Diabetes Prediction

(Multiclass classification)

Harsimran Kaur, Saumya Sinha, Shruthi Sathish, Shwetarani Shwetarani

Department of Applied Data Science, San Jose State University

DATA 240: Data Mining

Dr. Seung Joon Lee

March 31, 2023

Diabetes Prediction

Background and Motivation

Diabetes is a long-lasting metabolic disorder that affects how the human body turns food into energy. Insulin is a hormone produced naturally in the body. However, when the body doesn't make enough insulin or synthesize the produced insulin, it causes the blood sugar to rise abnormally. Diabetes can lead to various other diseases affecting different body parts. A person with diabetes is more prone to developing heart disease, eye problems, kidney damage, and nerve damage. The causes of diabetes include factors such as a sedentary lifestyle, poor eating habits, age, family history, and stress.

According to the Centers for Disease Control and Prevention (CDC), 37.3 million (1 out of 10) Americans have diabetes, and around 96 million American adults (1 out of 3) are prediabetic. People diagnosed with diabetes have been reported to suffer from high blood pressure, cholesterol, chronic kidney diseases, vision impairments, and blindness (*The Facts, Stats, and Impacts of Diabetes*, 2022). The disease affects not only the patient's physical but also their mental well-being. Even with medical science advancements, no therapy can stop diabetes from progressing.

Recently, the emergence of the machine and deep learning techniques has paved new paths to predict the chances of a person developing diabetes. An early prediction of diabetes can help identify the triggers and suggest changes to lifestyle, eating habits, etc., to slow down or stop the disease in its tracks.

Goal and Methodology

This research aims to identify the crucial factors that lead to diabetes using a data-driven approach with machine learning models. For the study, the data rich in information such as

demographics, diet, lab results, medications, and survey data about family history is collected from NHANES 2013-2014 (Centers for Disease Control and Prevention (CDC), n.d.). The nutritional status and health of adults in the United States are assessed by National Health and Nutrition Examination Survey (NHANES). The dataset is approximately 50 MB in size and contains over 1800 features. However, not all features are related to diabetes.

Diabetes can be predicted using the average blood sugar level in the body, medically known as Glycohemoglobin. A typical range of glycohemoglobin percentage is between 0% and 5.7%, 5.7% and 6.4% for prediabetes, and 6.4% and above for diabetes (*All About Your A1C*, 2018). A mix of supervised, unsupervised, statistical, and feature importance methods, namely feature correlation, Recursive Feature Elimination (RFE), ANOVA, Chi-Squared, Random Forest logistic regression (p-test), extra-tree, and mutual information, are used to identify the best set of features that can predict diabetes linked to the glycohemoglobin levels.

The best features selected are used to train and test machine learning models. Based on the literature survey conducted, the most popular classification models, decision tree (baseline model), SVC, CatBoost, Multinomial Naïve Bayes, and Multinomial logistic regression, are trained to classify the data into three target categories, namely, 0 (normal), 1 (prediabetes), and 2 (diabetes).

Data Collection

The data collected from the NHANES website consists of data in .xpt files. These files are converted to CSV using the xport python library. Alternatively, the data in CSV format curated by CDC is also available on Kaggle (Centers for Disease Control and Prevention, n.d.). The dataset consists of a CSV file each for demographics, diet, examination, labs, medications,

and questionnaires. Each file contains data for a person represented by "#SEQN," which is the sequence number. This acts as a key to merge all the data related to a person.

Demographics data consists of 47 features such as age, gender, race, country of birth, number of children, etc., as shown in Table 1.

Table 1

Sample Demographic Data

SEQN	SDDSRVYR	RIDSTATR	RIAGENDR	RIDAGEYR	RIDAGEMN	RIDRETH1	RIDRETH3	RIDEXMON	RIDEXAGM	DMQMILIZ	DMQADFC	DMDE
73557	8	2	1	69		4	4	1		1	1	
73558	8	2	1	54		3	3	1		2		
73559	8	2	1	72		3	3	2		1	1	
73560	8	2	1	9		3	3	1	119			
73561	8	2	2	73		3	3	1		2		
73562	8	2	1	56		1	1	1		1	2	
73563	8	2	1	0	5	3	3	2	6			
73564	8	2	2	61		3	3	2		2		
73565	8	1	1	42		2	2			2		
73566	8	2	2	56		3	3	1		2		
73567	8	2	1	65		3	3	2		2		
73568	8	2	2	26		3	3	2		2		
73569	8	1	2	0	10	5	7					
73570	8	2	2	9		5	7	1	109			
73571	8	2	1	76		3	3	1		2		
73572	8	2	2	10		4	4	1	125			

Dietary data consists of 168 features with information about total nutrient intake in terms of carbohydrates, protein, fiber, iron, and total energy, as shown in Table 2.

Table 2

Sample Dietary Data

SEQN	WTDRD1	WTDR2D	DR1DRSTZ	DR1EXMER	DRABF	DRDINT	DR1DBIH	DR1DAY	DR1LANG	DR1MNRSP	DR1HELPD
73557	16888.327864	12930.890649	1	49	2	2	6	2	1	1	13
73558	17932.143865	12684.148869	1	59	2	2	4	1	1	1	13
73559	59641.81293	39394.236709	1	49	2	2	18	6	1	1	13
73560	142203.069917	125966.366442	1	54	2	2	21	3	1	1	12
73561	59052.357033	39004.892993	1	63	2	2	18	1	1	1	13
73562	49890.828664	0	1	49	2	1	11	3	1	1	13
73563	31417.217097	40735.782424	4	54	1	2	2	3	1	2	13
73564	78988.755072	52173.157754	1	54	2	2	12	7	1	1	13
73566	30697.88078	0	1	49	2	1	3	2	1	1	13
73567	44503.03602	0	1	61	2	1	16	7	1	1	13
73568	56404.073021	43955.549706	1	87	2	2	12	7	1	1	13
73570	31176.247118	24410.40789	1	22	2	2	15	2	1	2	1
73571	59076.733601	39020.994051	1	25	2	2	12	7	1	1	13
73572	32965.832195	25811.618932	1	61	2	2	8	4	1	1	2

Lab data contains 424 features that cover various test results, albumin, glycohemoglobin, insulin, vitamins, urine tests, and many other lab work results. Table 3 shows the data contained in the lab dataset.

Table 3

Sample Lab Data

SEQN	URXUMA	URXUMS	URXUCR.x	URXCRS	URDACT	WTSAF2YR.x	LBXAPB	LBDAPBSI	LBXSAL	LBDSALSI	LBXSAPSI	LBXSASSI
73557	4.3	4.3	39	3447.6	11.03				4.1	41	129	16
73558	153	153	50	4420	306				4.7	47	97	18
73559	11.9	11.9	113	9989.2	10.53	142196.890197	57	0.57	3.7	37	99	22
73560	16	16	76	6718.4	21.05							
73561	255	255	147	12994.8	173.47	142266.006548	92	0.92	4.3	43	78	36
73562	123	123	74	6541.6	166.22				4.3	43	95	24
73563												
73564	19	19	242	21392.8	7.85	134054.10976	77	0.77	3.9	39	72	20
73566	1.3	1.3	18	1591.2	7.22				4.1	41	93	23
73567	35	35	215	19006	16.28				4	40	67	29
73568	25	25	31	2740.4	80.65	216002.505861	59	0.59	4.5	45	65	20

As shown in Table 4, the examinations dataset with 224 features contains results of physical examinations conducted on a person, such as body measures, blood pressure, oral health, grip test, taste, smell, etc.

Table 4

Sample Examination Data

SEQN	PEASCST1	PEASCTM1	PEASCCT1	BPXCHR	BPAARM	BPACSZ	BPXPLS	BPXPULS	BPXPTY	BPXML1	BPXSY1	BPXDI1	BPAEN1
73557	1	620			1	4	86	1	1	140	122	72	2
73558	1	766			1	4	74	1	1	170	156	62	2
73559	1	665			1	4	68	1	1	160	140	90	2
73560	1	803			1	2	64	1	1	130	108	38	2
73561	1	949			1	3	92	1	1	170	136	86	2
73562	1	1064			1	5	60	1	1	180	160	84	2
73563	1	90		152				1					
73564	1	954			1	5	82	1	1	150	118	80	2
73566	1	625			1	4	86	1	1	140	128	74	2
73567	1	932			1	3	70	1	1	170	140	78	2
73568	1	585			1	3	70	1	1	120	106	60	2
73570	1	710			1	2	78	1	1	130	102	44	2

The questionnaire data set contains 953 features that answer questions collected through a survey, such as family history, existing conditions, etc., as shown in Table 5.

Table 5

Sample Questionnaire Data

SEQN	ACD011A	ACD011B	ACD011C	ACD040	ACD110	ALQ101	ALQ110	ALQ120Q	ALQ120U	ALQ130	ALQ141Q	ALQ141U	ALQ151	ALQ160
73557	1					1		1	3	1	0		1	
73558	1					1		7	1	4	2	1	1	0
73559	1					1		0					2	
73560	1													
73561	1					1		0					2	
73562				4		1		5	3	1	0		2	0
73563														
73564	1					2	1	2	3	1	0		2	
73565				5										
73566	1					1		1	1	1	0		2	
73567	1					1		4	1	3	0		2	
73568	1					1		2	1	2	1	3	2	0

Lastly, the medications dataset with 13 features contains details about any current or past medications a person takes, including prescribed, over-the-counter, and multivitamins. Table 6 shows a sample of the medication dataset.

Table 6

Sample Medication Data

SEQN	RXDUSE	RXDDRUG	RXDDRGID	RXQSEEN	RXDDAYS	RXDRSC1	RXDRSC2	RXDRSC3	RXDRSD1
73557	1	99999							
73557	1	INSULIN	d00262	2	1460	E11			Type 2 diabetes melli
73558	1	GABAPENTIN	d03182	1	243	G25.81			Restless legs syndro
73558	1	INSULIN GLARGINE	d04538	1	365	E11			Type 2 diabetes melli
73558	1	OLMESARTAN	d04801	1	14	E11.2			Type 2 diabetes melli
73558	1	SIMVASTATIN	d00746	1	61	E78.0			Pure hypercholestero
73559	1	INSULIN ASPART	d04697	1	365	E11			Type 2 diabetes melli
73559	1	INSULIN GLARGINE	d04538	1	4380	E11			Type 2 diabetes melli
73559	1	PANCRELIPASE	d01002	1	365	K86.9			Disease of pancreas,
73559	1	SIMVASTATIN	d00746	1	2920	E78.0			Pure hypercholestero
73559	1	VALSARTAN	d04113	1	3650	I10			Essential (primary) hy

As seen in the data samples, some features are discreet, some categorical, some textual, and others continuous. Most of these features are measured in different units, and since it is

survey data, there are a lot of missing values. One of the most critical takes is to clean, transform, and normalize the dataset to use various data mining and machine learning techniques. Also, the column names must be renamed to an understandable format using the metadata provided on the NHANES website.

Literature Survey

Data mining and machine learning techniques for predicting diabetes have garnered significant attention recently. An overview of factors affecting diabetes and existing practices used to predict diabetes that supports the selection of data mining techniques and machine learning models used in this research is provided in this section.

According to the dissertation submitted by Ouyang (2007) at Tufts University, diabetes is a serious, severe, and expensive public health issue. In Taiwan, it has been considered the fourth most common cause of death among people. Hence, education about self-care measures is essential for the clinical management of people with diabetes. Adopting these measures and implementing the guidelines enhances life expectancy and quality of life.

The Project implemented three main techniques to improve self-care measures among people. The first technique was to identify the frequency at which patients participated in seven self-care activities related to their diet. An evaluation was conducted to analyze how demographic, psychological, and environmental aspects influenced these behaviors. The Project was conducted by examining 180 patients who were 40 years and above in the country and had at least one prescribed nutrition session. The patients then took a questionnaire survey, and the effectiveness of self-care habits based on nutritional intake was predicted using logistic regression.

As part of the second technique, the Factors Affecting Diabetes Self-Care (FADSC) questionnaire's psychometric properties were validated for the patients. This tool comprises eight subscales that assess various aspects such as knowledge, attitudes, self-efficacy, psychological difficulties, understanding, environmental barriers, family, and medical assistance. Cronbach's alpha was employed to evaluate the internal consistency, while reliability was assessed using Spearman correlations.

The third factor aimed to investigate the influence of various factors such as background traits, the frequency at which five diabetes self-care behaviors were performed, and their impact on clinical outcomes. The survey used a scale from never to always to evaluate patients engaging in five diabetic self-care behaviors such as diabetes meal plans, exercising, taking medications, testing blood glucose levels, and inspecting one's feet. The result showed 79% of patients took their medication on time. Apart from influencing how these actions were executed, demographic, psychological, and environmental factors also played a role in determining the clinical and health outcomes.

The research by Woldemichael and Menaria (2018) discusses the links of diabetes to other health problems such as kidney disease, heart issues, and blindness. Data mining has been proven to be a practical technique for supporting medical decisions making for better accuracy of predictions. This has helped to reduce the load on professionals. The backpropagation algorithm is utilized to know about a patient is diabetic or not. The other techniques used to determine the classification are J48, Naive Bayes, and SVM. The dataset used is PIMA India and was conducted using R language. Overall performance was observed with 83% accuracy, 86% sensitivity, and 76% specificity. These results were acquired by comparing each of these models. The chi-square test feature selection technique was utilized to select significant features. The

features glu, age, bmi, serum, Npreg, and skin were the most important attributes for predicting diabetes disease and which was selected during the Chi-squared test.

The study conducted by Zou et al. (2018) aimed to develop a machine learning-based model to predict the occurrence of diabetes mellitus in patients accurately. The data used in the study comprised 145,060 patient records collected from a clinical laboratory in Iran, including demographic information, laboratory test results, and medical history. The methodology used in the study involved feature selection using statistical analysis techniques and machine learning algorithms such as logistic regression, SVM, and decision trees to develop the predictive model.

The study reported an AUC score of 0.90, indicating that the developed model had good discriminative power and could accurately predict the occurrence of diabetes mellitus in patients. The authors concluded that machine learning techniques could be effectively used to predict the occurrence of diabetes mellitus, aiding in early diagnosis and treatment of the disease. However, the study had some limitations, including the lack of external validation of the developed model and using a relatively small dataset from a single clinical laboratory in Iran. Therefore, further studies using larger and more diverse datasets are needed to validate the findings of this study and improve the generalizability of the developed model.

The goal of research conducted by Abbas et al. (2019) is to develop a machine learning model that can predict the risk of diabetes in healthy individuals. To achieve this goal, the authors used electronic health records of 75,054 individuals, which included demographic information, laboratory test results, and medical histories. The dataset was collected from a large healthcare organization in the United States, and the team used statistical analysis techniques to select relevant features from the dataset. The authors employed various machine learning algorithms, such as logistic regression and decision trees, to create a predictive model with the

highest AUC score of 0.89, indicating machine learning models can aid in early diagnosis and prevention of the disease. However, the authors also noted that the dataset used in the study was limited to a single healthcare organization in the United States, which may affect the model's generalizability to other populations or healthcare settings. Therefore, further studies using more diverse and extensive datasets are needed to validate the findings of this study.

The paper by Kavakiotis et al. (2017) aims to highlight the potential of machine learning and data mining techniques for improving diabetes diagnosis and management. The authors used various datasets from different sources, including electronic health records, clinical trials, population studies, and publicly available datasets, such as the National Health and Nutrition Examination Survey (NHANES), to illustrate the application of machine learning techniques in diabetes research.

They provided an overview of various machine learning algorithms, including decision trees, random forests, support vector machines (SVM), and neural networks, applied in diabetes research. The authors highlighted using machine learning techniques in personalized medicine, such as developing customized treatment plans based on individual patient characteristics. As a future scope, the research suggested that future studies should focus on enhancing machine learning and data mining techniques in diabetes research.

Gupta and Goel (2020) used KNN classifiers for predicting diabetes where the K value is optimized with the relevant feature selection process. In this study, It was observed that KNN classifiers worked better for small datasets. A classifier's effectiveness is assessed using various measures such as accuracy, precision, recall, specificity, and F1 score. This paper aims to improve the efficiency of KNN classifiers with the help of the feature selection process, optimizing K value, and normalization. KNN has the advantage of not requiring a training phase,

as the whole dataset is used to classify the model during the testing phase. Its accuracy will not be affected if new data samples are added to the existing dataset. Another element that affects the KNN classifier is the value of the number of neighbors. The dataset used in this study is PIMA Indian diabetes dataset. Feature selection steps involved is ANOVA which is a correlation-based feature selection process. ANOVA determines the relationship between each dataset feature and the target variable. It uses Chi-square and F1 scores to determine the value. The most appropriate features are determined by ANOVA, which computes the test scores for each characteristic by comparing their test results. In this research, KNN performance is done within a range of 1 to 60 as the K value depending on performance.

A research paper by R et al. (2020) discusses how feature selection is essential for a better-performing model. This paper uses a cancer dataset where a limited number of features are found helpful in identifying cancer patients. For predicting with accuracy, there is a need for a perfect model per the dataset, so the author found that ensemble models will perform ideally for this dataset. The models are built for predicting cancer by combining supervised machine learning algorithms and the currently available ensemble methodologies. For this task, the author implemented the following classifiers: Naive Bayes, Support Vector Machine (SVM), Decision Tree, Multi-Layer Perceptron (MLP), Logistic Regression, and K-Nearest Neighbors (KNN), along with feature selection techniques such as f-test and variance thresholding.

F-test, a statistical test, is used to determine the F score. This score is calculated by taking ratios of variation between groups to variance within a group. A higher f-score suggests a more significant distance between groups and a smaller distance within groups. ANOVA uses the F-test technique in this paper. The author talks about three ensemble techniques, i.e., bagging, boosting, and stacking. Wrapper methods are used with the subset of features, and then the

model is trained based on selected features. This method is primarily used when models don't perform better, even after using an ensemble model and feature selection. The three datasets used are Wisconsin, WDBC, and microRNA datasets. The results showed tremendous improvement in accuracy by using F-test while variance thresholding either retained or reduced accuracies. The mean calculation of the F Score and variance wasn't that impactful. Hence F-test was used for further experiments.

References

Abbas, H. T., Alic, L., Rios, M., DeFronzo, R. A., & Qaraqe, K. A. (2019). Predicting Diabetes in Healthy Population through Machine Learning. *Computer-Based Medical Systems*. https://doi.org/10.1109/cbms.2019.00117

All About Your A1C. (2018, August 21). Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/managing/managing-blood-sugar/a1c.html

Centers for Disease Control and Prevention (CDC). (n.d.). NHANES 2013-2014 [dataset]. In *National Health and Nutrition Examination Survey*. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2013

Centers of Disease Control and Prevention. (n.d.). *Kaggle NHANES datasets from 2013-2014* [dataset]. https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey

Gupta, S., & Goel, N. (2020). Performance enhancement of diabetes prediction by finding optimum K for KNN classifier with feature selection method. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*. https://doi.org/10.1109/icssit48917.2020.9214129

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, *15*, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

Ouyang, C. (2007). *Factors affecting diabetes self-care among patients with type 2 diabetes in Taiwan* [Dissertation]. Tufts University.

R, D., Paul, I. R., Akula, S. K., Sivakumar, M., & Nair, J. J. (2020). F-test feature selection in Stacking ensemble model for breast cancer prediction. *Procedia Computer Science, 171*, 1561–1570. https://doi.org/10.1016/j.procs.2020.04.167

The Facts, Stats, and Impacts of Diabetes. (2022, June 20). Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/library/spotlights/diabetes-facts-stats.html

Woldemichael, F. G., & Menaria, S. (2018). Prediction of Diabetes Using Data Mining Techniques. *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*. https://doi.org/10.1109/icoei.2018.8553959

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Zhang, J. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. *Frontiers in Genetics*, *9*. https://doi.org/10.3389/fgene.2018.00515