# Machine learning and balanced techniques for diabetes prediction

Liliana Narvaez
*Universidad Técnica Particular de Loja*
Loja, Ecuador
lenarvaez@utpl.edu.ec

Ruth Reátegui
*Universidad Técnica Particular de Loja*
Loja, Ecuador
rmreategui@utpl.edu.ec

***Abstract*—*Diabetes mellitus is a metabolic disorder characterized by high blood glucose levels, resulting from defects in insulin secretion, insulin action, or both. This study applied some supervised learning such Support Vector Machine, Random Forest and Gradient Boosting to predict diabetes mellitus. Additionally, a comparative analysis of two balanced data techniques, namely SMOTE and RandomUnderSampler, is presented. Results show that Gradient Boosting yielded the most favorable outcomes in terms of accuracy and precision when utilizing SMOTE technique. Furthermore, the inclusion of insulin variable and the exclusion of SkinThickness and BloodPressure variables led to improve the results.*

***Keywords*—*diabetes; machine learning; imbalance data*.**

## I. INTRODUCTION (*HEADING 1*)

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels, resulting from defects in insulin secretion, insulin action, or both. It encompasses two primary forms: type 1 diabetes and type 2 diabetes, each with distinct underlying causes and mechanisms.

Globally, around 422 million people suffer from diabetes, and 1.5 million deaths are directly attributed to diabetes each year [1]. Poor blood sugar control in diabetic patients can lead to various complications, including cardiovascular disease, diabetic neuropathy, kidney damage, retinopathy, skin, foot and hearing problems [2]

In the field of machine learning (ML) and deep learning (DL) algorithms, essential characteristics such as age, insulin production, body mass index, and family history are utilized to predict diabetes. Numerous studies have employed the PIMA Indian Diabetes dataset for diabetes prediction, some of which are outlined below.

Reference [3] implemented an age adaptation algorithm to DM prediction. This study utilized Linear Regression (LR), Logistic Regression, Polynomial Regression (PR), Neural Network (NN), Support Vector Machines (SVM), Random Forest (RF), and XGboost (XGB) algorithms. The incorporation of compensated features and age adaptation methods improved the results, allowing models trained on one age group to be adapted to another, thereby addressing data scarcity in specific age ranges.

In another investigation [4], Logistic Regression, Naïve Bayes, and K-nearest Neighbor were employed. The results demostrated that Logistic Regression is the most efficient. The authors in reference [5] applied various algorithms, such as Random Forest, Light Gradient Boosting Machine, Gradient Boosting Machine, Support Vector Machine, Decision Tree, and XGBoost. The findings indicated that Light Gradient Boosting Machine yielded the best results. Additionally, Data Augmentation and Sampling techniques were utilized.

Furthermore, some studies applied ensemble techniques to improve the results. Reference [6] conducted a comparison of several ML algorithms: Logistic Regression, Linear Discriminant Analysis, K-nearest Neighbor, Decision Tree, Support Vector Machine, AdaBoost classifier, Gradient Boosting Classifier, Random Forest classifier and extra tree classifier. These algorithms were evaluated using both the PIMA Indian Diabetes dataset and Early Stage Diabetes Risk Prediction Dataset. The ensemble machine learning algorithms provide better classification accuracy compared to other machine learning algorithms in both datasets. In other study [7], Decision Tree, SVM, Random Forest, Logistic Regression, KNN, and various ensemble techniques were utilized. The study employed the Pima Indian diabetes dataset and 203 samples of females patients from Bangladesh. Additionally, SMOTE and ADASYN approaches were employed to address the class imbalance problem. The XGBoost classifier, in combination with the ADASYN approach, yielded the best results, achieving an accuracy of 81%, an F1 coefficient of 0.81, and an AUC of 0.84.

Given the widespread use of the PIMA Indian Diabetes dataset in various studies, this work adopts the same dataset for comparison purposes. To assess balancing data techniques, SVM, Random Forest, and XGBoost algorithms were evaluated. Furthermore, different scenarios were created based on the presence or absence of the insulin variable as part of the predictor variables.

## II. METHODOLOGY

### *A. Data*

For the development of this study, the researchers utilized the "Pima Indians Diabetes Dataset" obtained from the National Institute of Diabetes and Renal Digestive Diseases [8]. The dataset represents the Pima Indians, an indigenous group residing in Arizona (USA) and Sonora, Chihuahua (Mexico), and exclusively comprises women aged 21 years and above. This dataset consists of 768 samples and includes eight variables: Pregnancies (number of pregnancies), Glucose (glucose concentration), BloodPressure (blood pressure), SkinThickness (tricep skin fold thickness), Insulin (insulin concentration), BMI (body mass index), Age (age),

and the target variable (outcome), which determines the presence or absence of diabetes in the individuals.

As shown in Table I, some variables contain zero values. In the case of BMI and blood pressure, these zero values need to be corrected to ensure accurate analysis. Additionally, the objective variable is imbalanced, with 65% of the data corresponding to non-diabetics and 34.8% to diabetics.

To explore the correlation between variables and the "Outcome," the researchers constructed a correlation matrix. As depicted in Fig. 1, "Glucose," "BMI," and "Age" exhibited the highest correlations with the outcome, while "SkinThickness" and "BloodPressure" displayed the lowest correlations. Furthermore, Fig. 2 illustrates the mutual information between variables and the "Outcome," indicating that "SkinThickness" and "BloodPressure" have the lowest information content in relation to the target variable.

### *B. Preprocessing*

As previously mentioned, certain variables such as "Glucose," "BloodPressure," "BMI," "SkinThickness," and "Insulin" contain incorrect zero values. Notably, approximately 30% and 49% of the data contain zero values for the "SkinThickness" and "Insulin" variables, respectively (refer to Fig. 3). To address this issue, the researchers employed the Simple Imputer class in Python for imputing the incorrect zero values. Additionally, the data underwent scaling using the MinMaxScaler class in Python to scaler the variables.

### *C. Imbalance Techniques*

To address the issue of imbalanced data, two techniques were employed:

**Oversampling**: This method technique increase data from the minority sample, in this work we will apply Synthetic Minority Oversampling Technique (SMOTE). Reference [9] proposed SMOTE in which the minority class is over-sampled by introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. For generate the synthetic samples the difference between a sample feature vector and its nearest neighbor is calculated and multiply by a random number between 0 and 1, and add it to the feature vector.

**Undersampling**: This technique eliminates records randomly from the majority class, the disadvantage of this method is that important information is lost, in this work we will use RandomUnderSampler. This technique randomly selects the instances from the majority class without replacement until the desired number of instances is reached and combine them with instances of the minority class.

### *D. Scenarios for experiments*

In the context of applying balancing techniques to address imbalanced data, two general scenarios are derived: one with oversampling and another with undersampling.

Given the significance of the insulin hormone in maintaining normal blood glucose levels and its role in controlling glucose levels in the body [10], it is considered as an important variable for diabetes prediction. Accordingly, the researchers devised various combinations of experiments based on the inclusion or exclusion of the insulin variable, resulting in four distinct scenarios, as presented in Table II.

Furthermore, taking into account the relatively low correlation and mutual information values for "SkinThickness" and "BloodPressure," the researchers developed an additional four scenarios, as described in Table III.

TABLE I. DATA

| | **Pregnancies** | **Glucose** | **Blood Pressure** | **Skin Thickness** | **Insulin** | **BMI** | **Diabetes Pedigree Function** | **Age** |
|---|---|---|---|---|---|---|---|---|
| **count** | 657 | 763 | 733 | 541 | 394 | 757 | 768 | 768 |
| **mean** | 3.83 | 120.89 | 69.11 | 20.54 | 79.8 | 31.99 | 0.47 | 33.24 |
| **std** | 3.34 | 31.97 | 19.36 | 15.95 | 115.24 | 7.88 | 0.33 | 11.76 |
| **min** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.08 | 21.0 |
| **25%** | 1.0 | 99.0 | 62.0 | 0.0 | 0.0 | 27.3 | 0.24 | 24.0 |
| **50%** | 3.0 | 117.0 | 72.0 | 23.0 | 30.5 | 32.0 | 0.37 | 29.0 |
| **75%** | 6.0 | 140.25 | 80.0 | 32.0 | 127.25 | 36.6 | 0.63 | 41.0 |
| **max** | 15.0 | 199.0 | 122.0 | 99.0 | 846.0 | 67.1 | 2.42 | 81.0 |

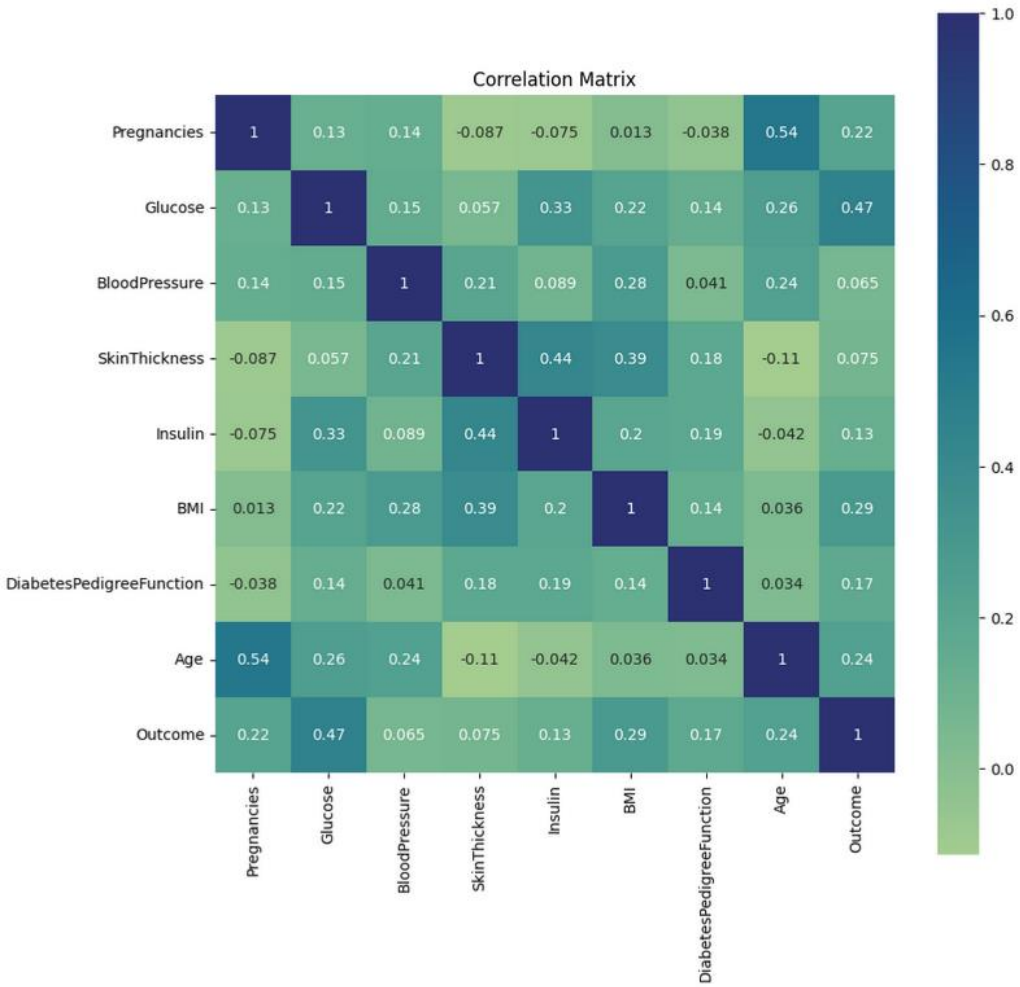


Fig. 1. Correlation Matrix

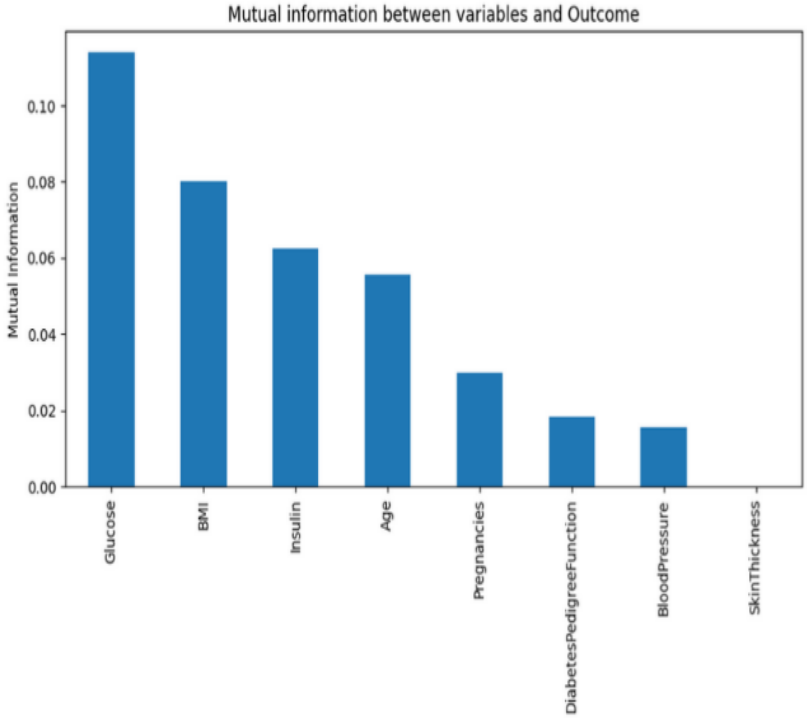


Fig 2. Mutual Information

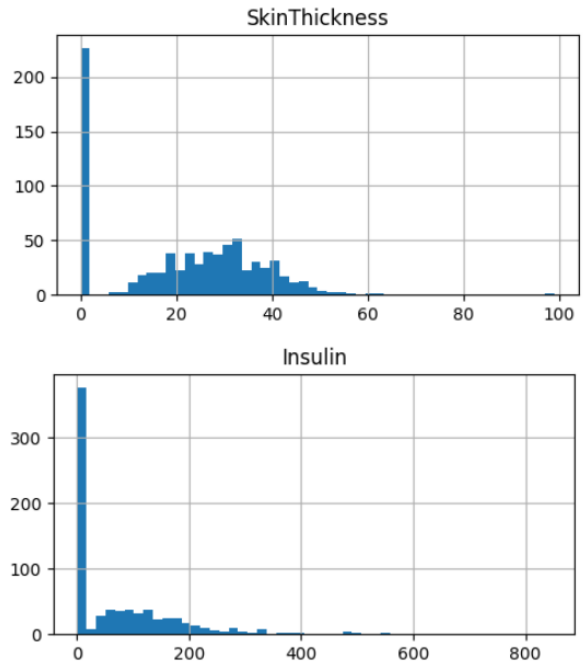


Fig. 3. Zero values in SkinThickness and Insuline

### *E. Algorithms and evaluation metrics*

In this study, the classification algorithms employed are geared towards binary classification, which predicts the label or output based on learned observations. The three classification methods used are Support Vector Machine (SVM), Random Forest, and Gradient Boosting.

**Support Vector Machine (SVM):** It is generally used to solve classification and regression problems. Its advantages include its ability to handle high-dimensional data, effectiveness on small data sets, and resistance to overfitting. SVM works by finding the optimal hyperplane that best separates the data points of different classes in the feature space [11].

**Random Forest:** It is a collection of many decision trees trained independently with a subset of data from the initial dataset [11]. This algorithm is generally trained through bagging method. Random Forest is effective in reducing overfitting and improving the overall accuracy and robustness of the model

**Gradient Boosting:** It is an algorithm that sequentially adds predictors to an ensemble, each one correcting its predecessor [12]. It builds a stepwise model step by step with the goal of minimizing a loss function. Gradient Boosting is particularly useful for handling complex datasets and achieving high predictive performance

The evaluation of the results obtained from these classifiers will be conducted using various metrics listed in Table IV. These metrics will provide valuable insights into the performance and effectiveness of the classification models, enabling a comprehensive assessment of their capabilities in predicting diabetes outcomes.

## III. RESULTS

For each scenario described above, the 70% of the data was used for training and the 30% for testing.

TABLE II. SCENARIOS FOR EXPERIMENTS

| Scenarios | Oversampling | Undersampling |
|---|---|---|
| Including "Insulin" variable | x | x |
| Excluding "Insulin" variable | x | x |

TABLE III. SCENARIOS EXCLUDING SKINTHICKNESS AND BLOODPRESSURE

| Scenarios | Oversampling | Undersampling |
|---|---|---|
| Including "Insulin" and excluding "SkinThickness" and "BloodPressure" variables | x | x |
| Excluding "Insulin" "SkinThickness" and "BloodPressure" variables | x | x |

TABLE IV. EVALUACION METRICS

| Metric | Formula | Description |
|---|---|---|
| Accuracy | $\frac{f_{11} + f_{00}}{N}$ | Proportion of correctly predicted classifications out of the total number of instances. |
| Recall | $\frac{f_{11}}{f_{11} + f_{01}}$ | Proportion of correctly classified positive cases. |
| Precision | $\frac{f_{11}}{f_{11} + f_{10}}$ | Proportion of true positive predictions out of the total positive predictions made. |
| F1-Score | $2 * \frac{precision * recall}{precision + recall}$ | Harmonic mean of the accuracy and recall metrics. |

### *A. Scenario including and excluding insulin and 7 variables*

The Table V shows the results of the scenarios including and excluding insulin and with the others 7 variables showed in Table I.

Using oversampling technique and including the insulin variable, Random Forest had the best result with values of 74, 62, 65 and 64 for accuracy, precision, recall and F1-score respectively. With the same balanced technique and excluding the insulin variable, SVM has the best result with values of 75, 70 and 66 for accuracy, recall and F1-score respectively. However, Random Forest had the best result with a value 64 for precision.

Using the undersampling technique and including the insulin variable, Gradient Boosting had the best results with values of 76, 63, 69 for accuracy, precision and F1-score respectively. The best result for recall was obtained with SVM algorithm with a value of 78. With the same balanced technique and excluding the insulin variable SVM had the best results with values of 75, 62, 78 and 69 for accuracy, precision, recall and F1-score respectively. Gradient Boosting also shows the same high value for recall. As we can see, with the undersamplig technique, the best results for accuracy and F1were obtained with the Gradient Boosting algorithm and including the insulin variable.

### B. *Scenario including and excluding insulin and excluding SkinThickness" and "BloodPressure" variables*

The Table VI shows the results of the scenarios with and without insulin. Also, the results of this table do not consider the SkinThickness" and "BloodPressure" variables.

Using oversampling technique and including the insulin variable, Gradient Boosting had the best result with values of 77.92, 66.30, 75.31 and 70.52 for accuracy, precision, recall and F1-score respectively. With the same balanced technique and excluding the insulin variable, SVM had the best result with values of 75.76, 62.63 and 68.89 for accuracy, precision and F1-score respectively, but Gradient Boosting has the best of recall with a value of 77.78. As we can see, including the insulin variable helps to improve the results.

Using the undersampling technique and including the insulin variable, Gradient Boosting had the best results with values of 76.62, 62.86, 81.48 and 70.97 for accuracy, precision, recall and F1-score respectively. With the same balanced technique and excluding the insulin variable, SVM had the best results with values of 76.19, 62.74 and 69.94 for accuracy, precision, and F1-score respectively, but Gradient Boosting has the best of recall with a value of 81.48. As we can see, again the inclusion of the insulin variable helps to improve the results.

## IV. CONCLUSIONS

The present study conducted a comparison between two balancing techniques, SMOTE and RamdomUnderSample, in the context of predicting diabetes outcomes. The best results were obtained with the Gradient Boosting, this algorithm is an ensembled algorithm that had been used in other works with acceptable results. In [7] obtained an accuracy of 86% as a highest result, but they extend the PIMA dataset with samples obtained from a private dataset.

Furthermore, the experiments in this study highlighted the significance of certain variables. Excluding "SkinThickness" and "BloodPressure" while including "Insulin" led to improved results, indicating the importance of the latter in diabetes prediction.

For future research, it is recommended to explore other ensemble and optimization methods to further enhance predictive performance. Additionally, as the aforementioned study indicated, adding more samples to the dataset could be considered as a means to improve the model's accuracy and precision. However, this extension should be done with caution, considering data quality, representativeness, and potential implications of using external data sources.

By exploring different ensemble techniques, optimization methods, and considering data augmentation approaches, future studies can continue to advance the field of diabetes prediction and contribute to the development of more accurate and reliable models for clinical decision-making.

TABLE V. RESULTS WITH ALL VARIABLES

| Insulin | Algorithm | Oversampling | Undersampling | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|---|---|---|---|
| Yes | **SVM** | x | | 71.429 | 58.241 | 65.432 | 61.628 |
| | **Random Forest** | x | | 74.025 | 62.353 | 65.432 | 63.855 |
| | **Gradient Boosting** | x | | 73.160 | 60.919 | 65.432 | 63.095 |
| No | **SVM** | x | | 74.892 | 62.637 | 70.370 | 66.279 |
| | **Random Forest** | x | | 74.891 | **63.529** | 66.667 | 65.060 |
| | **Gradient Boosting** | x | | 74.459 | 63.095 | 65.432 | 64.242 |
| Yes | **SVM** | | x | 74.459 | 60.577 | **77.778** | 68.108 |
| | **Random Forest** | | x | 74.459 | 60.784 | 76.543 | 67.760 |
| | **Gradient Boosting** | | x | **76.190** | 63.265 | 76.543 | **69.273** |
| No | **SVM** | | x | 75.325 | 61.765 | **77.778** | 68.852 |
| | **Random Forest** | | x | 71.861 | 57.692 | 74.074 | 64.865 |
| | **Gradient Boosting** | | x | 73.160 | 58.879 | **77.778** | 67.021 |

TABLE VI. RESULTS WITHOUT VARIABLES SKINTHICKNESS AND BLOODPRESSURE

| Insulin | Algorithm | Oversampling | Undersampling | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|---|---|---|---|
| Yes | **SVM** | x | | 75.76 | 64.37 | 69.14 | 66.67 |
| | **Random Forest** | x | | 75.32 | 63.63 | 69.14 | 66.27 |
| | **Gradient Boosting** | x | | **77.92** | **66.30** | 75.31 | 70.52 |
| No | **SVM** | x | | 75.76 | 62.63 | 76.54 | 68.89 |
| | **Random Forest** | x | | 73.59 | 61.11 | 67.90 | 64.33 |
| | **Gradient Boosting** | x | | 74.46 | 60.58 | 77.78 | 68.11 |
| Yes | **SVM** | | x | 75.76 | 62.14 | 79.01 | 69.56 |
| | **Random Forest** | | x | 76.19 | 62.75 | 79.01 | 69.95 |
| | **Gradient Boosting** | | x | 76.62 | 62.86 | **81.48** | **70.97** |
| No | **SVM** | | x | 76.19 | 62.74 | 79.01 | 69.94 |
| | **Random Forest** | | x | 74.03 | 60.40 | 75.31 | 67.03 |
| | **Gradient Boosting** | | x | 74.89 | 60.55 | **81.48** | 69.47 |

## REFERENCES


[1] World Health Organization. "Diabetes." WHO https://www.who.int/health-topics/diabetes#tab=tab_1 (accessed June, 13, 2023)

[2] Mayo Clinic. "Diabetes." MFMER https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444 (accessed June, 13, 2023)

[3] Y. Su, C. Huang, W. Yin, X. Lyu, L. Ma and Z. Tao, "Diabetes Mellitus risk prediction using age adaptation models", *Biomed. Signal Process. Control*, vol. 80, pp. 104381, Febrary 2023.

[4] F. Alaa Khaleel and A. M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms", *Mater. Today: Proc.*, July 2021.

[5] B. S. Ahamed, M. S. Arya, S. K. B. Sangeetha and N. V. Auxilia Osvin, "Diabetes Mellitus Disease Prediction and Type Classification Involving Predictive Modeling Using Machine Learning Techniques and Classifiers", *Appl. Comput. Intell. Soft Comput.*, vol. 2022, pp. 1–11, Dicember 2022.

[6] S. Saxena, D. Mohapatra, S. Padhee and G. K. Sahoo, "Machine learning algorithms for diabetes detection: a comparative evaluation of performance of algorithms", *Evol. Intell.*, November 2021.

[7] I. Tasin, T. U. Nabil, S. Islam and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques", *Healthcare Technol. Lett.*, Dicember 2022.

[8] J.W. Smith, J.E. Everhart, W.C. Dickson, W.C. Knowler and R.S. Johannes. "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus", Proc Annu Symp Comput Appl Med Care. pp. 261–265, 1998.

[9] G. Wilcox, "Insulin and insulin resistance", Clin Biochem Rev, vol. 26, pp. 19-39, May 2005. PMID: 16278749; PMCID: PMC1204764.

[10] N.V. Chawla, K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer. "SMOTE: Synthetic Minority Over-Sampling Technique." Journal of Artificial Intelligence Research, vol. 16, pp. 321–57, June 2002.

[11] S. Ben-David y S. Shalev-Shwartz, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[12] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Incorporated, 2022.