Prediction Models for Type 2 Diabetes Mellitus and Prediabetes Using Subgroup Discovery, Random Forest and Decision Tree Algorithms

Master Thesis MSc Cognitive Science & Artificial Intelligence (2018-2019)

Luisa Kapucianova

Master Thesis CSAI

Thesis Committee:
Dr. Martin Atzmüller
Dr. Emmanuel Keuleers

Word Count: 9953

Academic Year 2018/2019

Tilburg University

School of Humanities and Digital Sciences

Tilburg, the Netherlands

Monday, 10 June 2019

Table of Contents

Nomenclature

ANN: Artificial Neural Network

ATC: Anatomical Therapeutic Chemical Classification System

AUC: Area Under Curve

BSD: Bitset Based Subgroup Discovery

CHD: Coronary Heart Disease

DT: Decision Tree

DV: Dependent variable

E10: Type 2 Diabetes

E10-E14: Diabetes Mellitus

E11: Type 1 Diabetes

FP: False Positive

IFG: Impaired Fasting Glycemia

IGT: Impaired Glucose Tolerance

IHD: Ischaemic Heart Disease

IV: Independent variable

R73: Prediabetes

RF: Random Forest

SD: Subgroup Discovery

SN: Sensitivity

SP: Specificity

T2D: Type 2 Diabetes

TN: True Negative

TP: True Positive

Abstract

The current study has been carried out in cooperation with Dôvera healthcare insurance company, to aid in developing a model for predicting high-risk individuals affected by type 2 diabetes and prediabetes, and in detecting the indicators of the diseases. The application of machine learning algorithms has been widely used in epidemiological studies for predicting the incidence of diabetes. The current study introduces an original approach to the landscape of state-of-the-art healthcare research by conducting an exploratory data analysis followed by a prediction analysis. The research uses subgroup discovery, decision tree, and random forest algorithms. The exploratory data analysis uses the subgroup discovery algorithm aimed at exploring patient groups at high risk for the diseases. Moreover, the method proposes a way for a unique feature selection method, applied during the predictive analysis. For predictions, random forest, and decision tree baseline were compared and evaluated based on sensitivity, specificity, area under the ROC curve, accuracy, and F1-score. The study demonstrates the importance of using evaluation metrics other than accuracy and F1-score. For instance, as skewed classes cause a bias towards the minority class, the results might often be misleading, and thus, the emphasis has been put on sensitivity and area under the ROC curve. Due to a severe class imbalance, oversampling methods SMOTE, and down-sampling were applied. Main findings of this research were related to the predictive power of the approach. Generally, the random forest model outperformed the decision tree baseline in all cases. The subgroup discovery feature selection has been successful in improving the models' sensitivity, while in almost all cases, it decreased the specificity. However, the application of feature selection did not outperform the models of random forest trained on down-sampled sets that used all predictor variables in the predictions of both diseases. The final random forest model predicting diabetes achieved an AUC of 68.34%, a sensitivity of 73.11%, a specificity of 63.58%, an F1-score of 71%, and an accuracy of 64.80%. The final random forest model predicting prediabetes achieved an AUC of 56.10%, a sensitivity of 61.40%, a specificity of 50.80%, an F1-score of 57%, and an accuracy of 54.59%.

Keywords: Type 2 Diabetes, Prediabetes, Random Forest, Decision Tree, Subgroup Discovery

Introduction

The present research attempts to build a model for predicting type 2 diabetes and prediabetes and determine risk factors of both diagnoses. Diabetes is one of the most prevalent chronical diseases in the 21st century, having a detrimental impact on individuals and societies with expected global rise over the next few decades (Barber, Davies, Khunti & Gray, 2014; DÔVERA Health Insurance Company, 2017a). There is a growing body of literature that recognizes different data mining algorithms in diabetic research, to improve clinical predictions and find patterns, previously unknown to medics. The results are varying across the studies, depending on social, environmental, and individual influences, as well as data types available to researchers (Bhopal, 2002). Dôvera is a healthcare insurance company based in Slovakia that launched screening for Type 2 Diabetes (henceforth T2D) and prediabetes. With the use of patients' medical data, their effort lies within developing a useful tool in detecting high-risk individuals, when the diseases are yet unknown to them.

Type 2 Diabetes and Prediabetes

Chronic illnesses pose significant challenges for the health care system, both in the socioeconomic and clinical sphere (Stock et al., 2019). The World Health Organization estimates that diabetes was the seventh leading cause of death in 2016 and is currently the most common disease in the 21st century (WHO, 2018). According to data from the National Health Information Centre, the prevalence of diabetes in Slovakia is about 8.6%, from which around 27% of individuals are not aware of the illness (National Health Information Center [NCZI], 2018). The medics recognize three types of diabetes: Type 1, Type 2, and Gestational diabetes (DÔVERA Health Insurance Company, 2017b).

T2D, comprises the majority of people affected by the disease (91% in Slovakia) (NCZI, 2018), and will be researched in the course of this study. Physical inactivity and excess body weight cause ineffective use of insulin. This is one of the most common triggers for T2D. The condition is linked to severe malfunctions of the bodily system including heart disease, blindness, kidney disease, amputations and even shorter life expectancy (Barber et al., 2014). The estimated burden on the health-care system was around 9% out of the total healthcare expenses in Europe in 2015 (International Diabetes Federation, 2015).

Moreover, we will focus on the transition state between normality and diabetes called impaired glucose tolerance (IGT) and impaired fasting glycemia (IFG), also known as a prediabetic stage. Patients with prediabetes are at a high risk of developing T2D, although preventive measures could be taken. As the evidence suggests, in around 20% of the cases, TD2 is preventable, and thus targeting those individuals becomes essential (WHO, 2016), (Ramazenkhani et al., 2014).

In detecting affected individuals, Dôvera launched screening for their patients in general

practice as a part of their initiative “Dôvera Helps Diabetics”. The screening consists of three tests patients can take: random glycemia, fasting glucose and oral glucose tolerance test (OGTT), Figure 1 presents detailed information (Szalay, Jankó, Mužík, Melo, & Benková, 2017). Currently, the patients are targeted randomly, leaving a significant number of affected patients unidentified (B. Benková, M. Poliak, personal communication, October 06, 2018). Therefore, they focus their efforts on developing a strategy, essential for early illness detection and treatment, with the help of the current research (DÔVERA Health Insurance Company, 2018).

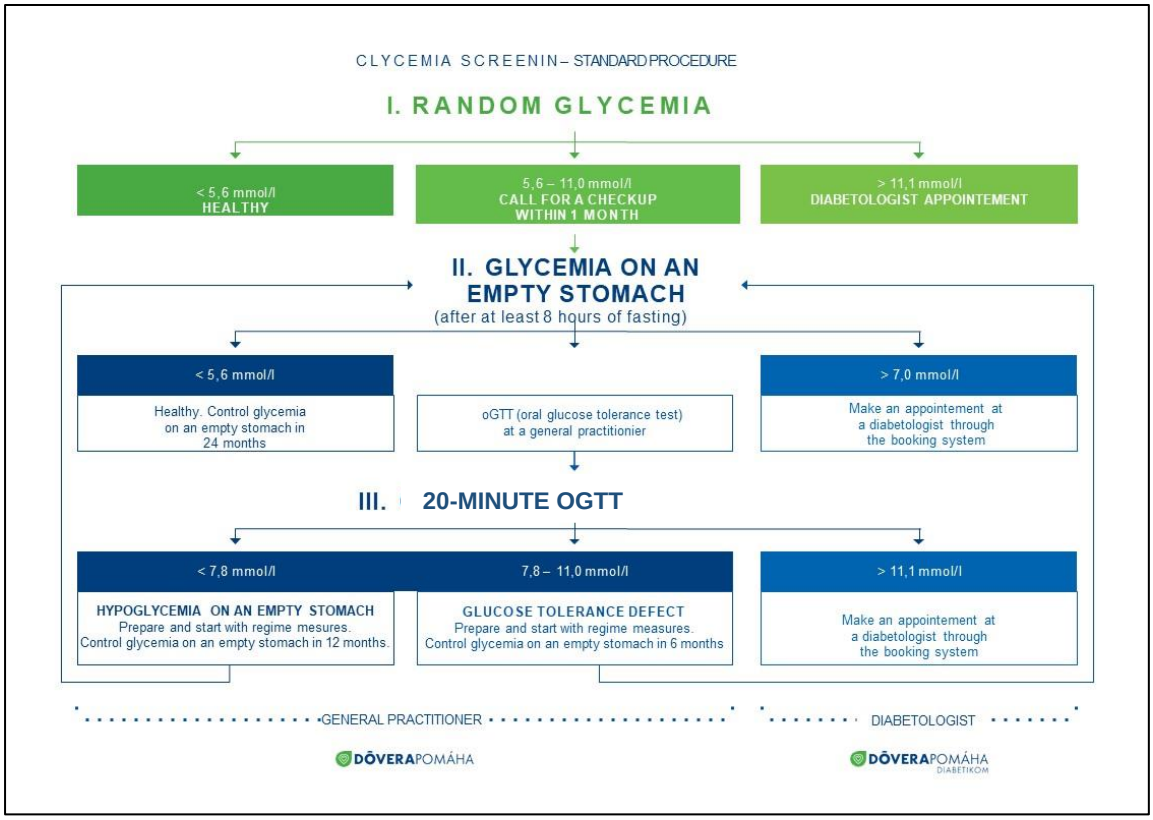

Figure 1. The standard procedure for glycemia screening. The program developed by Dôvera Insurance Company with its initiative "Dôvera Helps Diabetics".

Problem Formulation and Research Questions

In the quest of identifying risk factors of diabetes and prediabetes, we propose a model in which we apply subgroup discovery (henceforth SD) algorithm to explore the relationships in the data, and predictive tree-based algorithms, to establish a model with the best predictive power. Commonly, when building models, researchers choose exploratory variables included in a model in advance. For example, subgroup analyses are aimed at measuring specific treatments and outcomes, where a specific hypothesis is developed (Rui Wang, Lagakos, Ware, Hunter, & Drazen, 2007). In other research settings, such as ours, the exploratory variables are not predetermined, and identifying them becomes a part of an analysis. Therefore, we conducted an exploratory data analysis (henceforth EDA), using the SD algorithm. In comparison to other algorithms such as

frequent pattern mining or rule association, SD allows for a flexible definition of an applied quality function or interestingness measure, which determines the quality of the induced rules and is suitable for the medical domain setting. Furthermore, the method scrutinizes all data by searching for relations between independent variable and dependent variable (henceforth, IV and DV, respectively) and can lead to novice knowledge discovery (Atzmüller, 2015). For example, SD has been implemented onto the discovery for coronary heart disease risk detection and has been successful in identifying rules suitable for representation of the affected individuals (Gambergera, Lavrac, & Krstačić, 2003). Therefore, we formulated the research questions as follow:

RQ1: *Which features contribute most to detecting high-risk individuals affected by T2D and prediabetes?*

RQ2: *To what extent can the Subgroup Discovery algorithm contribute to successfully detecting high-risk individuals affected by T2D and prediabetes?*

In light of the prediction analysis, extensive research covers the application of various models for diabetes incidence and analyses of which courses of action prove useful. Data from several studies suggest that tree-based algorithms achieved the best results in epidemiological research. In a study predicting diabetes and prediabetes, researchers compared logistic regression, decision tree, and Artificial Neural Network (henceforth LR, DT, and ANN, respectively), to determine the best classification performance. The models were evaluated based on selected criteria (accuracy, sensitivity, and specificity). Following the criteria, the decision tree algorithm achieved the highest results with an accuracy of 77.87%, sensitivity of 80.68%, and specificity of 75.13% (Meng, Huang, Rao, Zhang, & Liu, 2013). However, the literature suggests that when comparing algorithms of the random forest (henceforth RF) to those of a decision tree, they appear to achieve better classification predictive power (James, Witten, Hastie & Ibshirani, 2013). In a large longitudinal study, Esmaily et al. (2018) investigated the incidence of diabetes, by comparing DT and RF models. The RF model surpassed the results of the DT by the difference in accuracy, sensitivity, specificity and area under the ROC curves (henceforth AUC) of 6.2%, 6.8%, 3.1%, and 8.7% respectively (Esmaily, Tayefi, Doosti, Ghayour-Mobarhan, & Amirabadizadeh, 2018). Consequently, we implemented a random forest algorithm for classification and a decision tree algorithm as a baseline comparison for our study. Hence, the final research question sounds as follow:

RQ3: *To what extent can Logistic Regression, Decision Tree, and Random Forest algorithms successfully detect high-risk individuals affected by T2D and prediabetes?*

Related Work

Exploratory Analysis

Exploratory data analysis is an approach that analyzes datasets beyond formal modelling or hypothesis testing, inevitable in exploring raw data and findings patterns. SD is one of the techniques used for identifying the most interesting relationships inside the data statistically, given specific interestingness criteria, formalized by a quality function (Atzmüller, 2015). The method allows researchers to identify relationships between several IVs and DV, and thus contributes to the understanding of the groups and their forms. The subgroup discovery setting is mostly based on defining the target variable and the search space, the quality function that corresponds to statistical or other user-defined quality criteria, and choosing the search strategy such as heuristic or beam search algorithms (Atzmueller & Puppe, 2006).

Up to this date, SD has not been applied onto the diabetic research. However, its algorithms were successfully used for other chronical diseases such as coronary heart disease (CHD) (Gamberer, Lavrac & Krstac, 2003). During this research, the data has been collected from patients screening at the Institute for Cardiovascular Prevention and Rehabilitation. The database consists of a variety of clinical data such as amnestic parameters, laboratory test results, or echocardiography results. The induction process has been conducted by implementing an if-then rule form, *Class*←*Cond*, where *Cond* is a conjunction of conditions (conjunction of features describing the illness). This rule was induced by using the generality parameter, which constructs rules with high specificity. For constructing the rules describing a subgroup, a combination of machine learning based induction and statistical analysis of the detected subgroups has been developed, guided by a domain expert. For individual rule construction heuristic Algorithm SD with pseudocode and a covering algorithm involving example weighting for the rule set construction Algorithm DMS with pseudocode were used. Each of the five discovered subgroups A1, A2, B1, B2, and C1 were evaluated based on the sensitivity (TP) and specificity (FP) and were successful in discovering CHD patients. Table 1 presents the performance of each rule, evaluated by specificity and sensitivity (Gambergera, Lavrac, & Krstačić, 2003).

Subgroup	Training set (Pos = 111, Neg = 127)		Test set (Pos = 50, Neg = 20)		Employee set (Pos = 30, Neg = 170)	
	TPr (%)	FPr (%)	TPr (%)	FPr (%)	TPr (%)	FPr (%)
A1	47	27	85	78	95	28
A2	48	7	41	27	60	6
B1	29	9	36	20	37	5
B2	32	13	42	15	83	48
C1	23	5	82	40	27	8
Expert	–	–	–	–	97	18

Table 1. Results of the five discovered subgroups applied onto the test set and independent employee set. The final line represents results for domain expert classification for employee data.

Reprinted from "Active subgroup mining: a case study in coronary heart disease risk group detection," by D. Gambergera, Nada Lavrac, G. Krstačić,2003, *Artificial Intelligence in Medicine*, p. 48. Copyright 2003 by Elsevier.

Similarly to the previous research, SD has been applied in the discovery process of relevant coexisting risk factors of brain ischemia (Gamberger, Lavrač, Krstačić, & Krstačić, 2007). The dataset consists of patients records treated at the department of neurology at the University Hospital Centre, including amnestic data, physical examination data, laboratory test data, and ECG data and information of previous therapies. The process has been consulted numerously by experts, mainly for rule selection and interpretation. The subgroup mining was used in the same way as in the aforementioned research, using if-then rules of the form *Class*←*Cond*, by application of SD and DMS algorithms, inducing 15 rules. Table 2 represents the results of the evaluation. The results indicate that the subgroups are suitable representations of characteristics of groups of individuals suffering from brain ischemia, providing a deeper understanding of the disease and its forms (Gamberger, Lavrač, Krstačić, & Krstačić, 2007).

Ref.	Rule	Sens.	Spec.	Overlap
generalization parameter $g = 5$				
g5a	$(fibr > 4.55)$ and $(str = no)$	25%	100%	–
g5b	$(fibr > 4.45)$ and $(age > 64.00)$	41%	100%	94%
g5c	$(af = yes)$ and $(ahyp = yes)$	28%	95%	36%
generalization parameter $g = 10$				
g10a	$(fibr > 4.45)$ and $(age > 64.00)$	41%	100%	–
g10b	$(af = yes)$ and $(ahyp = yes)$	28%	95%	34%
g10c	$(str = no)$ and $(alcoh = yes)$	28%	95%	67%
generalization parameter $g = 20$				
g20a	$(fibr > 4.55)$	46%	97%	–
g20b	$(ahyp = yes)$ and $(fibr > 3.35)$	65%	73%	71%
g20c	$(sys > 153.00)$ and $(age > 57.00)$ and $(asp = no)$	45%	88%	80%
generalization parameter $g = 50$				
g50a	$(ahyp = yes)$	74%	54%	–
g50b	$(fibr > 3.35)$ and $(age > 58.00)$	79%	63%	76%
g50c	$(age > 52.00)$ and $(asp = no)$	64%	63%	96%
generalization parameter $g = 100$				
g100a	$(age > 52.00)$	96%	20%	–
g100b	$(dya > 75.00)$	98%	8%	98%
g100c	$(ahyp = yes)$	74%	54%	100%

Table 2. Induced rules per generalization parameter, measured based on specificity and sensitivity
Reprinted from "Clinical data analysis based on iterative subgroup discovery: experiments in brain ischemia data analysis" by D. Gambergera, Nada Lavrac, A. Krstačić, G. Krstačić, 2007, *Applied Intelligence*, p. 212. Copyright Springer Science+Business Media, LLC 2007.

Predictive Analysis

The use of prediction models became a conventional method in medical research for estimating a risk that a specific disease or condition is present (Collins, Johannes B. Reitsma, Altman,

& Moons, 2015) and many have been successful in predicting diabetes mellitus. The choice of classification algorithms varies across studies, with most frequently used and achieving the highest predictive power being tree-based algorithms (Esmaily, Tayefi, Doosti, Ghayour-Mobarhan & Amirabadizadeh, 2018; Ment et al., 2013; Nai-aruna & Rungruttikarn, 2015, Ramezankhani et al., 2016a). The use of multiple algorithms has been studied by Meng et al. (2013), who implemented three predictive models of DT, LR, and ANN. The models used 12 predictive variables based on generally known risk factors and one output variable. The predictive variables were selected based on the Chi-square feature selection. For example, some of the predictive variables are body mass index, alcohol consumption, physical activity, or family history of diabetes. The researchers used accuracy, specificity, and sensitivity for evaluating the results. The results established that the DT model produced the best-achieved classification accuracy of 77.87%, with a sensitivity of 80.68% and specificity of 75.13% (Meng et al., 2013).

The decision tree algorithm has also proven to be useful in identifying low-risk individuals for T2D. The model was applied to 6647 individuals without diabetes during a 12 years follow-up study, by analyzing diverse patient data ranging from clinical to laboratory data. In total, 60 input variables and one output variable were used in the research. The identified risk factors are fasting plasma glucose, body mass index, triglycerides, mean arterial blood pressure, family history of diabetes, educational level, and job status. Model's attributed were measured based on information gain, Gini index, and gain ratio, and the evaluation criteria were accuracy, specificity, sensitivity, precision, and F1-score. The overall classification accuracy was 90.5%, with 31.1% sensitivity and 97.9% specificity (Ramezankhani et al., 2014). Notably, the lower sensitivity rate, demonstrates a result of class imbalance, causing classifiers to produce high accuracy over that of the majority class (Ramezankhani et al., 2016b).

Various solutions can be applied to neutralize the influence of unbalanced classes, such as synthetic oversampling method (henceforth SMOTE). In medical research, Ramezanhkami et al. (2015), evaluated the impact of the SMOTE algorithm on the performance of the probabilistic neural network (PNN), Naïve Bayes (NB) and DT, for predicting diabetes. The data was collected in a cohort of the Tehran Lipid and Glucose study, from non-diabetic patients. The models were built with 21 common risk factors. Both original and oversampled sets were used to establish the models' power. The results indicate that a wholly balanced sample increased the accuracy of PNN, DT, and NB by 64%, 51% ad 5%, respectively. Finally, the model of NB achieved the best results both before and after oversampling. However, the research concludes that DT is an optimal classifier in predicting diabetes, especially when the class is imbalanced (Ramezankhani et al., 2016b).

	Classifiers	Sensitivity	Specificity	F-Measure	Precision	Accuracy	Youden's index
Original training dataset	PNN	0.027	0.999	0.053	0.857	0.893	0.026
	DT	0.215	0.992	0.336	0.77	0.907	0.207
	NB	0.721	0.825	0.459	0.337	0.814	0.546
Balanced training dataset with SMOTE (700%)	PNN	0.667	0.80	0.405	0.291	0.785	0.467
	DT	0.726	0.802	0.436	0.312	0.794	0.528
	NB	0.776	0.784	0.440	0.307	0.783	0.56

Table 3. Results of original training datasets and over-sampled training dataset, using SMOTE.
Reprinted from "The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of Type 2 Diabetes" by A. Ramezankhani, O. Pournik, J. Shahrabi, F. Azizi, F. Hadaaegh, D. Khalili., 2014, Medical Decision Making, 36, p.142. Copyright SAGE Journals 2016.
Note DT= decision tree, PNN= probabilistic neural network, NB= Naïve Bayes, SMOTE= synthetic oversampling technique

Using the predictive power of decision trees has been a suitable method for diabetes prediction. However, evidence suggests, that robustness of the model can be improved when building random forests models. In broad terms, RF models build several random decision trees, resembling "forests", which are later combined to yield a single consensus. As decision trees grow very deep and can overfit their training sets, introducing low bias but a very high variance, random forest average multiple decision trees trained on different parts of the same training set, aiming to reduce the variance. This might come at the expense of a slight increase in the bias but it generally enhances the final model performance (Esmaily et al., 2018). Various studies have assessed the predictive power of RF over the DT. Notable examples include studies where DT and RF were directly compared. In a study comparing DT and RF, the RF model achieved higher results when determining risk factors associated with T2D. Specifically, the model achieved accuracy of 71.1%, sensitivity of 71.3%, specificity of 69.9%, and AUC of 77.3% (Esmaily, Tayefi, Doosti, Ghayour-Mobarhan, & Amirabadizadeh, 2018).

The comparison has also been conducted in an empirical research study by Nai-arun and Moungmai (2015). In their study, they firstly analyzed DT, ANN, LR, Naïve Bayes, and RF. Next, the models have been tuned based on boosting and bagging. The clinical data were collected from Primary Care Units in Sawanprachrak Regional Hospital in 2013-2014. After selecting a subset of 10 input variables, based on a consultation with medical specialists, the models have been built and compared. The evaluation criteria, such as AUC, sensitivity, and specificity, were used in the analysis, similar to the current evaluation design. Finally, the random forest produced the best results, with an accuracy of 85.56% and AUC of 91.20%. Figure 2 shows that the use of bagging and boosting improved the performance of AUC, in all models used. The RF model was eventually chosen as a diabetes risk assessment tool (Nai-aruna & Rungruttikarn, 2015).

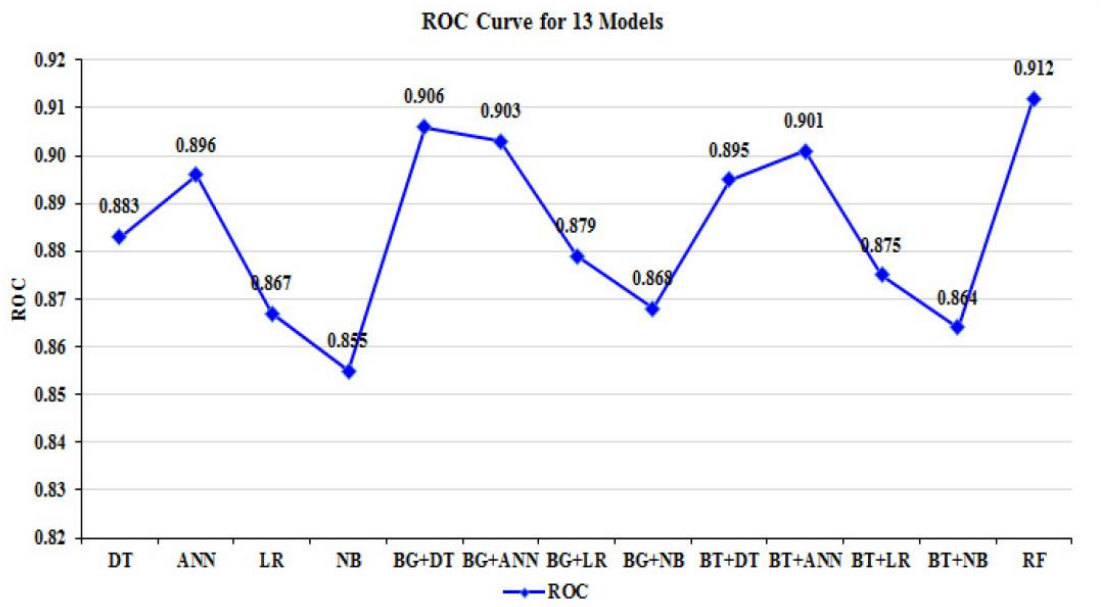

Figure 2. Comparison of ROC values of all models,
Reprinted from "Comparison of classifiers for the risk of diabetes prediction" by N. Nai-Aruna, M. Rungruttikarnvrac, 2015, *Procedia Computer Science*, 69, p.140. Copyright Elsevier 2015.

Contribution of the Current Research

In the current research, we propose a methodological model comprising of two main parts: EDA and predictive analysis, with methods stemming from the established literature review. In the first part, we applied SD algorithms to answer the RQ1 and RQ2. Despite the remote findings in the literature, concerning the diabetes, SD is a promising tool for exploring the target variables and sub-populations inside the data. Moreover, as we did not predetermine any set of predictive variables, the inducted subgroups were used instead of a feature selection method, for building prediction models. For the predictive analysis, we implemented algorithms of random forest and decision tree. As the literature review established, the RF model is a promising and robust algorithm, and thus, the DT served as a baseline comparison in the analysis.

Essentially, an establishment of a relevant model for detecting T2D and prediabetes opens opportunities for developing risk assessment tools, which could aid in designing preventive measures against the diseases. For instance, the detection of risk factors could aid health care professionals to locate affected individuals at an early stage, and prevent the diseases from intensifying. Moreover, a risk assessment tool can be also beneficial for Dôvera, to make their procedures more effective and precise.

Methods

Procedure

The procedure of this research consisted of extensive pre-processing of two datasets, exploratory analysis conducted with the SD algorithm and predictive analysis for which we built RF and DT algorithms. In the analysis, the DT provided a baseline comparison. During the exploratory

analysis, we inducted sub-populations within diabetic and prediabetic patients. Further in the research, we extracted features explored during the SD and applied them in the predictive models, instead of other feature selection techniques. Therefore, we constructed the models using two sets of predictors: (1) all 103 predictor variables, and (2) selected variables from induced subgroups. All models were trained on training sets, optimized for precision, and the hyperparameters determined and tuned by grid search with a 5-fold validation (Razavian et al., 2015). Finally, the models were evaluated on the test set, according to the chosen metrics. Figure 3 depicts a representation of the models' procedure.

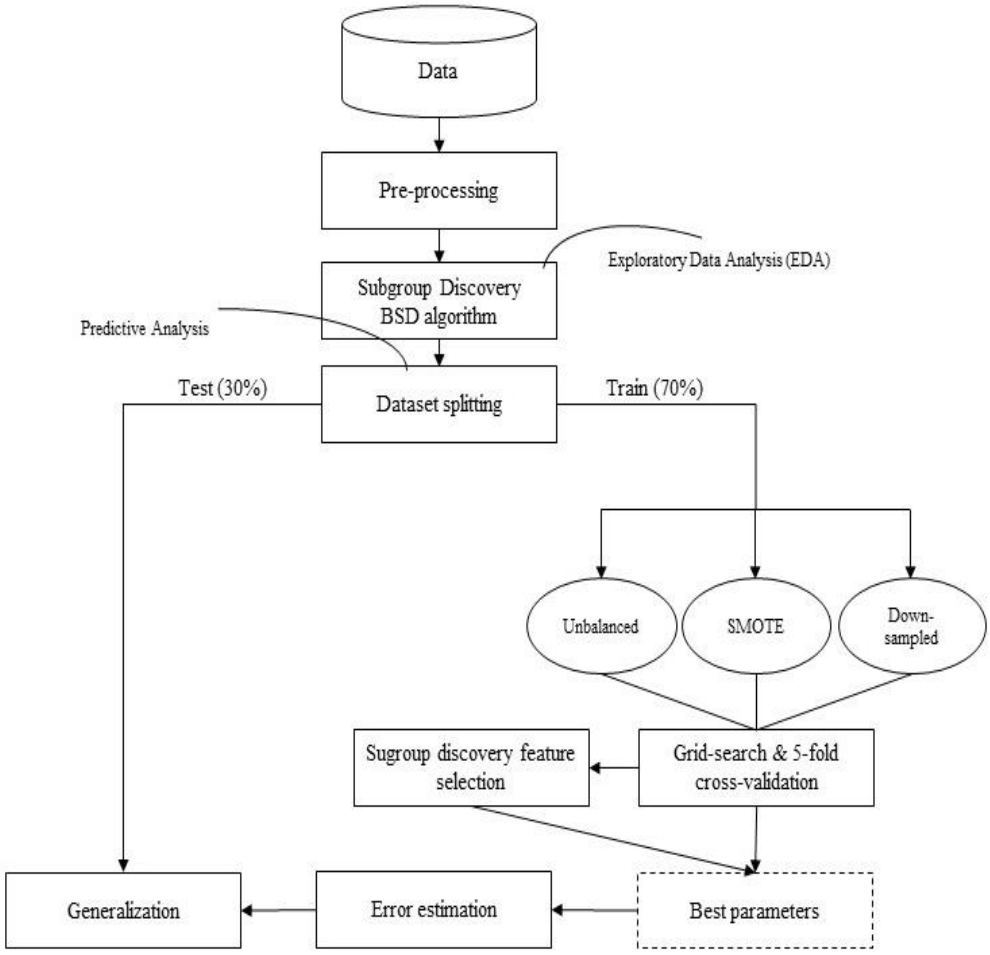

Figure 3. Representation of the model's design and procedures

Dôvera Datasets

The exploratory and predictive analyses were performed on longitudinal data from Dôvera health insurance company. Two datasets IKAP and VSZP comprise data recorded from September 2015 to September 2018, of patients insured under Dôvera, who were screened for T2D and prediabetes in between November 2017 until June 2018 at the general practice in Slovakia. The study scrutinized data of 5482 patients aged between 6 and 94 (*M*= 55.38, *Mdn*=57, *SD*=15.03). The former dataset contains demographic information pertaining subject id, age, gender, location, economic

activity, insurance type, information on the screening frequency and the screening date. The latter dataset pertains utilization data such as time of the check-up, screening date, ICD-10 codes (the International Statistical Classification of Diseases and Related Health Problems), subgroup code, ATC (Anatomical Therapeutic Chemical Classification System), specific care (type of visited medical specialist), healthcare costs and points derived from the level of difficulty of medical examinations.

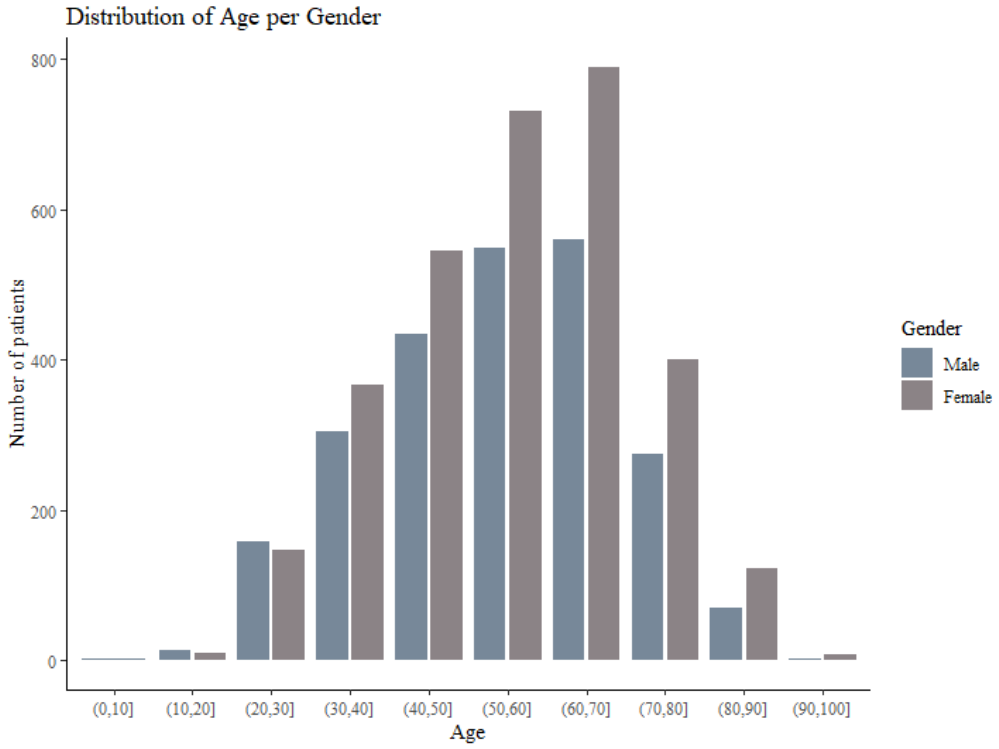

Figure 4. Distribution of age per gender of patients in Dôvera

Data Pre-processing

The first step in the process was to compose a new dataset with predictive and response features selected and re-coded from the two datasets. We transformed the data by conducting binarization and discretization of variables. Predictor variables in the analysis were created by combining IKAP (age, gender, location, and economic activity) and VSZP datasets (ICD10 codes, ATC codes), and were selected after consultation with Dôvera specialists. In the first step, the ATC codes and ICD10 codes were re-coded from two multiclass variables into 318 binary variables each corresponding to a specific code, with classes "1= presence of an event" and "0= no event". The ICD10 codes variables were grouped according to the ICD10 codes index categories (e.g., all codes between A30 and A49, resulting in A30.A49 implying "other bacterial diseases"). The ATC codes were grouped based on the first three signs (e.g., A02BC01 and A02BC02, resulting in A02). Furthermore, we assigned the variables to patients uniquely and considered them present (i.e., 1= presence of an event) only under the condition that they were recorded during two different months (e.g., a patient diagnosed with G00-G99 at two different check-ups). Due to this, some variables had 0 occurrences, and thus, were removed from the analysis (e.g., A75.A79 implying rickettsioses). To

reduce the models' dimensionality, we filtered out all variables that had an occurrence of 1 below 150, resulting in 103 explaining variables. Additionally, the categorical variables “location”, “sex”, and “economic activity” were encoded into numeric quantities. There were no missing values in the dataset.

The target variables were re-coded based on ICD10 index classification, that defines both diabetes and prediabetes (i.e., E10-E14 for T2D and R73 for prediabetes). Moreover, in case a patient has been diagnosed with prediabetes, which later progressed into diabetes, we assigned diabetes to those patients only.

Training and Test Set

For the analysis, we divided the dataset into train and test set, with a ratio of 30:70. The train set (70%) served for fitting the models and parameter tuning with grid search and 5-fold cross-validation. The results were tested on the unseen data in the test set (30%) (Esmaily, Tayefi, Doosti, Ghayour-Mobarhan, & Amirabadizadeh, 2018). Figure 5 represents the class distribution of both target variables in the dataset.

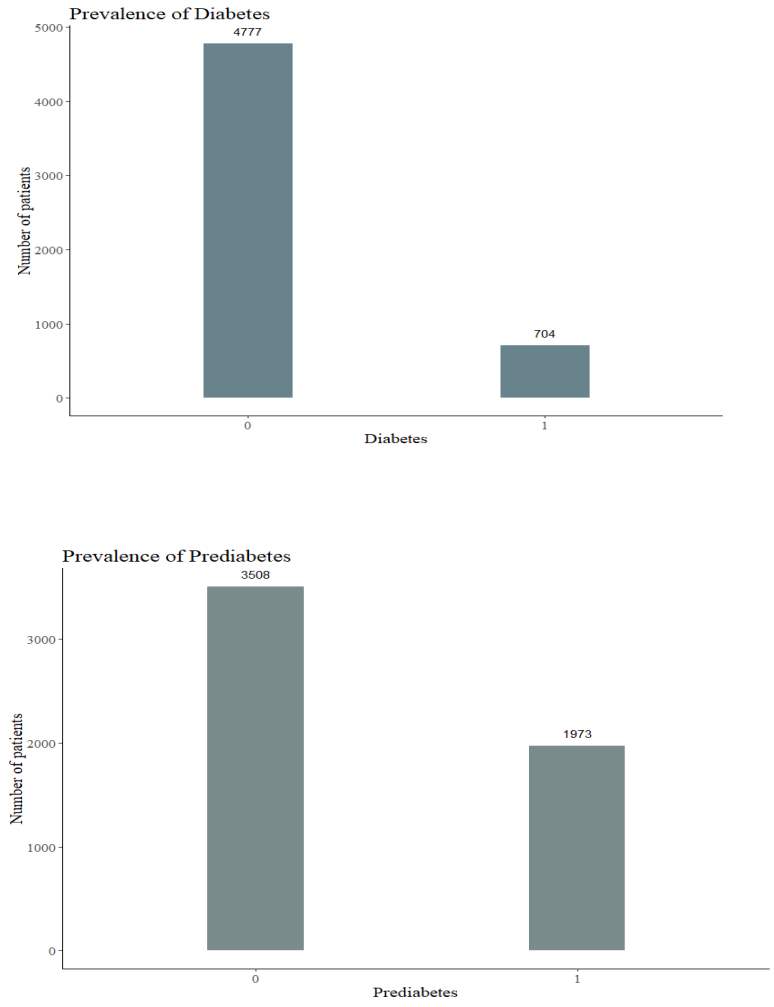

Figure 5. Class distribution of Diabetes and Prediabetes

The unbalanced distribution of classes, known as class imbalance, has a significant influence on the predictive power of the algorithms (Menardi & Torelli, 2012). In the current dataset, both targets have a majority class "0 = no event". The imbalanced classes caused data to overfit and disabled classifiers to correctly detect the minority class "1= presence of an event" (Ramezankhani et al., 2016b). In medical settings such as ours, the minority class is usually the critical class, which is preferred to be predicted with higher accuracy (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). The characteristics of the data led us to apply resampling methods such as down-sampling and SMOTE, in order to improve the predictive power of algorithms.

We used down-sampling to eliminate data from the majority to match the minority class, and SMOTE to oversample the minority class at 100%, by creating "synthetic" examples. Both resampling techniques were used based on the prior research (Merandi & Torelli, 2012; Ramezankhani et al., 2016b). Rather than copying the instances, such as in down-sampling, the SMOTE algorithm resamples the data by taking each minority class and introducing synthetic examples along with the line segments, which joins the k minority class nearest neighbors. Afterward, it discomposes one attribute of an instance at a time by a random number within the range of k neighboring instances (Chawla et al., 2002). Table 4 represents class distribution in all of the training sample sets.

	Diabetes	Prediabetes
Original set	1= 492, 0= 3344	1=1385, 0=2451
SMOTE set at 100%	1= 3344, 0=3344	1=2451, 0=2451
Down-sample set	1= 492, 0=492	1=1385, 0=1385

Table 4. Class distribution of Diabetes and Prediabetes response variables on original oversampled and down-sampled training sets.

Subgroup Discovery Algorithm BSD

We conducted the subgroup discovery aiming to discover k best subgroups in between predictor variables by considering the whole subspaces of the search space. The subgroup description is defined by a combination of selectors or expressions: $sd = \{e1, e2, \ldots, en\}$, which are selections on domains of attributes $ai \in \Omega A$, $Vi \subseteq dom(ai)$. The Ωsd is defined as a set of all possible subgroup descriptions.

In the current research, we used bitset based subgroup discovery algorithm (henceforth BSD), which is a vertical mining algorithm, that combines vertical bit-set based representation of the data in the search-space, with advanced pruning strategies and an efficient relevance check (Atzmüller, 2015), (Atzmueller and Lemmerich 2009). The search consists of two main phases. In the first phase (line1-17 in Figure 6**)** all relevant selectors sel_{rel}, consisting of sel_{cond}, the current conditioned

selectors and s_{curr}, the new selector, are being considered. The bit-set is computed, when all positive labeled classes tp, fulfill the new description. The quality function is always given by the subset which holds the positive cases, and the amount is sufficient to compute an optimistic estimate for the current combination of selectors.

The estimate is used in two ways. Firstly, only if the optimistic estimate reaches a certain point, the bit-set of negatives for the current combination of selectors is computed. After that, positives and negatives are counted and the quality of the subgroup is computed. By doing this, the negatives are considered only for promising selector combinations, essentially leading to significantly shorter runtime. Secondly, only if the optimistic estimate indicates an improvement in the current subgroup, with a new selector that has a sufficient quality to be added into the final results, the selector is added to the list for the next level of search. In a relevance check, subgroup relevance is tested, and only the most relevant subgroups are stored. In phase two, the list of all relevant selectors is sorted according to their optimistic estimate (line 18 Figure 6), allowing the algorithms to evaluate more promising paths first. Finally, every relevant selector is added to the list of conditioned selectors, resulting in a recursive search using the respective bit-set and the conditioned selectors (Lemmerich, Rohlfs, & Atzmueller, 2010).

We applied the algorithm in subgroup search of both target variables. Moreover, except implementing the SD for exploring the targets, we constructed a new feature selection method that uses selectors found during the SD task. The feature selection was implemented into models of RF and DT trained on the three different sample sizes, each run on grid search and 5-fold cross-validation. Consequently, all results were cross-validated on the test set.

Algorithm 1 function bsd

Require:
sel_{cond}: List of conditioned selectors,
sel_{rel}: List of relevant selectors,
$c_{condPos}$: Bitset of positive instances for sel_{cond}
$c_{condNeg}$: Bitset of negative instances for sel_{cond}
depth: Current search depth,
res: The result set of the best k found subgroups
Ensure:
res as a set of the best k relevant subgroups

```
1: newSel_rel := new List()
2: for all Selector s_curr in sel_rel do
3:    c_currPos = c_condPos AND(s_curr.bitsetPos)
4:    tp = c_currPos.cardinality()
5:    if optEstimate(tp) > res.getMinQuality() then
6:       c_currNeg = c_condNeg AND(s_curr.bitsetNeg)
7:       n = tp + c_currNeg.cardinality()
8:       newSel_rel.add(s_curr)
9:       s_curr.attach(c_currPos, c_currNeg, optEstimate(tp))
10:      if quality(tp, n) > res.getMinQuality() then
11:         r = checkRel(res, c_currPos, c_currNeg)
12:         if r then
13:            sg = createSubgroup(sel_cond, s_curr)
14:            res.add(sg, c_currPos, c_currNeg)
15:            res.checkRelevancies(sg)
16:            if res.size > k then
17:               result.removeLowestQualitySubgroup()
18: sort (newSel_rel)
19: if depth < MAXDEPTH then
20:    for all Selector s: newSel_rel do
21:       if s.optEstimate > res.getMinQuality() then
22:          newSel_rel(s)
23:          sel_cond.add(s)
24:          c_new = getCurrentBitSetFor(s)
25:          bsd (sel_cond + s, newSel_rel, s.getPositives(),
             s.getNegatives(), depth + 1, res)
```

Figure 6. Representation of BSD algorithm by F. Lemmerich, M. Rohlfs, M. Atzmueller, 2010, *Fast Discovery of Relevant Subgroup Patterns. In Proc. 23rd International FLAIRS Conference*, p. 430. Copyright Mary Ann Liebert 2015., AAAI Press.

Decision Tree

A decision tree algorithm was built as a baseline and benchmark to compare the performance of the RF model, on all training samples. Classification decision trees consist of "if-then-else" rules, generally used in predicting qualitative tasks. By recursive partitioning, the algorithm breaks down a dataset into smaller subsets with a decrease in depth of a tree. Classification trees predict that each observation belongs to the most frequently occurring class of training observations (James, Witten, Hastie, & Ibshirani, 2013). Initially, all features are assigned to the root, from which the most important features are selected. The root-node represents an entire population and is further divided into two or more homogenous sets, by taking one feature at a time and test a binary condition. For the binary splits, we either used Gini index or entropy, depending on the selection of best parameters, of grid search and 5-fold cross validation (James, Witten, Hastie, & Ibshirani, 2013). Grid search is an exhaustive brute-force search that uses a specified list of values for different hyperparameters and

subsequently evaluates the model performances for each combination of hyperparameters (Raschka, 2016).

The Gini index is defined by:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

The formula represents a measure of total variance across K classes. Here the $\hat{p}mk$ represents the proportion of training observations in the mth region that are from the kth class. Gini takes values that are close 0 or 1 and thus, is referred to as node purity, which indicates that a node contains mainly observations from a single class.

Entropy is defined by:

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

Here entropy will take a value near zero or one. Thus, similarly to Gini index, the entropy will also take a small value if the mth node is pure (Bruce & Bruce, 2017), (James, Witten, Hastie, & Ibshirani, 2013).

Random Forest

For the binary classification task, we build a RF model as the algorithm is believed to substantially improve the predictive performance of trees (James, Witten, Hastie, & Ibshirani, 2013). The two main tasks of RF are bagging and random subspace method. In the random subspace method, the algorithm generates many classification trees by selecting subsets of given datasets where subsets of predictor variables are selected randomly to create any tree. A newly generated bootstrapped datasets are later combined to yield a single consensus prediction. For each bootstrap, the algorithm creates a new tree, and when classifying the input data, the data is passed through each tree and produces an output which can be denoted by $Y = \{y1, y2 \ldots . ys\}$, where Y represents the output. The final prediction represents a majority vote on the final set (James, Witten, & Hastie, Ibshirani, 2013). The model's prediction is estimated by out of bag error (OOB), which represents the error rate for the trained models, applied to the data left out of the training set for a particular tree (Bruce & Bruce, 2017).

We tuned the model's parameters by grid search and 5-fold cross-validation. All different sample sizes were fit to the RF models and repeated the process with feature selection, to test for increase in the models' performance. Generally, the feature selection was not necessary as the algorithm already performs an embedded variables selection (Olivera et al., 2017). The RF has a built-in function that estimated the importance of each variable, by measuring the degree of association

between a given variable and the classification result (Ramezankhani et al., 2016a). Two measures of variable importance are mean decrease accuracy and a total decrease in node impurity. The former measures the mean decrease of accuracy in predictions on the out of bag samples when a certain variable is excluded from the model. The latter measures the total decrease in node purity that results from splits over a variable, averaged over all trees, by deviance (James, Witten, Hastie, & Ibshirani, 2013).

Evaluation Methods

To assess the performance of the predictive power of the DT baseline and the RF model, we applied several evaluation methods, as accuracy has been found insufficient evaluation metric, especially in medical data (Chawla et al., 2002). Each model was 5-fold cross-validated and generalized on the test set. In our model we used conventionally used evaluation methods such accuracy, sensitivity, specificity, and AUC, as suggested by prior research (Bhopal, 2002). Those methods were also chosen in the current research.

The classification accuracy measured the proportion of cases which were correctly classified. Moreover, the measure of sensitivity (TP) represents a fraction of positive cases that are correctly classified as positive, and the measure of specificity (TN) represents a fraction of negative cases that are correctly identified as negative. Followed are corresponding formulas, where TP represent true positive, TN true negative, FP false positives, and FN false negatives rate:

1. $Accuracy = \frac{TP+TN}{(TP+FP+TN+FN)}$
2. $Sensitivity = TP/(TP + FN)$
3. $Specificity = TN/(FP + TN$

We also evaluated the models based on AUC, which can only be used on binary target variables. The AUC measures how well a parameter can distinguish between two diagnostic groups and gives a probability that a predicted risk for a participant with an event is higher than for a participant without an event. ROC plots the TP rate on y-axis and FP rate on the x-axis for a particular decision threshold. Therefore, the TP rate represents fraction of cases which are correctly identified as "diabetic" or "prediabetic", while the FP rate represents fraction of classes which were falsely identified as positive classes. In addition, ROC curves are useful for comparing different classifiers, since they take into account all possible thresholds (James, Witten, Hastie, & Ibshirani, 2013)

Finally, the results were evaluated based on F1-score, which is a combination of precision and recall.

1. $Precision = \frac{TP}{\mathrm{TP} + \mathrm{FP}}$
2. $Recall = \frac{TP}{\mathrm{TP} + \mathrm{FN}}$

3. $F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$

Precision determines how many of the predicted positives are actual positives and recall determines the ability of a model to find all actual true positives (diabetic and prediabetic patients). The F1-score represents a harmonic mean of precision and recall (Raschka, 2016)

Programming Software and Packages

For merging, transforming and further pre-processing the dataset, Python version 3.7.2 (Python Software Foundation, 2019) with preprocessing, pandas and numpy packages, and R Studio version 1.1.456 (Anaconda Documentation, 2019) with caret and stats packages. Both EDA and predictive analysis were conducted in Python. For EDA we used rsugroup package. The datasets were split and resampled using train_test_split, SMOTE, and RandomUnderSampler packages. Moreover, the classifiers were built and trained by using RandomForestClassifier, DecisionTreeClassifier and GridSearchCV. The results were evaluated using metrics, roc_curve, auc, cross_val_predict, confusion_matrix, and accuracy_score, and classification report packages (Python Software Foundation, 2019). Finally, the results were visualized by using Python package matplotlib (Python Software Foundation, 2019), and R Studio package ggplot (Anaconda Documentation, 2019).

Results

This section scrutinizes the results of EDA analysis with the SD algorithm and predictive model of the random forest compared to the baseline model of the decision tree. Due to the class imbalance, we resampled the original dataset using SMOTE and down-sampling algorithms. We constructed the models using three different training sets for each target variable, with two sets of predictors: (1) all 103 predictor variables, and (2) selected variables from induced subgroups. In total, the SD algorithm generated ten different subgroup conjunctions for the response variable “Diabetes” and 10 for the response variable “Prediabetes”. Out of those, we used 12 selectors for predicting diabetes and 14 for predicting prediabetes. The model's parameters were optimized on precision by grid search and 5-fold cross-validation, run on all samples. Finally, we evaluated and compared the models' performance.

In general, 704 (12.84%) people had diabetes, 1973 (36%) had prediabetes, and 2804 (51.16%) had neither of the diagnoses. Of the diabetic individuals, 395 (56.11%) were female and 305 (43.32%) male, with the majority (92.47%) between 50 to 80 years old. From the prediabetic patients, 106 (56.06%) were female and 867 (43.94%) male participants, most of them (70.16%) between 40 to 70 years old. Moreover, out of the diabetic people, 140 (19.89%) were economically active, and 564 (80.11%) were economically inactive patients, and of the 1937 prediabetics 692 (35.73%) were

economically active, and 1281 (66.13%) were economically inactive. Finally, out of the studied sample, the highest incidence of both diabetes and prediabetes is in Kosice.

Exploratory Data Analysis

In the exploratory part of the analysis, we conducted subgroup discovery, a broadly applicable descriptive data mining technique, that determines interesting subgroups, regarding the target variable of interest (Atzmüller, 2015). Moreover, in this research, we proposed subgroup discovery as a feature selection technique, extracting unique induced selectors of both target variables and applied them on the predictive models. The first phase consisted of selecting subgroup objects, such as response variables and the search space. For both targets, we used the full search space, using all 103 predictor variables. The next step defined the SD task, with result size, depth of the conjunction of selectors, and simple binomial quality function. These values were set up manually for different searches, in order to achieve different conjunctions, with varying sizes. For simplifying the rule interpretation and improvement of actionability, the quest was aimed at finding rules between 2-4 patterns. Finally, the process was completed at subgroup description with a stored list of selectors, interpreted as a conjunction.

For constructing the rules of the target variable diabetes, a BSD algorithm was used. In total, ten different subgroups were found, consisting of 12 unique selectors, reaching the target share below 50%, signifying a limited amount of available information (Gamberger, Lavrac, & Krstac, 2003). All selected conjunctions form subgroups where the binary selectors have the class 1. In general, we run the code in 4 different settings. The variable that has been predominant in the majority of the conjunctions was “Economic activity= inactive patients”. Firstly, the two rules A1 and A2 both achieved a target share below 50%. Both subgroups consist of 4 conditions. A1 consists of economically inactive patients, diagnosed with E70.E90 (metabolic disorders other than diabetes), G40.G47 (episodic and paroxysmal disorders) and H49.H52 (disorders of ocular muscles, binocular movement, accommodation, and refraction), with target share 35%. A2 represents economically inactive patients diagnosed with F09 (mental and behavioral disorders), G40.G47 (episodic and paroxysmal disorders, and disorders of ocular muscles, binocular movement, accommodation, and refraction), with 40% target share.

B1 and B2 were built based on two conditions: In B1 economically inactive patients diagnosed with M80.M85 (disorders of bone density and structure) were grouped, while in B2 economically inactive patients diagnosed with G40.G47 (episodic and paroxysmal disorders were grouped). Subgroups C1-C4 were established based on two to three pattern conditions. C1 and C2 achieved a target share above 38%, while both C3 and C4 reached below 35%. The subgroup C1 represents patients diagnosed with I10.I15 (hypertensive diseases) and prescribed with drugs C03 (diuretics) and C10 (lipid modifying gents), commonly prescribed for patients with cardiovascular diseases (Nordqvist, 2017). Similarly, the C2 group consists of patients prescribed with C03

(diuretics) and C10 (lipid modifying agents). C3 subgroup consists of patients with 10.I15 (hypertensive diseases), and I20.I25 (ischaemic heart diseases) prescribed with C03 (diuretics). In C4 patients with I20.I25 (ischaemic heart diseases) prescribed with C03 (diuretics) and C10 (lipid-modifying agents) are located. The final SD group D consists of 3 rules, all with two different conditions achieving ±20% target share. The ruling class found in all groups is Economic activity= inactive patients, under state aid. The second condition in D1 are patients diagnosed with I60.I69 (cerebrovascular diseases), in D2 patients diagnosed with K70.K77 (diseases of liver) and in D3 with L80.L99 (other disorders of the skin and subcutaneous tissue). Table 5 presents all the induced rules.

	Rules	Subgroup Size	Target Share
A1	Economic activity= state aid AND E70.E90= 1 and G40.G47=1 AND H49.H52=1	20	35%
A2	Economic activity= state aid AND F0.F09=1 AND G40.G41=1 AND H49.H52=1	10	40%
B1	Economic activity= state aid AND G40.G47=1	311	19.30%
C1	I10.I15=1 AND C03=1 AND C10=1	276	38.77%
C2	C03=1 AND C10=1	284	38.03%
C3	I10.I15=1 AND C03=1 AND I20.I25=1	357	34.27%
C4	I20.I25=1 AND C03=1	378	34.13%
D1	Economic activity= state aid AND I60.I69=1	263	25.10%
D2	Economic activity= state aid AND K70.K77=1	279	23.66%
D3	Economic activity= state aid AND L80.L99=1	162	18.53%

Table 5. Rules induced during subgroup discovery analysis concerning diabetes

We used the same procedure for exploring the response variable prediabetes. The rules were selected based on selector conjunctions in which all binary variables corresponded to the class 1. In total, ten conjunction rules were constructed with 14 unique selectors. The first set of selectors in A1 represents patients living in Kosice who are between 52 and 61 years old and are diagnosed with M5.M14 (inflammatory polyarthropathies). In the second construction of subgroup objects, groups B1-B6 were induced, with a dominant variable Location= Kosice. In B1 patients from Kosice who are above 68 years old and are prescribed with A02 (drugs for acid-related disorders) were located. Group B2 consists of patients from Kosice who are above 68 years old and are diagnosed with F30.F39 (mood affective disorders). Group B3 consists of female patients living in Kosice. Similarly, group B4 consists of female patients living in Kosice and are above 68 years old. In B5 female patients from Kosice diagnosed with L60.L75 (disorders of skin appendages) were grouped and in B6 patients living in Kosice who are above 68 years old, diagnosed with M5.M14 (inflammatory polyarthropathies) were grouped.

	Rules	Subgroup Size	Target Share
A1	Location= Kosice AND Age=52-61 AND M5.M14=1	9	66.67%
B1	Location= Kosice AND Age >= 68 AND A02=1	62	66.13%
B2	Location= Kosice AND Age >= 68 AND F30.F39=1	23	52.18%
B3	Location= Kosice AND Sex = 1 AND F30.F39=1	45	53.33%
B4	Location= Kosice AND Age >= 68 AND Sex= 1	126	51.59%
B5	Location= Kosice AND Sex=1 AND L60.L75=1	18	50%
B6	Location= Kosice AND Age>=68 AND M5.14=1	30	60%
	K80.K87=1 AND R03=1	38	
	H90.H95=1 AND M05=1	22	
	K80.K87=1 AND C04=0 AND R03=1	37	

Table 6. Rules induced during subgroup discovery analysis, concerning prediabetes

Diabetes Classification

For predicting the incidence of diabetes, we constructed the models of RF and DT using three different training sets, with two sets of variables. In order to achieve the best possible predictive power, we optimized the models for precision by grid search and 5fold cross-validation. All results provided in this section were established during the generalization step on the test data.

Importantly, the uneven distribution of classes (492 diabetic and 3344 nondiabetic patients in the training set), caused a bias towards majority class and thus, higher misclassification rate for the minority class (Ramezankhani et al., 2016b; Lopez, Fernández, García, Palade, & Herrera, 2013). As can be seen in Table 7, when the models were fitted on unbalanced samples using all predictor variables, classifiers were sensitive towards class imbalance, and while in both instances the accuracy achieved more than 85% and F1-score above 80%, it failed to identify diabetic patients achieving a sensitivity of ± 1%. Therefore, it is important to note that measures such as accuracy and F1-score are often insufficient for evaluation, introducing misleading outcomes. By conducting the SD feature selection on the unbalanced training set on both classifiers, the true positive rate slightly improved. The RF sensitivity rose by 3.28% (1.42% - 4.7%) and by 1.5% (0.9% - 2.4%) on the DT baseline. However, such low results on sensitivity do not pose a solid representation for making predictions. Table 7 provides full results of the evaluation in terms of the selected metrics of RF models and its DT baselines.

	Classifiers	AUC	SEN	SP	F1	Accuracy
Unbalanced	DT	50.22%	0.9%	99.51%	81%	86.80%
	RF	50.67%	1.42%	99.93%	82%	87.23%
	DT, SD features	51.07%	2.4%	99.80%	82%	87.23%
	RF, SD features	51.56%	4.7%	98.40%	82%	86.32%
SMOTE	DT	52.36%	26.41%	78.30%	75%	71.61%
	RF	55.23%	33.02%	77.46%	75%	71.73%
	DT, SD features	52.36%	26.41%	78.30%	75%	71.61%
	RF, SD features	62%	57.55%	66.43%	71%	65.29%
Down-sampled	DT	61.09%	69.81%	54.36%	63%	56.35%
	RF	68.34%	73.11%	63.58%	71%	64.80%
	DT, SD features	67.15%	65.57%	68.74%	73%	68.33%
	RF, SD features	67.25%	69.81%	64.69%	71%	65.35%

Table 7. Results of random forest and decision tree baseline models using all predictor variables compared to results of random forest and decision tree baseline using subgroup discovery feature selection — results provided in %.

Due to a class imbalance, we implemented ensemble resampling techniques SMOTE and down-sampling and used new sample sizes for building the RF and baseline models. In general, the random forest models outperformed the decision tree baseline on all sample sets, which is consistent with previous studies (Esmaily et al., 2019; Chawla, Bowyer, Hall & Kegelmeyer, 2012; Nai-aruna & Rungruttikarn, 2015). Most notably, the RF trained on SMOTE sample set achieved better results in AUC of 2.87% (55.23%- 52.36%) and SEN of 6.61% (33.02%- 26.41%), when compared to the baseline identifying 70 diabetic patients correctly and misclassifying 148 patients. The difference in accuracy and the F1-score was negligible. The RF using selected features finally enhanced the model's performance by 6.77% in AUC (62%- 55.23%), and by 24.53% in sensitivity (57.55%- 33.02%), when compared to the model using all predictors. Still, with the selected features the rate of true negative decreased by 11.03% (77.46%- 66.43%), which lead to decrease in F1-score by 4% (75%-71%) and 6.44% in accuracy (71.73%-65.29%). However, the results of sensitivity are a rather unexpected outcome, as the literature suggests steep improvements in performance when oversampling with SMOTE algorithm (Chawla et al., 2002; Ramezankhani et al., 2016b). The rate of correctly classified patients was finally higher with the RF model trained on down-sampled sets, with 155 correctly classified and 57 incorrectly classified diabetic patients. Similarly to other models, the RF model produced higher results than the DT baseline by 7.25% on AUC (68.34%-61.09%), 3.3% on SEN (73.11%-69.81%), 9.22% on SP (63.58%-54.36%), 8% on F1-score (71%-63%) and 8.45% on accuracy (64.80%-56.35%).

Inconsistently with previous results of applying selected features, the method caused the classifiers to perform worse on AUC by 1.09% (68.34%-61.09%), and sensitivity by 3.3% (73.11%-69.81%). The specificity (tn) was slightly higher by 1.11% (64.69%-63.58%) and accuracy by 0.55%

(65.35%-64.80%). However, none of these differences were significant.

Overall, the SD feature selection increased the performances of models trained on unbalanced and SMOTE samples, while decreased the performance of a model trained on the down-sampled training set. Moreover, comparison of the ensemble models (using selected features) to the models using all predictors indicates, that by application of feature selection, the performance improved in the identification of the true positive cases with higher sensitivity and AUC while achieving lower performance in specificity, F1-score, and accuracy. In general, all random forest models outperformed its decision tree baselines, when being compared. Overall, out of all models, the random forest model trained on down-sampled set using all 103 predictor variables, achieved the best performance on AUC and sensitivity and therefore, was chosen for predicting the incidence of diabetes.

Besides measuring the model performances, we also focused on identifying explaining features of T2D, for which we used RF built-in function to select important features (Esmaily etl al., 2018). The function was run on the initial unbalanced sample set, as resampled sets do not reflect a real representation of the patient's data. Figure 7 shows feature importance computed with mean decrease Gini, measuring the average gain of purity by splitting induced variables (Bruce & Bruce, 2017). The model generated 10 variables: Age, Location, C08 (calcium channel blockers), C03 (diuretics), I20.I25 (ischaemic heart diseases), C07 (beta blocking agents), E70.E90 (metabolic disorders), H30.H36 (disorders of choroid and retina), Sex, C10 (lipid modifying gents) and C02 (antihypertensives).

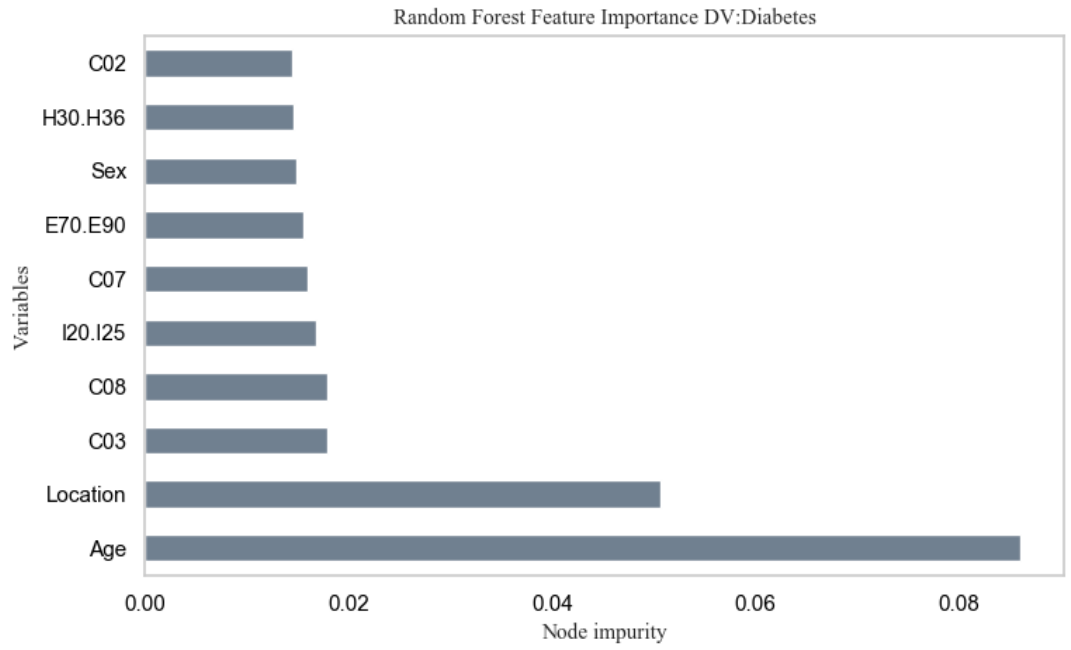

Figure 7. Ten most important variables when predicting diabetes in ascending order, generated by the random forest built-in function feature importance.

Out of these variables, the majority are directly linked to diagnoses caused by diabetes. For example, the ICD 10 code variable I20.I25 represents different forms of heart diseases which have been previously researched and confirmed as a strong predictor of the disease (Barber et al., 2014; Olivera et al., 2017, Razavian et al. 2015). Moreover, all the variables C02, C03, C07, C08 fall under the drugs used for the cardiovascular system, for lowering blood pressure, protecting against heart attacks, and improving the outlook for people with heart failure (Nordqvist, 2017). Other variables such as H30.H36 has also been previously subjected to diabetes in several studies, as T2D is one of the leading causes of open-angle glaucoma (eye diseases in which the optic nerve degenerates) (NIDDK, 2017).

From the demographic variables, Age, Location, and Sex were found significant, which is consistent with previous studies (Oliviera et al., 2017; Meng et al., 2013). Further inspection revealed that with increasing age, the incidence of diabetes increased as well. These results are somewhat in line with subgroups established during EDA, in which economically inactive patients were grouped. Furthermore, Figure 9 compares the distribution of age among diabetic and prediabetic patients, and as the curve suggests, prediabetes occurs at a younger age, in comparing to diabetes (i.e., 30-40 and steeply rises to 60-70 years). Figure 8 reveals that the majority of diabetic people are living Kosice, Komarno, and Snina.

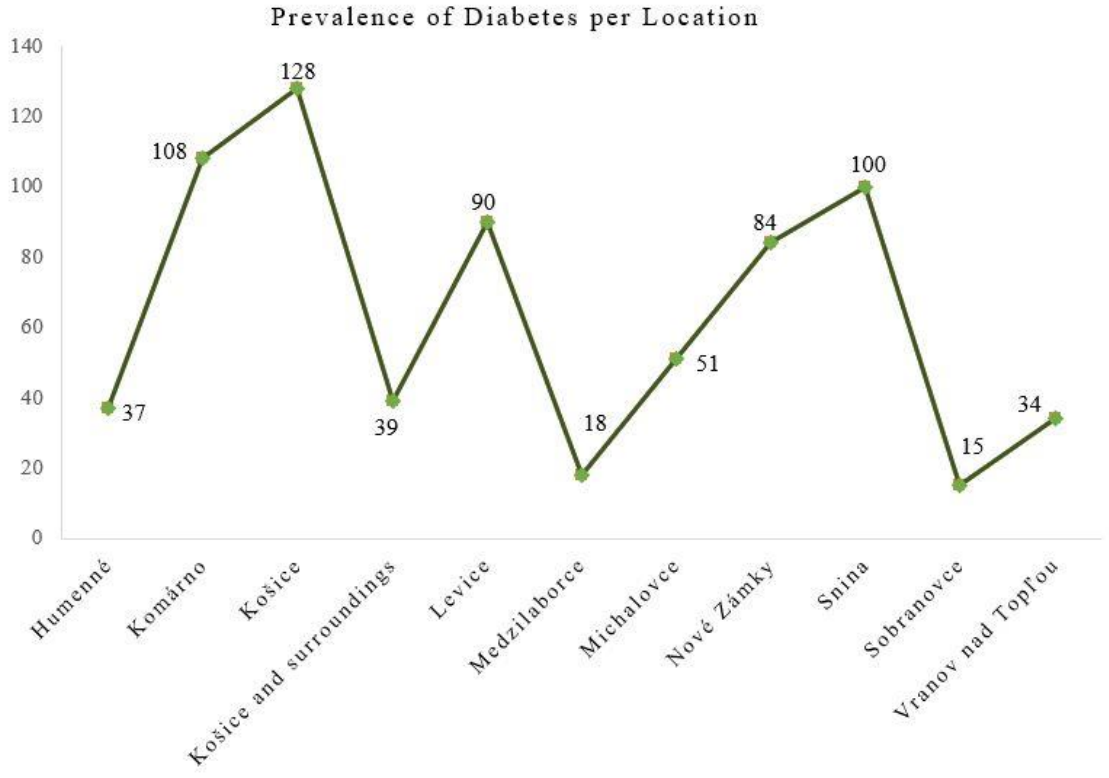

Figure 8. 15 Prevalence of diabetes among eastern cities in Slovakia.

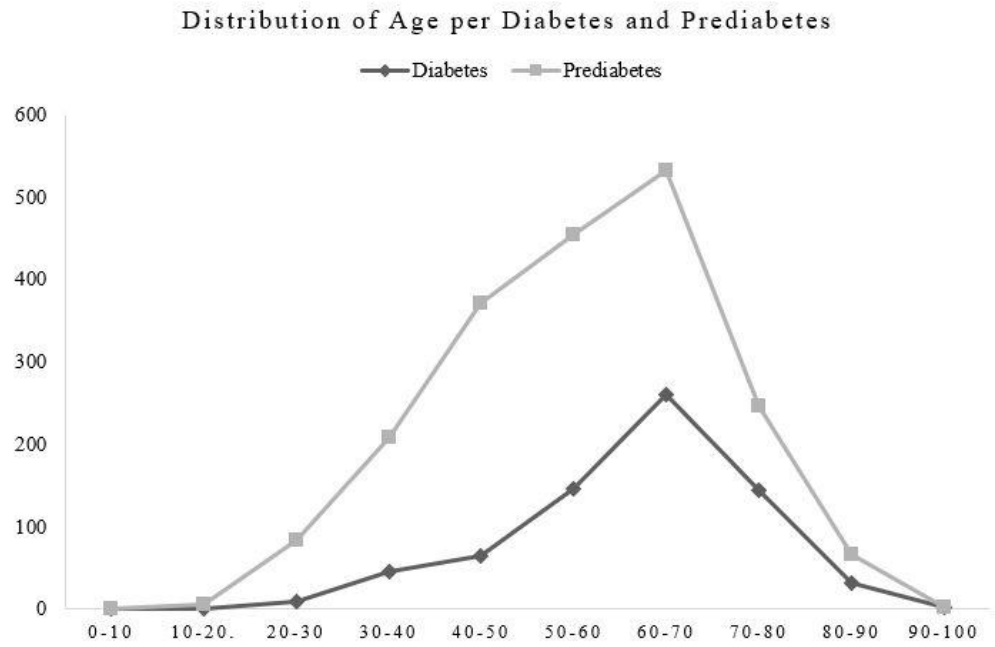

Figure 9. 15 Prevalence of diabetes among eastern cities in Slovakia.

Prediabetes Classification

Prediction of prediabetes followed the same sets of procedures as when predicting diabetes. We constructed the RF and DT baseline models using the unbalanced training set, balanced with SMOTE and balanced by down-sampling training sets, using all 103 predictor variables and selected variables during the subgroup discovery. In general, the RF model achieved better predictive power using all three sample sets, when compared to the DT baseline, which is in line with the previous findings (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

As the classes of the target variable were severely imbalanced (1=1385, 0=2451), we encountered the same bias towards the minority class, as when predicting diabetes. The RF trained on the unbalanced sample set hardly outperformed the DT baseline by only a few percentages on all metrics except specificity, which was higher for the DT. While on both models of RF and DT, the accuracy was ± 64% and F1-score ±54%, the sensitivity did not exceed 10%. The feature selection improved the model´s accuracy only slightly, most notably on sensitivity by 3.2% (9.7%-6.5%), while it decreased the performance on specificity, F1-score, and accuracy.

The RF model trained on the SMOTE sample compared to the DT baseline, produced AUC better by 0.87% (54.72%-53.85%), sensitivity by 2.89% (46.43%-43.54%), and F1-score by 1% (58%-57%). The difference in accuracy was almost indistinguishable. By implementing feature selection, the sensitivity rose by 5.61% (46.43%-40.82%). On other metrics, the improvement was only marginal. The balanced sample set with down-sampling achieved the best predictive power, while outperforming the DT by 5.19% on AUC (56.10%-50-91%), by 12.03% on sensitivity (61.40%-48.47%), by 2% on the F1-score (55%-53%), and by 2.98% on accuracy (54,59%-51.61%).

Feature selection improved the model's sensitivity by 2.20% (63.60%-61.40%). Surprisingly on all other metrics, the performance decreased. Overall, the feature selection performed on all sample sets, increased sensitivity (tp), while in all cases, decreased specificity (tn). Moreover, despite the fact, the down-sampled model using selected features achieved the highest performance on sensitivity, the down sampled model using all features performed higher on all other evaluation metrics and was therefore chosen for predicting prediabetes. Table 8 presents the full results of the RF and DT models using the SD feature selection.

	Classifier	**AUC**	**SEN**	**SP**	**F1**	**Accuracy**
Unbalanced	DT	50.89%	5.2%	96.50%	53%	63.89%
	RF	51.53%	6.5%	96.60%	54%	64.38%
	DT, SD features	50.36%	5.4%	95.27%	53%	63.16%
	RF, SD features	51.95%	9.7%	94.03%	55%	63.95%
SMOTE	DT	52.50%	37.24%	67.74%	57%	56.84%
	RF	53.38%	40.82%	65.94%	57%	56.97%
	DT, SD features	53.85%	43.54%	64.14%	57%	56.78%
	RF, SD features	54.72%	46.43%	63%	58%	57.08%
Down-sampled	DT	50.91%	48.47%	53.36%	53%	51.61%
	RF	56.10%	61.40%	50.80%	55%	54.59%
	DT, SD features	54.90%	59.86%	49.96%	54%	53.49%
	RF, SD features	55.54%	63.60%	47.69%	54%	53.38%

Table 8. Results of random forest and decision tree baseline predicting response variable Prediabetes, compared to results of random forest and decision tree baseline using subgroup discovery feature selection.

The important features were evaluated based on the Gini index. In Figure 10, two variables Location and Age are visibly more significant, while the node impurity remains almost steady for the rest: Sex, J0.J6, J01, M50.M54, M01, K20.K31, C07. As already established, C07 (beta blocking agents) are drugs commonly used for individuals with cardiovascular diseases. As research suggests, early detection of prediabetes can reduce the risk of developing cardiovascular diseases (Brannick & Dagogo-Jack, 2018), meaning that those are symptoms that usually arise when the prediabetic stage is already present. M50.M54 (other dorsopathies) represents any disorder of back or spine and seems to be an interesting outcome, as the comorbidity between back pain and diabetes have been previously a subject of medical research, where diabetes and back pain have been developed simultaneously. In addition, patients with T2D usually have more severe symptoms (Iskra, 2018).

Moreover, a link has been established between J01 (antibacterials for systemic use) and prediabetes, as infectious diseases are more severe and frequent at diabetic and prediabetic patients (Casqueiro & Alves, 2012). Notably, diabetes also increases the propensity for both chronic and acute

infections seemingly as a part of the impaired immunity and therefore strongly linked to J0.J6 (acute upper respiratory infections) as established in a prior research (Aburawi, Liuba, Pesonen, Ylä-Herttuala, & Sjöblad, 2004). K20.K31 (diseases of oesophagus, stomach and duodenum) has been previously found as a strong predictor of diabetes, as it affects almost every part of the gastrointestinal tract from the esophagus to the rectum (Yarandi & Srinivasan, 2014). However, the link to prediabetes has not been previously established. Similarly, M01 (anti-inflammatory and antirheumatic products) has not been directly linked to prediabetes in the literature.

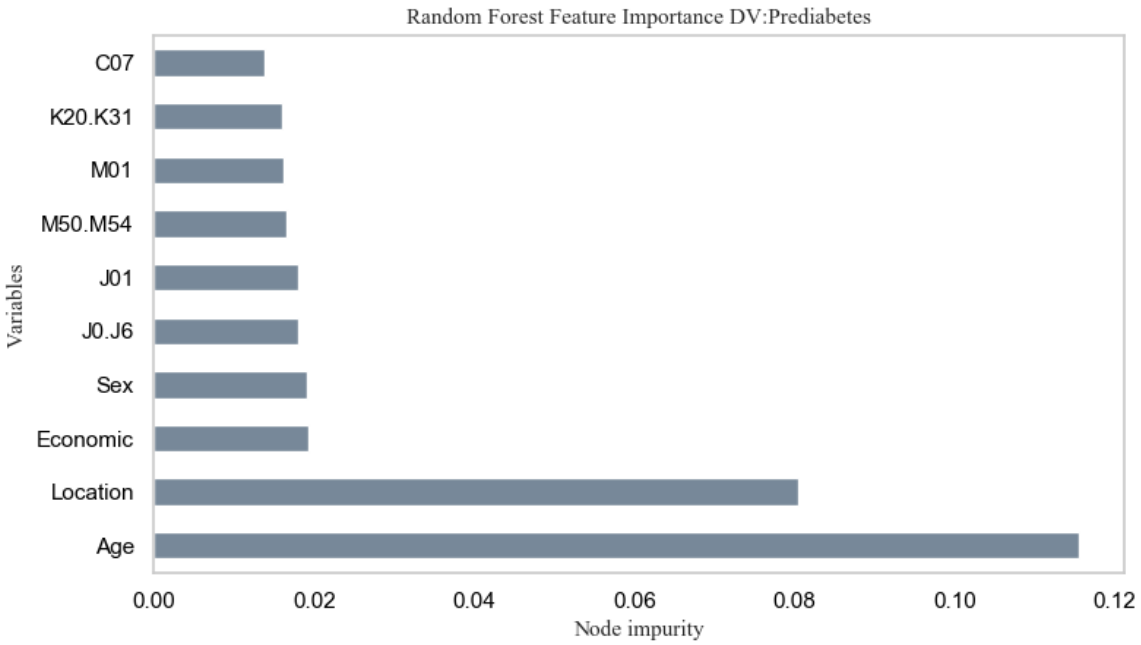

Figure 10. Random Forest feature selection results of 10 most important variables when predicting prediabetes.

Further research revealed that the majority of the people affected by both diagnoses are between 60-70 years old, as shown in Figure 9. Interestingly, in comparison to diabetic patients, prediabetes affects more people who are between 30-50 years old. In Figure 11, the incidence of prediabetes among eastern cities in Slovakia can be observed, with the highest incidence in Kosice.

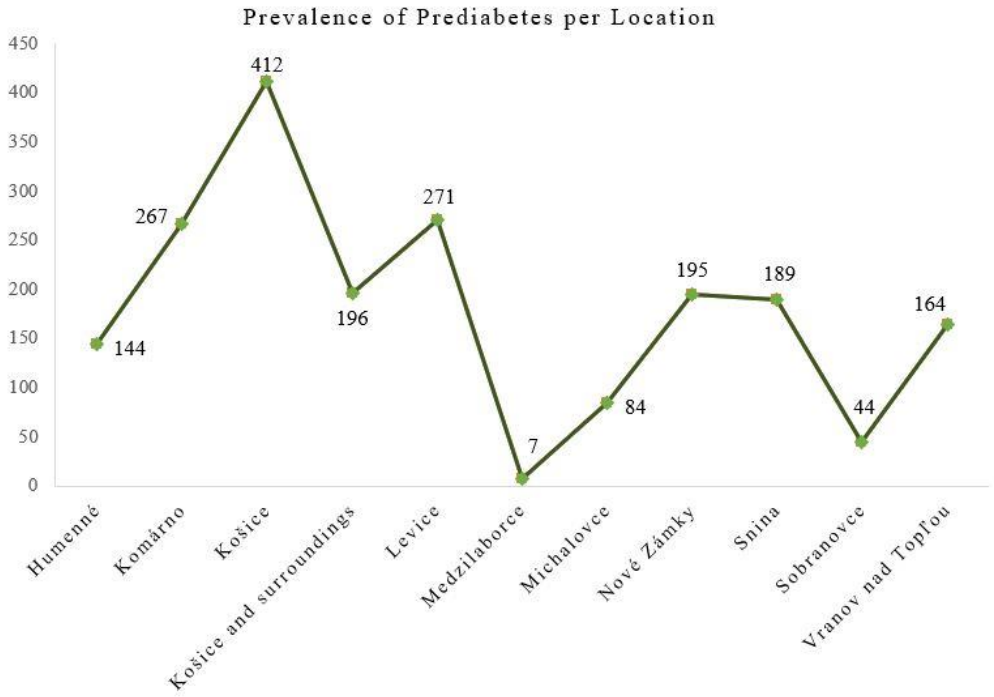

Figure 11. Incidence of prediabetes among eastern cities in Slovakia

Discussion

The study was aimed at developing the best performing predictive model to estimate the probability of Type 2 Diabetes Mellitus and prediabetes incidence and help in high-risk individual identification. In this chapter results of the study are being evaluated, regarding their implications to the prediction of diabetes and prediabetes. The methodological approach in designing the predictive models, combined with the EDA analysis, is somewhat original to the landscape of state-of-the-art healthcare research. During EDA analysis (subgroup discovery), interesting subgroups pertaining a unique set of conjunction selectors were found, regarding both target variables, describing the diagnoses and their forms. By conducting the exploratory analysis, we laid a unique approach in selecting explaining variables for further prediction analysis and explored whether models achieved any improvement in their performances. The models used were RF and DT baseline. We used the RF built-in function to identify the most important models' features. Those were compared to the SD selectors, and variables reoccurring in both methods were selected for consideration. The current study found that for predicting diabetes, variables E70.E90, C03, and I20.I25 are suitable. Age, Location, and Sex were found significant for predicting prediabetes. Thus, those features contribute most to predicting diabetes and prediabetes incidence from the patient demographic and utilization data, available to the analysis.

For the research, SD was postulated as a tool for exploring the targets and for selecting features when building predictive models. In this study, it was found that the SD feature selection

technique caused a successful improvement of the model's sensitivity and AUC. We mostly focused on improving sensitivity and AUC, as we aimed at identifying the critical class (tp). It is important to note that different research goals and settings require different evaluation metric. Therefore, an essential finding, in line with the previous studies, is that accuracy, and F1-score as performance criterions are not sufficiently robust measures, mostly when classes are skewed (Chawla et al., 2002). The most notable improvement of SEN and AUC was observed on the RF model using SMOTE when the SEN rose from 55.23% to 62% and AUC from 33.02% to 57.55%.

Surprising is that, while the SD feature selection secured higher rates of identified diabetic and prediabetic patients, it decreased the rates of correctly classified negative cases, which resulted in lower performance of F1-scores and accuracy. Another surprising outcome is that feature selection performed well on all models, but those trained on down-sampled sets. Even though the general findings suggest that feature selection is an effective method in the light of improving performance of predictive models, it did not suffice in outperforming the RF models trained on down-sampled sets using all 103 explaining variables when predicting both diabetes and prediabetes. Hence, it could conceivably be hypothesized that subgroup discovery could be a promising tool for feature selection, but it did not improve the performance of the RF trained on down-sampled sets, using all 103 predictor variables.

Finally, the present study was designed to determine the extent to which models of random forest and decision tree contribute to diabetic and prediabetic identification. In the analysis, the DT served the purpose of a baseline comparison, which was surpassed in all instances by the RF model. In accordance with the present results, previous studies have demonstrated that RF algorithm boosts the robustness of the model (Esmaily et al., 2018). Due to a severe class imbalance of both targets, that lead to a bias towards the minority class (true positive class), the models were resampled by using SMOTE and down-sampling algorithms. Some studies have shown satisfactory improvement in predictive models when using SMOTE (Chawla et al., 2002; Ramezankhani et al., 2016b). This does not appear to be the case in the current analysis. Regardless of the fact, that comparison of the RF using SMOTE and unbalanced set improved sensitivity by 31.06%, achieving 32.02% for predicting diabetes and by 34.32% resulting into 40.82% when predicting prediabetes, the results are not sufficient for making predictions.

Moreover, while resampling seems to improve the sensitivity, it decreases the specificity and accuracy of models, which has been previously sketched in other studies as well (Chawla et al., 2002). Finally, the RF model using down-sampling achieved satisfactory predictive power in both instances, and thus is considered suitable for predicting diabetes and prediabetes. With the prediction of diabetes, the model achieved an AUC of 68.34%, a sensitivity of 73.11%, a specificity of 63.58%, an F1-score of 71%, and an accuracy of 64.80%. With the prediction of prediabetes, the model achieved an AUC of 56.10%, a sensitivity of 61.40%, a specificity of 50.80%, an F1-score of 57%, and an accuracy of 54.59%. Due to the fact, models were fitted on the down-sampled data; the results need to

be interpreted with caution, as some substantial variables may have been cut-out of the analysis (Menardi & Torelli, 2012).

The implications for further research stem mainly from the design limitations. Skewed classes of the target variables posed one of the biggest pitfalls and resampling them becomes an inevitable part of the analysis, mainly in the medical domain. Results of the SMOTE models suggest that exploring SMOTE at different percentages (e.g., 200%, 300%, 700%) of its original size, could potentially achieve higher prediction results in the analysis. However, exploring different sampling techniques is limited by a time constraint and usually forms a central topic of research (Ramezankhani et al., 2016b). Future research could also explore the usage of other machine learning models applied to the predictions of diabetes and prediabetes.

Furthermore, another pitfall of using healthcare data is that records can be often misleading and inaccurate. Practitioners are prone to record patient's data inaccurately, lacking details or distinction in between closely related diagnoses. This may result in confusing diagnoses or leaving out important information (B. Benková, M. Poliak, personal communication, November 05, 2018). For example, the target variable representing T2D had to be built by a combination of all ICD10 diabetes codes (E10-E14), which comprise all diabetes types rather than only using an ICD code corresponding to T2D. Therefore, the analysis was unable to retain detailed information specific to T2D only. A possible solution to diminish the consequences is setting up an experiment, such as a longitudinal study, in which practitioners would be provided thorough guidelines for recording patients' data.

Moreover, in further studies, researchers could conduct a time-varying analysis. However, it is suggested that before starting such an analysis, a considerable amount of data should be collected. This could be especially beneficial for analyzing the transition from prediabetic to the diabetic stage. A thorough inspection of the diagnoses could reveal patterns in diagnoses over time, and potential indicators of the transition. Furthermore, a time-varying analysis could be also beneficial for inspecting the diagnoses independently by building several time frames, with each representing a different stage of the diagnoses. Such an analysis could unveil changes in the diseases over time and provide a better understanding of its forms.

Conclusion

In this research, we proposed a methodological design consisting of exploratory and predictive analyses. During the EDA, by using the subgroup discovery algorithm BSD, we induced sets of interesting rules for diabetes mellitus and prediabetes. In the predictive analysis, we examined the predictive power of two models, to predict the incidence of T2D and prediabetes. Due to the fact that both target variables had skewed classes, we resampled the training set by using SMOTE and down-sampling algorithms. In order to improve the models' performance, we used several ensemble techniques such as parameter tuning by grid search and 5-fold cross-validation and feature selection.

We used selectors explored in the subgroup discovery as a feature selection method. This study has identified that the RF model, using down-sampling for skewed classes and all 103 predictor variables achieved the highest results, predicting both diabetes and prediabetes. The second major finding was an identification of predictors of both diseases, which could assist Dôvera in identifying high-risk individuals and creating a valuable tool to help practitioners in screening for T2D and prediabetes.

References

Aburawi, E., Liuba, P., Pesonen, E., Ylä-Herttuala, S., & Sjöblad, S. (2004). Acute respiratory viral infections aggravate arterial endothelial dysfunction in children with Type 1 Diabetes. *Diabetes Care, 27*(11), 2733-2735. http://dx.doi.org/10.2337/diacare.27.11.2733

Atzmueller, M., & Lemmerich, F. (2009). Fast Subgroup Discovery for continuous target concepts. In J. Rauch, W. Z. Ras, P. Berka, & T. Elomaa, *Foundations of Intelligent Systems* (pp. 35-41). Prague: Springer.

Atzmueller, M., & Puppe, F. (2006) SD-Map – A Fast Algorithm for Exhaustive Subgroup Discovery. In: Fürnkranz J., Scheffer T., Spiliopoulou M. (eds) *Knowledge Discovery in Databases: PKDD 2006.* PKDD 2006. Lecture Notes in Computer Science, vol 4213. Springer, Berlin, Heidelberg. http://dx.doi.org/10.1007/11871637_6

Atzmüller, M. (2015). Subgroup Discovery. *WIREs Data Mining Knowl Discovery, 5*(1), 35-49. http://dx.doi.org/10.1002/widm.1144

Barber, S. R., Davies, M. J., Khunti, K., & Gray, L. J. (2014). Risk assessment tools for detecting those with pre-diabetes: A systematic review . *Diabetes Research and Clinical Practice*, 1-13. http://dx.doi.org/10.1016/j.diabres.2014.03.007

Bhopal, R. S. (2002). *Concepts of Epidemiology.* New York: Oxford University Press.

Brannick, B., & Dagogo-Jack, S. (2018). Prediabetes and cardiovascular disease: Pathophysiology and interventions for prevention and risk reduction. *Endocrinology and Metabolism Clinics of North America, 47*(1), 33-50. http://dx.doi.org/10.1016/j.ecl.2017.10.001

Bruce, P., & Bruce, A. (2017). *Practical statistics for data scientists: 50 essential concepts.* Sebastopol: O'Reilly Media Inc.

Casqueiro, J., & Alves, C. (2012). Infections in patients with diabetes mellitus: A review of pathogenesis. *Indian Journal of Endocrinology and Metabolism*, 27-36. http://dx.doi.org/10.4103/2230-8210.94253

Collins, G. S., Johannes B. Reitsma, J. B., Altman, D. G., & Moons, K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement free. *Annals of Internal Medicine*, 1-73. http://dx.doi.org/10.7326/M14-0697

DÔVERA Health Insurance Company. (2017a, 03 23). *Lepsi zivot s cukrovkou*. Retrieved from Dôvera: https://www.dovera.sk/lepsizivotscukrovkou/zivot-s-cukrovkou/o-cukrovke

DÔVERA Health Insurance Company. (2017b, 05 10). *Novinky o cukrovke*. Retrieved from Dôvera: https://www.dovera.sk/lepsizivotscukrovkou/o-nas/novinky-o-cukrovke/64_za-20-rokov-na-slovensku-pribudlo-200-tisic-diabetikov-kazdy-den-pribudaju-dalsi

DÔVERA Health Insurance Company. (2018, 07 04). *O nás*. Retrieved from Dôvera: https://www.dovera.sk/o-nas/english-info

Esmaily, H., Tayefi, M., Doosti, H., Ghayour-Mobarhan, M. N., & Amirabadizadeh, A. (2018). A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes. *Journal in Research in Health Sciences, 18*(2), 2-7.

Fortmann-Roe, S. (2012, 06). *Bias variance*. Retrieved from Understanding the bias-variance tradeoff: http://scott.fortmann-roe.com/docs/BiasVariance.html

Gamberger, D., Lavrač, N., Krstačić, A., & Krstačić, G. (2007). Clinical data analysis based on iterative subgroup discovery: Experiments in brain ischaemia data analysis. *Applied Intelligence, 27*(3), 205-217. http://dx.doi.org/10.1007/s10489-007-0068-9

Gambergera, D., Lavrac, N., & Krstačić, G. (2003). Active subgroup mining: a case study in coronary heart disease risk group detection. *Artificial Intelligence in Medicine*, 27-57. http://dx.doi.org/10.1016/S0933-3657(03)00034-4

Gonzalez-Abril, L., Cuberos, F. J., Velasco, F., & Ortega, J. A. (2009). Ameva: An autonomous discretization algorithm. *Expert Systems with Applications, 36*, 5327–5332. http://dx.doi.org/doi:10.1016/j.eswa.2008.06.063

Guillausseau, P. J., & Dupuy, E. (1996). Antithrombotic agents and diabetes: Benefits and recommendations for use. *Archives des Maladies du Cœur et des Vaisseaux - Pratique*, 1557-1561. http://dx.doi.org/10.1016/S0003-3928(10)70011-1

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 321–357. http://dx.doi.org/10.1613/jair.953

International Diabetes Federation. (2015). *IDF Diabetes Atlas.* Karakas Print.

Iskra, D. A. (2018). Comorbidity of type 2 diabetes mellitus and low back pain. *Zhurnal nevrologii i psikhiatrii imeni S.S. Korsakova*, 126-130. http://dx.doi.org10.17116/jnevro2018118081126

James, G., Witten, D., Hastie, T., & Ibshirani, R. (2013). *An Introduction to Statistical Learning.* New York: Springer Science.

Jin, X., Xu, A., Bie, R., & Guo, P. (2006). Machine learning techniques and Chi-square feature selection for cancer classification using SAGE gene expression profiles. *Data Mining for Biomedical Application*, 106-114. http://dx.doi.org/10.1007/11691730_11

Koh, H. C., & Tan, G. (2005). Data mining applications in healhcare. *Journal of Healthcare Information Management, 19*(2), 61-72. http://dx.doi.org/10.1007/978-1-4614-7138-7

Köster, I., von Ferber, L., Ihle, P., Schubert, I., & Hauner, H. (2006). The cost burden of diabetes mellitus: the evidence from Germany-the CoDiM study. *Diabelotogia, 49*(7), 1498-1504. http://dx.doi.org/10.1007/s00125-006-0277-5

Lemmerich, F., Rohlfs, M., & Atzmueller, M. (2010). Fast discovery of relevant subgroup patterns. *In Proc. 23rd International FLAIRS Conference* (pp. 428–433). Palo Alto, CA, USA: AAAI Press.

Lopez, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 113-141. http://dx.doi.org/10.1016/j.ins.2013.07.007

Menardi, G., & Torelli, N. (2012). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery, 28*(1), 92-122. http://dx.doi.org/10.1007/s10618-012-0295-5

Meng, X., Huang, Y., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or pridiabetes by risk factors. *The Kaohsiung Journal of Medical Sciences*, 93-99. http://dx.doi.org/10.1016/j.kjms.2012.08.016

Mwaniki, A. (2017, 04 25). *World Atlas*. Retrieved from Biggest Cities In Slovakia: https://www.worldatlas.com/articles/biggest-cities-in-slovakia.html

Nai-aruna, N., & Rungruttikarn, M. (2015). Comparison of classifiers for the risk of diabetes prediction. *Procedia Computer Science, 69*, 132-142. http://dx.doi.org/10.1016/j.procs.2015.10.014

NCZI. (2018, 11 14). *Činnosť diabetologických ambulancií v SR 2017*. Retrieved from Národné centrum zdravotníckych informácií : http://www.nczisk.sk/Aktuality/Pages/Cinnost-diabetologickych-ambulancii-v-SR-2017.aspx

NIDDK. (2017, 05). *Diabetic eye disease*. Retrieved from Health Information: https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/diabetic-eye-disease

Nordqvist, C. (2017, 06 30). *Articles*. Retrieved from What you need to know about beta blockers : https://www.medicalnewstoday.com/articles/173068.php

Olivera, A. R., Roesler, V., Iochpe, C., Schmidt, M. I., Vigo, A., Barreto, S. M., & Duncan, B. D. (2017). Comparison of machine-learning alhorithms to build a predictive model for detecting undiagnosed diabetes- ELSA- Brasil: accuracy study. *Sao Paulo Medical Journal, 135*(3), 234-246. http://dx.doi.org/10.1590/1516-3180.2016.0309010217

Priya, R. J., & Umamaheswari, K. (2014). Type 2 diabetes prediction using multinominal logistic regression. *Australian Journal of Basics and Applied Sciences, 8*(10), 31-37.

Ramazenkhani, A., Pournik, O., Shahrabi, J., Khalili, D., Azizi, F., & Hadaegh, F. (2014). Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study. *Diabetes Research and Clinical Practice, 105*(3), 391-398. http://dx.doi.org/10.1016/j.diabres.2014.07.003

Ramezankhani, A., Hadavandi, E. P., & Shahrabi, J. (2016a). Decision tree-based modelling for identification of potential interactions between type 2 diabetes risk factors: A decade follow-up in a Middle East prospective cohort study. *BMJ Open, 6*(12), 1-13. http://dx.doi.org/10.1136/bmjopen-2016-013336

Ramezankhani, A., Pournik, D., Shahrabi, J., Azizi, F., Hadaegh, F., & Khalili, D. (2016b). The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes. *Medical Decision Making, 36*(1), 137-144. http://dx.doi.org/10.1177/0272989X14560647

Raschka, S. (2016). *Python machine learning.* Birmingham: Packt Publishing.

Razavian, N., Blecker, S., Schmidt, A. M., Smith-McLallen, A., Nigam, S., & Sontag, D. (2015). Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data, 3*(4), 277-287. http://dx.doi.org/10.1089/big.2015.0020

Reunanen, A., Kangas, T., Martikainen, J., & Klaukka, T. (2000). Nationwide survey of comorbidity, use, and costs of all medications in Finnish diabetic individuals. *Diabetes Care, 23*(9), 1265-1271. http://dx.doi.org/10.2337/diacare.23.9.1265

Román-Pintos, L. M., Villegas-Rivera, G., Rodríguez-Carrizalez, A. D., Miranda-Díaz, A. G., & Cardona-Muñoz, E. G. (2016). Diabetic polyneuropathy in type 2 diabetes Mellitus: Inflammation, cxidative stress, and mitochondrial function. *Journal of Diabetes Research*, 1-16. http://dx.doi.org/10.1155/2016/3425617

Rui Wang, M. S., Lagakos, S. W., Ware, J. H., Hunter, D. J., & Drazen, J. M. (2007). Statistics in medicine: Reporting of subgroup analyses in clinical trials. *The New England Journal of Medicine*, 2189-2194. Retrieved from https://www.nejm.org/doi/full/10.1056/NEJMsr077003

Scikit-learn. (2019). *sklearn.tree.DecisionTreeClassifier*. Retrieved from Sci-kit Learn: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Simmons, R. K., Griffin, S. J., Lauritzen, T., & Sandbæk, A. (2017). Effect of screening for type 2 diabetes on risk of cardiovascular disease and mortality: a controlled trial among 139,075 individuals diagnosed with diabetes individuals diagnosed with diabetes. *Diabetologia*, 2192-2199. http://dx.doi.org/10.1007/s00125-017-4299-y

Stock, S., Drabik, A., Büscher, G., Graf, C., Ullrich, W., & Gerber, A.; Lauterbach, K., W.; Güngen, M. (2010). German diabetes management improve quality of care and curb costs. *Health Affairs, 29*(12), 2019-2205. http://dx.doi.org/10.1377/hlthaff.2009.0799

Szalay, T., Jankó, V., Mužík, R., Melo, M., & Benková, B. (2017). *Programy riadenej zdravotnej starostlivosti. Aká je rola všeobecného lekára?* Bratislava: Dôvera Zdravotná Poisťovňa.

WHO. (2016). *Global Reports on Diabetes.* France: World Health Organization 2016. https://apps.who.int/iris/bitstream/handle/10665/204871/9789241565257_eng.pdf;jsessionid=91DEDC680CF6A60C5552A9942245F2F7?sequence=1

WHO. (2018, 10 30). *Diabetes*. Retrieved from News: http://www.who.int/news-room/fact-sheets/detail/diabetes

Yarandi, S. S., & Srinivasan, S. (2014). Diabetic gastrointestinal motility disorders and the role of enteric nervous system: Current status and future directions. *Neurogastroentology and Motility* , 611–624. http://dx.doi.org/10.1111/nmo.12330

Appendix A

Table A1: The importance of the 12 input variables in three models

Order	Logistic regression	Artificial Neural Network (B-P)	Decision tree (C5.0)
1.	Age	Age	Age
2.	Family history of diabetes	Family history of diabetes	Education level
3.	Marital status	Duration of sleep	Family history of diabetes
4.	Education level	Preference for salty food	Marital status
5.	Work stress	Marital status	Preference for salty food
6.	Duration of sleep	Education level	Drinking coffee
7.	Physical activity	Work stress	Duration of sleep
8.	Preference for salty food	Physical activity	Body mass index
9.	Gender	Drinking coffee	Work stress
10.	Eating fish	Gender	Eating fish
11.	Drinking coffee	Body mass index	Physical activity
12.	Body mass index	Eating fish	Gender

Table A1. The order according to importance, from the most to the least important. Reprinted from "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors" by X. H. Meng, Y. X. Huang, D. P. Rao, Q. Zhang, Q. Liu, 2013, *The Kaohsiung Journal of Medical Sciences*, 29, p.97. Copyright Elsevier 2012.

Appendix B

Table B1: Groups identified by decision tree models for men (Tehran Lipid and Glucose Study 1999–2012)

Models	Groups	Definition	Probability	Predicted class
Model (1)	1	FPG<4.9 and 2h-PCPG<7.7	0.90	Non-diabetic
	2	9<FPG<5.3 and 2h-PCPG<7.7 and WHtR<0.6	0.72	Non-diabetic
	3	FPG>5.3 and 2h-PCPG<4.4 and age <43	0.67	Non-diabetic
	4	4.9<FPG<5.3 and 2h-PCPG<7.7 and WHtR>0.6	0.53	Diabetic
	5	FPG<5.3 and 2h-PCPG>7.7	0.70	Diabetic
	6	FPG>5.3 and 2h-PCPG<4.4 and age >43	0.68	Diabetic
	7	FPG>5.3 and 2h-PCPG>4.4 0.79 Diabetic	0.79	Diabetic
Model (2)	1	FPG<4.9	0.86	Non-diabetic
	2	4.9<FPG<5.3 and WHItR<0.56	0.70	Non-diabetic
	3	4.9<FPG<5.3 and WHtR>0.56 and FHD='no'	0.58	Non-diabetic
	4	FPG>5.3 and 0.4<WHtR<0.49and MAP<92	0.75	Non-diabetic
	5	4.9<FPG<5.3 and WHtR>0.56 and FHD='yes'	0.78	Diabetic
	6	FPG>5.3 and WHtR<0.45	0.56	Diabetic
	7	FPG>5.3 and 0.45<WHtR<0.49 and MAP>92	0.67	Diabetic
	8	FPG>5.3 and<0.49<WHtR<0.56	0.74	Diabetic
	9	FPG>5.3 and WHtR>0.56	0.84	Diabetic

Table B1. Model (1) was developed based on 15 variables which included 2h-PCPG.
Model (2) was developed based on 14 variables (2h-PCPG was excluded).
*The percentage of population in the defined subgroup, which can be interpreted as probability of an outcome. +Predicted outcome for men who belong to the defined subgroup.
2h-PCPG, 2-hour postchallenge plasma glucose (mmol/L); FHD, family history of diabetes; FPG, fasting plasma glucose (mmol/L); MAP, mean arterial blood pressure (mm Hg); WHtR, waist-to-height ratio.
Reprinted from "Decision tree-based modelling for identification of potential interactions between type 2 diabetes risk factors: a decade follow-up in a Middle East prospective cohort study." by A. Ramezankhani, E. Hadavandi, O. Pournik, J. Shahrabi, F. Azizi, F. Hadaegh, 2016, *BMJ Open*, 6, p.11. BMJ Publishing Group 2016.

Table B2: Groups identified by decision tree models for women (Tehran Lipid and Glucose Study 1999–2012)

Models	Groups	Definition	Probability	Predicted class
Model (1)	1	FPG≤5.2 and WHtR≤0.55	0.88	Non-diabetic
	2	2 FPG≤5.2 and 0.55<WHtR≤0.66 and 2h-PCPG≤7.4	0.72	Non-diabetic
	3	3 FPG≤5.2 and WHtR>0.66 and 2h-PCPG≤6.9	0.57	Non-diabetic
	4	4 FPG>5.2 and WHtR≤0.52 0.74	0.74	Non-diabetic
	5	5 FPG≤5.2 and 0.55<WHtR≤0.66 and 2h-PCPG>7.4	0.69	Diabetic
	6	6 FPG≤5.2 and WHtR>0.66 and 2h-PCPG>6.9	0.75	Diabetic
	7	7 FPG>5.2 and WHtR>0.52	0.81	Diabetic
Model (2)	1	Model (2) 1 FPG≤5.2 and WHtR≤0.55	0.88	Non-diabetic
	2	2 FPG≤4.9 and 0.55<WHtR≤0.66	0.73	Non-diabetic
	3	3 4.9<FPG≤5.2 and 0.55<WHtR≤0.66 and MAP≤97	0.64	Non-diabetic
	4	4 FPG≤5.2 and WHtR>0.66 and MAP≤99	0.59	Non-diabetic
	5	5 FPG>5.2 and WHtR≤0.52	0.74	Non-diabetic
	6	6 4.9<FPG≤5.2 and 0.55<WHtR≤0.66 and MAP>97	0.67	Diabetic
	7	7 FPG≤5.2 and WHtR>0.66 and MAP>99	0.66	Diabetic
	8	8 FPG>5.2 and WHtR>0.52	0.81	Diabetic
	9	9 FPG>5.2 and WHtR>0.56	0.84	Diabetic

Table B2. Model (1) was developed based on 20 variables which included 2h-PCPG.
Model (2) was developed based on 19 variables (2h-PCPG was excluded).
*The percentage of population in the defined subgroup, which can be interpreted as probability of an outcome.
†Predicted outcome for women who belong to the defined subgroup. 2h-PCPG, 2-hour postchallenge plasma glucose (mmol/L); FPG, fasting plasma glucose (mmol/L); MAP, mean arterial blood pressure (mm Hg); WHtR, waist-to-height ratio.
Reprinted from "Decision tree-based modelling for identification of potential interactions between type 2 diabetes risk factors: a decade follow-up in a Middle East prospective cohort study." by A. Ramezankhani, E. Hadavandi, O. Pournik, J. Shahrabi, F. Azizi, F. Hadaegh, 2016, *BMJ Open*, 6, p.12. BMJ Publishing Group 2016.

Appendix C

Table C1: The rules extracted through the random forest and decision tree models

Random Forest Model
Rl: FF TG=204.5 and bs-CRP=1.32 and occupation=employment, THEN class: a person without diabetes (187.236 or 79.2%)
R2: FF TG=204.5 and bs-CRP-:1.32 and occupation=retired and TC=:257, THEN class: a person without diabetes (43:72 or 59.7%)
R3: IF TG:204.5 and bs-CRP-:1.32 and occupation=retired and TC:-257 and LDL=:110.9, THEN class: a person without diabetes (22:23 or 91.3%)
R4: IF TG:=204.5 and hs-CRP-:1.32 and occupation=retired and TC:+257 and LDL--110.9, THEN class: a person with diabetes (5:9 ar 55.5%)
R5: IF TG=204.5 and bs-CRP-1.32 and occupation=unemployment and hs-CRP:-4.66, THEN class: a person with diabetes (9:10 or 90%)
R6: IF TG=204.5 and hs-CRP-:1.32 and occupation=unemployment and hs-CRP-4.66 and BPD~:57.9, THEN class: a person without diabetes (138'199 or 69.3%)
RU: IF TG:204.5 and bs-CRP-1.32 and occupation=mnemployment and bs-CRP-4.66 and BPD:-57.9 and FHD=yes, THEN class: a person with diabetes (14:16 or 87.5%)
R&: IF TG=204.5 and bs-CRP-:1.32 and occupation=unemployment and bs-CRP=4.66 and BPD: 57.9 and FHD=no, THEN class: a person without diabetes (25.32 or 78.1%)
RO: IF TG=204.5 and hs-CRP- 1.81 and age=46.10, THEN class: person without diabetes (569'753 or 79.1%)
R10: IF TG=204.5 and hs-CRP-1.81 and age=46.10 and HDL--67.5 and TG>227 and BMI» 24.61, THEN class: a person with diabetes (8:9 or 88.8%)
R1t: IF TG=204.5 and hs-CRP-=1.81 and age=46.10 and HDL--67.5 and TG=227 and BMI<=:24.61. THEN class: a person without diabetes (5'9 or 55.5%)

R12: IF TG=204.5 and hs-CRP-1.81 and age=46.10 and HDL: 67.5 and TG=227, THEN class: a person without diabetes (11,12 or 91.6%)
R13: IF TG=204.5 and hs-CRP--1.81 and age=46.10 and HDL=67.5 and PAL=2.18, THEN class: a person without diabetes (129/136 or 94.8%)
R14: IF TG=204.5 and hs-CRP--1.81 and age=46.10 and HDL=67.5 and PAL=:2.18 and BPS=128.16, THEN class: a person without diabetes (48 or 50%}
R15: IF TG=204.5 and hs-CRP--1.81 and age=46.10 and HDL=67.5 and PAL=2.18 and BPS:>128.16, THEN class: a person with diabetes (8'12 or 66.6%)
Decision Tree Model
Rl: IF FHD=no and TG-184, THEN class: a person without diabetes (3604/3921 or 92%)
RQ: IF FHD=no, TG::184 and age-48, THEN class: a person without diabetes (340/386 or 88%)
R3. IF FHD=no, TG=184, age-48, and hs-CRP-:2 2, THEN class: a person without diabetes (272/307 or 88%0)
R4: IF FHD=no, TG=184, age-48, and hs-CRP2.2, THEN class: a person with diabetes (100/198 or 51%)
RS. IF FHD=ves, age--48 and SBP=140. THEN class: a person without diabetes (809894 or 90%)
R6: IF FHD=yes, age-48 and SBP::140, THEN class: a person with diabetes (72/133 or 54%)
R77. IF FHD=ves, age48, SBP::130, DBP=81 and PAL--1.6, THEN class: a person without diabetes (1629 or 55%)
R8: IF FHD=yes, age48, SBP:=130, DBP=81 and PAL=:1.6, THEN class: a person with diabetes (37'47 or 79%)
RO: IF FHD=yes, age:-48, SBP:=130, DBP281, HDL=:29, THEN class: a person with diabetes (1113 or 85%)
R10: IF FHD=yes, age:48, SBP:2130, DBP::81, HDL:229, LDL<:148 and hs-CRP-6.8, THEN class: person without diabetes (96'138 or 70%)
R1t: IF FHD=yes, age48, SBP=130, DBP81, HDL=29, LDL=148, and hs-CRP:6.8, THEN class: a person with diabetes (17/33 or 52%)
R12: IF FHD=yes, age:48, SBP:-130, DBP:2831, HDL=29, LDL: 148 and occupation=employed, THEN class: a person without diabetes (7'9 or 78%)
R13: IF FHD=yes, age:48, SBP-=}30, DBP-31, HDL:=29, LDL=148 and occupation=other, THEN class: person with diabetes (34,58 or 59%)
R14: IF FHD=yes, age=48, SBP=130, BMI--23. THEN class: a person without diabetes (324'442 or 73%)
R15: IF FHD=yes, age::48. SBP=130, BMI:23 and education=low, THEN class: a person with diabetes (15/20 or 75%)
R16: IF FHD=yes, age>48, SBP=130, BMI::23 and education=high &moderate, THEN class: person without diabetes (15/26 or 58%)

Table C1. (FHD: family history of diabetes; TG: Triglycerides; DBP: diastolic blood pressure; hs-CRP: high sensitivity C-reactive protein; BMI: Body mass index; SBP: systolic blood pressure; HDL: high-density lipoprotein; LDL: low-density lipoprotein). Reprinted from "A Comparison Between Decision Tree and Random Forest in Determining the Risk Factors Associated with Type 2 Diabetes" by H. Esmaily, M. Tayefi, H. Doosti, M. Ghayour-Mobarhan, H. Nezami, A. Amirabadizadeh, *Journal of Research in Health Sciences*, 18, p.5. Open Journal Systems 2018.