Knowledge Discovery in Healthcare databases: feature selection in diabetes classification

MSc Research Project
Data Analytics

Svetlana Nicolenco
X01419950

School of Computing
National College of Ireland

Supervisor: Dr. Eugene O'Loughlin

National College of Ireland
Project Submission Sheet – 2015/2016
School of Computing

Student Name:	Svetlana Nicolenco
Student ID:	X01419950
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Dr. Eugene O'Loughlin
Submission Due Date:	22/08/2016
Project Title:	Knowledge Discovery in Healthcare databases: feature selection in diabetes classification
Word Count:	8338

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	12th September 2016

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

Knowledge Discovery in Healthcare databases: feature selection in diabetes classification

Svetlana Nicolenco
X01419950
MSc Research Project in Data Analytics

12th September 2016

Abstract

Artificial Intelligence may enhance and complement Medical Intelligence to deliver Healthcare of the 21 century. Medical databases accumulate vast amounts of data, holding potential remedies to many diseases. Data mining opens new dimensions and opportunities to the existing statistical approaches in medical domain. A better set of predictors is needed for diagnostics and classification of medical conditions and that is where feature selection become indispensable. The National Health and Nutrition Examination Surveys (NHANES) data was utilised in this research, with demographic data, details of laboratory tests and food components data accessed for knowledge discovery. Diabetes is one of the main causes of disabilities and deaths in the world and one of the disorders where causes are poorly understood. This was the main motivation for exploration of the data from a data scientist perspective. Glucose blood level (*serum glucose*) was selected as the target feature, as it is the main factor in identification of diabetes. Algorithms with feature selection may work with predictors that were never considered before but could help improve accuracy. ”Gbm”, ”glmnet”, ”svmRadial” and ”nnet” packages were applied for feature selection in this research within the R environment. Comparative analysis of features selected by different algorithms estimated with Receiver Operating Characteristic (ROC), sensitivity, specificity and kappa. Complex analysis revealed important features for predicting glucose level: blood osmolality, blood sodium and blood phosphorus level. Validation of the predictive accuracy using a ROC curve has been done on a test set with accuracy almost 87% with all modelling techniques. Pima Indian Diabetes data has been chosen as a reference against the proposed model, accuracy of 76% attained in comparison with the same methods.

1 Introduction

Data mining is a prominent field of research that potentially can change the way people extract knowledge from Big Data. Powerful computers, advanced algorithms and unlimited storage contribute towards the generation of vast amounts of data from one side, and to assist the public to manage it from the other side. Categorization, routing, filtering and searching for relevant material from complex datasets is challenging for humans, but has a great potential for machine learning. Big data in healthcare encompass additional challenges. Privacy issues, data governance and domain knowledge are constraints to hold back data from release for research and delay new discoveries. Resistance to change is another important factor of human nature. Ramesh et al. (2004) clarify that one of the reasons is the level of confidence of medical practitioners towards the technological advancement for decision making. (Smith et al.; 1988) explains that the nature of some rare diseases, its small ratio in total population, might not provide enough evidence to be statistically significant. Lately this view has been supported by Groves et al. (2013). In this case data mining might be particularly useful for knowledge discovery. Consolidation of patients records into a single file and analysing it with some advanced machine learning algorithms improves predictive accuracy in many cases. It is a source of innovation and future of scientific, evidence based medicine.

Hospitals, pharmaceutical industry, research centres and laboratories are conglomerate of Big Data production. Vast amounts of data are stored in clinical repositories, concerning patient details such as demographic, payment, as well as laboratory and treatment data (Musen and van Bemmel; 1997);(Groves et al.; 2013). The utilisation of these records became popular in prediction of insurance claims abuse (Joudaki et al.; 2014), ranking hospitals and identifying high-risk patients (Obenshain; 2004). However even now this field is considered as ”data rich and knowledge poor” domain (Raghupathi and Raghupathi; 2014). Powerful new algorithms running on high performance machines could crunch massive amounts of existing medical data in near-real time to deliver long awaited concept of personalised treatment.

According to World Health Organisation 2016 report the number of people affected by diabetes almost doubled from 1980 to 2014 and approximates to 422 million people. A condition where a main symptom is elevated glucose level was the cause of 2.2 million deaths on top of 1.5 million of people diagnosed with diabetes. Serious health complications are associated with the disease, like heart attacks, loss of vision, amputations, strokes, kidney failures result in disability and a loss of employment. As a result, people with diabetes and their families suffer from significant economic loss and overall it is a serious burden to national economies (Organization et al.; 2006).

The objective of this research is formulated as follows:

1. Exploration of large healthcare data repository for discovery of features that may contribute towards better classification accuracy of glucose level;

2. To select methods that not only give a superior accuracy, but also provide opportunities to expose significant features and its relative importance;

3. Development of feature selection process flow from complex healthcare data repositories for data mining purposes;

4. Comparative analysis of the proposed methodology with other research performed on Pima Indian Diabetes dataset.

This research paper is structured as follows. In the next chapter revision of related work within the context of feature selection and diabetes prediction is explored, common pitfalls in data analytics domain are reviewed from the medical perspective. Methodology, selection of induction algorithms and justification of evaluation are discussed in chapter 3. Chapter 4 includes implementation process flow, statistical findings and rationale for investigation. Four case studies are presented in chapter 5, along with discussion on evaluation metrics and research findings. Conclusions and efficiency of the research are presented in the last section, with indications of future work and research limitations.

2 Related Work

The size of datasets imposes additional challenges for the researchers and analysts. It is not unusual now to have datasets with millions of features (Zhai et al.; 2014). Multidimensionality has an impact on the quality and complexity of the data, where noise (Tan et al.; 2009); (Xiong et al.; 2006), outliers (Ben-Gal; 2009);(Escalante; 2005), class imbalance (Ling and Sheng; n.d.); (Galar et al.; 2012);(Weiss; 2004) became prevalent issues. Application of data mining in the field of bioinformatics and healthcare reveals an additional issue of disparity: small number of samples to enormous amount of features, such as in DNA microarray classification (Tabus and Astola; 2005), chemical structures, (Buydens et al.; 1999), medical imaging (Saeys et al.; 2007);(Fu et al.; 2005). Dimensionality reduction in this field became fundamental for success. Bellman in 1961 introduced the notion of "curse of dimensionality", where larger amount of features depreciates the generalization capability. Dimensionality reduction meant to reduce multivariate data to a subset with a smaller set of features, preserving the predictive power. Feature selection is one of the techniques for the efficient resolution of multidimensionality problem (Motoda and Liu; 2002).

2.1 Feature Selection Methods

Guyon and colleagues (Guyon et al.; 2002), (Nikravesh et al.; 2006) carried out considerable research in the field of feature selection methods for machine learning. Feature selection and feature extraction are important steps in data exploration and preparation for predictive model building, and also used as approaches for reduction of dimensionality. Differentiation between two techniques is important, as they have different objectives in model building. The purpose of feature selection is to select the best relevant predictors for building a model. Removing redundant and irrelevant features without major loss of information is the main principle behind feature selection. Conceptually, *redundancy* and *irrelevance* are distinctive notions in terms of feature selection, as correlation between two relevant attributes may return one of them to be redundant.

Feature extraction constructs new features from a given subset. Replacement of an attribute by a logarithm, cube root, reciprocal function or binning of features is considered a transformation. Certain objectives can be achieved with such transformation: scale adjustment (without the effect on the shape of the distribution), transformation of complex relationships into linear. Generation of new features from the existing ones can be done by creation of derived features or by introduction of dummy features for model optimisation. This technique could make a number of features in a dataset larger or more condensed (Nikravesh et al.; 2006).

Guyon and Elisseeff (2003) examined motivations and benefits for choosing subsets of features that are suitable to build good predictors. Among them are model simplifications that may enhance algorithm performance, reduce computational cost, cut down the storage space, minimise the cost at deployment stage, etc. Last but not least reason is knowledge discovery, where information from the model can be interpreted and analysed. Classification of methods into three categories: *filter*, *wrapper* and *embedded* are now widely referenced in the literature (Kuhn; 2012).

2.1.1 Filters

Review of feature selection methods and related references was done in the study of Saeys et al. (2007). Special emphasis in the paper is given to filter methods, with additional division of filter methods into univariate and multivariate. The attractiveness of filter methods is based on their relatively economical computational cost and usually applied as a pre-processing step in model building. In principle, filters methods estimate predictors performance without a reference to the induction algorithm. Wilcoxon tests, t-tests and ANOVA models can be used to estimate the mean variance and for modelling use the predictors that have statistically significant differences between the groups. Research of Kira and Rendell (1992); Kononenko (1994) revealed the flaw of the filter approach, where the concept of minimal subset of features of FOCUS model may have severe implications in clinical studies. As an example of such bias the Social Security Number (SSN) was selected as the only relevant feature for medical diagnosis by FOCUS method. This choice of attribute poorly performed by most algorithms and was considered irrelevant by the domain experts. Medical induction problems that were mentioned in the research of Bratko's group, prove that the attribute selected by the gain criterion ('age of patient') was judged by specialists to be less relevant than other attributes (Kononenko et al.; 1997).

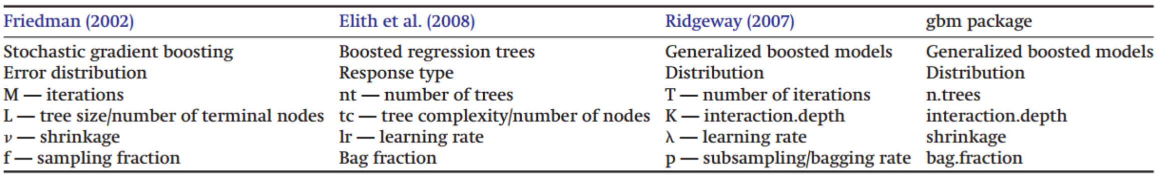
2.1.2 Wrapper

Fundamentally wrappers are algorithms that automatically optimise the output from a given set of features. It continually fits the model with various variables and the best set is governed by classification accuracy or other performance metrics. Forward, backward, stepwise, genetic algorithms, simulated annealing are feature selection methods that hypothetically may be used to assemble the optimum set of variables (Saeys et al.; 2007). Wrapper as a method for feature selection was proposed by John, Kohavi and Pfleger (John et al.; 1994). In their research they advocate that selection of features must be governed by consolidation of an algorithm and target concept. Their claim of replacement of filter models by wrapper was based on the analysis and experiment setup of FOCUS (Almuallim and Dietterich; 1991) and Relief algorithms (Kira and Rendell; 1992); (Kononenko; 1994). The evaluation of the wrapper method for feature selection was performed on nine datasets with two greedy search heuristic methods. Employing backward stepwise elimination and forward stepwise on training dataset with 25-fold cross validation did not improve the algorithm performance considerably except in a few cases, however it reduced the tree size using C4.5 algorithm. *Relief* algorithm assigns weight on the relevant features but redundant features are not targeted and thus are not eliminated from the subset. (Kira and Rendell; 1992) introduced the concept of feature relevance with two degrees of importance weak and strong, on top of irrelevant attributes for model performance.

2.1.3 Embedded Methods

Guyon in collaboration with Weston (Guyon et al.; 2002) established that ranking features goes beyond building a predictor, allowing identification relevant to cancer genes. Experimental research on gene expression data for classification outperformed the baseline method based on correlation. Innovative approach was based on Recursive Feature Elimination (RFE) within Support Vector Machine (SVM) algorithm that automatically constructs a compact subset of predictors. Proposed method is appropriate for application to DNA micro-arrays, as well as drug discovery. This method is now considered as embedded, as it tackles feature selection and learning as a whole. SVM with Radial Basis Function Kernel (RBFK) is fit for regression and classification tasks. Two parameters may be adjusted towards the model improvement: sigma and cost function. The cost (C) of the radial kernel has been stipulated to random values of five in the tuneLength function. The complexity of the borderline between support vectors is controlled by this parameter. Sigma is a smoothing parameter, that radial kernel also requires to regulate. Default setting of the train function derived estimates for sigma=0.09 and c=0.25. Additional sensitivity analysis around this two values has been done with expand.grid of carets train. Tuning the model parameters with the tuneGrid function to sigma = c(.05, 0.01, 0.005), C=c(0.10, 0.05, 0.20, 0.25, 0.15)) produced a data-frame of values that has been built with all the combination of the parameter settings.

The general approach for model fitting is a construction of a single model. Alternative approach is to build a stronger model of predictors on the basis of the combination of many models - *ensemble* technique. The ordinary approach to draw an average of models for ensemble is used in Random Forest (Qi; 2012). Different boosting models are gaining in popularity in data mining competitions for their high accuracy rate and become more and more popular in various practical applications (Hutchinson et al.; 2011). Schapire (2003) has precisely recapped the core idea of boosting method. The improvement of the model accuracy is easier to achieve by averaging all good fits than finding a single perfect prediction lead. Furthermore consecutive, forward stepwise process gives an advantage over related techniques (bagging, stacking and averaging). Predictive performance can be increased with the joined powers of two algorithms: regression trees that run as recursive binary splits and combination of numerous simple models known as boosting. Single trees that are fitted in a stepwise forward mode result in an ensemble, known as boosted regression trees. In general, the predictive performance of simple trees is mediocre. However tree based methods have considerable advances in terms of handling complex nonlinear relationships, as they are robust to outliers and missing data and do not require data transformation - time consuming pre-processing step. Building multiple trees overcomes the biggest drawback of a single tree and outperform many conventional techniques. Selection of important features in boosted regression trees is based on a relative importance of predictors and recursive feature elimination method is applied for dropping the least important (Elith et al.; 2008). Generalised Boosted Models (GBM) could be considered to be a methodological framework that can efficiently acquire complex non-linear function association. GBM, also known as "gbm" packages are the adaptations of Adaboost function of exponential loss (misclassification rate assured) with gradient descent algorithm presented by Friedman in 2000 and 2010. In the literature, algorithm is referenced under several different names. See Figure 1 for details.



Friedman (2002)	Elith et al. (2008)	Ridgeway (2007)	gbm package
Stochastic gradient boosting	Boosted regression trees	Generalized boosted models	Generalized boosted models
Error distribution	Response type	Distribution	Distribution
M — iterations	nt — number of trees	T — number of iterations	n.trees
L — tree size/number of terminal nodes	tc — tree complexity/number of nodes	K — interaction.depth	interaction.depth
ν — shrinkage	lr — learning rate	λ — learning rate	shrinkage
f — sampling fraction	Bag fraction	p — subsampling/bagging rate	bag.fraction

Figure 1: Stochastic Gradient Boosting (SGB) terms by the source of origins and model parameters (Freeman et al., 2015).

Accuracy of the target feature in GBM, is achieved by the process of consecutive fits of a new learning models. The principle idea behind this algorithm is to construct the new base-learners to be maximally correlated with the negative gradient of the loss function, associated with the whole ensemble (Natekin and Knoll; 2013). In implementation of GBM users could tune several settings. (Freeman et al.; 2015) recommends the use of Bernoulli distribution for classification tasks, as a logistic regression with binary outcome, for optimisation of the loss function. Flexibility of the GBM is achieved with their high level of customisation functions, like different loss functions, and the ideal set can be achieved experimentally. By regulating the optimization speed, initiating low-variance regression methods, and applying appreciations from hardy regression allow creation of non-parametric regression systems that allow straightforward learning from massive datasets (Freeman et al.; 2015). The simplicity of the model tree for interpretation is appreciated, however boosting combines hundreds, or even thousands of trees for the final model, making model interpretation challenging. The implementation of built-in relative importance function in gbm packages helps to prevail over this inconvenience (Elith et al.; 2008).

Shrinkage value is recommended to be set at a range 0.01 - 0.001, the lowest rate generally gives an improvement in performance of the learning algorithm. Adjusting the learning rate to a smaller value has downsides associated with computational costs and memory utilisation. Another regularization technique is controlling for the optimal number of iterations. This is done by estimating *n.trees*, which is recommended by the (Elith et al.; 2008) to be set between 3,000 to 10,000 iterations. Increasing the number may lead to model overfitting. Cross validation showed the best relative performance tested on 13 real datasets methods, so choosing 5 or 10-fold is considered a good choice with the considerateness on the run time (Ridgeway; 2012). Complexity of the model can be regulated with the penalty function. Bernoulli function is suitable for two class classification task, as it is a logistic loss function.

Friedman et al. (2010) in their study provide a comprehensive explanation of Regularization Paths for Generalized Linear Models and variation between the three approaches. The difference between ridge, lasso and elastic net is based in a distinct way how they treat correlated features. Adjustment of penalties setting gives the possibility to apply the lasso (1), ridge (2) and elastic net (fusion of two models) for valuation of generalized linear models). Collective settings of ridge regression force Regularization Paths to pull back coefficients of correlated attributes and make use of power from each other. From a Bayesian perspective, ridge penalty is particularly suitable in the case of predominance of non-zero coefficients from a Gaussian distribution. Lasso uses different approach to the same problem. It selects one correlated predictor and does not take into account the rest. Laplace prior is a mode in which two subsets has a different ratio of zero/nonzero coefficients, and it prefers the bigger one to have majorities close to zero. This principle is consistent with the Lasso penalty. Elastic net constructs a convenient compromise between ridge and lasso. In a way, elastic net acts in a similar manner to the lasso, but corrupted instances that are derived from severe correlation are removed. The increase of from 0 to 1, for a given prevalence of zero coefficients of the solution (the sparsity) increases monotonically from 0 to the sparsity of the lasso solution. One of the many real life datasets challenges is the presence of correlation between predictor attributes or where number of observations is less then number of features (genomic studies). In this situation elastic net penalty might be particularly useful. Friedman et al. (2010) suggests that a wide selection of estimates can be made from the fitted models, 'glmnet' showed a good performance with many medical datasets (Tibshirani et al.; 1997); (Ustun and Rudin; 2016).

Package 'glmnet' brings in linear, logistic, Poisson and Cox regression under one modelling 'roof'. Gaussian and grouped regression models are recent additions to the package, allowing it to handle multiple responses. The enormous size and sparse features are characteristic of many datasets, this fact poses challenges to many algorithms, but 'glmnet' is able to handle it efficiently. (Ruddock et al.; 2015) suggests that evaluation of features with non-zero coefficients gives 'glmnet' a substantial advantage in time. That is one of the reasons for computational advantage over similar algorithms. The optimization of the process originates from the cyclical coordinate descent calculated alongside a regularization path. Package glmnet implements lasso or elasticnet regularization into Generalized Linear Model (glm) for regression and classification assignments. Cyclic coordinate descent is used to turn model parameters, the regularization path is computed for the lasso or elasticnet penalty at a grid of values for the regularization parameter *lambda*. An ordinary approach for explaining convex problems is with ridge-regression penalty (alpha = 0) and the lasso penalty (alpha = 1) or elastic net mingling parameter among 0 and 1 (Friedman et al.; 2010). The order of *lambda* values can be specified by the user or there is an option that the system computes its own lambda series based on nlambda and lambda.min.ratio. Manual specification of custom values was suggested by (Friedman et al.; 2010) to sequentially decrease *lambda* i.e. (0.1, 0.01, 0.001).

Artificial Neural Network has been used in classification of human tumors in the study of Ball et al. (2002). Feature selection has been used in quick identification of biomarkers, correlating strongly to disease progression. The weights of trained Artificial Neural Network has been used as relative importance values to identify masses that accurately predict tumor grade. Neural Networks with Feature Extraction is another method of caret package that is designed for classification and regression types of problems (Tantithamthavorn et al.; 2016). The process flow designed as follows: first Principal Component Analysis (PCA) runs on a dataset and then outcome is used for modelling by a neural network. The number of components that has to be retained to capture the variance in the predictors is regulated by the thresh argument. At least two distinct values must appear in each predictor in order to run the analysis, otherwise predictor is automatically removed by the algorithm. By default the system output supports the logistic function, this setting is brought in this research considering the data characteristics. Number of units in the hidden layer has been set to size 10. Weight decay default parameter is 0, has been changed to 0.1 for this study. Default parameters have been accepted for case weights for each sample as 1 and maximum number of iterations 100.

2.2 Diabetes Prediction

A considerable amount of research has been done in diabetes prediction with the implementation of different data mining algorithms. Pima Indian diabetes dataset from University of California, Irvine (UCI) Repository of Machine Learning databases is a popular choice for testing different induction algorithms (Bache and Lichman; 2013). It is a small dataset of only 8 predictors, 768 instances and the binary outcome feature. Dataset was presented for exploration by National Institute of Diabetes and Digestive and Kidney Diseases in 1990 with an accompanying research paper. That was an earliest attempt to predict diabetes from the databases. Smith et al. (1988) explained their reasoning behind selection of attributes. Only the female population of Pima origin, aged

21 years and older have been selected on the basis that they have a substantial risk to contract diabetes within a five year timeframe. They estimated diagnosis according to criteria recommended by World Health Organization. Plasma glucose concentration after two hour carbohydrate solution intake, greater than 200 mg/dl was classified as diabetes. ADAP, an early model of neural network, has been tested on the dataset for its predicting ability on 8 input attributes to forecast the disease. ROC, sensitivity, specificity have been used as measures of effectiveness, and 0.76 was reported as the crossover point for last two parameters (Smith et al.; 1988).

Recently, Iyer et al. (2015) used this dataset in the study and two algorithms were tested: Nave Bayes and decision tree algorithm (J48) with 10 fold cross validation. Authors indicated that at pre-processing step some features were omitted. However the justification for this choice is missing, and they fail to present what attributes were excluded, and what techniques were used for feature selection. Authors claimed that results were reasonable: with pruned tree of J48 algorithm correctly classified almost 77%, Kappa statistics 0.47. Application of Nave Bayes algorithm with the same parameters gives better prediction accuracy: correctly classified 80% and Kappa 0.50. J48 is the realisation of C4.5 algorithm (Quinlan; 1993).

Another study with the same dataset with the application of wide-ranging learning algorithms, such as SVM, K-Nearest Neighbours (KNN), Nave Bayes, ID3, C4.5, C5.0, and CART was done by Farahmandian et al. (2015). According to authors, almost 82% accuracy was obtained with SVM, precision 83%, recall 90%, F-measure nearly 87%.

Parashar et al. (2014) have tried an approach for improving model efficiency through feature selection. Although there are only 8 features in this dataset, authors claim that further reducing number of predictors in the dataset results in performance improvement. Aggregation of Linear Discriminant Analysis (LDA) with Support Vector Machine gave almost 76% accuracy, and had a better execution over Feed Forward Neural Network and LDA with Feed forward networks in this study.

An extensive study of the early detection of type 2 diabetes using machine learning has been published in 2015 by Razavian and group of researchers. They considered features selection for prediction model from the insurance claims perspective. More than 42000 features were considered for model building and this data was obtained from existing health and pharmacy records and administrative files. Authors have built a model that does not require additional examinations on top of the existing data, but the full model employed hundreds of features. Experimental study demonstrated that the expanded selection of predictors and artificial intelligence tools improved accuracy. Authors claimed the enhanced model had an AUC of 0.80 in comparison of the baseline model with AUC of 0.75, but do not provide the accuracy values. Their study confirmed that L1-regularized logistic regression model that used a comprehensive number of predictors showed better results that thoroughly tuned algorithms such as Random Forests, GBM and neural networks. Their baseline model relied on a small set of features (21) area under curve of 0.75 and enhanced model with 900 features with AUC increased to 0.80. The feasibility and efficiency of the model has been proved by the fact that it has been deployed at Independence Blue Cross for the commitment of intervention allocation (Razavian et al.; 2015).

2.3 Logical Concepts in medical domain

Foundation of scientific testimony often is based on correlation of attributes. In statistical context correlation coefficient does not imply cause and effect but rather the association between features and their direction (Kenny; 1979);(Pearl and Verma; 1995). Observational and experimental clinical studies are traditionally used by researchers to confirm the correlation and establish causality between attributes. Controlling for features is feasible in experimental setups thus assumptions of cause and effect are legitimate in properly framed trials. Splitting groups on those who receive an active treatment and placebo groups is a common practice to assess the effectiveness of the therapy or medication (Little and Rubin; 2000). Caruana and Sa (2003) in their study on feature selection comment on the bias that in disease prediction (output) it is a common practice to select features for supervised learning from symptoms (inputs), despite the fact that disease initiates the symptoms. Thus in classification of medical diagnosis reversing cause and effect logic brings a doubtful usefulness for clinical research (Lucas; 1995). In medical diagnosis the risk associated with certain disorders is well examined (Wu et al.; 2014); (Joseph et al.; 2010), however the cause of the problem is not always clearly understood by the professionals. Feature selection for this task may be challenging. Lack of domain knowledge poses limitations on the scope of feature selection for the data scientist, heavily depending on the expertise of the consultant (McQueen and Thorley; 1999).

3 Materials and methods

3.1 Methodology

Cross Industry Standard Process for Data Mining (CRISP-DM) has been chosen as a comprehensive data mining guide for this research (Chapman et al.; 2000). CRISP-DM presents data mining project as an iterative process that consists of six stages: 1 - business understanding, 2 - data understanding, 3 - data preparation, 4 - modelling, 5 - evaluation and 6 - deployment.

3.2 Development environment

Details on the hardware and software, including packages that has been in used in this research are presented in configuration manual.

3.3 Evaluation Metrics

In clinical diagnostics sensitivity, specificity, false positive rate and false negative rate are common measures of accuracy of a disease (Altman; 1990). ROC is useful when comparing multiple methods. The main challenge lies outside the statistical field, it is to decide which test is clinically beneficial. Sensitivity and specificity do not provide the test probability of correct diagnosis, but their advantage is that they are not biased by prevalence of abnormality. The importance of the measure differs from the objective of the research, in epidemiological studies, screening the population for the signs of a serious disease requires high specificity and negative predictive value. However for the diagnostic of the disease high sensitivity and positive predictive value is more critical .

Kappa is another popular measure to report classification accuracy. However tte scientific community votes against the use of this measure. Several studies including (Foody; 2009), (Pontius Jr and Millones; 2011) and (Stehman; 1997) evidenced the inefficiency of its application for remote censoring and mapping. The controversy around the use of Kappa was discussed by Olofsson and team (Olofsson et al.; 2014) and suggested to use alternative measures of accuracy. Redundancy of kappa is the consequence of correlation with overall accuracy, thus limited practicality.

Understanding and calculating kappa statistics is essential in clinical trials and evidence based medicine (Sim and Wright; 2005). A common deliberation in medical domain is the agreement of observers (raters) to classify subjects into groups. 100% agreement between clinicians is considered as maximum value of 1 and kappa is used as a measure. (Altman; 1990) clarifies that using Pearsons correlation and chi2 test of association is not an appropriate measure to judge agreement. He suggests that kappa is an appropriate test and defined it as a chance-corrected proportional agreement. There is no absolute definition of generally accepted level of agreement, but values behind 0.8 and 1 is considered very good, and 0.41 and 0.60 as moderate. However limitations in this type of analysis are based on the absence of true values.

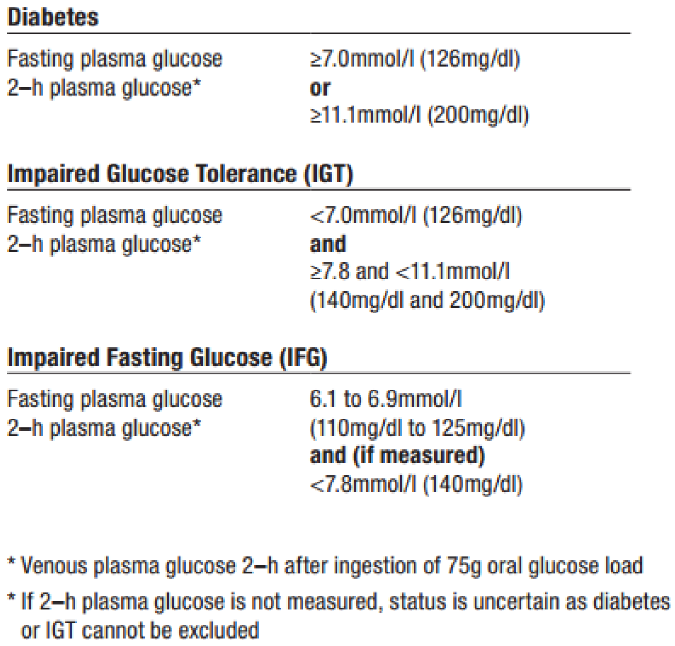
Kappa paradoxes were discovered by (Feinstein and Cicchetti; 1990) where low level of kappa can present even with high level of concordance. Data normality does not guarantee high kappa, in fact asymmetric imbalanced tables may have higher values. Another limitation of kappa statistics is its sensitivity to the distribution of the marginal sums. Since there is no agreed opinion on the use of kappa, but it used in a medical field, presenting its value alongside other measures of accuracy is considered appropriate for this study. ROC, sensitivity, specificity has been chosen as a measure of classification accuracy for all models.

3.4 Domain overview

Diabetes is a serious disorder that is associated with decreased insulin production by the pancreas or inappropriate utilisation by the body. Insulin is a hormone that regulates glucose (blood sugar). Complexity of the disorder which has two types, Diabetes type 1, Diabetes type 2 and complicated laboratory tests are required to set them apart, is an obstruction to an early diagnosis. It is approximated that type 2 diabetes among adult population is prevalent and more and more children are affected by it, however the number of undiagnosed cases is not even anticipated (Organization et al.; 2006).

World Health Organization (WHO) has published guidelines for the diagnosis and classification of diabetes. WHO issued a report definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia where venous plasma glucose should be the standard method for measuring and reporting glucose concentrations in blood. Recommendation on the fasting plasma glucose cut-point for Impaired Fasting Glucose (IFG) should remain at 6.1mmol/l and overall instructions is summarised in Figure 2 (Organization et al.; 2006).

Diagnosing of diabetes is a complex task and has controversial reference interval on glucose level to be considered normal. One important requirement for diagnostic of diabetes is fasting prior to analysis. Standard Biochemistry Profile (SBP) of blood results from the NHANES patients who's fasting status is unknown, is considered as non-fast in this study. However high level of glucose in people with undiagnosed diabetes puts them at risk too. As a consequence of these guidelines, the scope of this research is to



Diabetes

Fasting plasma glucose 2–h plasma glucose*	≥7.0mmol/l (126mg/dl) **or** ≥11.1mmol/l (200mg/dl)

Impaired Glucose Tolerance (IGT)

Fasting plasma glucose 2–h plasma glucose*	<7.0mmol/l (126mg/dl) **and** ≥7.8 and <11.1mmol/l (140mg/dl and 200mg/dl)

Impaired Fasting Glucose (IFG)

Fasting plasma glucose 2–h plasma glucose*	6.1 to 6.9mmol/l (110mg/dl to 125mg/dl) **and (if measured)** <7.8mmol/l (140mg/dl)

* Venous plasma glucose 2–h after ingestion of 75g oral glucose load

* If 2–h plasma glucose is not measured, status is uncertain as diabetes or IGT cannot be excluded

Figure 2: Guidelines for the diagnostic criteria for diabetes and intermediate hyperglycaemia. (World Health Organization, 2006).

investigate relationship between glucose level and other predictors, omitting the medical deliberations.

3.5 Methods

The choice of models has been influenced by the nature of the input data: sparse continuous input features with complex non-linear relationship for two-class classification problem require some robust algorithms. In this regard it seems reasonable to build a model directly from the dataset, with the application of non-parametric machine learning techniques. Two feature selection approaches have been applied in this study. Wrapper method has been applied with fscaret package within R environment (Szlek; 2015). it is a special package developed on top of caret package for feature selection (abbreviated from *Feature Selection CARET*. Neural networks, SVM, GBM and generalised linear models are promising as they have built-in feature selection option within caret package (Kuhn; 2012).

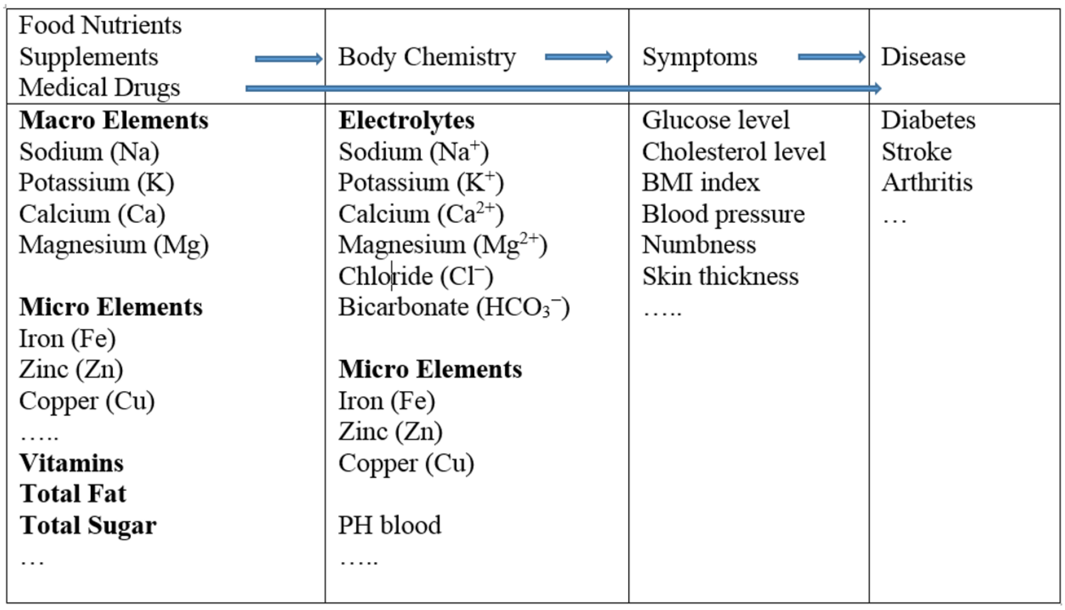
4 Implementation

4.1 Dataset narrative

Target feature in this research is *serum glucose* level from the laboratory data file (Organization et al.; 2006);(Balkau et al.; 2008). Dataset formation was dictated by the selection of the main features of interest, as the objective of the research was to investigate significant predictors for diabetes by the most advanced feature selection algorithms.

Laboratory blood analysis were performed on the population aged 12 years and older. The inclusion of main blood electrolytes and some other micro-elements from laboratory data file was dictated by the generally accepted practice of clinical experimental study design (Festing and Altman; 2002); (Nyirongo et al.; 2008).

The overview of main food, laboratory components and proposed rationale are presented in Figure 3. Initial selection criteria of Standard Biochemistry Profile was based on the correspondence of certain elements between blood samples and nutritional specifics of food. The structure of the NHANES database compels the logical sequence of clinical data used in most clinical studies under experimental settings. Similarly, the treatment of the disease should follow a reverse path from laboratory data (extended symptoms) to corrections of food / supplements / drugs.



Food Nutrients Supplements Medical Drugs	Body Chemistry	Symptoms	Disease
Macro Elements	**Electrolytes**	Glucose level	Diabetes
Sodium (Na)	Sodium (Na^+)	Cholesterol level	Stroke
Potassium (K)	Potassium (K^+)	BMI index	Arthritis
Calcium (Ca)	Calcium (Ca^{2+})	Blood pressure	…
Magnesium (Mg)	Magnesium (Mg^{2+})	Numbness	
	Chloride (Cl^-)	Skin thickness	
Micro Elements	Bicarbonate (HCO_3^-)	…..	
Iron (Fe)			
Zinc (Zn)	**Micro Elements**		
Copper (Cu)	Iron (Fe)		
…..	Zinc (Zn)		
Vitamins	Copper (Cu)		
Total Fat			
Total Sugar	PH blood		
…	…..		

Figure 3: Proposed Logical Sequence for study of diseases prediction in medical domain.

4.2 Exploratory Data Analysis (EDA)

Preliminary analysis of the data has been done with the univariate analysis and included the examination of the distribution and data normality, identification of outliers, association between features has been inspected visually. Most of the data is continuous, except demographic features, race and gender which are categorical.

General demographic characteristic of the population such as age, race, and sex were included for the purpose of general surveillance, reference point and for future work. Age feature has been binned at source to total respondents over age of 80 in one group. Gender features are weighted and has approximately equal number of males and females.

Laboratory data contains 9 features and includes blood serum glucose, osmolality, calcium, sodium, potassium, chloride, bicarbonate, iron, phosphorus, total protein. Some blood results have been reported in both (mg/dL) and (mmol/L), however main electrolytes are presented only in mmol/L. Values with the same level of measurements has been selected for comparison - mmol/L in this study. Laboratory data has missing values in iron (5 cases), total protein (7 cases), and potassium (1 case). Cases with missing data have been removed, 5497 observations were left.

Food attributes include total amount of nutrients calculated for one day, such as sodium, potassium, magnesium, calcium, iron, phosphorus, total protein, total fats etc. These values were calculated by joining the nutritional value of food table with the results of NHANES survey. Food values contain total amounts for some groups, i.e. total fat included total saturated fatty acids (gm), total monounsaturated fatty acids (gm) and total polyunsaturated fatty acids (gm), and each group has its own values. Attributes with the lowest levels of granularity have been selected for this research.

4.3 Data preparation

Mukaka (2012) recommends to use Pearson's product moment correlation coefficient in the presence of normally distributed data. However, when distribution is not normal, Spearman's correlation coefficient is more robust to the presence of outliers. Experimental results on a DNA gene reveals the effect of normalisation methods on a correlation by reducing its strength in terms of the related t-statistics. However (Qiu et al.; 2005) emphasize that the biological data has natural tendency towards associations, even feature ranking is incapable to remove correlation entirely. Given these suggestions both methods were applied, using z-transformation as a pre-processing step. Normalisation reduces the influence of scale on the results. Correlation analysis has revealed that two features osmolality and age have particularly strong relationships with the target feature blood glucose, given the size of the dataset. Osmolality and level of glucose were positively correlated r=.346, p=.000. Age and Glucose level correlated r=.269, p=.000. Spearman non parametric test reveals that age and glucose level correlated r=.379, p=.000, and osmolality with glucose r=.340, p=.000.

Correlation test has also been done in R Studio, to indicate what features are correlated more than 0.75. Analysis revealed that 24 features are strongly correlated, its nearly one third of the features. Careful examination of correlation matrix exposed an interesting fact, that feature food protein has a consistently high correlation with other features. The value of protein in the NHANES database has been given as a total, and it is a compound value of many different proteins such as arginine, cysteine, glycine, glutamine, proline, tyrosine, etc. At least 20 different amino acids are consumed with the food and minimal daily intake recommendations are given by World Health Organisation as it cannot be synthesized by the body (Joint et al.; 2007). Other attributes in the dataset have lower level of granularity, design of this survey does not provide full information for this complex feature. The importance of this feature does not seems right to remove it from the analysis.

Spearmans non parametric test reveals that food phosphorus is correlated with many food features. Test of multicollinearity statistics in SPSS confirmed the results of the finding. Two features food protein has a tolerance of .066 and VIF=15, food phosphorus tolerance 0.58 and VIF=17. Statistical approach suggests to remove highly correlated features from the analysis (Paul; 2006). However, the concepts of correlation and semantic relevance are not the same. Height and weight almost always correlated to some extent in humans, but the difference in their semantics is obvious. Semantic conflict is a challenging notion from statistical and data mining perspective (Lu et al.; 1998).

Outcome feature Glucose has been checked for normality. Glucose level data is not normally distributed, skewness = 5.00 and kurtosis 38. Histogram of frequencies distribution and Normal Q-Q plot of glucose serum level are presented in Fig.4. Common types of transformation such as, LOG10, SQRT, INVERSE, have been tested on the data, aiming at standardising the distribution. Reciprocal transformation (INVERSE) improves data normality, however this transformation obscures the interpretation capability. According to WHO recommendations (2006) it was decided in this research to bin the target feature values into 2 groups, min glucose value up to 6.1 mmol/l is group 0, and above 6.1 mmol/l to maximum is group 1 (Demchuk et al.; 1999). Most of the food data has highly skewed distribution to the right. Automatic feature transformation has been applied to the dataset and analysed for appropriateness of such transformation. After applying Box-Cox transformation method for skewed features, some of the features still were skewed to the right. The decision was made to leave features without transformation.

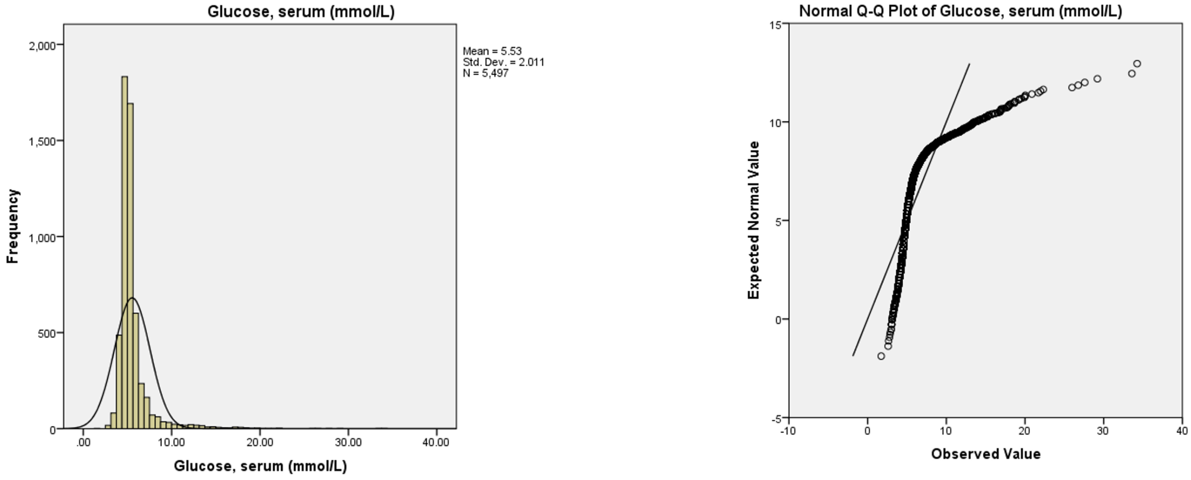


Figure 4: Histogram and Normal Q-Q plot of glucose serum level.

Dataset has been checked for outliers, based on the observations of relevant boxplots and descriptive statistics, several extreme cases have been removed as probable errors of recording. In multidimensional space it is suggested by (Mahalanobis; 1936) to check outliers with Cooks distance and Mahalanobis distance. Linear regression in SPSS provides both statistical values. There were no values below 1 in Cooks distance. Maximum value for Mahalanobis distance is 1832, just to prove the point that data distribution is not normal. One case with highest distance was removed, and all other values were below 1000. After removal of extreme cases and applying binning transformation to the glucose feature splitting it into two groups, data was checked for normality once again. A Shapiro-Wilks test (Shapiro and Wilk; 1965) was performed resulting in p=.000 and a visual inspection of their histograms, normal Q-Q plots and box plots showed that data is not normally distributed for both glucose groups. Kolmogorov-Smirnov test of normality in SPSS validates the results (p=.000) for all predictors in the dataset. After initial exploration in SPSS software data was exported into .csv format for modelling in R environment.

5 Evaluation

5.1 Experiment / Case Study 1

The idea of the fscaret package is to get the automated feature selection with the least amount of user specification utilising default settings, its average estimates of many included methods and balance calculations. Caret engine is used to build models for fscaret and to extract features importance from them directly or implicitly according to generalization error. Just ranking of features would be of little worth, since comparisons of results in the raw form could not be obtained. Application of *scale* function allows further models assessment. The final output produced a data-frame of feature importance rankings. Package fscaret is equipped with 227 different packages to perform analysis for feature selection, 129 of them are able to work on classification tasks. "glm", "gbm", "treebag", "ridge", "lasso", "rpart", "svmRadial" have been selected to do the feature selection and the top 10 features are presented in Figure 5 . Full list of features with relative influence factors is available in supplementary material. The advantages of the *wrapper* are the speed and choice of the models to test before learning the model. However wrapper does not test the accuracy of the models with chosen features.

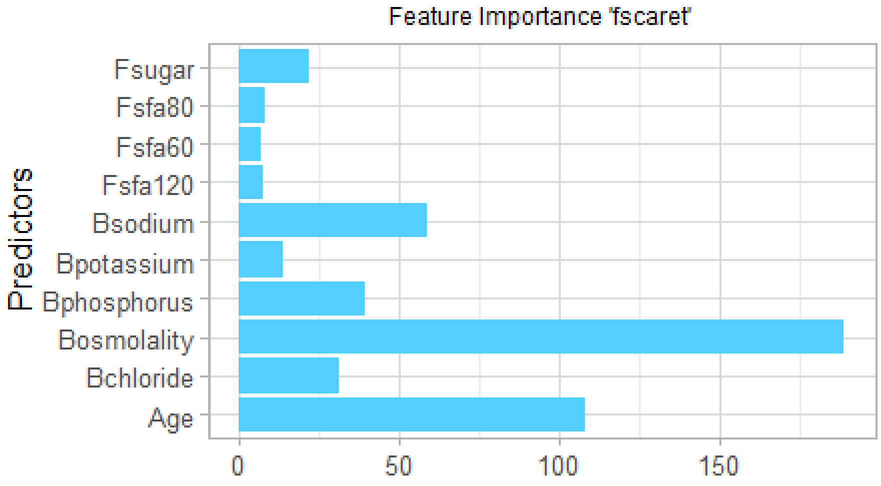


Figure 5: Feature importance with fscaret top 10 features.

5.2 Experiment / Case Study 2

Classification and regression training package caret has many build-in packages in R environment to streamline the process of model development (Kuhn; 2008). Caret functions predictors and train return a list of models along with features that were used in building the selected models. A case study on developing prediction models (Tantithamthavorn et al.; 2016) proves the caret is an efficient automated parameter optimization technique. Results presented in the paper claim that Area Under the Curve (AUC) performance has been improved by 40%, caret optimised classifiers are more stable and performance improvement of more than 80% was achieved. Recommendations on experimenting with the parameter settings have been implemented in this research. Each modelling technique was compared on the same set of evaluation metrics and the assessment of selected attributes in terms of glucose level prediction of each model. All selected models provide estimated values of relative importance of each feature. Comprehensive model assessment parameters are available from confusion matrix executed for each method. Models statistics with 95% confidence interval are presented in Figure 6, together with the plot of 10 most influential predictors based on ROC measure.

5.3 Experiment / Case Study 3

Laboratory dataset is a subset of the master dataset and consists of 11 features in total, 9 predictors, sequence number and binary outcome (0/1). Same methods and model settings have been applied on a subset of blood results with the same outcome feature - glucose group. Models statistics with 95% confidence interval are presented in Figure 7, together with the plot of relative feature importance based on ROC measure.

5.4 Experiment / Case Study 4

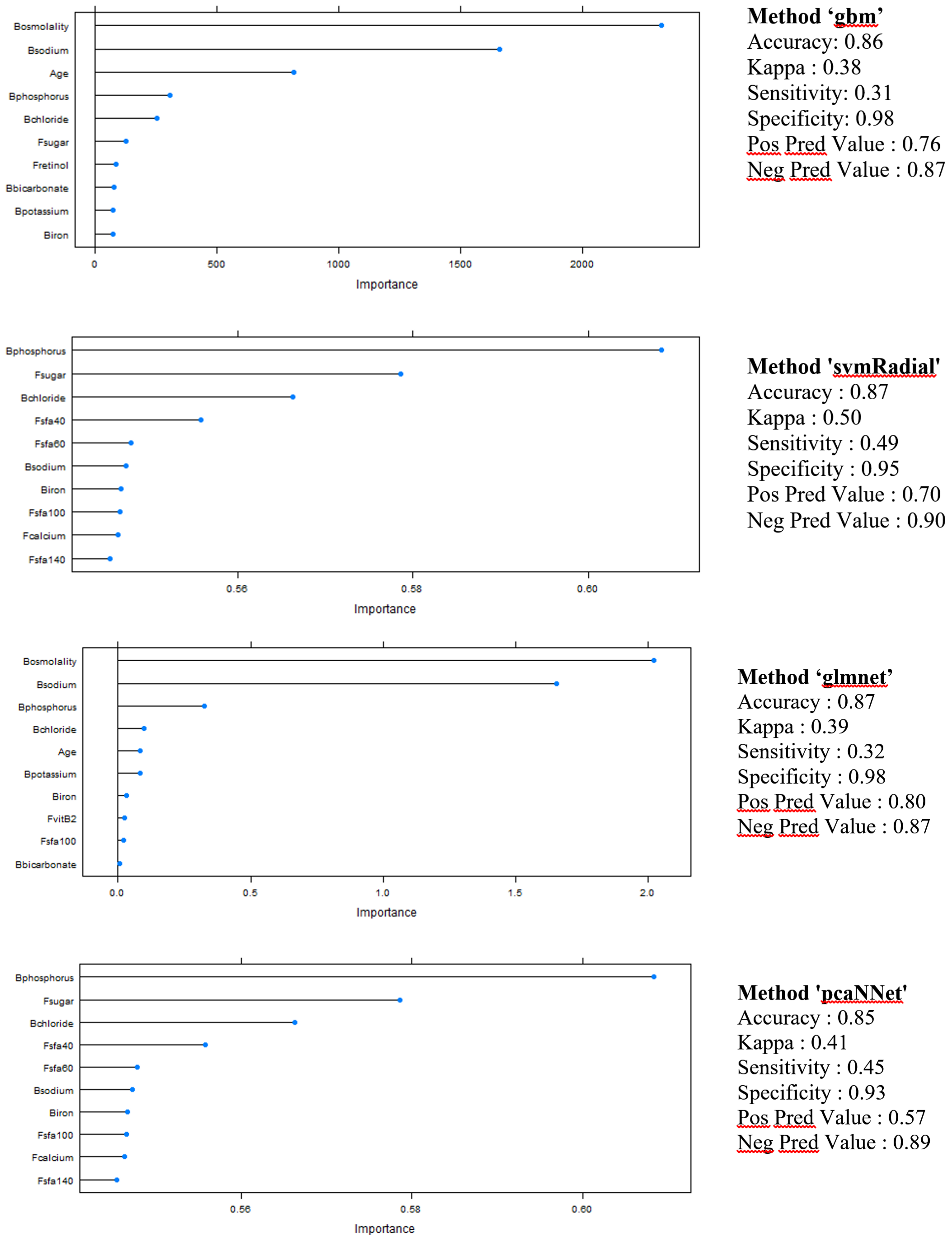
Pima Indian Diabetes dataset from UC-Irvine Machine Learning repository (Blake and Merz; 1998) has been selected for comparative analysis. The dataset has n=768 cases, with 9 numeric attributes. The target variable, diabetes onset within 5 years, has a binary outcome (0, 1). Class 0 indicate healthy patients, and class 1 cases with diabetes. Models statistics with 95% confidence interval and plot of influential predictors based on ROC measure are presented in Figure 8.

5.5 Discussion

Feature selection is a fundamental step in data mining projects and can be split in two stages. Initial feature selection has to be considered on the business understanding step and hugely depends on the objectives of the study. For the same problem, different types of features may be obtained, i.e. predicting diabetes for insurance industry may be more concerned with demographic data, and potentially would benefit from large number of attributes (Razavian et al.; 2015). However for disease diagnostics biochemical features and clinical parameters are considered an industry standard (Altman; 1990). The common objective for selection of important attributes is improvement of the learning algorithm, as it does not need all features for a good predictive model.

Second stage of feature selection is model dependent and often used as a dimensionality reduction step. Smaller set of predictors were used to improve accuracy and reduce computational costs. Caret package train function has access to many sophisticated models with built-in feature selection methods. Nonparametric methods do not require the exclusion of non-significant features, but they do involve the optimization of model settings. Parameter tuning may improve accuracy, in particular sensitivity of the final model. Running few different algorithms on the same dataset may result in differences in selected features. A typical approach to assess performance in clinical studies is to measure sensitivity, specificity, positive predictive value, negative predictive value and kappa. Same accuracy measures could be used in data mining.

Comparative analysis of Pima Indian Diabetes and NHANES derived datasets were used in this research and have been tested for accuracy. Same methods, gbm, glmnet,



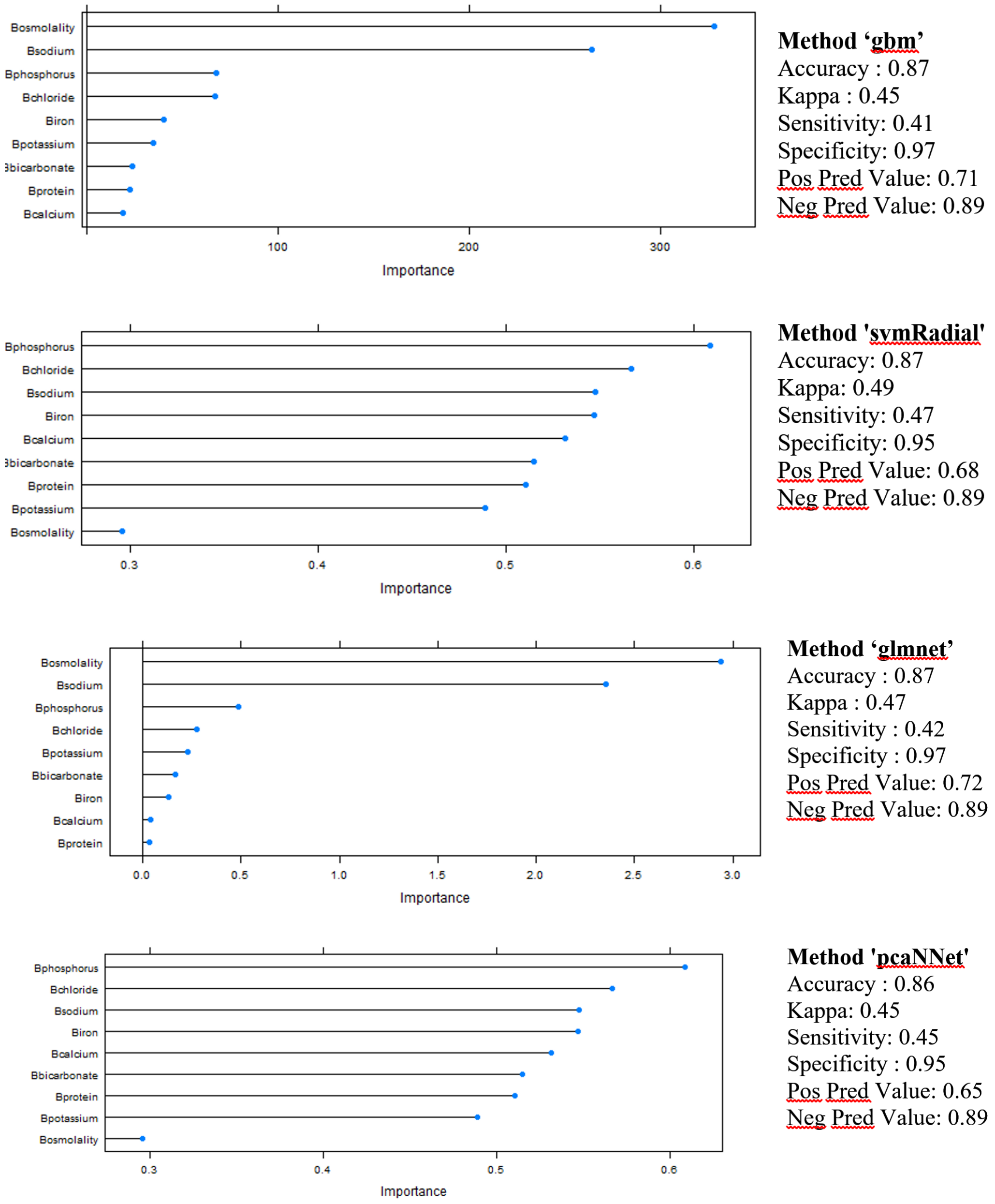
Method ‘gbm’
Accuracy: 0.86
Kappa : 0.38
Sensitivity: 0.31
Specificity: 0.98
Pos Pred Value : 0.76
Neg Pred Value : 0.87

Method 'svmRadial'
Accuracy : 0.87
Kappa : 0.50
Sensitivity : 0.49
Specificity : 0.95
Pos Pred Value : 0.70
Neg Pred Value : 0.90

Method ‘glmnet’
Accuracy : 0.87
Kappa : 0.39
Sensitivity : 0.32
Specificity : 0.98
Pos Pred Value : 0.80
Neg Pred Value : 0.87

Method 'pcaNNet'
Accuracy : 0.85
Kappa : 0.41
Sensitivity : 0.45
Specificity : 0.93
Pos Pred Value : 0.57
Neg Pred Value : 0.89

Figure 6: Feature importance with fscaret top 10 features.



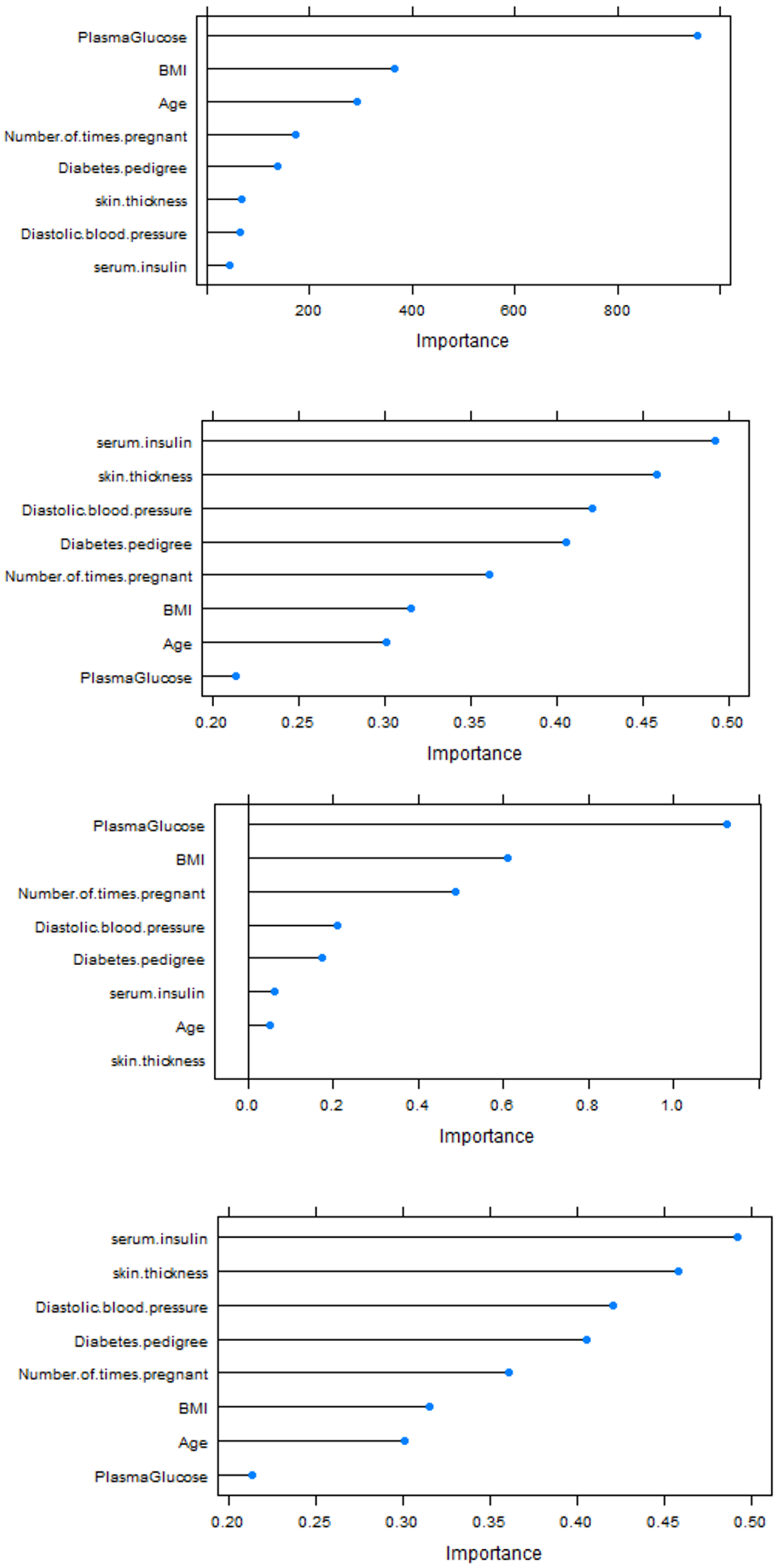
Method 'gbm'
Accuracy : 0.87
Kappa : 0.45
Sensitivity: 0.41
Specificity: 0.97
Pos Pred Value: 0.71
Neg Pred Value: 0.89

Method 'svmRadial'
Accuracy: 0.87
Kappa: 0.49
Sensitivity: 0.47
Specificity: 0.95
Pos Pred Value: 0.68
Neg Pred Value: 0.89

Method 'glmnet'
Accuracy : 0.87
Kappa : 0.47
Sensitivity : 0.42
Specificity : 0.97
Pos Pred Value: 0.72
Neg Pred Value: 0.89

Method 'pcaNNet'
Accuracy : 0.86
Kappa: 0.45
Sensitivity: 0.45
Specificity : 0.95
Pos Pred Value: 0.65
Neg Pred Value: 0.89

Figure 7: ROC feature importance plots and model statistics on Case study 3.



Method ‘gbm’
Accuracy : 0.76
Kappa : 0.45
Sensitivity : 0.61
Specificity : 0.83
Pos Pred Value : 0.66
Neg Pred Value : 0.80

Method 'svmRadial'
Accuracy : 0.76
Kappa : 0.46
Sensitivity : 0.60
Specificity : 0.85
Pos Pred Value : 0.68
Neg Pred Value : 0.80

Method ‘glmnet’
Accuracy : 0.74
Kappa : 0.40
Sensitivity : 0.54
Specificity : 0.85
Pos Pred Value : 0.65
Neg Pred Value : 0.77

Method 'pcaNNet'
Accuracy : 0.73
Kappa : 0.38
Sensitivity : 0.56
Specificity : 0.81
Pos Pred Value : 0.62
Neg Pred Value : 0.78

Figure 8: ROC feature importance plots and model statistics on Case study 4.

svm, nnet and same configuration parameters have been applied to both datasets, with considerable improvement in accuracy, 76% against 87% for HNANES data. Such an improvement can be attained with the different, more relevant set of features, as proposed in this research. Improvement in accuracy is not the only advantage of the use of stronger predictors. It allows a change of approach for disease management from reactive to proactive. Changes in the diet may prevent of the escalation of glucose level, and in the long run the containmnet of the disease. In comparison with the studies of (Iyer et al.; 2015), (Farahmandian et al.; 2015), (Parashar et al.; 2014), (Razavian et al.; 2015) where accuracy was claimed at 77%, 82%, 76%, 80% respectively, the predictive accuracy with the selected set of features in this study is considerably higher with almost 87% accuracy achieved by most methods. This result was possible to achieve only with the inclusion of the different set of features, with the access to healthcare database.

For this research all population 12-80+ years of age has been included, in order to see overall trend and let the most advanced algorithms to select the important features that may influence level of glucose. Despite the fact that demographic features *age, race* and *sex* were selected initially as features if interest, only age has been selected by fscaret, gbm and glmnet. *Race* and *Sex* do not seem to have a serious impact on glucose level. Other two packages did not select any demographic features at all. This is consistent with the study of (Emir et al.; 2015), where they get similar results with the inclusion of demographic features. Transformations in the body happen with age. Altman (1990) recommends to explore potential associations with age, in order to overcome false discovery of age associated prevalence of abnormality.

All methods have selected the blood results as the main predictor, which is in line with the suggested process flow mentioned in section Dataset narrative. Blood osmolality and blood phosphorus have been selected as the most important features in predicting glucose level with almost 87% accuracy for all methods. 'pcaNNet' and 'svmRadial' revealed blood phosphorus level as another predictor. The best sensitivity (0.47) and best Kappa (0.49) has been achieved with 'pcaNNet'. Model should include assumptions about how data was generated (i.e. whether data is of experimental or observation nature) and its correlation inferences are subject to the choice of statistical methods (Kenny; 1979). Business understanding, knowledge of the research design and its objectives are important to make the results of a study credible and useful. Correlation and causality is an important concept in medical domain where logic engineering is essential.

Exploratory Data Analysis (EDA) has identified positive correlation between glucose and blood osmolality. glmnet, gbm and fscaret have confirmed it as a top important feature. As part of EDA, a classification tree of laboratory subset has been presented in the configuration manual. Osmolality, sodium, chloride in the blood serum are main features in node splits of classification tree. In medical context, plasma osmolality is implied to evaluate the status of hyponatremia (Morley; 2015); (Drake et al.; 2015); (Gupta et al.; 2015).

6 Conclusion and Future Work

Medicine of the future could change its approach from reactive to proactive with the help of advanced data analytics tools and techniques. Evidence based medicine is a prerequisite towards scientific therapy and personalised treatment. Some government bodies foster the exploration of electronic health records and repositories for data mining, as the burden of chronic diseases has a serious negative impact on states finances (Organization et al.; 2016). Data mining allows the shift of diagnostics focus from the secondary symptoms (BMI, skin thickness, etc.) to primary symptoms (blood composition). Consequently, it provides medical professionals with information to intervene at earlier, nutrient intake stage with micro-element granularity. Benefits of proactive healthcare for the public are painless interventions in the diet along with substantial reduction of reliance on drugs. Overall proactive healthcare may result in economy growth and population prosperity.

Tuning machine learning algorithms and creation of model ensembles often brings cutting edge improvements in prediction of a pre-built datasets, i.e. Kaggle competitions. However identifying the right mix of strong relevant features when building a dataset could make a greater impact on the prediction accuracy. This research proposes an innovative way to predict average level of serum glucose based on features selected in Case study 3 (Fig 7). These features predict patients pre-disposition of diabetes irrespective of its current serum glucose results, since one of the limitations of glucose blood tests is high temporal variation of serum glucose levels, relative to other blood parameters (Moskovitch and Shahar; 2009). *Blood osmolality, blood sodium* and *blood phosphorus* have been identified as important features in predicting the level of glucose in this research, while *race* and *sex* features proved to be relatively redundant. Results of Case study 3 prove that 9 features from laboratory data, such as blood macro and microelements, were able to predict glucose level with 87% accuracy (Fig 7). Two important blood serum measures are missing in the NHANES survey data, *magnesium* and *blood PH* (Wang et al.; 2013). Inclusion of these 2 elements in the future analysis may further improve accuracy of the model, since *magnesium* is one of the four blood macro-elements (Na, K, Ca and Mg). Final conclusions on the results of this research are set aside for further analyses, clinical trials and to the judgement of medical professionals.

Based on this research, several aspects require further investigation. Data aggregation over several years may result in accuracy improvement, with special emphasis on sensitivity of the final model. Furthermore, based on the nutritional intake data collected, identification of food that contributes towards elevated glucose levels is needed. NHANES database provides data for nutrition supplements taken. Combination of supplements and food datasets may explain reasons for high osmolality levels. Limitation of the study is that survey design has an observational nature.

References

Almuallim, H. and Dietterich, T. G. (1991). Learning with many irrelevant features., *AAAI*, Vol. 91, Citeseer, pp. 547–552.

Altman, D. G. (1990). *Practical statistics for medical research*, CRC press.

Bache, K. and Lichman, M. (2013). Uci machine learning repository. university of california, irvine, school of information and computer sciences, 2013.

Balkau, B., Lange, C., Fezeu, L., Tichet, J., de Lauzon-Guillain, B., Czernichow, S., Fumeron, F., Froguel, P., Vaxillaire, M., Cauchi, S. et al. (2008). Predicting diabetes: clinical, biological, and genetic approaches data from the epidemiological study on the insulin resistance syndrome (desir), *Diabetes care* **31**(10): 2056–2061.

Ball, G., Mian, S., Holding, F., Allibone, R., Lowe, J., Ali, S., Li, G., McCardle, S., Ellis, I. O., Creaser, C. et al. (2002). An integrated approach utilizing artificial neural networks and seldi mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers, *Bioinformatics* **18**(3): 395–404.

Ben-Gal, I. (2009). Outlier detection, *Data mining and knowledge discovery handbook*, Springer, pp. 117–130.

Blake, C. and Merz, C. J. (1998). {UCI} repository of machine learning databases.

Buydens, L. M., Reijmers, T. H., Beckers, M. L. and Wehrens, R. (1999). Molecular data-mining: a challenge for chemometrics, *Chemometrics and intelligent laboratory systems* **49**(2): 121–133.

Caruana, R. and Sa, V. R. d. (2003). Benefitting from the variables that variable selection discards, *Journal of machine learning research* **3**(Mar): 1245–1264.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). Crisp-dm 1.0 step-by-step data mining guide.

Demchuk, A. M., Morgenstern, L. B., Krieger, D. W., Chi, T. L., Hu, W., Wein, T. H., Hardy, R. J., Grotta, J. C. and Buchan, A. M. (1999). Serum glucose level and diabetes predict tissue plasminogen activator–related intracerebral hemorrhage in acute ischemic stroke, *Stroke* **30**(1): 34–39.

Drake, K., Nehus, E. and Goebel, J. (2015). Hyponatremia, hypo-osmolality, and seizures in children early post-kidney transplant, *Pediatric transplantation* **19**(7): 698–703.

Elith, J., Leathwick, J. R. and Hastie, T. (2008). A working guide to boosted regression trees, *Journal of Animal Ecology* **77**(4): 802–813.

Emir, B., Masters, E. T., Mardekian, J., Clair, A., Kuhn, M. and Silverman, S. L. (2015). Identification of a potential fibromyalgia diagnosis using random forest modeling applied to electronic medical records, *Journal of pain research* **8**: 277.

Escalante, H. J. (2005). A comparison of outlier detection algorithms for machine learning, *Proceedings of the International Conference on Communications in Computing*, pp. 228–237.

Farahmandian, M., Lotfi, Y. and Maleki, I. (2015). Data mining algorithms application in diabetes diseases diagnosis: A case study, *Technical report*, MAGNT Research Report.

Feinstein, A. R. and Cicchetti, D. V. (1990). High agreement but low kappa: I. the problems of two paradoxes, *Journal of clinical epidemiology* **43**(6): 543–549.

Festing, M. F. and Altman, D. G. (2002). Guidelines for the design and statistical analysis of experiments using laboratory animals, *ILAR journal* **43**(4): 244–258.

Foody, G. M. (2009). Classification accuracy comparison: hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority, *Remote Sensing of Environment* **113**(8): 1658–1663.

Freeman, E. A., Moisen, G. G., Coulston, J. W. and Wilson, B. T. (2015). Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance 1, *Canadian Journal of Forest Research* **46**(3): 323–339.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent, *Journal of statistical software* **33**(1): 1.

Fu, J., Lee, S.-K., Wong, S. T., Yeh, J.-Y., Wang, A.-H. and Wu, H. (2005). Image segmentation feature selection and pattern classification for mammographic microcalcifications, *Computerized Medical Imaging and Graphics* **29**(6): 419–429.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. and Herrera, F. (2012). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **42**(4): 463–484.

Groves, P., Kayyali, B., Knott, D. and Van Kuiken, S. (2013). The big datarevolution in healthcare, *McKinsey Quarterly* **2**.

Gupta, E., Kunjal, R. and Cury, J. D. (2015). Severe hyponatremia due to valproic acid toxicity, *Journal of clinical medicine research* **7**(9): 717.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection, *Journal of machine learning research* **3**(Mar): 1157–1182.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines, *Machine learning* **46**(1-3): 389–422.

Hutchinson, R. A., Liu, L.-P. and Dietterich, T. G. (2011). Incorporating boosted regression trees into ecological latent variable models., *AAAI*, Vol. 11, pp. 1343–1348.

Iyer, A., Jeyalatha, S. and Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques, *arXiv preprint arXiv:1502.03774* .

John, G. H., Kohavi, R., Pfleger, K. et al. (1994). Irrelevant features and the subset selection problem, *Machine learning: proceedings of the eleventh international conference*, pp. 121–129.

Joint, W. et al. (2007). Protein and amino acid requirements in human nutrition., *World health organization technical report series* (935): 1.

Joseph, J., Svartberg, J., Njølstad, I. and Schirmer, H. (2010). Incidence of and risk factors for type-2 diabetes in a general population: The tromsø study, *Scandinavian journal of public health* **38**(7): 768–775.

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M. and Arab, M. (2014). Using data mining to detect health care fraud and abuse: a review of literature., *Global journal of health science* **7**(1): 194–202.

Kenny, D. A. (1979). Correlation and causation.

Kira, K. and Rendell, L. A. (1992). The feature selection problem: Traditional methods and a new algorithm, *AAAI*, Vol. 2, pp. 129–134.

Kononenko, I. (1994). Estimating attributes: analysis and extensions of relief, *European conference on machine learning*, Springer, pp. 171–182.

Kononenko, I., Bratko, I. and Kukar, M. (1997). Application of machine learning to medical diagnosis, *Machine Learning and Data Mining: Methods and Applications* **389**: 408.

Kuhn, M. (2008). Caret package, *Journal of Statistical Software* **28**(5).

Kuhn, M. (2012). Variable selection using the caret package, *URL {http://cran. cermin. lipi. go. id/web/packages/caret/vignettes/caretSelection. pdf}* .

Ling, C. and Sheng, V. (n.d.). Cost-sensitive learning and the class imbalance problem. 2008.

Little, R. J. and Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches, *Annual review of public health* **21**(1): 121–145.

Lu, H., Fan, W., Goh, C. H., Madnick, S. E. and Cheung, D. W. (1998). Discovering and reconciling semantic conflicts: a data mining perspective, *Data Mining and Reverse Engineering*, Springer, pp. 409–427.

Lucas, P. J. (1995). Logic engineering in medicine, *The knowledge Engineering review* **10**(02): 153–179.

Mahalanobis, P. C. (1936). On the generalized distance in statistics, *Proceedings of the National Institute of Sciences (Calcutta)* **2**: 49–55.

McQueen, G. and Thorley, S. (1999). Mining fool's gold, *Financial Analysts Journal* **55**(2): 61–72.

Morley, J. E. (2015). Dehydration, hypernatremia, and hyponatremia, *Clinics in geriatric medicine* **31**(3): 389–399.

Moskovitch, R. and Shahar, Y. (2009). Medical temporal-knowledge discovery via temporal abstraction., *AMIA*.

Motoda, H. and Liu, H. (2002). Feature selection, extraction and construction, *Communication of IICM (Institute of Information and Computing Machinery, Taiwan) Vol* **5**: 67–72.

Mukaka, M. (2012). A guide to appropriate use of correlation coefficient in medical research, *Malawi Medical Journal* **24**(3): 69–71.

Musen, M. A. and van Bemmel, J. H. (1997). *Handbook of medical informatics*, Bohn Stafleu Van Loghum Houten.

Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial, *Frontiers in neurorobotics* **7**.

Nikravesh, M., Guyon, I., Gunn, S. and Zadeh, L. (2006). Feature extraction: Foundations and applications.

Nyirongo, V., Mukaka, M. and Kalilani-Phiri, L. (2008). Statistical pitfalls in medical research, *Malawi Medical Journal* **20**(1): 15–18.

Obenshain, M. K. (2004). Application of data mining techniques to healthcare data, *Infection Control & Hospital Epidemiology* **25**(08): 690–695.

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E. and Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change, *Remote Sensing of Environment* **148**: 42–57.

Organization, W. H. et al. (2006). Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a wh.

Organization, W. H. et al. (2016). Global report on diabetes.

Parashar, A., Burse, K. and Rawat, K. (2014). A comparative approach for pima indians diabetes diagnosis using lda-support vector machine and feed forward neural network, *International Journal of Advanced Research in Computer Science and Software Engineering* **4**: 378–383.

Paul, R. K. (2006). Multicollinearity: Causes, effects and remedies, *IASRI, New Delhi* .

Pearl, J. and Verma, T. S. (1995). A theory of inferred causation, *Studies in Logic and the Foundations of Mathematics* **134**: 789–811.

Pontius Jr, R. G. and Millones, M. (2011). Death to kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment, *International Journal of Remote Sensing* **32**(15): 4407–4429.

Qi, Y. (2012). Random forest for bioinformatics, *Ensemble machine learning*, Springer, pp. 307–323.

Qiu, X., Brooks, A. I., Klebanov, L. and Yakovlev, A. (2005). The effects of normalization on the correlation structure of microarray data, *BMC bioinformatics* **6**(1): 1.

Quinlan, J. R. (1993). C4. 5: Programming for machine learning, *Morgan Kauffmann* p. 38.

Raghupathi, W. and Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential, *Health Information Science and Systems* **2**(1): 1.

Ramesh, A., Kambhampati, C., Monson, J. and Drew, P. (2004). Artificial intelligence in medicine., *Annals of The Royal College of Surgeons of England* **86**(5): 334.

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

Ridgeway, G. (2012). Generalized boosted models: A guide to the gbm package. r package vignette.

Ruddock, M. W., de Matos Simoes, R., ORourke, D., Duggan, B., Stevenson, M., OKane, H. F., Curry, D., Abogunrin, F., Emmert-Streib, F., Reid, C. N. et al. (2015). Biology and medicine, *Biol Med (Aligarh)* **8**(1): 1000260.

Saeys, Y., Inza, I. and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics, *bioinformatics* **23**(19): 2507–2517.

Schapire, R. E. (2003). The boosting approach to machine learning: An overview, *Nonlinear estimation and classification*, Springer, pp. 149–171.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples), *Biometrika* **52**(3/4): 591–611.

Sim, J. and Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements, *Physical therapy* **85**(3): 257–268.

Smith, J. W., Everhart, J., Dickson, W., Knowler, W. and Johannes, R. (1988). Using the adap learning algorithm to forecast the onset of diabetes mellitus, *Proceedings of the Annual Symposium on Computer Application in Medical Care*, American Medical Informatics Association, p. 261.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy, *Remote sensing of Environment* **62**(1): 77–89.

Szlek, J. (2015). A short fscaret package introduction with examples.

Tabus, I. and Astola, J. (2005). Gene feature selection, *Genomic Signal Processing and Statistics* pp. 67–92.

Tan, K. C., Teoh, E. J., Yu, Q. and Goh, K. (2009). A hybrid evolutionary algorithm for attribute selection in data mining, *Expert Systems with Applications* **36**(4): 8616–8630.

Tantithamthavorn, C., McIntosh, S., Hassan, A. E. and Matsumoto, K. (2016). Automated parameter optimization of classification techniques for defect prediction models, *Proceedings of the 38th International Conference on Software Engineering*, ACM, pp. 321–332.

Tibshirani, R. et al. (1997). The lasso method for variable selection in the cox model, *Statistics in medicine* **16**(4): 385–395.

Ustun, B. and Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems, *Machine Learning* **102**(3): 349–391.

Wang, S., Hou, X., Liu, Y., Lu, H., Wei, L., Bao, Y. and Jia, W. (2013). Serum electrolyte levels in relation to macrovascular complications in chinese patients with diabetes mellitus, *Cardiovascular diabetology* **12**(1): 1.

Weiss, G. M. (2004). Mining with rarity: a unifying framework, *ACM SIGKDD Explorations Newsletter* **6**(1): 7–19.

Wu, Y., Ding, Y., Tanaka, Y. and Zhang, W. (2014). Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention, *International journal of medical sciences* **11**(11): 1185.

Xiong, H., Pandey, G., Steinbach, M. and Kumar, V. (2006). Enhancing data analysis with noise removal, *IEEE Transactions on Knowledge and Data Engineering* **18**(3): 304–319.

Zhai, Y., Ong, Y.-S. and Tsang, I. W. (2014). The emerging" big dimensionality", *IEEE Computational Intelligence Magazine* **9**(3): 14–26.