Université Jean Monnet Saint-Etienne

Data Mining Final Project

Diabetes Health Indicators

Presented by:
Alejandro Carvajal Montealegre

Presented to:
Pr. Fabrice Muhlenbach

April 25, 2024

Contents

1 Introduction

Data Mining is defined as the application of computational techniques to extract useful pattern or knowledge from the given data. The output of a data mining algorithm is typically a pattern or a set of patterns that are valid in the dataset. [5]. The two most used techniques for data mining are: supervised machine learning and pattern discovery.

In this project, The aim is to analyze a real-life dataset, providing a comprehensive description of its characteristics including origin, number of features, and observations. Subsequently, A data analysis will be conducted thorough the dataset to identify outliers, determine statistical measures, and uncover other significant insights. Finally, supervised machine learning algorithms will be applied to the labeled dataset, thereby constructing classifiers. These classifiers will then be compared, and their performance will be evaluated using various metrics such as accuracy, F1-score, among others.

2 Data description

This section will describe how the dataset can be accessed, how it was created and compiled by the author, and which features were incorporated, including their type and range of values.

2.1 Data origin

Diabetes mellitus is a general term for heterogeneous disturbances of metabolism for which the main finding is chronic hyperglycaemia. The cause is either impaired insulin secretion or impaired insulin action or both [7]. This can lead to serious damage to the heart, blood vessels, eyes, kidneys and nerves.

Diabetes in the United States is a serious problem. According to the American Diabetes Association, in 2021, 38.4 million Americans, or 11.6% of the population, had diabetes. Diabetes was also the eighth leading cause of death in the United States in 2021 based on the 103,294 death certificates in which diabetes was listed as the underlying cause of death [2].

The Behavioral Risk Factor Surveillance System (BRFSS) is a large state-based telephone survey made in the United States and collected anually by the Center of Disease Control and Protection (CDC). It's designed to monitor the leading risk factors for morbidity and mortality in the United States at the local, state, and national levels. Each year, more than 400,000 Americans answer the survey via telephone [9]. The historical annual survey data can be found in the Center for Disease Control and Prevention (CDC) website: https://www.cdc.gov/brfss/annual_data/annual_data.htm

The dataset used for this project is a refined version of the 2015 dataset. The original dataset contained 330 features, where some of them are redundant, noisy or dependent from other features. The most relevant features were selected and transformed to numerical values by [Teboul, Alex] [6]. The resulting dataset contains 253,680 observations without missing values, 21 features and distinct output classes: 'diabetes' or 'no diabetes'. Among these observations, 86.07% are categorized as 'no diabetes', while 13.93% belong to the 'diabetes' class, indicating an imbalance within the dataset.

2.2 Feature description

The dataset comprises 21 features along with a binary class label. Among these features, all 21 are numerical variables, with 14 of them taking on binary values. Below is a table detailing the definition of each feature, its type, possible values, and range. Information used to fill out this table was extracted from [4].

Table 1: Variable Roles and Types

Variable Name	Role	Type	Description
Diabetes_binary	Target	Binary	0 = no diabetes, 1 = pre-diabetes or diabetes
HighBP	Feature	Binary	0 = no high BP, 1 = high BP
HighChol	Feature	Binary	0 = no high cholesterol, 1 = high cholesterol
CholCheck	Feature	Binary	0 = no cholesterol check in 5 years, 1 = yes cholesterol check in 5 years
BMI	Feature	Integer	Body Mass Index

Table 1 (continued)

Variable Name	Role	Type	Description
Smoker	Feature	Binary	Have you smoked at least 100 cigarettes in your entire life? 0 = no, 1 = yes
Stroke	Feature	Binary	Have you had a stroke. 0 = no, 1 = yes
HeartDiseaseorAttack	Feature	Binary	coronary heart disease (CHD) or myocardial infarction (MI) 0 = no, 1 = yes
PhysActivity	Feature	Binary	physical activity in past 30 days - not including job 0 = no, 1 = yes
Fruits	Feature	Binary	Consume Fruit 1 or more times per day 0 = no, 1 = yes
Veggies	Feature	Binary	Consume Vegetables 1 or more times per day 0 = no, 1 = yes
HvyAlcoholConsump	Feature	Binary	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no, 1 = yes
AnyHealthcare	Feature	Binary	Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no, 1 = yes
NoDocbcCost	Feature	Binary	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no, 1 = yes
GenHlth	Feature	Integer	Would you say that in general your health is: scale 1-5, 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
MentHlth	Feature	Integer	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? scale 1-30 days
PhysHlth	Feature	Integer	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? scale 1-30 days
DiffWalk	Feature	Binary	Do you have serious difficulty walking or climbing stairs? 0 = no, 1 = yes
Sex	Feature	Binary	Sex 0 = female, 1 = male
Age	Feature	Integer	13-level age category, 1 = 18-24 9 = 60-64 13 = 80 or older
Education	Feature	Integer	Education Level, scale 1-6. 1 = Never attended school, 2 = Grades 1 through 8 (Elementary), 3 = Grades 9 through 11, 4 = Grade 12 or GED (High school graduate), 5 = College 1 year to 3 years, 6 = College 4 years or more (College graduate)
Income	Feature	Integer	Income scale 1-8, 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more

3 Data analysis

In this section, the dataset will be analyzed to detect outliers, possible correlation between variables, dimensional reduction (PCA).

3.1 Correlation matrix

A correlation matrix is a square matrix that contains correlation coefficients between variables. It provides a comprehensive overview of how each variable in a dataset relates to every other variable, making it a valuable tool for understanding patterns and associations within the data.

The Phi correlation coefficient (ϕ) measures the association between two binary variables by calculating the frequency of concordant and discordant pairs in a contingency table. The Phi coefficient ranges from -1 to 1, where -1 indicates a perfect negative association, 0 indicates no association, and 1 indicates a perfect positive association [1].

The dataset features were categorized into two groups: numerical and binary. The correlation matrix was computed for the numerical features, while the Phi correlation coefficient was employed for the binary features. The resulting correlations are as follows:

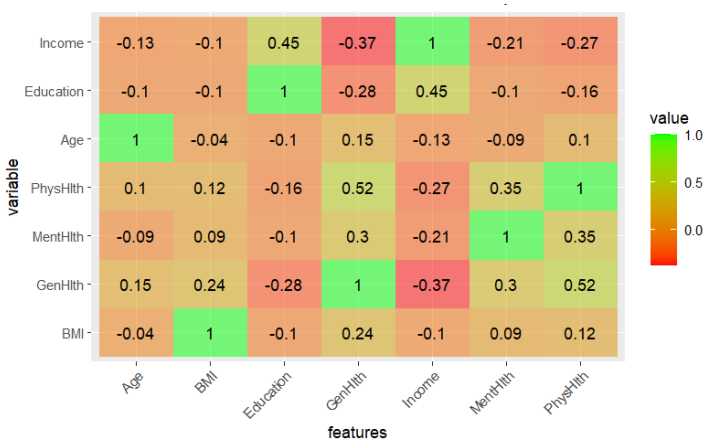

Figure 1: Correlation matrix of numerical features.

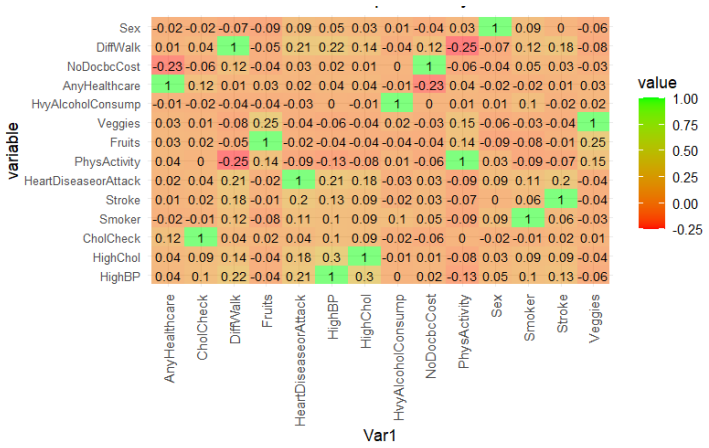

Figure 2: Phi coefficient matrix (ϕ) for binary features.

From the correlation matrices obtained, it ca be concluded the following:

- The features Age and Body Mass Index (BMI) don't show a strong correlation with other numerical features.
- Education and Income have a positive correlation.
- Income and GenHlth have a negative correlation. The lower the value of GenHlth, the higher the Income is. It's important to remember that GenHlth scales 1-5, being 1 excellent health. This means that people with higher income also have better health.

- The strongest correlation of the numerical matrix is between GenHlth and PhysHlth. This can lead to the conlcusion that a good physical health and exercising leads to a general good health condition.
- The correlation matrix for binary features doesn't show strong correlation among features. This can also be because of the nature of the variables.
- The most remarkable correlations (even though they are not strong) are: DiffWalk and HighBP, DiffWalk and HeartDissease with correlation coefficients (ρ) $0.2 \leq \rho \leq 0.3$. In other words, it means that there is a slight correlation between having difficulties to walk and having a high blood pressure and a hearth dissease.
- The same situation happens with negative correlation, where there are no strong negative-correlated features. However some remarkable negative correlations are: PhysActivity and DiffWalk. Meaning that people who exercise, don't have difficulties to walk.

3.2 Outliers

To identify outliers within the dataset's seven numerical variables, box plots will be employed. These plots offer valuable insights by utilizing the inter-quartile range. Specifically, the lower whisker corresponds to the first quartile, while the box represents the second and third quartiles. The median is indicated by the line within the box, and the upper whisker pertains to the fourth quartile [8]. The results obtained were the following:

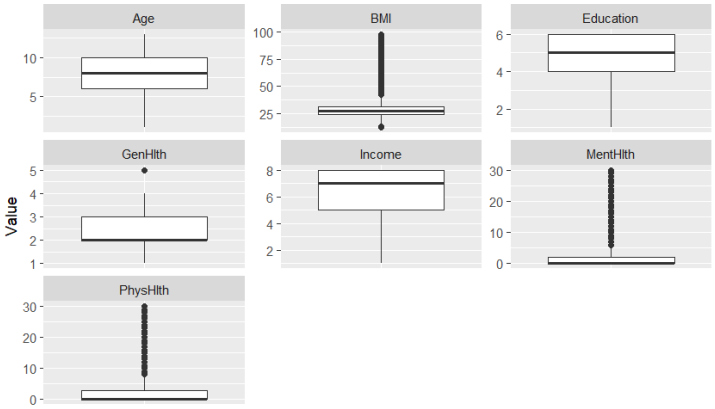

Figure 3: Box plots of numerical features.

By analyzing the box, it can be concluded the following:

- The BMI feature has a lot of outliers. According to the Center for Disease Control and Prevention, a healthy weight typically falls within the range of 18.5 to 25. Values below 18.5 indicate underweight, while those exceeding 25 suggest overweight [3]. However, values close to 0 or 100 are likely errors, as they are unrealistic.
- he GenHlth variable indicates outliers for observations with a value of five. However, due to the limited range of possible values (1-5), these observations may not be true outliers.
- The variable MentHlth shows that the majority of the population appears not to have experienced emotional problems in the last 30 days. Instances where individuals reported 10 days or more are considered outliers.
- A similar pattern is observed with PhysHlth. Individuals who exercise one day or more are identified as outliers, given that approximately 70% report no exercise at all.
- The remaining variables present in the box plot show a normal behavior without many outliers.

The Z-score is a standardized measure that allows you to compare data points from different distributions. It measures how many standard deviations a data point is away from the mean of a dataset

[11]. In the context of this dataset, an observation is identified as an outlier if its Z-score is more than 2, indicating a deviation of two standard deviations from the mean. The amount of data before and after the outliers reductions is the following:

Table 2: Amount of samples in the dataset

Original dataset	Outliers reduction
253,680	192,824

3.3 PCA

Principal Component Analysis (PCA) is a dimensionality reduction technique commonly used in machine learning and statistics. Its primary objective is to simplify the complexity of high-dimensional data. This is achieved by converting the original variables into a set of orthogonal components, which capture the maximum variance present in the data [12].

PCA is useful to reduce the complexity of large datasets as the one used in this project (253,680 observations) and the aim was to apply it. However after doing further investigation was made, it was discovered that PCA calculates the principal components from the eigenvalues of the covariance matrix. PCA is not well-suited for categorical or binary variables as it relies on numerical information. Consequently, this powerful technique could not be applied in this project due to the absence of such numerical data.

3.4 Balanced dataset

Balancing a very large imbalanced, can be a way of improving the performance and training time. However it has to be carefully considered as down-sampling could also lead to lower accuracy, over-fitting and adding bias to the model.

The dataset used for the project contains 253,680 observations with 86.07% belonging to the negative class and 13.93% to the positive class. In order to address the class imbalance, a balanced dataset was generated by down-sampling the majority class of the cleaned dataset, thereby eliminating noisy or irrelevant observations. The resulting dataset size is as follows:

Table 3: Number of samples per dataset

Original dataset	Outliers reduction	Cleaned and balanced
253,680	192,824	61,146

This dataset will be taken into account when developing supervised machine learning algorithms. Subsequently, a comparative analysis of time and accuracy performance will be conducted between the original dataset and the balanced and cleaned dataset.

4 Machine Learning algorithms

Supervised machine learning algorithms are different type of techniques used to train models that can predict or classify new data based on labeled examples in a training dataset. The algorithm learns the relationship between input features and target labels by iteratively adjusting model parameters to minimize the discrepancy between predicted and actual outcomes [10]. Common types of supervised algorithms include Support Vector Machine, Logistic Regression, Random Forest, Decision Trees, and K-Nearest Neighbors.

For this project, the dataset was divided into two parts: 80% for the training set and 20% for the test set. The objective was to evaluate various supervised classification algorithms and compare their performance. Considering the dataset's characteristics, four different classifiers were explored: Logistic Regression, Random Forest, and K-Nearest Neighbor.

Initially, the idea was to use the 80% training set to conduct cross-validation and test different hyper-parameter values. The intention was to select the optimal parameters for a final evaluation on the test set to measure accuracy. However, due to the dataset's large size, this process was computationally expensive. Instead, fixed values were chosen for the hyper-parameters, prioritizing simplicity while aiming for high accuracy. The values selected for the hyper-parameters are the following:

Table 4: Hyper-parameters used in both datasets

Classifier	Hyper-parameters
Random Forest	ntree = 100, mtry = 4
KNN	k = 5

Where ntree is the number of decision trees, mtry is the number of variables randomly sampled as candidates at each split in each tree. in KNN, K is the number of neighbors considered to make the prediction. Logistic Regression is a simple linear model and doesn't have hyper-parameters.

Three distinct classifiers were applied to both the original and balanced datasets with the same hyper-parameters values. The goal is to compare their performance and execution times. The results obtained are as follows:

Table 5: Performance on the original dataset

Classifier	Accuracy	F1-Score	Execution time
Logistic Regression	86.46%	24.22%	2 s
Random Forest	86.5%	22.41%	75 s
KNN	85.34%	24.52%	215 s

Table 6: Performance on the cleaned and balanced dataset

Classifier	Accuracy	F1-Score	Execution time
Logistic Regression	74.78%	75.45%	0.3 s
Random Forest	74.68%	75.76%	14 s
KNN	71.77%	72.82%	12 s

The results obtained reveal that the imbalanced dataset has a higher accuracy compared to the balanced one. However, it's important to note that accuracy can be influenced by the majority class, potentially leading to a skewed evaluation. To address this issue, the F1-Score is the preferred for evaluating classifier performance. Unlike accuracy, the F1-Score considers both precision and recall, providing a more comprehensive understanding of overall performance.

Initially, the results indicate that algorithms trained on the imbalanced dataset achieve an accuracy of approximately 86%, while those trained on the balanced dataset achieve around 75%. However, the greatest difference is in the F1-score metric, where the average difference between the datasets is approximately 50%. This result shows that the imbalanced dataset significantly influences the performance of the classifiers and most of the predictions done in the minority class are incorrect.

For the scope of the project, where the goal is to predict whether a person has diabetes or not, the minority class is the one that contains the positive instances (diabetes). Therefore, despite exhibiting lower accuracy in the majority class compared to the imbalanced dataset, the balanced dataset has shown to be more suitable for the project.

Another important consideration is the model training time. Training times for all three classifiers were measured on both datasets. Table 5 and Table 6 shows the significant difference0 in training times between the imbalanced and balanced datasets. For instance, Logistic Regression training time is approximately 10 times longer, Random Forest 5 times longer, and KNN around 20 times longer on the imbalanced dataset. In certain scenarios, particularly those involving computational constraints, such differences in computational cost must be carefully considered

In summary, classifiers trained on the balanced dataset demonstrate superior performance in terms of both balance and execution time. This shows the importance of doing data analysis in the dataset, before applying machine learning. Beyond the analysis made in this project, real-world scenarios often involve additional pre-processing steps such as Principal Component Analysis (PCA), conversion of categorical variables to numerical ones, etc. The dataset available online that was used for this project already had some pre-processing steps and for some reason some of the critical steps were not applied in the dataset.

References

[1] Haldun Akoglu. User's guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3):91–93, 2018.

[2] American Diabetes Association. Statistics about diabetes — ada. https://www.diabetes.org/about-diabetes/statistics/about-diabetes. Accessed: 16/04/2024.

[3] Centers for Disease Control and Prevention. CDC - About Adult BMI. https://www.cdc.gov/healthyweight/assessing/bmi/index.html. Accessed: 18/04/2024.

[4] Centers for Disease Control and Prevention. Cdc diabetes health indicators. http://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators. Accessed: 16/04/2024.

[5] Krzysztof J Cios, Witold Pedrycz, and Roman W Swiniarski. *Data mining methods for knowledge discovery*, volume 458. Springer Science & Business Media, 2012.

[6] Centers for Disease Control and Prevention. Diabetes health indicators. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset, 2015. Kaggle dataset.

[7] Wolfgang Kerner and J Brückel. Definition, classification and diagnosis of diabetes mellitus. *Experimental and clinical endocrinology & diabetes*, 122(07):384–386, 2014.

[8] Robert McGill, John W Tukey, and Wayne A Larsen. Variations of box plots. *The american statistician*, 32(1):12–16, 1978.

[9] Ali H Mokdad. The behavioral risk factors surveillance system: past, present, and future. *Annual review of public health*, 30:43–54, 2009.

[10] FY Osisanwo, JET Akinsola, O Awodele, JO Hinmikaiye, O Olakanmi, J Akinjobi, et al. Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, 48(3):128–138, 2017.

[11] SGOPAL Patro and Kishore Kumar Sahu. Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*, 2015.

[12] Markus Ringnér. What is principal component analysis? *Nature biotechnology*, 26(3):303–304, 2008.