Prediction of the occurrence of type II diabetes

Sergei Bakaleinik
Innopolis University
Innopolis, Russia
s.bakaleinik@innopolis.university

Supervisor: Yaroslav Kholodov
Innopolis University
Innopolis, Russia
ya.kholodov@innopolis.ru

Abstract

This work is about effective usage of machine learning methods and data engineering for prediction of type II diabetes mellitus (T2DM) and contributes to development of methods of prediction T2DM occurrence. The goal of this research is to show effective work with big data, some analytics, and models for classification. In the results of this research, I collect different metrics (such as accuracy, precision, recall, f1-score, AUC, and ROC-curve) on tested models.

Introduction

Type II diabetes mellitus (T2DM) is one of the major causes of mortality in the world. According to WHO information:

- The number of people with diabetes rose from 108 million in 1980 to 422 million in 2014.
- The global prevalence of diabetes among adults over 18 years of age rose from 4.7% in 1980 to 8.5% in 2014.
- Diabetes prevalence has been rising more rapidly in low- and middle-income countries than in high-income countries.
- Diabetes is a major cause of blindness, kidney failure, heart attacks, stroke, and lower limb amputation.
- In 2016, an estimated 1.6 million deaths were directly caused by diabetes. Another 2.2 million deaths Ire attributable to high blood glucose in 2012.
- Almost half of all deaths attributable to high blood glucose occur before the age of 70 years. WHO estimates that diabetes was the seventh leading cause of death in 2016.
- A healthy diet, regular physical activity, maintaining normal body weight, and avoiding tobacco use are ways to prevent or delay the onset of type 2 diabetes.
- Diabetes can be treated and its consequences avoided or delayed with diet, physical activity, medication, and regular screening and treatment for complications.

Over the past three decades, a large number of potential prognostic biomarkers have been investigated in the context of T2DM risk prediction. Among them are traditional biomarkers belonging to the canonical signaling and metabolic pathways associated with T2DM etiology and pathogenesis (their relationship with T2DM is Ill established) and new biomarkers obtained using modern high-performance methods such as mass spectrometry / liquid chromatography-tandem. DNA sequencing, and gene expression differential analysis. The following biomarkers have the highest predictive power according to a broad meta-analysis of various predictors and predictive models [1]: fast glucose concentration, fructosamine concentration, glucose tolerance test results, and concentrations of glycated albumin, glycated hemoglobin, and uric acid. Other biomarkers may also be used, including biochemical liver function indicators (ALT, AST, bilirubin, and others), immunological biomarkers, biomarkers for metal metabolism, and so on.

The use of non-invasive biomarkers (body mass index, wrist thickness, history, smoking, etc.), classical glycemic biomarkers (glycated hemoglobin, blood glucose concentration, glucose tolerance test), as Ill as new biomarkers (metabolic and genetic) are reviewed by Herder et al [2]

Single nucleotide polymorphisms in more than 60 loci of the genome are associated with the risk of developing type 2 diabetes [3–5]. The presence of each of these variants separately increases the risk of developing diabetes by an amount of 5 to 40% (OR 1.05–1.4). The use of 40 SNPs for predicting type 2 diabetes makes it possible to achieve an indicator of the AROC model from 0.55 to 0.63 [6]. A recent study of the MTNR1B locus encoding the melatonin 1B receptor showed that rare mutations could be present in this locus, leading to a greater risk of developing type 2 diabetes (OR 5.7, 95% CI 2.2, 14.8) [7]. It is shown that the use of genetic data allows making more accurate predictions of the risk of developing type 2 diabetes.

The decrease in plasma adiponectin concentration of less than 11.54 g/L is associated with the risk of developing type 2 diabetes. Other factors included in the analysis: HOMA-IR, AIR, IFG, IGT, number of cigarettes per year, [8]

To predict the risk of developing type 2 diabetes, profiling of metabolites was used by Gimble et al [9]. 9 metabolites (sorbitol, galactitol, mannose, galactose, uric acid, oxalic acid, glucaric acid-1,4-lactone, 3-methyl-2-oxopentanoic acid, 2-hydroxybutyric acid), combined with non-invasive methods can improve the quality of prediction by 9% compared with the adiponectin model

Three metabolites (glycine, lysophosphatidylcholine, and acetylcarnitine) are associated with the risk of developing insensitivity to glucose and type 2 diabetes. Wang-Sattler et al. 2012 identified 7 genes (PPARG, TCF7L2, HNF1A, GCK, IGF1, IRS1, and IDE), that change expression in the phenotype of developing type 2 diabetes [10]

Moreover, at this point, researchers are ready to introduce various machine-learning algorithms to try to improve predictions using routine clinical data [11]. These risk prediction models of different types include classic generalized linear regression, distributed random forest, gradient boosting machine, and artificial neural networks. For instance, data obtained from metabolic studies using machine learning to interpret results show that the decrease in the concentration of alpha-tocopherol is associated with the risk of developing type 2 diabetes

This exploratory study pursues two main objectives to be reached via machine-learning analysis: First, investigate whether biochemical biomarkers widely used in routine clinical practice can be employed for the T2DM risk prediction without any prior knowledge about biometrical parameters of the individual. Second, assess the predictive ability of a derived parameter (rate of a biomarker change over time) to be employed for T2DM risk prediction.

Data description

This study used a dataset from the MedExpert Medical Center. MedExpert is one of the largest non-state medical groups in the Voronezh region, Russia, which can collect most of the medical data of its patients (up to half a million) in digital form. The database contains main for our usage tables: measurement results, patient data, and anamnesis. Firstly I collect data in the table with rows as each measurement and columns timestamp; result id; patient id; gender; birthday; age; real target; glucose in blood and urine; height; weight; bilirubin common, direct, indirect, diastolic blood pressure, with size (2M)x(15). But the data was too sparse and working with it seems impossible. So, I decide to reduce dimensionality. I do this by collecting the last data (in terms of time) on each feature for each patient. So final data representation was each row is patient data, columns are the same as in first ones but without timestamp and result id with size (186K)x(14). These Ire examined during the years from 2012 to 2020, among which 165K persons obtained their measurements more than once. Features in this dataset: gender; age;

glucose in blood and urine; height; weight; bilirubin common, direct, indirect, and diastolic blood pressure.

Work with data

I found that body mass index is applicable as a diabetes indicator. I calculated it by the next formula:

$$I = \frac{m}{h^2} \; (1)$$

m – weight in kg; h – a height in meters.

By calculating BMI, I find that some height and weight could be misplaced. Also, I find in [12] that bilirubin direct must not be bigger than 30% of bilirubin common, and here also could be misplaced. I dropped unnormal data (data with very height BMI and diastolic blood pressure, very strange height, and weight and normalize somehow bilirubin).

Statistics

In our dataset men count with diabetes bigger than women. The biggest age group is 60 to 69 years old. Here are distributions by age and gender.

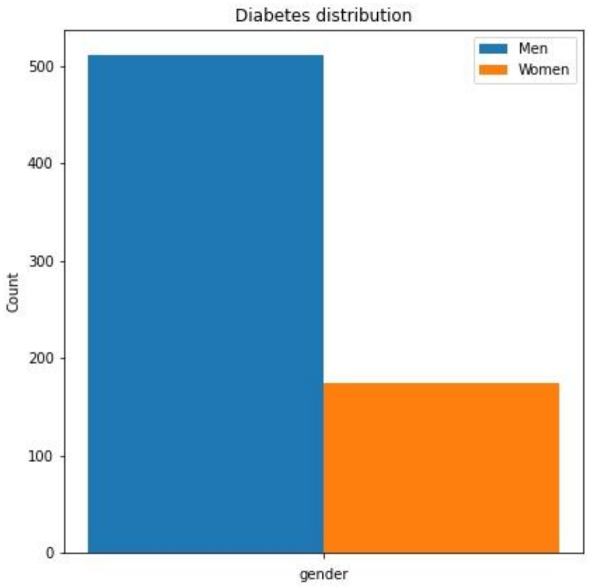


Figure 1. Diabetes distribution by gender

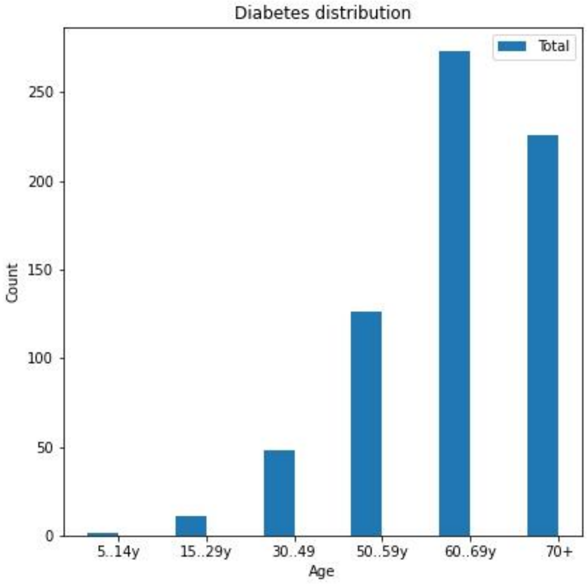


Figure 2. Diabetes distribution by age

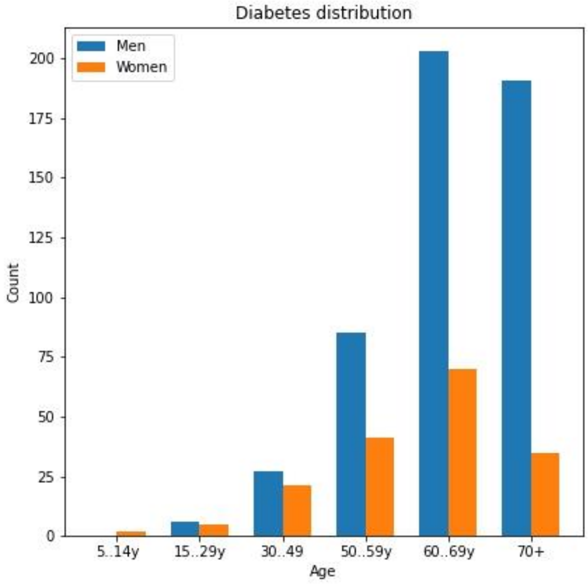
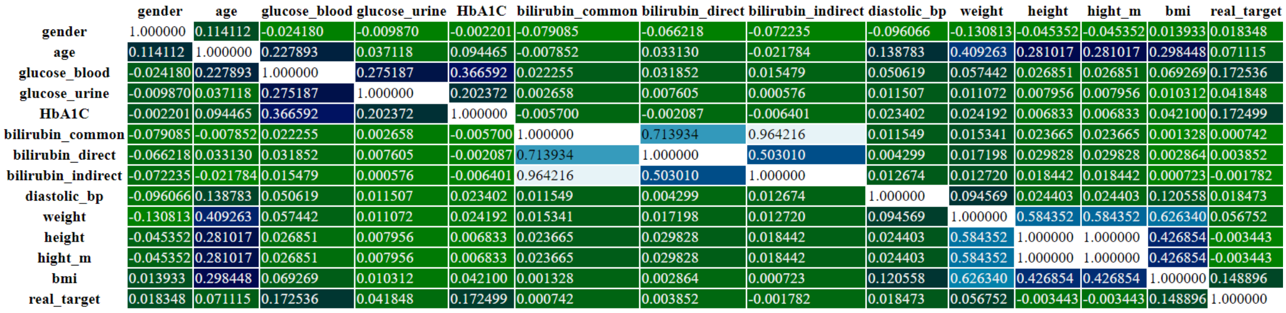


Figure 3. Diabetes distribution by age and gender

I also collected the correlation of features and include the target column. I understand that glucose in blood correlates with glucose in urine, age, and glycated hemoglobin; age correlates with diastolic blood pressure, weight, height, and BMI; the target correlates with glucose in blood, glycated hemoglobin, and BMI; all correlation is below.



	gender	age	glucose_blood	glucose_urine	HbA1C	bilirubin_common	bilirubin_direct	bilirubin_indirect	diastolic_bp	weight	height	hight_m	bmi	real_target
gender	1.000000	0.114112	-0.024180	-0.009870	-0.002201	-0.079085	-0.066218	-0.072235	-0.096066	-0.130813	-0.045352	-0.045352	0.013933	0.018348
age	0.114112	1.000000	0.227893	0.037118	0.094465	-0.007852	0.033130	-0.021784	0.138783	0.409263	0.281017	0.281017	0.298448	0.071115
glucose_blood	-0.024180	0.227893	1.000000	0.275187	0.366592	0.022255	0.031852	0.015479	0.050619	0.057442	0.026851	0.026851	0.069269	0.172536
glucose_urine	-0.009870	0.037118	0.275187	1.000000	0.202372	0.002658	0.007605	0.000576	0.011507	0.011072	0.007956	0.007956	0.010312	0.041848
HbA1C	-0.002201	0.094465	0.366592	0.202372	1.000000	-0.005700	-0.002087	-0.006401	0.023402	0.024192	0.006833	0.006833	0.042100	0.172499
bilirubin_common	-0.079085	-0.007852	0.022255	0.002658	-0.005700	1.000000	0.713934	0.964216	0.011549	0.015341	0.023665	0.023665	0.001328	0.000742
bilirubin_direct	-0.066218	0.033130	0.031852	0.007605	-0.002087	0.713934	1.000000	0.503010	0.004299	0.017198	0.029828	0.029828	0.002864	0.003852
bilirubin_indirect	-0.072235	-0.021784	0.015479	0.000576	-0.006401	0.964216	0.503010	1.000000	0.012674	0.012720	0.018442	0.018442	0.000723	-0.001782
diastolic_bp	-0.096066	0.138783	0.050619	0.011507	0.023402	0.011549	0.004299	0.012674	1.000000	0.094569	0.024403	0.024403	0.120558	0.018473
weight	-0.130813	0.409263	0.057442	0.011072	0.024192	0.015341	0.017198	0.012720	0.094569	1.000000	0.584352	0.584352	0.626340	0.056752
height	-0.045352	0.281017	0.026851	0.007956	0.006833	0.023665	0.029828	0.018442	0.024403	0.584352	1.000000	1.000000	0.426854	-0.003443
hight_m	-0.045352	0.281017	0.026851	0.007956	0.006833	0.023665	0.029828	0.018442	0.024403	0.584352	1.000000	1.000000	0.426854	-0.003443
bmi	0.013933	0.298448	0.069269	0.010312	0.042100	0.001328	0.002864	0.000723	0.120558	0.626340	0.426854	0.426854	1.000000	0.148896
real_target	0.018348	0.071115	0.172536	0.041848	0.172499	0.000742	0.003852	-0.001782	0.018473	0.056752	-0.003443	-0.003443	0.148896	1.000000

Figure 4. Correlation table

Models

In this task I try to use classical models: SVM (with weight for class “1” equals 10), Random Forest, Gradient boosting from sklearn, XGBoost, and CatBoost from Yandex. All models without SVM was used with basic parameters. Even after collecting data in a table that I collected, there were NA-values. Before training models, I split data by target and feel NA-values by the median for each feature and concatenate it to one dataset. The results of the work of models that I used below.

Table 1. Models results on 80% data for train

model	precision		recall		f1-score		accuracy
	on class 1	on class 0	on class 1	on class 0	on class 1	on class 0	
RandomForest	0.99	1	0.92	1	0.96	1	0.96
SVM (weight for class 1 - 10)	0.35	1	0.71	1	0.47	1	0.855
GradientBoostingClassifier	0.97	1	0.92	1	0.95	1	0.96
XGBoost	0.99	1	0.93	1	0.96	1	0.965
CatBoost	0.94	1	0.90	1	0.92	1	0.95

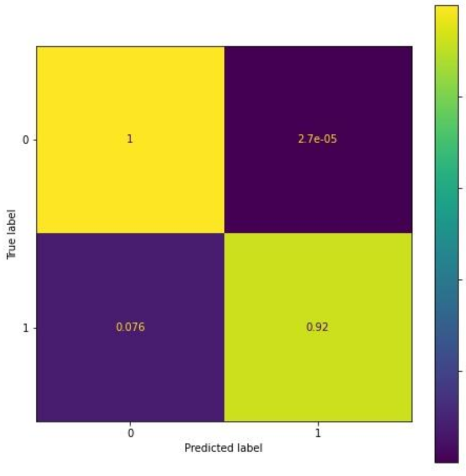


Figure 5. Confusion matrix for RandomForest

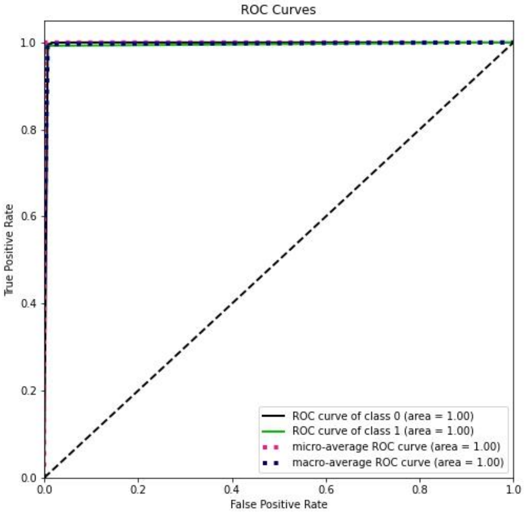


Figure 6. ROC-curve for RandomForest

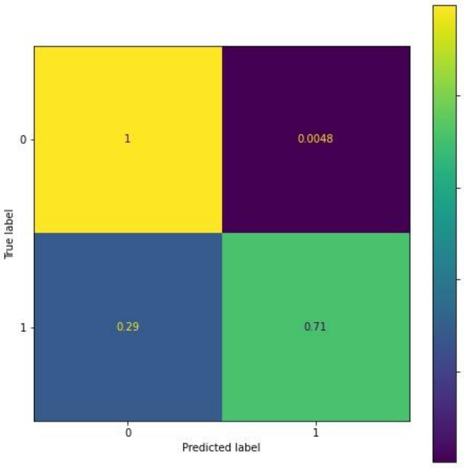


Figure 7. Confusion matrix for SVM

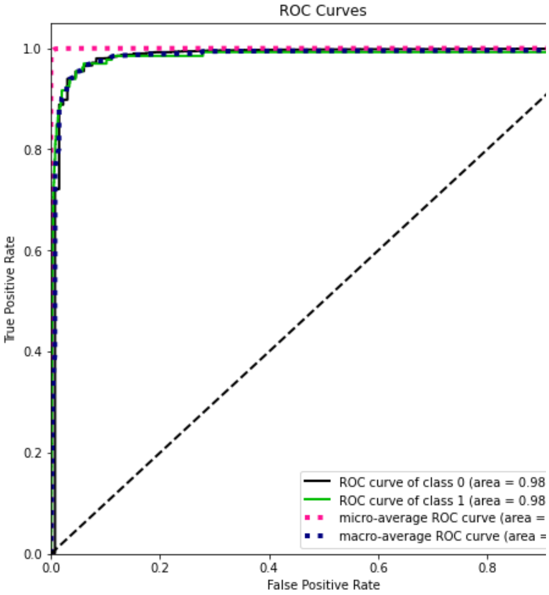


Figure 8. ROC-curve for SVM

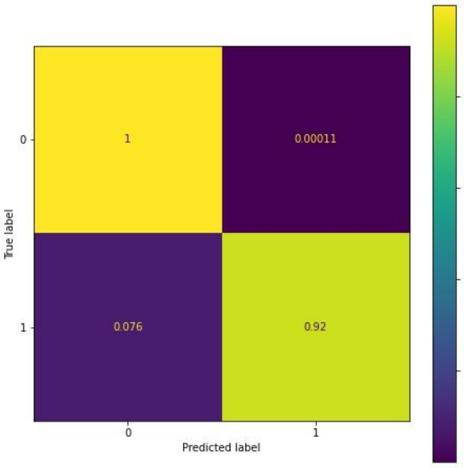


Figure 9. Confusion matrix for Gradient Booster Classifier

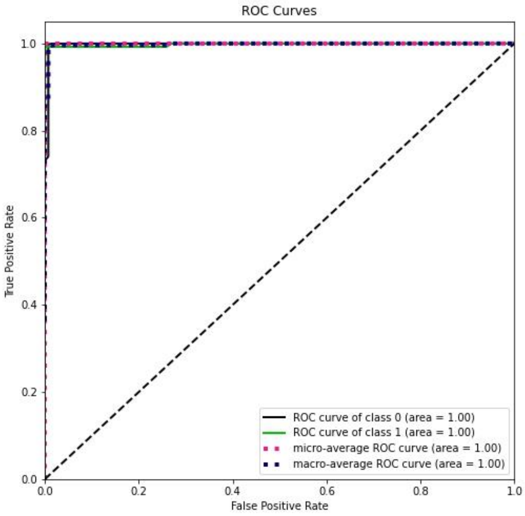
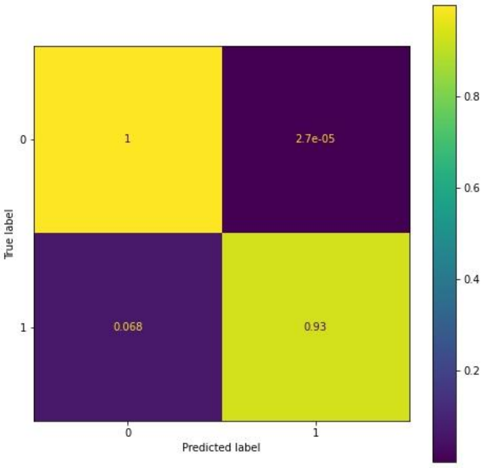


Figure 10. ROC-curve for Gradient Booster Classifier



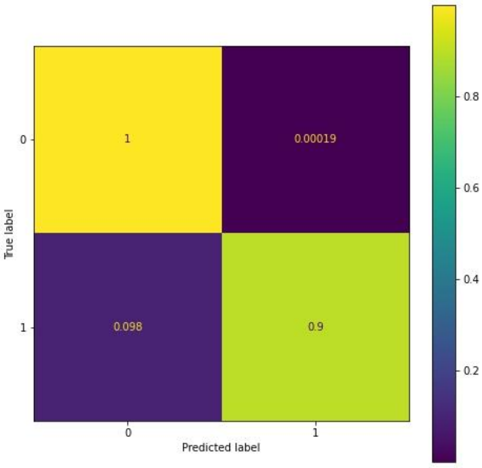


Figure 11. Confusion matrix for XGBoost

Figure 13. Confusion matrix for CatBoost

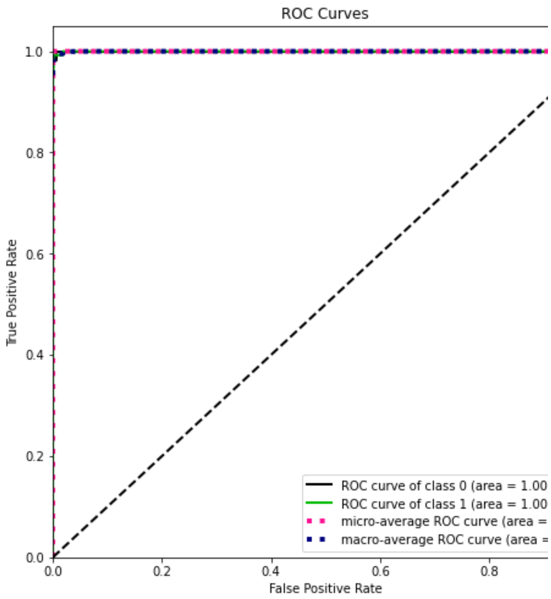


Figure 12. ROC-curve for XGBoost

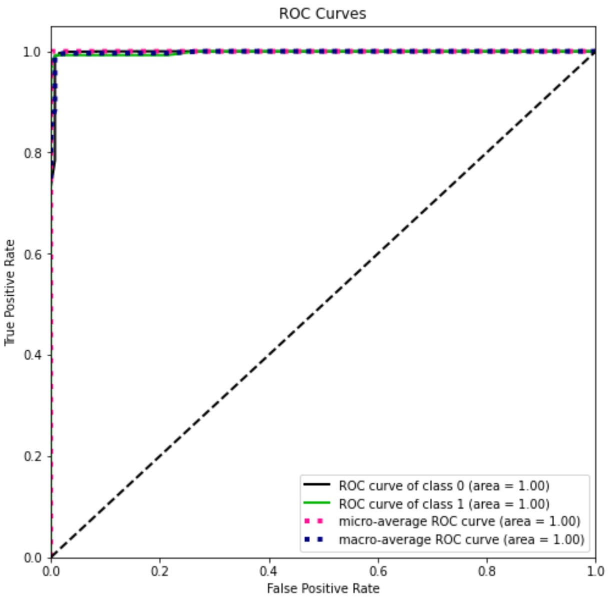


Figure 14. ROC-curve for CatBoost

Conclusion

From models results, we know that gradient boosting and tree-based models show good results. But more good results shown by XGBoost – model based on trees and gradient boosting. CatBoost that also based on trees and gradient boosting not showed very good results.

A variety of studies show that genetic factors significantly contribute to the risk of developing various forms of diabetes. Therefore, the model 's output can be improved by introducing new variables that represent the existence of certain genetic polymorphisms in the patient's genome.

Additionally, the development of the presented model and associated mathematical tools (including advanced machine learning methods) set the basis for risk assessment in other disease scenarios, including cardiovascular, oncological, and prenatal chromosome disorders.

References

[1] Abbasi A. et al. A systematic review of biomarkers and risk of incident type 2 diabetes: an overview of epidemiological, prediction and aetiological research literature //PloS one. – 2016. – Т. 11. – №. 10.

[2] C. Herder, B. Kowall, A.G. Tabak, W. Rathmann, The potential of novel biomarkers to improve risk prediction of type 2 diabetes., Diabetologia. 57 (2014) 16–29. doi:10.1007/s00125-013-3061-3.

[3] C. Herder, M. Karakas, W. Koenig, Biomarkers for the prediction of type 2 diabetes and cardiovascular disease., Clin. Pharmacol. Ther. 90 (2011) 52–66. doi:10.1038/clpt.2011.93.

[4] A. Pal, M. McCarthy, The genetics of type 2 diabetes and its clinical relevance, Clin. Genet. 83 (2013) 297–306. doi:10.1111/cge.12055.

[5] N.N. Mehta, Large-Scale Association Analysis Provides Insights Into the Genetic Architecture and Pathophysiology of Type 2 Diabetes Mellitus, Circ. Cardiovasc. Genet. 5 (2012) 708–710. doi:10.1161/CIRCGENETICS.112.965350.

[6] H. Zarkoob, S. Lewinsky, P. Almgren, O. Melander, H. Fakhrai-Rad, Utilization of genetic data can improve the prediction of type 2 diabetes incidence in a Swedish cohort, PLoS One. 12 (2017) e0180180. doi:10.1371/journal.pone.0180180.

[7] A. Bonnefond, N. Clément, K. Fawcett, L. Yengo, E. Vaillant, J.-L. Guillaume, A. Dechaume, F. Payne, R. Roussel, S. Czernichow, S. Hercberg, S. Hadjadj, B. Balkau, M. Marre, O. Lantieri, C. Langenberg, N. Bouatia-Naji, G. Meta-Analysis of Glucose and Insulin-Related Traits Consortium (MAGIC), G. Charpentier, M. Vaxillaire, G. Rocheleau, N.J. Wareham, R. Sladek, M.I. McCarthy, C. Dina, I. Barroso, R. Jockers, P. Froguel, Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes., Nat. Genet. 44 (2012) 297–301. doi:10.1038/ng.1053.

[8] B. Fagerberg, D. Kellis, G. Bergström, C.J. Behre, Adiponectin in relation to insulin sensitivity and insulin secretion in the development of type 2 diabetes: a prospective study in 64-year-old women., J. Intern. Med. 269 (2011) 636–43. doi:10.1111/j.1365-2796.2010.02336.x.

[9] O. Savolainen, B. Fagerberg, M. Vendelbo Lind, A.-S. Sandberg, A.B. Ross, G. Bergström, J. Gimble, Biomarkers for predicting type 2 diabetes development—Can metabolomics improve on existing biomarkers?, PLoS One. 12 (2017) e0177738. doi:10.1371/journal.pone.0177738.

[10] R. Wang-Sattler, Z. Yu, C. Herder, A.C. Messias, A. Floegel, Y. He, K. Heim, M. Campillos, C. Holzapfel, B. Thorand, H. Grallert, T. Xu, E. Bader, C. Huth, K. Mittelstrass, A. Döring, C. Meisinger, C. Gieger, C. Prehn, W. Roemisch-Margl, M. Carstensen, L. Xie, H. Yamanaka-Okumura, G. Xing, U. Ceglarek, J. Thiery, G. Giani, H. Lickert, X. Lin, Y. Li, H. Boeing, H.-G. Joost, M.H. de Angelis, W. Rathmann, K. Suhre, H. Prokisch, A. Peters, T. Meitinger, M. Roden, H.-E. Wichmann, T. Pischon, J. Adamski, T. Illig, Novel biomarkers for pre-diabetes identified by metabolomics., Mol. Syst. Biol. 8 (2012) 615. doi:10.1038/msb.2012.43.

[11] Ing SF, Reps J, Kai J, Garibaldi JM, Qureshi N (2017) Can machine-learning

improve cardiovascular risk prediction using routine clinical data? PLoS ONE 12(4): e0174944. https://doi.org/10.1371/journal.pone.0174944

[12] Mohammad Ihbi (2019) Bilirubin. Medscape [online] https://emedicine.medscape.com/article/2074068-overview

[13] How to Handle Imbalanced Classes in Machine Learning [online] https://elitedatascience.com/imbalanced-classes

[14] World Health Organization, WHO (2020) Diabetes [online] https://www.who.int/news-room/fact-sheets/detail/diabetes