DIABETES PREDICTION COMPETITION

Team 16:
Alessandro Belotti 896985, Luca Cilloni 898109, Camilla Fracchia 898235

Abstract

The goal of this research work is to look into the dataset and classify whether a person is sick with Diabetes. This is a binary classification problem where the labels of class attribute are: 0 which means that the person is not affected by Diabetes, 1 that means the opposite situation.

This report wants to present in a clear way the process of building a Machine Learning model, from Data Exploration to Model Selection. The aim is to accurately predict the target variable “Diabetes” provided the given data and evaluate it on the basis of a loss function called logarithmic loss function.

CONTENTS

Introduction 1
1 Data exploration 2
2 Preprocessing 3
2.1 Unsupervised discretization 3
2.2 Feature construction 3
3 Models 4
3.1 Introduction of the problem and evaluation metrics 4
3.2 Presentation of the selected models 4
3.3 Hold-out and features selection 4
4 Model selection 6
4.1 Improvement and choice of the model 6
Conclusion 7
References 7

INTRODUCTION

Diabetes occurs when there is an increase in glycaemia levels due to a deficit in the quantity and, often, in the biological effectiveness of insulin, an hormone produced by the pancreas that controls glucose in the blood.

According to doctors, the most common patterns for Diabetes are overweight, obesity and sedentary lifestyle, insulin resistance, genes and family history, genetic mutations and hormonal diseases.

Diabetes is a widespread disease: about 3.5 million people in Italy have been diagnosed with it, an increase of about 60 per cent in the last 20 years. Unfortunately, so many more are thought to be suffering from the condition without knowing it, so that the actual number of Italian diabetics is thought to exceed 4 million. The most common is type 2 Diabetes and only about 5% of the population has type 1 Diabetes. Given this massive diffusion, it is interesting to analyze the factors that are contributing to its development.

Within our dataset, a total of 17 variables that are likely or not to contribute to Diabetes are covered.

It is composed of 40108 rows, each one with the following 17 explanatory attributes together with the class attribute:

- **Age** (categorical - ordinal): 1 = 18-24, 2 = 25-29, 3 = 30-34, 4 = 35-39, 5 = 40-44, 6 = 45-49, 7 = 50-54, 8 = 55-59, 9 = 60-64, 10 = 65-69, 11 = 70-74, 12 = 75-79, 13 = 80 or older.

  Age of the person.

- **Sex** (categorical - nominal/binary): 0 = female, 1 = male.

  Gender.

- **HighChol** (categorical - nominal/binary): 0 = absence of high cholesterol in blood, 1 = presence of high cholesterol.

  Level of cholesterol measured in blood.

- **CholCheck** (categorical - nominal/binary): 0 = no cholesterol check in 5 years, 1 = yes cholesterol check in 5 years.

  Whether the subject carried out at least one cholesterol check in the last five years.

- **BMI** (numeric): Body Mass Index.

  It is a biometric data, expressed as the ratio of weight to square of the height of an individual and is used as an indicator of the state of shape weight. In the dataset takes values ranging from 12 (severe underweight) to 98 (severe obesity).

- **Smoker** (categorical - nominal/binary): 0 = no, 1 = yes.

  Whether the subject smoked at least 100 cigarettes (5 packets) in his life.

- **HeartDiseaseorAttack** (categorical - nominal/binary): 0 = no, 1 = yes.

  The existence of a history of coronary heart disease (CHD) or myocardial infarction (MI).

- **PhysActivity** (categorical - nominal/binary): 0 = no, 1 = yes.

  The physical activity carried in past 30 days (not including job).

- **Fruits** (categorical - nominal/binary): 0 = no, 1 = yes.

  The consumption of fruit once or more times per day.

- **Veggies** (categorical - nominal/binary): 0 = no, 1 = yes.

  Consumption of vegetables once or more times per day.

- **HvyAlcoholConsump** (categorical - nominal/binary): 0 = no, 1 = yes.

  Depending on gender: for an adult male it is considered more than 14 drinks per week while for an adult female more than 7 drinks per week.

- **GenHlth** (categorical - ordinal): (scale 1-5) 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor.

  Self-assessment of individual physical health on a 1-to-5 likert scale.

- **MentHlth** (numerical): scale 1-30 days.

  Days of poor mental health during the last month.

- **PhysHlth** (numerical): scale 1-30 days.

  Physical illness or injury days during the last month.

- **DiffWalk** (categorical-nominal/binary): 0 = no, 1 = yes.

  Expresses a serious difficulty during walking or climbing stairs.

- **Hypertension** (categorical-nominal/binary): 0 = no, 1 = yes.

  Presence of elevated blood pressure.

- **Stroke** (categorical-nominal/binary): 0 = no, 1 = yes.

  Whether the subject experienced a stroke in his life.

- **Diabetes** (categorical-nominal/binary): 0 = no, 1 = yes.

  Indicates whether the subject is suffering from the illness or not (Target variable).

The goal of our analysis is to predict the presence of Diabetes using Machine Learning tools and using Log Loss function to evaluate the goodness of the predictions.
This report is organized as follows:

1. **Data exploration:** features analysis of the dataset.
2. **Preprocessing:** features construction and discretization of some variables in order to make the dataset more suitable for analysis.
3. **Models:** description of the different models used to predict the presence of Diabetes.
4. **Model selection:** comparation of the models described in the previous section.

1. Data Exploration

The starting point of a structured analysis entails the exploration on data through the tools of descriptive statistics in order to investigate the patterns of the data. Given the absence of missing values, the outliers or extreme values that potentially could distort the analysis are sought.

The boxplot shown in Figure 1 reveals the presence of outliers for the variables *BMI*, *MentHlth* and *PhysHlth*: in an effort to reduce the impact that these anomalous values could have on the model, they are discretized (refer to the next section for more details).

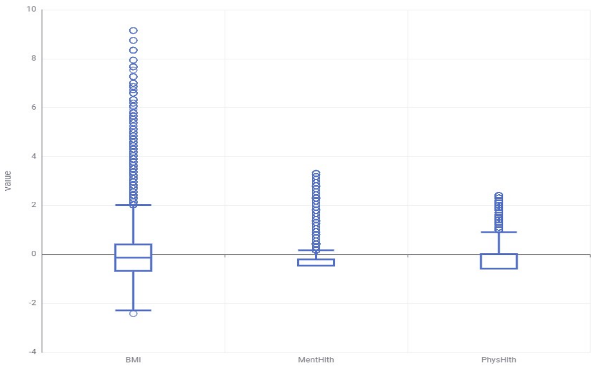


Figure 1: Box plot for the variables BMI, MentHlth, PhysHlth.

A multivariate analysis is performed employing the correlation matrix, which enables the observation of the interdependence degree that exists within the variables. The more saturated is the color, the higher is the correlation (direct -blue- or indirect -red-) between the values.

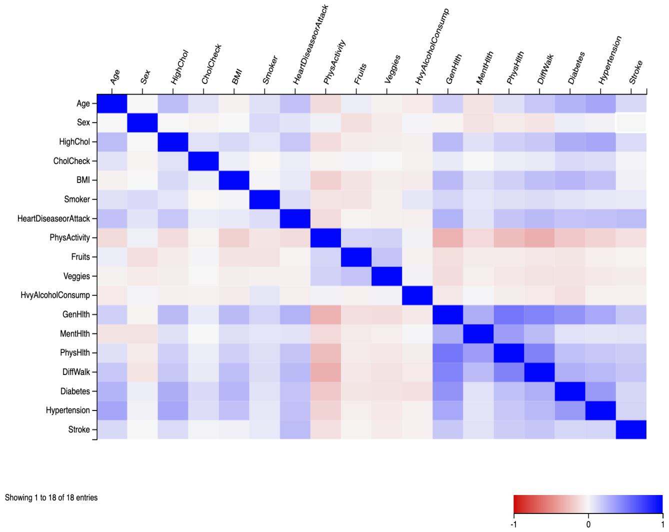


Figure 2: Correlation matrix of the variables.

Eventually, it is ensured that the dataset is balanced, in other words, that it contains equal occurrences (in terms of frequencies) in the two modalities of the target variable.

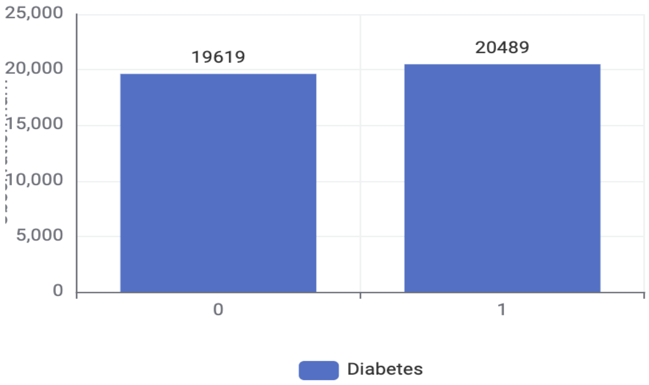


Figure 3: Bar Chart of the balanced dataset.

2. Preprocessing

The quality of the used dataset deeply impacts the chances of finding meaningful models. The most common problems that deteriorate the quality of the data and therefore need to be corrected are outliers, noise and missing values. Since there are no missing values, the problem of outliers is addressed by discretizing the variables that have them.

2.1 Unsupervised discretization

The BMI attribute is firstly discretized into 7 intervals according to the classes following the medical definition:

- 1: BMI less than 16 (Severe thinness)
- 2: BMI is 16 to < 19 (Underweight)
- 3: BMI is 19 to < 25 (Normal weight)
- 4: BMI is 25 to < 30 (Overweight)
- 5: BMI is 30 to < 35 (Obese class 1)
- 6: BMI is 35 to < 40 (Obese class 2)
- 7: BMI greater than 40 (Obese class 3).

Afterwards, it is further discretized into intervals of equal frequency using the *Autobinner* node (based on quartiles), resulting in 3 intervals: [1, 4] (4, 5] (5, 7].

The same arrangement is adopted for the variables *MentHlth* and *PhysHlth*: again by using the *AutoBinner* node, 3 intervals of equal frequency are selected, respectively equal to [0] (0, 2] (2, 30] and [0] (0, 6] (6, 30].

The choice of intervals with equal frequency rather than equal width can be explained by the desire to avoid potential distortions in the analysis due to inconsistent discretization.

2.2 Feature construction

Given the existing trade-off between performance and dimensionality, it has been opted for the creation of a new feature combining *Fruits* and *Veggies* into a single dichotomous variable, called *HealthyFood* with the aim of reducing the size of the data and the computational complexity of the algorithm, thus taking less time. By means of a *Rule Engine* node, it has been determined that if both variables have the value 1, the new variable also obtains the value 1, otherwise 0.

3. MODELS

3.1 INTRODUCTION OF THE PROBLEM AND EVALUATION METRICS

The purpose of the research is to predict the presence of Diabetes in a given subject: therefore it is necessary to develop a Supervised classification model which, functioning as a black box, manages to assign with a certain degree of correctness a class attribute label to an unknown record.

As previously seen, the reference dataset is not unbalanced since the distribution of the observations is equal among the modalities assumed by the variable: 51% of the observations assume class 1, corresponding to the existence of Diabetes, while 49% assume level 0.

Thus, it is possible to consider the Log Loss (or loss function) a reliable metric to assess the performance of a model. It measures the degree of proximity between the probability of the predicted class and the actual class.
It is expressed as:
Log loss (y,p) = - y log(p) - (1-y) log(1-p),
where y is the value of the actual class and p is the one of the predicted class. The name of this function is self-explanatory: the smaller its value, the better will be the model.

The absence of the class imbalance issue has also permitted to use Accuracy as a second performance indicator. This is calculated as a ratio between the sum of the values (positive and negative) correctly predicted and the sum of all observations. Being within the interval [0, 1], it can be interpreted in percentage points as the portion of correctly predicted observations by the model. It allows the selection of the best performing instance (namely higher accuracy values) on the records of the Test set, which are those to be predicted.

Accuracy = TN+TP/(TN+TP+FN+FP).

3.2 PRESENTATION OF THE SELECTED MODELS

The decision on the model is based on the analysis of each of the macro-categories of classification techniques:
- **Heuristic**: the *J48* model is picked, the algorithm returns a decision tree which, by exploiting the concept of 'node splitting', allows the tree to be cut on the basis of the chosen measures.
- **Probabilistic**: the *NBTree* model is selected, which develops a decision tree based on the *Naive Bayes* classifier, which itself uses Bayes' theorem and permits a post-computing probability of the class attribute.
- **Regression based**: the *Logistic Regression* model is chosen, which calculates the class attribute probability depending on the input attributes. This classification method is relying on the concept of multiple logistic regression (meaning with more than one explanatory variable) and allows the binary regression problem to be solved.
- **Separation models**: the *MultiLayerPerceptron* model is adopted which, being based on the separation of the attribute space, consists of artificial neurons communicating unidirectionally from input X to the class variable.

3.3 HOLD-OUT E FEATURES SELECTION

The starting point is the splitting of our dataset through the technique known in the literature as *Hold-out*, into two mutually exclusive partitions: the *Train* and *Test* set.
Respectively, they contain 67% and 33% of the total records extracted through the stratified sampling technique according to the target variable *Diabetes*.
This procedure is crucial as it allows the models to be trained on the Training set and to be validated via the Test set.

Since the dataset contains a total of 17 variables (16 + 1 target), it become necessary to make a selection of those most significant for predicting the presence of the disease.

The *Features Selection* process is performed using the Knime *AttributeSelectedClassifier* node. This allows univariate filtering (elimination of irrelevant/insignificant attributes) and multivariate filtering (elimination of irrelevant and redundant attributes).

An objective function is employed for univariate or multivariate attributes: in the first case, a measure of association between the class attribute and all explanatory variables is chosen. The values obtained are then sorted in descending order and the first R features (in this analysis, 3 5 or 10) are picked. The univariate objective functions used in the Knime features selection node and described in the Weka documentation are:

- ***CorrelationAttributeEval***: evaluation is performed by calculating the correlation (Pearson's) between the attribute and the class to be predicted.
- ***GainRatioAttributeEval*** : assesses the value of an attribute by measuring the gain ratio with respect to the class attribute.
- ***InfoGainAttributeEval*** : gets the attribute value by calculating the information gain in relation to the class attribute.
- ***SymmetricalUncertAttributeEval*** : evaluates the attribute value by determining the symmetrical uncertainty in comparison with the class attribute.

The downside of these measures is their failure to identify redundant attributes, in other words, those that are correlated together. In order to overcome this problem, multivariate functions were also analyzed:

- *ReliefFAttributeEval*: evaluation is performed by repeatedly sampling an instance and considering the given attribute value for the nearest instance of the same and different class. It can work on both discrete and continuous class.
- *CorrelationAttributeEval* : is assessed by calculating the correlation (Pearson's) between the attribute and the class to be predicted.
- *OneRAttributeEval* : estimates the value of an attribute using the OneR.
- *PrincipalComponents*: performs a Principal Component Analysis and transformation of the data and chooses the eigenvectors that take into account, i.e. are able to explain, the 95% of the variance.
- *CfsSubsetEval*: evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred.

The first step in the research is to identify 3 then 5 and finally 10 relevant variables, in an attempt to improve the model (increasing Accuracy and lowering the Log Loss) by using as few attributes as possible: this is done on the basis of the computational advantages provided by the use of faster and less memory-intensive algorithms.

Subsequently, thanks to the *AttributeSelectedClassifier* node, the classifier is trained on the Training dataset and the inducer is tested, with the *Weka Predictor* node, on the Test set.

The results of the analysis in terms of Accuracy (most familiar measure for assessing model goodness) and Log Loss (objective function for assessing model goodness) can be seen in the following charts and tables:

LOGLOSS	NBTree	J48	MLP	LOGISTIC
10_correlationAttributeEval	0,239	0,243	0,231	0,229
10_GainRatioAttrEval	0,243	0,25	0,232	0,232
10_infoGain	0,239	0,244	0,232	0,229
10_OneAttributeEval	0,239	0,248	0,234	0,229
10_PrincipalComponents	0,238	0,277	0,226	0,228
9_CfsSubsetEval	0,236	0,241	0,228	0,227
10_SymmetricalUncertAttributeEval	0,239	0,241	0,23	0,225
5_correlationAttributeEval	0,246	0,241	0,243	0,242
5_GainRatioAttrEval	0,251	0,244	0,247	0,244
5_infoGain	0,248	0,248	0,24	0,239
5_OneAttributeEval	0,248	0,248	0,24	0,239
5_PrincipalComponents	0,238	0,254	0,231	0,232
5_CfsSubsetEval	0,237	0,244	0,237	0,231
5_SymmetricalUncertAttributeEval	0,245	0,242	0,243	0,24
3_correlationAttributeEval	0,263	0,25	0,252	0,25
3_GainRatioAttrEval	0,252	0,254	0,253	0,252
3_infoGain	0,263	0,25	0,252	0,25
3_OneAttributeEval	0,257	0,252	0,254	0,252
3_PrincipalComponents	0,238	0,24	0,232	0,232
5_CfsSubsetEval	0,237	0,249	0,23	0,229
3_SymmetricalUncertAttributeEval	0,252	0,254	0,253	0,252

ACCURACY	NBTree	J48	MLP	LOGISTIC
10_correlationAttributeEval	0,733	0,735	0,735	0,735
10_GainRatioAttrEval	0,732	0,734	0,735	0,737
10_infoGain	0,735	0,735	0,737	0,736
10_OneAttributeEval	0,731	0,732	0,732	0,734
10_PrincipalComponents	0,729	0,726	0,741	0,74
9_CfsSubsetEval	0,739	0,74	0,743	0,745
10_SymmetricalUncertAttributeEval	0,736	0,736	0,733	0,737
5_correlationAttributeEval	0,718	0,717	0,721	0,716
5_GainRatioAttrEval	0,71	0,71	0,71	0,71
5_infoGain	0,719	0,721	0,72	0,72
5_OneAttributeEval	0,719	0,721	0,72	0,72
5_PrincipalComponents	0,731	0,726	0,737	0,733
5_CfsSubsetEval	0,732	0,729	0,74	0,736
5_SymmetricalUncertAttributeEval	0,719	0,718	0,713	0,72
3_correlationAttributeEval	0,7	0,7	0,703	0,698
3_GainRatioAttrEval	0,711	0,711	0,711	0,711
3_infoGain	0,7	0,7	0,703	0,698
3_OneAttributeEval	0,702	0,702	0,702	0,702
3_PrincipalComponents	0,731	0,728	0,734	0,732
3_CfsSubsetEval	0,732	0,731	0,732	0,738
3_SymmetricalUncertAttributeEval	0,711	0,711	0,711	0,711

Table 1 and 2: results of Feature Selection.

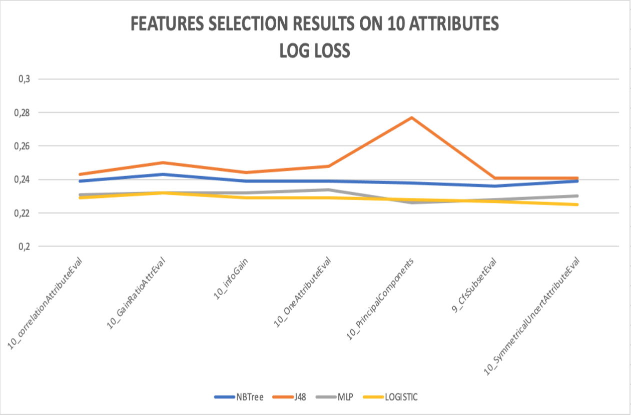


Figure 4: line chart of the results of the Features Selection over 10 variables in terms of Log Loss.

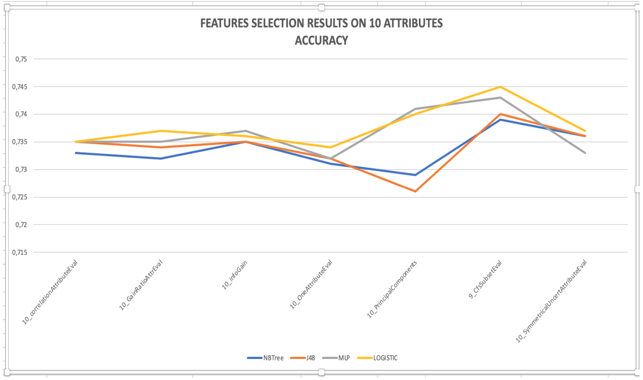


Figure 5: line chart of the results of the Features Selection over 10 variables in terms of Accuracy.

Given the existence of a trade-off between model accuracy and computational difficulty, based on the observation of the results obtained and the speed of execution of the algorithms, the most appropriate solution is the one obtained by applying the multivariate function *CfsSubsetEval*, which allows not 10 but 9 relevant variables to be selected (since the tenth candidate is not sufficiently correlated to be taken into consideration): *BMI binned, Age, HighChol, CholCheck, HeartDiseaserorAttack, HvyAlcholConsump, GenHlth, DiffWalk* and *Hypertension*.

4. Model selection

4.1 Improvement and choice of the model

The results obtained from the Features Selection can be further refined by means of Cross Validation: this Hold-out technique consists in dividing the dataset into K-folds, that is, into K disjointed and exhaustive subsets, containing approximately the same number of records (according to a stratified sampling).

Therefore, given D_1, D_2, ..., D_K partitions, in the K-th iteration the Training set will be formed by D_{K-1} folds, while D_K will constitute the Test set. This ensures that each observation appears the same number of times in the Train set and exactly once in the Test set.

Through the *Column Filter* and *Normalizer* nodes, the 9 variables resulting most significant are selected and normalized.

Cross Validation is employed (created via the *X-partitioning* and *Loop End* nodes) to prevent overfitting, the phenomenon whereby the model performs very well on the Training set but has a very high Generalization error: the 4 models described above are trained.

Next, by the use of the appropriate inducer node, each model is tested on the second partition or Test set: results in terms of Log Loss and Accuracy suggest the absence of overfitting, since the difference between performance on Train and Test set is not sufficiently large to raise any doubt.

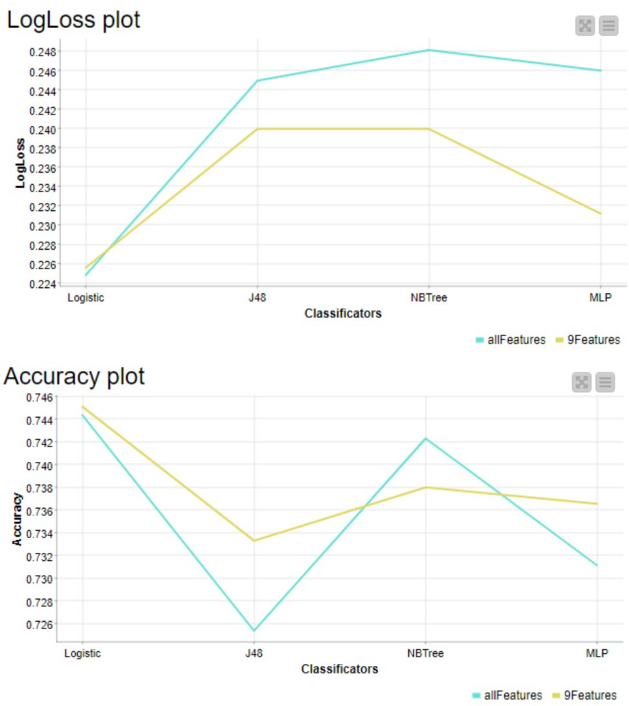
Lastly, for an additional confirmation of the decision taken and for comparative purposes, the same procedure is carried out by training and validating the models while using all the variables.

The findings achieved on the Test sets in terms of Accuracy and Log Loss can be seen in Tables 3 and 4, as well as Figures 5 and 6:

Row ID	allFeatures	9Features
Logistic	0.225	0.226
J48	0.245	0.24
NBTree	0.248	0.24
MLP	0.246	0.231

Row ID	allFeatures	9Features
Logistic	0.744	0.745
J48	0.725	0.733
NBTree	0.742	0.738
MLP	0.731	0.737

Tables 3 and 4: results in terms of Log Loss and Accuracy for the models trained and tested with Cross Validation using all variables and only the 9 selected with the Features Selection.



Figures 6 and 7: Knime charts of the results in terms of Log Loss and Accuracy for the models trained and tested with Cross Validation using all variables and only the 9 selected with the Features Selection.

As can be appreciated by observing the plots, the models constructed using 9 variables perform better in almost all cases.

In fact, in both plots the green line, which represents the performance of these models, outperforms the blue line by lying below it (in terms of Log Loss) and above it in terms of Accuracy.

In this analysis, the main measure of how well a model performs is the logarithmic loss function, so focusing on the graph in Figure 6, the following considerations can be made: the logistic regression model is the one that globally performs best in both cases, whether it is trained with some or all variables.

On the other hand, regarding the other models, the J48 and the NBTree show similar results for both lines; while the MLP is the model whose difference between the performances is the widest.

The lack of statistical tools such as confidence intervals and hypothesis tests and, above all, the desire to follow a conservative approach lead to assume that the minimal difference (equal to 0.001) between the behaviour of the logistic regression model with all variables and the one with only 9 variables is probably not great enough to favour the first one.
However, nothing can be said about the other models without risking to commit an error.

Following the considerations made, it is concluded that among the trained models, the one that proves to be the best, both in terms of performance and in terms of computational advantages and other desirable properties (such as speed), is the logistic regression model in the version trained with 9 variables, despite the fact that it is not the one that returned the lowest value of Log Loss.

The choice is principally driven by the computational advantages of using algorithms that work with a reduced number of variables. This in fact demonstrates the fact that the efficiency of a model does not depend on the number of available variables but rather on the relevant ones.
By eliminating the 'noise' that irrelevant and redundant parameters bring to the analysis, the results are improved.

Conclusion

The achievement of the objective established with the initial research question has been possible after an analysis of the set of possible models.

Following the exploratory phase of the dataset and the related statistics, decisions have been made concerning the filtering method, the number of variables and the supervised models to be used.

Possible future developments of this research involve the further optimization of the model parameters, in order to reduce the prediction error and the loss function and the implementation of confidence interval for different Log Loss values. The aim is to give statistical basis to our assumptions.

In addition, it is possible to pose a second research question: do individuals with Diabetes share certain characteristics? The answer can be found by performing a cluster analysis on the reference dataset.

References

[1] AttributeEvaluator. (2022, January 28). https://weka.sourceforge.io/doc.dev/weka/attributeSelection/AttributeEvaluator.html

[2] Diabetes Prediction Competition(TFUG Chd Nov 2022) | Kaggle. https://www.kaggle.com/competitions/diabetes-prediction-competitiontfug-chd-nov-2022/data

[3] Il diabete: cause, sintomi e cure | Mario Negri (2023). https://www.marionegri.it/magazine/diabete

[4] Ministero della Salute (2023)
https://www.salute.gov.it/