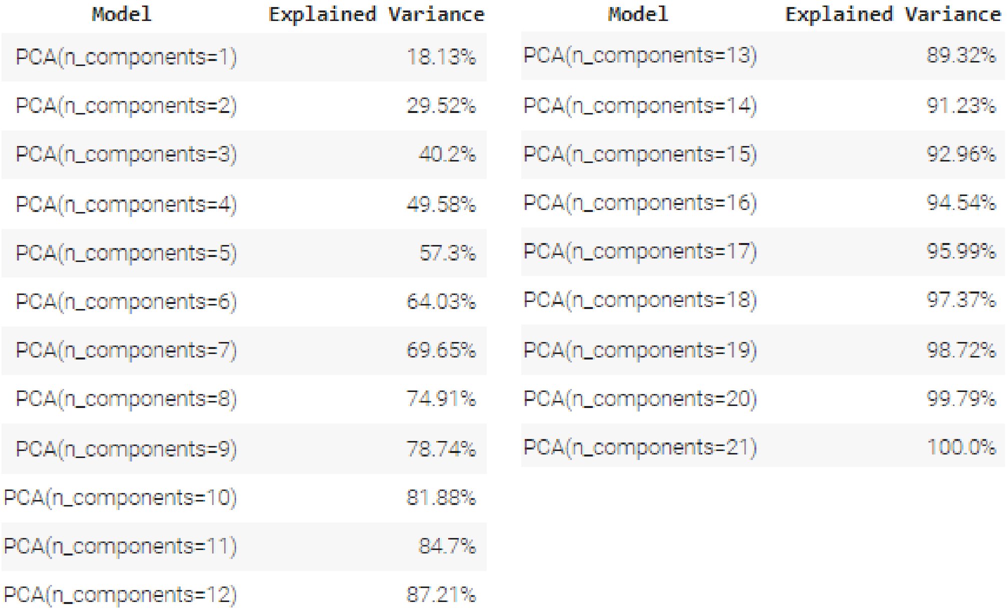


Fig. 7 Education and Income before and after left skewness adjustment

PCA

Large datasets are increasingly common and are often difficult to interpret. Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance. Finding such new variables, the principal components, reduces to solving an eigenvalue/eigenvector problem, and the new variables are defined by the dataset at hand, not *a priori*, hence making PCA an adaptive data analysis technique. It is adaptive in another sense too, since variants of the technique have been developed that are tailored to various different data types and structures.

In the table below, you can find the percentage of explained variance depending on the number of principal components.



Model	Explained Variance
PCA(n_components=1)	18.13%
PCA(n_components=2)	29.52%
PCA(n_components=3)	40.2%
PCA(n_components=4)	49.58%
PCA(n_components=5)	57.3%
PCA(n_components=6)	64.03%
PCA(n_components=7)	69.65%
PCA(n_components=8)	74.91%
PCA(n_components=9)	78.74%
PCA(n_components=10)	81.88%
PCA(n_components=11)	84.7%
PCA(n_components=12)	87.21%
PCA(n_components=13)	89.32%
PCA(n_components=14)	91.23%
PCA(n_components=15)	92.96%
PCA(n_components=16)	94.54%
PCA(n_components=17)	95.99%
PCA(n_components=18)	97.37%
PCA(n_components=19)	98.72%
PCA(n_components=20)	99.79%
PCA(n_components=21)	100.0%

Fig. 8 Retained variance based on the number of components

Polynomial features

We ran every possible combination of polynomial features in the format A^i * B^j with i and j ranging between -3 and 3 with a step of 0.5. We chose only the polynomial features which had a correlation with the target higher than 0.34.

Polynomial Feature	Base Features used
BMI^0.5 * GenHlth^0.5	BMI, GenHlth
HighBP^0.5 * GenHlth^1.5	GenHlth, HighBP
BMI^1.0 * GenHlth^1.0	BMI, GenHlth
HighBP^1.0 * GenHlth^1.5	GenHlth, HighBP
BMI^1.0 * GenHlth^1.5	BMI, GenHlth
BMI^1.0 * GenHlth^2.0	BMI, GenHlth
HighBP^1.5 * GenHlth^1.5	GenHlth, HighBP
HighBP^2.0 * GenHlth^1.5	GenHlth, HighBP
HighBP^2.5 * GenHlth^1.5	GenHlth, HighBP

Table 1. Polynomial Features

Metrics

For evaluating confusion matrices, various statistical methods can be employed in the binary case. Although accuracy and F1 scores are quite popular, they cannot be evaluated objectively – as they can show overoptimistic values for model performance – in the case of imbalanced datasets). In contrast, the Matthews correlation coefficient is a measure that relies on the four categories of a confusion matrix (true positive, true negative, false positive, false negative) in alignment with the associated proportions (Chicco & Jurman, 2020). Thus, it might be a more robust evaluation metric in cases where the class labels are highly unbalanced. Originally, it applies to the binary case, however its generalization to the multiclass case has been implemented in sklearn as well (*StackExchange*, n.d.; "Phi Coefficient," 2021). Together with our colleagues, we found rationale in its usage, principally because the task at hand is related to the medical field.

Given the nature of the problem, false negatives should be more focused on than false positives or the accuracy of a given model. Since a false positive could induce further checks – be it examinations or consultations by a specialist – the consequences of such a misclassification would be less impactful. In contrast, a false negative could produce a false belief that things are alright and could hinder a correct diagnosis, contributing to an increased risk factor for various complications.

Modelling Approaches

SVM

Dealing with a multiclass classification problem and having as input a dataset with a small number of dimensions relative to the sample size, it was no surprise that when applying the SVM algorithm, the best results were achieved using a polynomial kernel.

Given the highly dense space and the small number of classes, a low value of gamma was chosen in order to keep a large similarity radius between points. In our experiments, c is 0.

In order to deal with the class imbalance, class_weight was set to balanced. The "balanced" mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data

$$K(x, y) = (\gamma x^{\mathsf{T}} y + c)^d$$

All the variations of features are presented below.

Features v1 - base features
Features v2 - base features with skewness adjustment.
Features v3 - base features and polynomial features presented in Table 1.
Features v4 - base features and polynomial features with skewness adjustment
Features v5 - Applied PCA to base features and polynomial features and used the first 2 principal components
Features v6 - Applied PCA to base features and polynomial features and used the first 14 principal components, which cover at least 90% of total variance.

	Accuracy	MCC	F1 Score	C	Gamma	Kernel	Degree (d)
Features v1	0.639	0.272	71.45	0.1	0.1	polynomial	3
Features v2	0.649	0.288	72.26	0.1	0.2	polynomial	3
Features v1	0.654	0.291	72.36	0.1	0.1	polynomial	3
Features v2	0.656	0.301	72.43	0.1	0.2	polynomial	3
Features v3	0.644	0.278	71.86	0.1	0.1	polynomial	3
Features v4	0.65	0.284	72.24	0.1	0.2	polynomial	3
Features v3	0.66	0.294	72.55	0.1	0.1	polynomial	3
Features v4	0.661	0.296	72.69	0.1	0.2	polynomial	3
Features v5	0.792	0.244	79.77	0.1	0.1	polynomial	3
Features v5	0.752	0.266	78.58	0.1	0.2	polynomial	3
Features v6	0.759	0.261	78.96	0.1	0.1	polynomial	3
Features v6	0.695	0.281	69.56	0.1	0.2	polynomial	3

Fig. 9 Metrics and hyperparameters for every feature and parameter combination

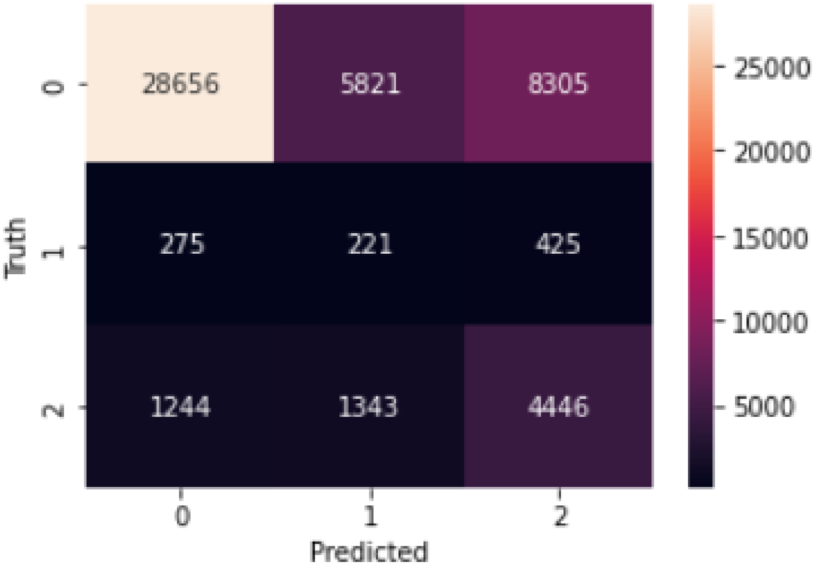


Fig. 10 Confusion Matrix for the model with the highest MCC.

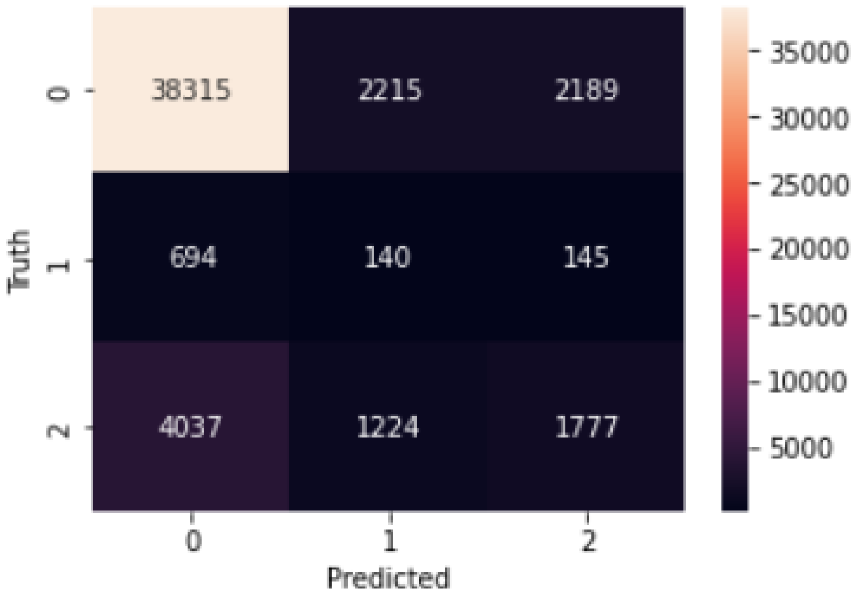


Fig. 11 Confusion Matrix for the model with the highest Accuracy.

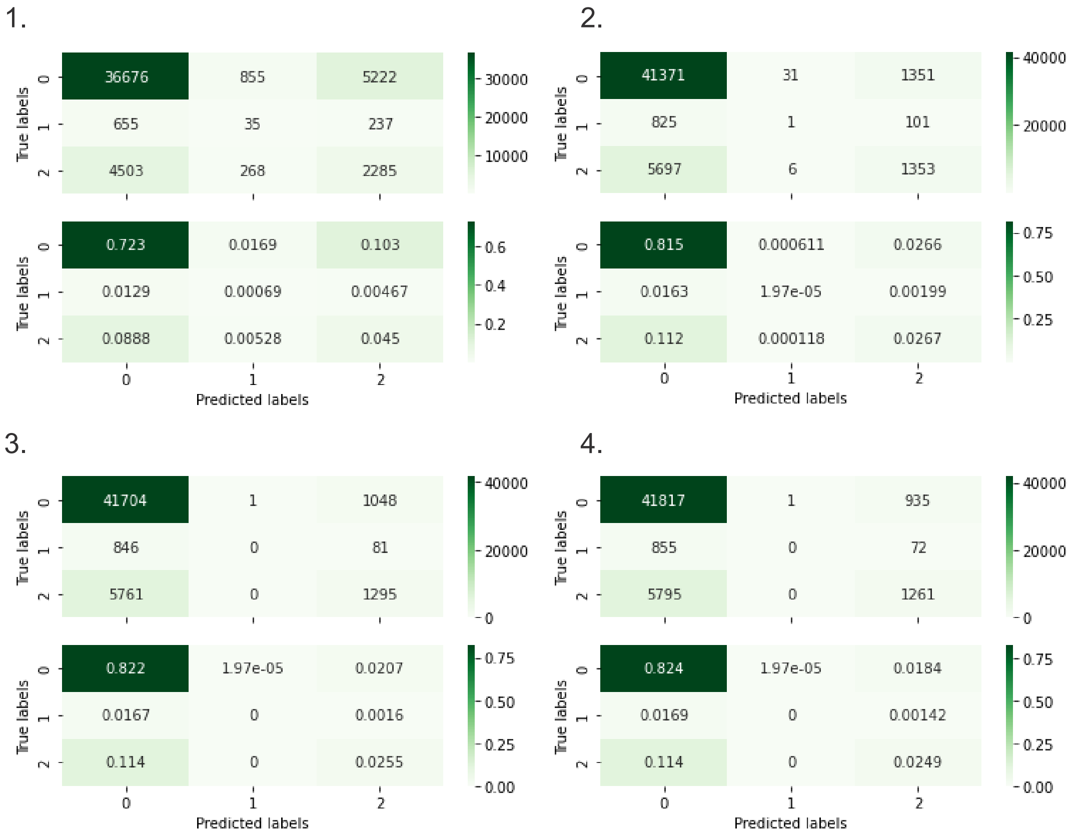
Tree based methods

For this section, four models were picked with various complexities; the list includes decision trees, random forests, XGB and LGBM. The first step was to establish a baseline for each model. Please find below the metrics for the base models:

#	Model type	Accuracy	MCC	F1
1	Decision tree	0.76	0.1843	0.77
2	Random forests	0.84	0.2303	0.84
3	XGB	0.8475	0.2453	0.85
4	LGBM	0.8490	0.2492	0.85

Fig. 12 Base models and the associated metrics

However, these initial values need to be interpreted with care, as the confusion matrices show it below. These models are rather bad, since the number of false positives and negatives is quite high, and thus the accuracy as performance metric is non-representative.



1.

True labels \ Predicted labels	0	1	2
0	36676	855	5222
1	655	35	237
2	4503	268	2285

True labels \ Predicted labels	0	1	2
0	0.723	0.0169	0.103
1	0.0129	0.00069	0.00467
2	0.0888	0.00528	0.045

2.

True labels \ Predicted labels	0	1	2
0	41371	31	1351
1	825	1	101
2	5697	6	1353

True labels \ Predicted labels	0	1	2
0	0.815	0.000611	0.0266
1	0.0163	1.97e-05	0.00199
2	0.112	0.000118	0.0267

3.

True labels \ Predicted labels	0	1	2
0	41704	1	1048
1	846	0	81
2	5761	0	1295

True labels \ Predicted labels	0	1	2
0	0.822	1.97e-05	0.0207
1	0.0167	0	0.0016
2	0.114	0	0.0255

4.

True labels \ Predicted labels	0	1	2
0	41817	1	935
1	855	0	72
2	5795	0	1261

True labels \ Predicted labels	0	1	2
0	0.824	1.97e-05	0.0184
1	0.0169	0	0.00142
2	0.114	0	0.0249

Fig. 13 Confusion matrices for each base model

As it can be observed the third and fourth models also struggle in terms of predicting class 1, however this can be also attributed to the fact that less than 2% of the entries belong to prediabetes. With the base models established, the workflow continued with experimenting with various parameter settings, exploring dimensionality reduction – for these models it didn't make that much sense as training time was quite reasonable – and finding ways to combat the class imbalance problem. After all, the most sensible approach consisted of undersampling the majority class, which helped considerably. Next, we have introduced the same polynomial features that were discussed before, however this time the top 95 features have been added to the dataset. After undersampling, the shape of the new dataframe ended up being (81994, 116). We have performed hyperparameter searches for the best models and selected the final ones based on their mcc score and the subsequent confusion matrices. Additionally, the related feature importance plots can be visualized as well.

#	Model type	Accuracy	MCC	F1
1	LGBM w/o tuning	0.7882	0.3749	0.79
2	XGB w/ tuning[1]	0.7876	0.3723	0.79
3	XGB w/o tuning	0.7860	0.3693	0.79
4	RF w/ tuning[2]	0.7872	0.3665	0.78

Fig. 14 Best models and the associated metrics

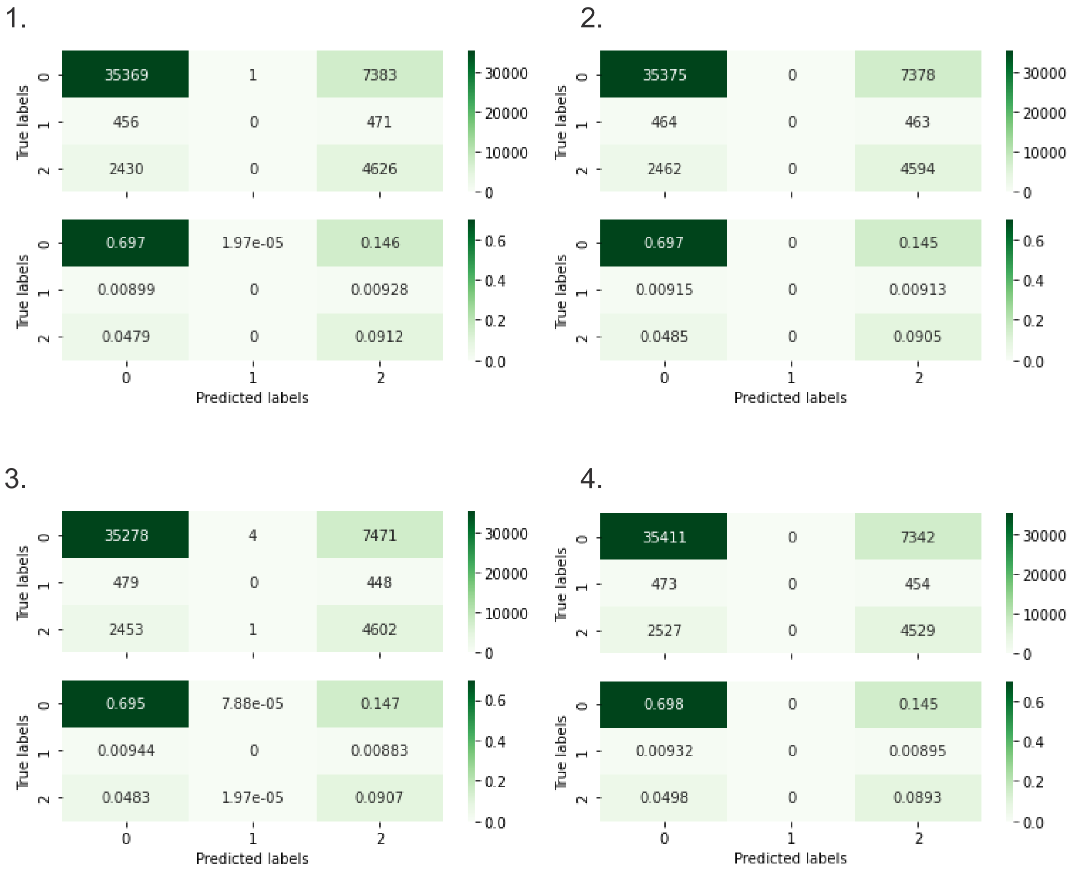


Fig. 14 Confusion matrices for each tuned models

[1] Search using RandomizedSearchCV; parameters: {'n_estimators': 1000, 'subsample': 0.5, 'min_child_weight': 0,'max_depth': 6,'max_delta_step': 2,'gamma': 2.5,'eta': 0.02526}

[2] Search using RandomizedSearchCV; parameters: {'n_estimators': 700, 'min_samples_split': 8, 'min_samples_leaf': 8, 'max_features': 'sqrt', 'max_depth': 10, 'criterion': 'gini', 'bootstrap': True}

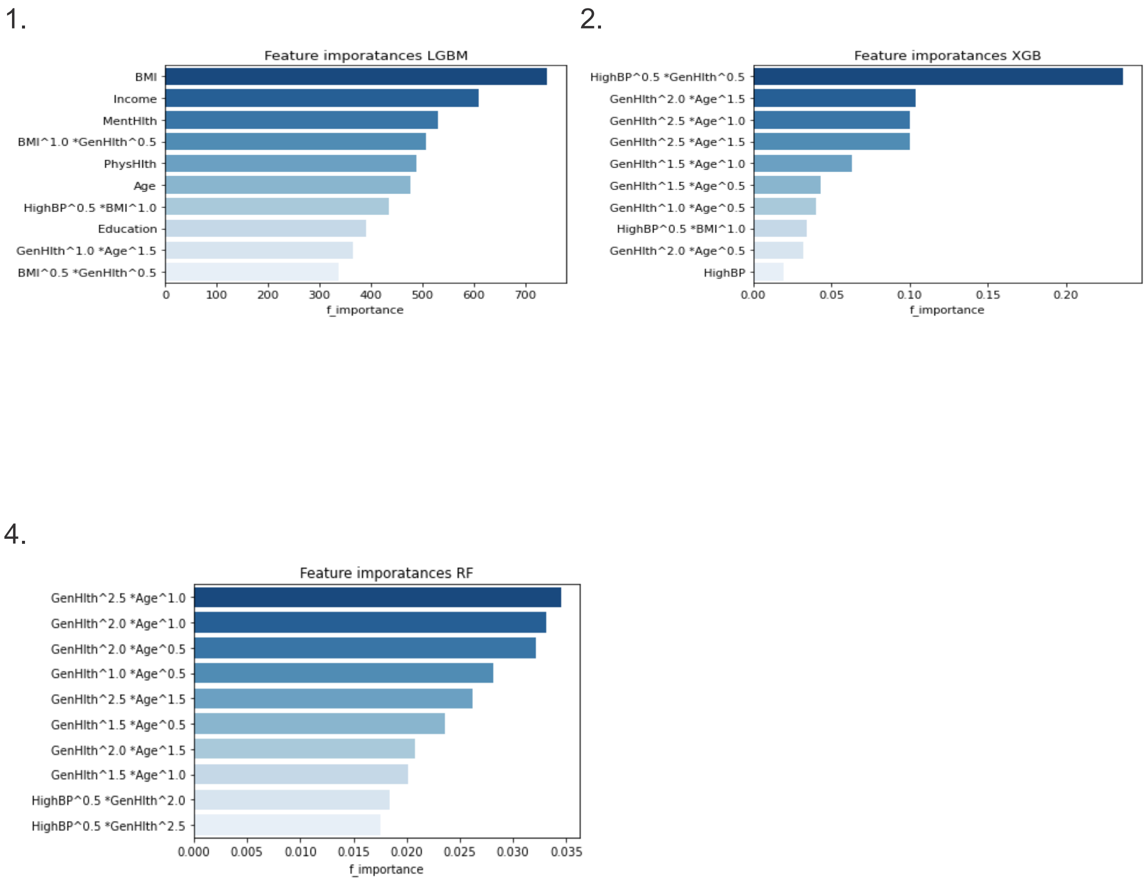


Fig. 15 Confusion matrices for each base model

We have tried manual tuning, however the lack of experience didn't enable us to improve our models. In the case of LGBM, hyperparameter tuning with Optuna[3], yielded no additional improvement. This can be attributed to probably two main things, we couldn't optimize directly for mcc, and secondly, we might have considered the wrong parameter. One more thing to note, is that we have also tried a couple of ensemble methods – hard voting classifiers – however it couldn't outperform the previous best.

[3] Optuna - A hyperparameter optimization framework : A bayesian optimization framework

KNN

For a multiclass classification, we thought that kNN could be an appropriate choice. The baseline model achieved the best accuracy score (82.24%), without preprocessing the data. After that, we tried to get better results, by scaling the data using MIN-MAX and standard scaler, but there were no major differences.

	Accuracy score	F1 score	MCC
1.No preprocessing	82.24	79.85	0.19
2.MIN-MAX scaler	81.79	79.58	0.19
3.Standard Scaler	82.08	79.85	0.20

Fig. 15 Metrics for various base models

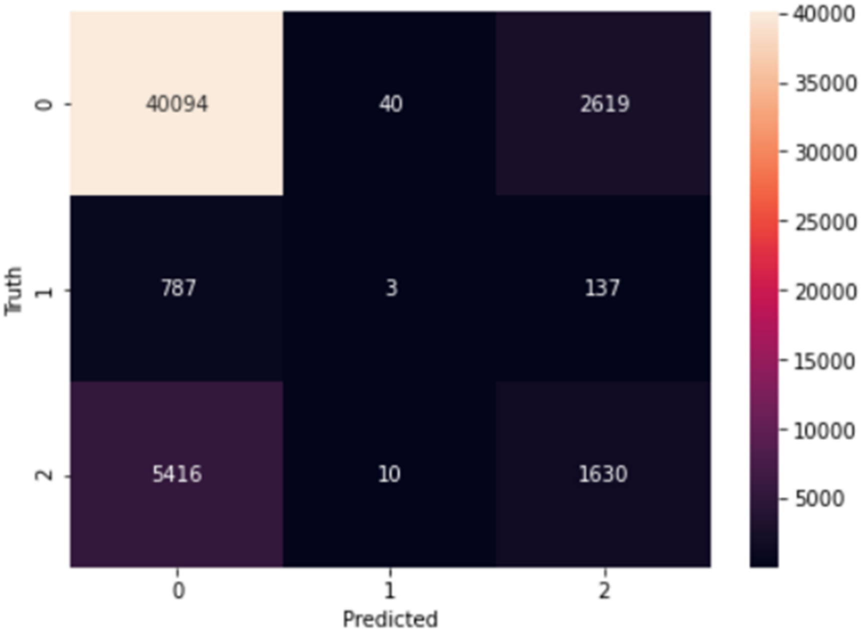


Fig. 16 No preprocessing, k = 3

After this we resampled the data, keeping only 50.000 entries from the original dataset and tried various hyperparams (n_neighbors, distance).
The best results for MCC were achieved using k = 19 and 21 (MCC = 0.33).

#	Hyperparams	Accuracy	MCC
1	k = 3	0.74	0.25
2	k = 5	0.83	0.20
3	k = 7	0.76	0.29
4	k = 9	0.77	0.30
5	k = 11	0.77	0.31
6	k = 13	0.77	0.31
7	k = 15	0.77	0.32
8	k = 17	0.77	0.32
9	k = 19	0.77	0.33
10	k = 21	0.77	0.33

Fig. 17 Trying various n_neighbors

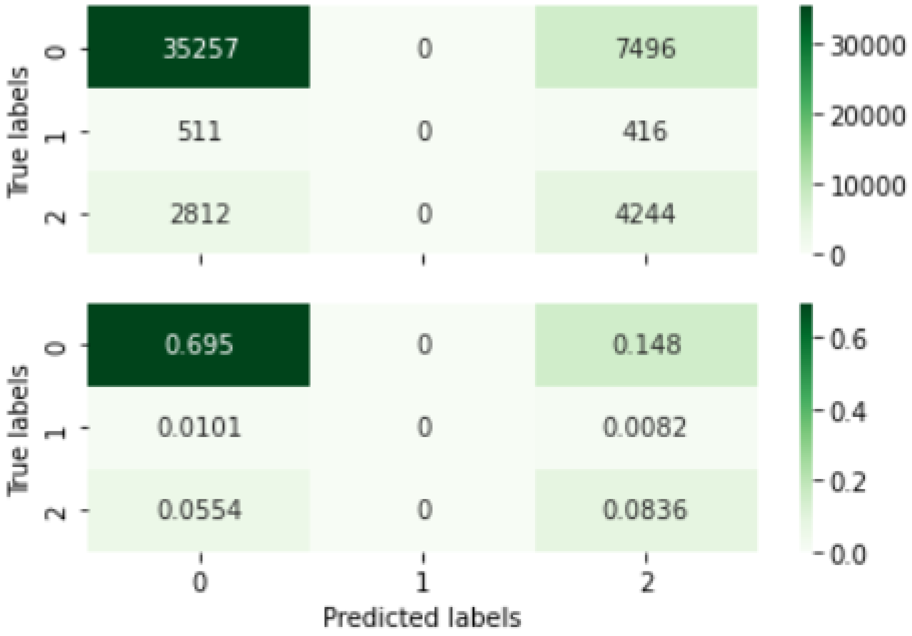


Fig. 18 Confusion matrix for n_neighbors = 19

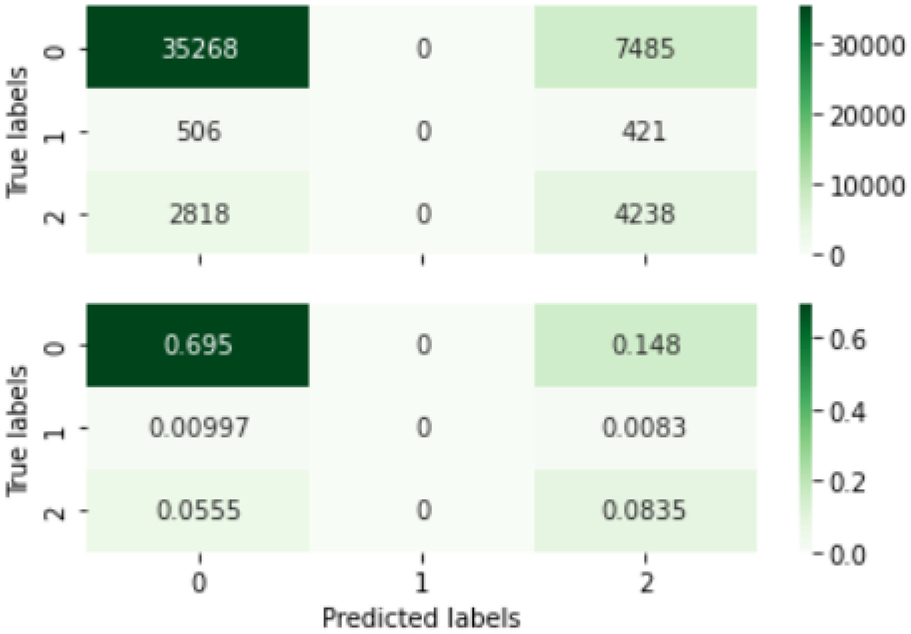


Fig. 18 Confusion matrix for n_neighbors = 21

Conclusion

Applying PCA to the base features and polynomial features and using the first 2 principal components ended up being the best trial for the svm class in terms of accuracy, however the combination of base features with skewness adjustment resulted in the best MCC. For the tree-based methods LGBM with no hyper-parameter tuning, 100 additional polynomial features and undersampling scored the highest value for MCC, closely followed by XGB with hyper-parameter tuning. Lastly, somewhat surprisingly a KNN model with 19 neighbors performed quite well.

APPENDIX

A detailed description of all the features and their corresponding types as described on the Kaggle page (*Diabetes Health Indicators Dataset*, n.d.).

#	Feature name	Meaning	Feature type
1	HighBP	High blood pressure	binary
2	HighChol	High cholesterol level	binary
3	CholCheck	Cholesterol check in the last 5 years	binary
4	BMI	Body Mass Index	continuous
5	Stroke	Respondent had a stroke in the last 5 years	binary
6	HeartDiseaseorAttack	If the respondent had coronary heart disease or myocardial infarction	binary
7	PhysActivity	If the respondent did any intense physical activity in the past 30 days	binary
8	Fruits	If the respondent consumes at least one fruit daily	binary
9	Veggies	If the respondent consumes at least one vegetable daily,	binary
10	HvyAlcoholConsump	If the respondent is an adult man who serves more than 14 drinks per week or a woman who serves more than 7 drinks per week	binary
11	AnyHealthcare	If the respondent has any kind of health care coverage, including health insurance or prepaid plans such as HMO	binary
12	NoDocbcCost	If the respondent needed to see a doctor in the last 12 months but couldn’t afford it	binary
13	GenHlth	How the respondents rated their health on a scale ranging from 1 to 5.	categorical
14	MentHlth	Quantification of stress, depression and emotion problems on a scale ranging from 1 to 30.	categorical

15	PhysHlth	Quantification of physical illnesses and injuries on a scale ranging from 1 to 30.	categorical
16	DiffWalk	If the respondent has serious difficulty walking or climbing stairs,	binary
17	Sex	0 represents female, 1 male	binary
18	Age	13-level age categories	categorical
19	Education	Education level on a scale ranging from 1 to 6.	categorical
20	Income	Income on a scale ranging from 1 to 8.	categorical

BIBLIOGRAPHY

CDC. (2021, November 16). *What is Diabetes?* Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/basics/diabetes.html

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Diabetes Health Indicators Dataset. (n.d.). Kaggle. Retrieved January 10, 2022, from https://kaggle.com/alexteboul/diabetes-health-indicators-dataset

Matthews correlation coefficient with multi-class. (n.d.). Stack Exchange. Retrieved January 15, 2022, from https://stats.stackexchange.com/questions/187768/matthews-correlation-coefficient-with-multi-class

Phi coefficient. (2021). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Phi_coefficient&oldid=1056280288

What is Diabetes? | NIDDK. (n.d.). National Institute of Diabetes and Digestive and Kidney Diseases. Retrieved January 15, 2022, from https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes