applied sciences

MDPI

Article

Evaluation of Synthetic Categorical Data Generation Techniques for Predicting Cardiovascular Diseases and Post-Hoc Interpretability of the Risk Factors

Clara García-Vicente [1,†], David Chushig-Muzo [1,†], Inmaculada Mora-Jiménez [1], Himar Fabelo [2,3], Inger Torhild Gram [4,5], Maja-Lisa Løchen [5], Conceição Granja [4,6] and Cristina Soguero-Ruiz [1,*]

[1] Department of Signal Theory and Communications, Telematics and Computing Systems, Rey Juan Carlos University, Fuenlabrada, 28943 Madrid, Spain
[2] Fundación Canaria Instituto de Investigación Sanitaria de Canarias, 35019 Las Palmas de Gran Canaria, Spain
[3] Research Institute for Applied Microelectronics, University of Las Palmas de Gran Canaria, 35017 Las Palmas de Gran Canaria, Spain
[4] Norwegian Centre for E-health Research, University Hospital of North Norway, 9019 Tromsø, Norway
[5] Department of Community Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway, 9019 Tromsø, Norway
[6] Faculty of Nursing and Health Sciences, Nord University, 8026 Bodø, Norway
[*] Correspondence: cristina.soguero@urjc.es
[†] These authors contributed equally to this work.

Abstract: Machine Learning (ML) methods have become important for enhancing the performance of decision-support predictive models. However, class imbalance is one of the main challenges for developing ML models, because it may bias the learning process and the model generalization ability. In this paper, we consider oversampling methods for generating synthetic categorical clinical data aiming to improve the predictive performance in ML models, and the identification of risk factors for cardiovascular diseases (CVDs). We performed a comparative study of several categorical synthetic data generation methods, including Synthetic Minority Oversampling Technique Nominal (SMOTEN), Tabular Variational Autoencoder (TVAE) and Conditional Tabular Generative Adversarial Networks (CTGANs). Then, we assessed the impact of combining oversampling strategies and linear and nonlinear supervised ML methods. Lastly, we conducted a post-hoc model interpretability based on the importance of the risk factors. Experimental results show the potential of GAN-based models for generating high-quality categorical synthetic data, yielding probability mass functions that are very close to those provided by real data, maintaining relevant insights, and contributing to increasing the predictive performance. The GAN-based model and a linear classifier outperform other oversampling techniques, improving the area under the curve by 2%. These results demonstrate the capability of synthetic data to help with both determining risk factors and building models for CVD prediction.

Keywords: synthetic categorical data generation; generative adversarial networks; imbalance learning; CTGAN; interpretable machine learning; cardiovascular disease

check for updates

Citation: García-Vicente, C.; Chushig-Muzo, D.; Mora-Jiménez, I.; Fabelo, H.; Gram, I.T.; Løchen, M.-L.; Granja, C.; Soguero-Ruiz, C. Evaluation of Synthetic Categorical Data Generation Techniques for Predicting Cardiovascular Diseases and Post-Hoc Interpretability of the Risk Factors. *Appl. Sci.* **2023**, *13*, 4119. https://doi.org/10.3390/app13074119

Academic Editor : Rasa Ruseckaite

Received: 3 February 2023
Revised: 15 March 2023
Accepted: 20 March 2023
Published: 23 March 2023

1. Introduction

With the ongoing development in information and communication technologies, unprecedented amounts of data have been generated in multiple fields of healthcare [1]. Data-driven models offer great research opportunities, allowing us to extract clinical knowledge and support decision-making. Along with data proliferation, the use of Artificial Intelligence (AI) has been intensified in recent years fostered by advances in computer processing, software platforms, and automatic differentiation [2]. Within AI, Machine Learning (ML) methods have attracted significant attention in both academia and industry, being used in multiple domains ranging from computer vision to natural language processing [3] and for different tasks such as classification, regression, and clustering, among others.

Despite the potential of ML methods, most of them are generally hampered by the class imbalance problem. It occurs when the proportion of samples of one class greatly outnumbers the others in the learning process [4,5]. Most ML algorithms are built to work properly in scenarios where the number of samples per class is fairly equal, since learning with imbalanced datasets biases the classifier towards the majority class. To deal with this, several methods have been proposed in the literature [4–7], which can be classified into two types, algorithmic-level and data-level approaches. The former adapts the loss function of the algorithm by assigning a higher weight to the misclassification of samples associated with the minority classes during the training process [8]. Examples of these approaches include cost-sensitive learning and ensemble methods [8]. In contrast, data-level approaches balance the class distribution by undersampling the majority class, by oversampling the minority classes, or by considering a hybrid approach that combines undersampling and oversampling approaches [9].

In this paper, we primarily focus on oversampling techniques. Among them, Synthetic Minority Oversampling Technique (SMOTE) [10] is one of the most used. SMOTE relies on the nearest neighbors technique, aiming to generate new samples that are mid-way between the nearest neighbors in any particular class. SMOTE has been used to generate numerical data and improve the generalization of predictive models in tasks such as regression and classification [11–13]. However, many real-world applications present high-dimensional and heterogeneous data (mixed-type) with numerical and categorical features. SMOTEN and its variants have been used in several clinical works for improving the predictive performance in identification or prevention of diseases, such as cervical cancer prevention using risk factor inputs [14], sepsis early prediction using both structured and unstructured data [15], COVID-19 identification using chest X-ray images [16], and smartphone-based cough recordings [17] or the cardiovascular monitoring of COVID-19 patients by using wearable medical devices [18].

Recently, generative models based on Artificial Neural Networks (ANNs) have revolutionized the outcomes in multiple knowledge areas due to their outstanding performance for creating synthetic data, particularly in computer vision and image applications [19,20]. Despite these outcomes, a few strategies have been studied for generating tabular (structured) data. For instance, a variant of the Variational Autoencoder (VAE) called Tabular VAE (TVAE) has been proposed, which uses two ANNs and trains using evidence lower-bound loss with the goal of creating mixed-type data [21]. Additionally, the techniques based on Generative Adversarial Networks (GANs) have emerged as a potential tool for creating synthetic data, frequently enhancing the model's performance in classification tasks, while also addressing data privacy issues. Although the application of GANs has been validated in different domains [19], they have not been well studied when considering electronic health records (EHRs) with structured and categorical and continuous data [22]. Because tabular data typically contain a mix of categorical and continuous features, generating realistic synthetic data is not an easy task. In this sense, Conditional Tabular GANs (CTGANs) have been created for modeling tabular data distributions and sampling entries from them [21] by employing a conditional generative antagonistic network. Furthermore, in several real-world datasets, CTGAN has outperformed Bayesian approaches [21].

In healthcare applications, class imbalance is a recurrent challenge to building predictive models with a reasonable generalization capacity, because a highly skewed distribution of training data will be prone to forcing the learning algorithm to be biased towards the majority class. In particular, we focused our research on cardiovascular diseases (CVDs) since they are the most significant cause of death worldwide [23]. Specifically, we analyzed data collected with a smartphone-based method from a population group in Norway [24]. The dataset comprised a series of survey questions related to socioeconomic factors, alcohol, and drug use, physical activity (PA), dietary intake, and one question indicating current/previous non-communicable diseases. Working with categorical features in ML is challenging due to most algorithms working adequately with numerical data, one-hot encoding being one of the most popular approaches to transforming categories into numbers [25]. However, this approach generally returns a sparse matrix, increasing the number of features (dimensions) that the model handles, and the risk of the curse of dimensionality problem [26]. Furthermore, when the feature includes an excessive number of categories, the majority of which are irrelevant to the prediction, this is amplified. To cope with these issues, a target encoding strategy is used [25], which aims to encode the categories by replacing them with a measurement of the impact that they may have on the target.

This research aimed to perform a comparative study of synthetic categorical data generation methods, with a special focus on GAN-based models. To this end, we have generated new samples with oversampling methods that seek to maintain the same feature categories as the original data, a similar probability distribution of attributes and the dependence between them, thus addressing the problem of data imbalance. All of this enables the enhancement of the effectiveness and accuracy of the developed classifiers. However, some ML methods use nonlinear transformations, leading to a lack of interpretability and creating black-box models [27]. Interpretation is defined as the process of generating human-understandable explanations of outcomes provided by computational models [27,28]. Several approaches have been proposed for gaining interpretability for improving model understandability and reliability, highlighting *model-specific* and *model-agnostic* methods [29]. The former is based on feature weighting, which seeks to identify the contributions of the features that determine the predictions of ML models. Although feature weighting is easy to apply to simple linear models, these models tend to have limited predictive performance and, therefore, also have limited interpretive relevance. The second approach outlined above, the model-agnostic approach, appears to address this limitation, which aims to extract post-hoc explanations by treating the original model as a black box.

Concerning the model interpretability, among the most popular interpretable models, the generalized linear and tree-based models are of great value to interpreting model predictions [30,31]. In this work, two linear models were considered: Least Absolute Shrinkage and Selection Operator (LASSO) [32] and Linear Support Vector Machine (SVM) [33]. The goal was to extract the most relevant features by analyzing the weights of the coefficients of each of the features to give information about their significance for predicting the output class. As a nonlinear model, a Decision Tree (DT) was considered since it provides the importance of each feature [34]. Additionally, the inherent characteristics of various ML models, due to nonlinear transformations in the learning process, make them powerful in terms of predictive performance, but they lack interpretability. In the case of the nonlinear ML classifier, such as K-Nearest Neighbors (KNN) [35], we focused on post-hoc interpretability called Shapley Additive Explanations (SHAP) [36], which is founded on game theory and local explanations. Since SHAP provides the contribution of each feature in the model's output, it can be considered a tool for model interpretability.

We next summarize our main contributions: *(i)* a comparison of different resampling and neural network generative models, highlighting oversampling techniques, for generating categorical data and their influence in the binary classification performance; and *(ii)* a methodology to interpret the more representative risk factors/features for identifying CVD subjects by using a dataset composed by sociodemographic, lifestyle and clinical categorical variables.

The remaining article is organized as follows. Section 2 describes the dataset and the pre-processing method. Additionally, we present the foundations of categorical encoding techniques, and the resampling techniques for addressing the imbalance learning problem. Section 3 presents the experimental setup, classification performance, and model interpretability outcomes of the linear and nonlinear models that were considered. Finally, the discussion and conclusions are presented in Sections 4 and 5, respectively.

2. Materials and Methods

First, this section presents the dataset description and the pre-processing method. Then, the workflow followed in the paper is introduced, where we present the foundations of categorical encoding techniques and resampling techniques, highlighting GAN-based models. Finally, we introduce several quality metrics for evaluating the synthetic data generated by oversampling methods.

2.1. Dataset Description and Pre-Processing

Data acquisition was carried out as part of a three-year project conducted by the Norwegian Centre for E-health Research, UiT The Arctic University of Norway, and Healthcom [24]. This study was approved by the Data Protection Section of the University Hospital of North Norway, and all participants signed the corresponding informed consent. The dataset was obtained through a survey questionnaire made to a population from Norway using a smartphone-based solution over a 2-year period [24]. The responses to the survey were anonymized after their submission. The aim was to track modifiable risk factors for four non-communicable diseases: CVD, diabetes, cancer, and chronic respiratory illness. However, during the data acquisition campaign, only individuals affected by CVD ($n = 465$), cancer ($n = 72$), and both CVD and cancer ($n = 46$) diseases were collected. Additionally, 1578 healthy individuals filled out the questionnaire. Due to the low number of individuals with cancer and both diseases, in this work, we only focused on CVD and healthy subjects. In summary, a total of 2043 individuals were included in the subsequent analysis. The dataset consists of 26 features and was organized into five different groups: *background, substance use, PA, dietary intake,* and *income*. Questions related to alcohol followed a disorders identification test [37], and information about PA was based on the International Physical Activity Questionnaire [38]. A summary of the categories associated with each feature is presented in Table 1.

Since all features in our study were categorical, for an exploratory analysis, we followed a one-hot encoding strategy to encode each category in a specific variable by assigning the value of ‘0’/‘1’. To better characterize and visually analyze the prevalence of certain categories compared to others, we built a representation named profile [39], showing the most prevalent risk factors for a specific disease. The profile is a one-dimensional visual representation, where the x-axis shows all the categories and the y-axis represents the frequency ratio of occurrence of each of the categories for a specific group of individuals. Figure 1 shows the profiles for both healthy and CVD subjects related to the five main groups of variables previously presented. Prior to building profiles and aiming to deal with missing data, we performed the following steps. First, we identified the number of missing values (represented by ‘NAN’ identifier) for each feature, which indicates that the answer is not available. Since the proportion of ‘NAN’ was low compared to the proportions of the rest categories, we adopted a simple imputation approach by replacing ‘NAN’ with a new category named ’NA’ (not available answer for a question). As shown in Figure 1, the values of ‘NA’ in each feature are lower compared to other categories, which could suggest that the imputation approach used to compensate for missing values will have a slight impact on the data distribution of features.

In Figure 1a, focusing on the features with the highest frequency of appearance (age), we observe that the population aged in the range between 60 and 69 years predominates for CVD subjects. By analyzing sex, we observe that there is a higher percentage of male subjects with CVD, while in the healthy group the frequency of occurrence is higher in women. Regarding BMI, there is a higher frequency of healthy people with a healthy weight (HW) with respect to the CVD group. Finally, note that presence of high cholesterol is clearly linked to the group of CVD subjects. In Figure 1b, related to the consumption of harmful substances, we observe that a high percentage of individuals are non-smokers, this percentage being higher among the healthy subjects. In general, individuals of both groups report alcohol consumption, though it is usually one or two units per week. Figure 1c shows how PA is performed. It has been found that both healthy and CVD subjects engage in severe and moderate PA on average one to two times per week, walk seven days per week, and spend three to five hours per day sitting down. The features related to dietary intake (see Figure 1d) show that there is a high number of CVD subjects who never ingest salt compared to the healthy group (who take it occasionally). The percentage of individuals who do not drink sugary drinks is similar in both groups, this being the majority case. As for fruit consumption, there is a higher frequency of one or two pieces per day in both groups. In Figure 1e, we can observe that the greatest wages are earned by healthy people, and two elderly people are the most common residents of the home, according to our observations. In the case of CVD subjects, there is a higher number of people with no minors living in the same house.

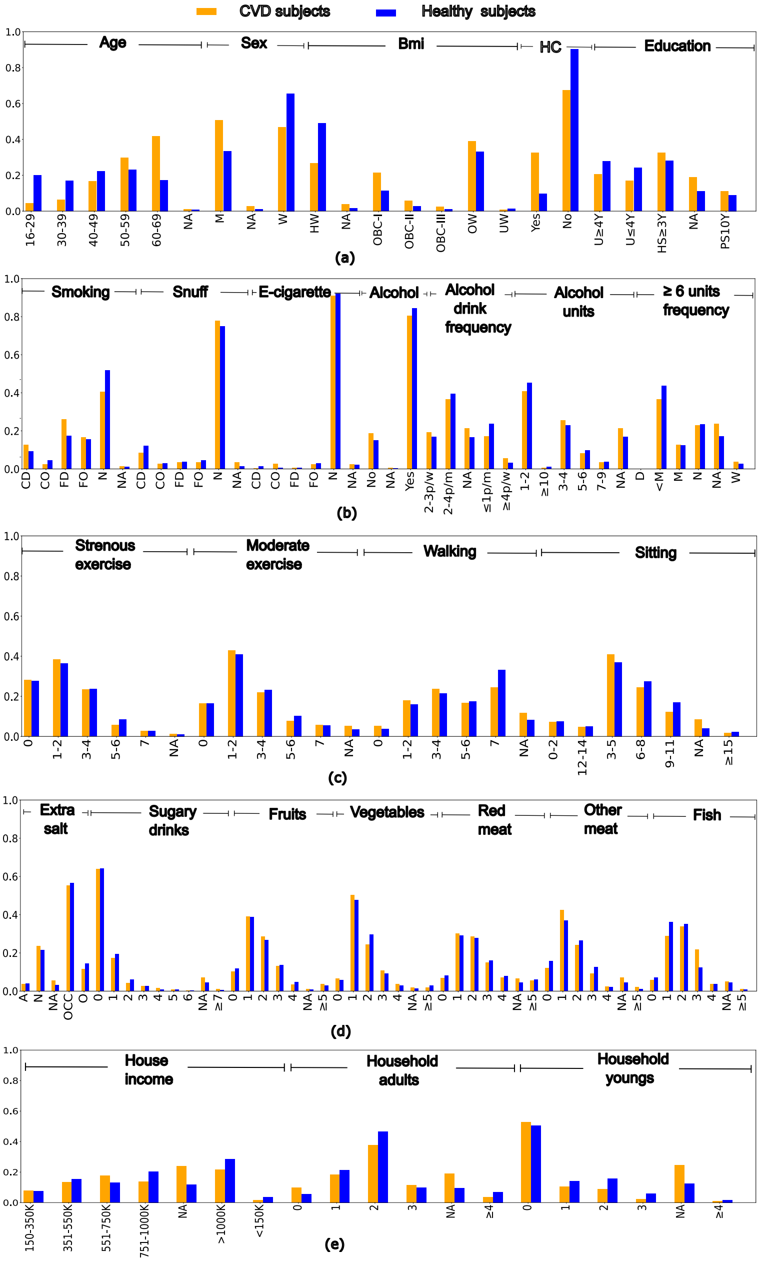


Figure 1. Profiles for healthy and CVD individuals considering: (**a**) background; (**b**) substance use; (**c**) PA; (**d**) dietary intake; and (**e**) income features.

Table 1. Summary of the features and categories in the dataset.

	Feature	Description	Categories
Background	Age	Indiviual's age	16–29, 30–39, 40–49, 50–59, 60–69, NA
	Sex	Indiviual's sex	Woman (W), Man (M), NA
	BMI [a]	Body mass index	HW, OBC-I, OBC-II, OBC-III, OW, UW, NA
	Education [b]	Education level achieved	U $<$ 4Y, U $\geq$ 4Y, PS10Y, HS $\geq$ 3Y, NA
	HC	Have cholesterol	Yes, No
Substance use	Smoking [c]	Cigarette use	CD, FO, FD, CO, N, NA
	Snuff [c]	Snuff use	CD, FO, FD, CO, N, NA
	E-cigarette [c]	E-cigarette use	CD, FO, FD, CO, N, NA
	Alcohol	Alcohol consumption	Yes, No
	Alcohol freq.	Alcohol drink frequency	2–3 p/w, 4 p/w, $\leq$1 p/m, 2–4 p/m, NA
	Alcohol units	# units consumed	1–2, 3–4, 5–6, 7–9, $\geq$10, NA
	$\geq$6 units	$\geq$6 units of consumed alcohol	D, W, $<$M, M, N, NA
PA	Strenuous PA	# days of strenuous PA	0, 1–2, 3–4, 5–6, 7, NA
	Moderate PA	# days of moderate PA	0, 1–2, 3–4, 5–6, 7, NA
	Walking	# days of walking $\geq$10 min	0, 1–2, 3–4, 5–6, 7, NA
	Daily sitting	# hours sitting on a weekday	0–2, 3–5, 6–8, 9–11, 12–14, $\geq$15, NA
Dietary intake	Extra salt [d]	Freq. extra salt added to food	N, OCC, O, A, NA
	Sugary drinks [f]	# sweetened drinks	0, 1, 2, 3, 4, 5, 6, $\geq$7, NA
	Fruits/Berries [f]	Fruit servings and berries	0, 1, 2, 3, 4, $\geq$5, NA
	Vegetables [f]	Lettuce and vegetable servings	0, 1, 2, 3, 4, $\geq$5, NA
	Red meat [e]	# consumed red meat	0, 1, 2, 3, 4, $\geq$5, NA
	Other meat [e]	# consumed processed meat	0, 1, 2, 3, 4, $\geq$5, NA
	Fish [e]	# consumed fish products	0, 1, 2, 3, 4, $\geq$5, NA
Income	House income	Gross household income	$\leq$150 K, 150–350 K, 351–550 K, 551–750 K, 751–1000 K, $\geq$1000 K, NA
	Household adult	# household members $\geq$ 18 years	0, 1, 2, 3, $\geq$4, NA
	Household young	# household members $\leq$ 18 years	0, 1, 2, 3, $\geq$4, NA

Description of categories. [a] For BMI: Underweight (UW), Healthy Weight (HW), Overweight (OW), and Obesity Class I (OBC-I), II (OBC-II), and III (OBC-III) [b] For education, Primary School up to 10 Years (PS10Y), High School (HS) minimum of 3 Years (HS $\geq$ 3Y), less than 4 years (U $<$ 4Y) and 4 years or more of University (U) (U $\geq$ 4Y). [c] For substance use: Currently Daily (CD), former occasional (FO), Former Daily (FD), Current Occasional (CO), and Never (N). [c] For consumption of $\geq$ six units of alcohol: Daily (D), less than once a month ($<$M), Monthly (M), Weekly (W), and Never (N). [d] For extra salt: Always (A), Often (O), Occasionally (OCC), and Never (N). [e] For red meat, other meat, and fish, they were measured per week (p/w). [f] For fruit servings and berries, vegetables, and sugary drinks were measured per day (p/d). NA stands for Not Available.

2.2. Workflow for Predicting Cardiovascular Diseases

The proposed pipeline for predicting CVD and identifying associated risk factors using ML methods is based on three stages (see Figure 2). First, the target encoding was carried out to transform categorical features into numerical values. Next, we present and evaluate two different approaches: (1) considering all features, and (2) considering a Feature Selection (FS) strategy based on a bootstrap resampling test to identify the most relevant features. Finally, using the training subset, we implemented different strategies of undersampling and oversampling methods for data balancing. Then, we trained several classifiers with the balanced datasets to predict whether an individual is affected by CVD or not. Finally, using only real data from the test subset that was not used for training, the performance was evaluated, analyzing the most important features using post-hoc interpretability models.

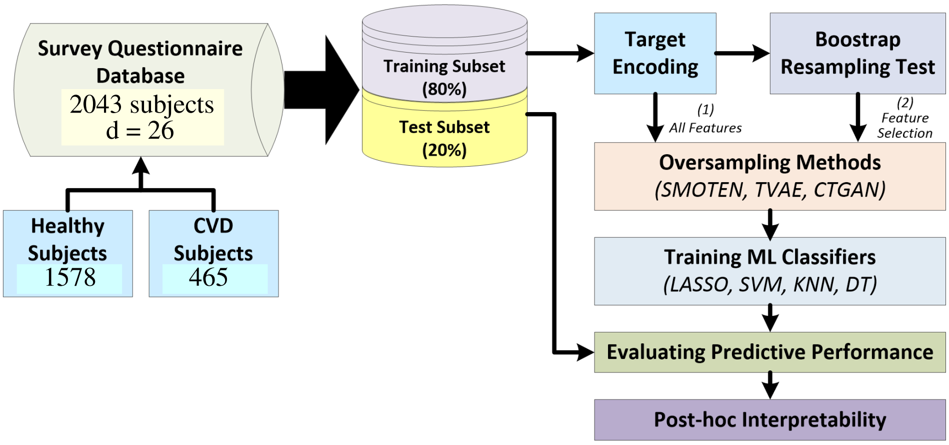

Survey Questionnaire Database
2043 subjects
d = 26
Healthy Subjects
1578
CVD Subjects
465
Training Subset (80%)
Test Subset (20%)
Target Encoding
Boostrap Resampling Test
(1) All Features
(2) Feature Selection
Oversampling Methods (SMOTEN, TVAE, CTGAN)
Training ML Classifiers (LASSO, SVM, KNN, DT)
Evaluating Predictive Performance
Post-hoc Interpretability

Figure 2. Proposed workflow using oversampling techniques (SMOTEN, TVAE, CTGAN) and different classifiers designed using ML approaches.

2.2.1. Encoding Categorical Data

Databases with categorical features are common in real-world applications and provide relevant information for predictive tasks. A categorical feature is composed of a discrete set of values called categories. In ML, working with categorical features is challenging because most algorithms work adequately with numerical data [25]. To make proper use of the information on these features, a pre-processing stage named encoding is needed, which consists in transforming all the categories of the feature into numerical values. In the literature, several techniques have been proposed to encode categorical data considering target-agnostic or target-based methods [25]. One of the most popular target-agnostic methods is one-hot encoding, in which for each individual arrays of '0'/'1' are created based on the presence/absence of a category in the corresponding feature [40]. Despite its extensive use, prior works that have analyzed categorical encoding techniques have pointed out two main shortcomings of the use of one-hot encoding techniques [40–44]. First, the dimension of the input space directly increases with the number of categories in the encoded features, thus substantially augmenting the dimensionality [25]. Second, the new features created are characterized to be sparse.

To overcome the limitations raised by one-hot encoding, a target-based method called target encoding has been proposed by [25]. In this technique, framed within Bayesian encoding techniques, each value of the categorical feature is mapped to a target mean conditional (target's posterior probability) on the value of the variable [25]. Broadly speaking, the target variable's mean for each category is computed and the original category is replaced by this value. The target encoding is obtained as follows:

$$\mu = \frac{N \times \overline{x} + m \times w}{N + m},$$

μ being the computed mean that will replace categorical data, N the number of samples, $\overline{x}$ the estimated mean for each category of a feature taking into account the target, w the smoothing parameter and m the overall mean of the target.

This encoding technique is not affected by the increase of dimensionality, is computationally less demanding than one-hot-encoding, and allows us to work with numerical data [25,45]. Despite these advantages, the main drawback of this technique is that it can be affected by overfitting since the encoding depends on the target [45]. Furthermore, a bias can be introduced in the ML model when categories with few samples are replaced by values close to the desired target feature since the model over-trusts the target encoded feature and the model is prone to overfitting. To overcome this problem, we applied a regularization-based approach that uses a smoothing parameter w aiming to shrink the effects toward the global mean [25].

2.2.2. Feature Selection Using Bootstrap Resampling Test

The performance of ML models can be affected by the number of input features [33]. To cope with this problem, FS techniques aim to find a subset of the input features that best describes the underlying structure of the data [46]. In this paper, we used a non-parametric technique called bootstrap resampling to estimate the distribution of a statistic (e.g., the mean) by obtaining samples from a population without replacement [47,48]. To do this, we resampled individuals of both classes 3000 times with the same number of samples each time (balancing the classes). For each resampling, we determined the mean difference Δ between the populations and the confidence interval (CI) computed at the 95% level (95% percentile) for each feature [48]. The null hypothesis H_0 holds true if $0 \in CI_{\Delta}$; the alternative hypothesis H_1 holds true when $0 \notin CI_{\Delta}$ (no overlapping of the CI_{Δ} with the zero value). Given a dataset $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ consisting of n patients, with i-th being the sample represented by a vector of d features, for a classification analysis, $\hat{d}$ features that fulfill H_1 will be selected (being $\hat{d} < d$).

2.2.3. Resampling Techniques for Imbalance Learning

Most ML algorithms work reasonably well when the number of training samples in the different classes is almost equal (balanced dataset) [49]. Nevertheless, in real-world scenarios, it is common that the distribution of training samples for each class is skewed. This hampers the application of ML algorithms since the learning process can be monopolized by samples of the majority class, impacting the generalization and performance of the classifiers [49]. To address this, class balancing strategies, which include oversampling, undersampling, and a hybrid approach are used to modify the class distribution of imbalanced datasets [8]. In undersampling, a random number of samples of the majority class is discarded [50]. However, the aim of oversampling techniques is to increase the number of samples by creating samples from the minority class. In this latter approach, exact copies of the training samples in the minority class can be created and therefore the generalization capabilities may decrease [10]. In the case of the hybrid strategy, the two approaches just exposed are used. The idea is to balance the number of the training samples in both classes by first oversampling the minority class and then undersampling the majority class.

In this paper, we compared the results when using different oversampling methods, using Random Under Sampling (RUS) as a benchmark. Among the most common oversampling methods, SMOTE [10] and variants such as SMOTE nominal (SMOTEN) have been applied in prior works for balancing datasets, improving the performance of predictive models [11,12]. The SMOTE algorithm is based on oversampling the minority class by adding random synthetic samples from the minority class [51]. Since SMOTE only deals with continuous features, SMOTEN [10] has been proposed for dealing with categorical features.

This section also presents the two generative neural network models implemented in this work: TVAE [52] and GANs [53]. TVAE is based on VAE, a latent generative model proposed by Kingma and Welling [52] which is composed of two parts, a generative and an inference model. In the generative part, given a sample $\mathbf{x}$, a probabilistic decoder produces a distribution over the latent values $\mathbf{z}$. In the inference part, a probabilistic encoder outputs a latent variable $\mathbf{z}$ into $\hat{\mathbf{x}}$. The variational lower boundary of the marginal likelihood of the input data is selected as an objective function of VAE. In recent years, VAE has been extensively used in applications such as image/text classification, anomaly detection or image generation [54]. To generate structured/tabular data, the variant TVAE proposed in [21] allows the generation of mixed-type tabular data.

GANs were proposed by Goodfellow [53] and are generative models composed of two ANNs: (i) A generator G that takes a random vector $\mathbf{z}$ from a distribution $F_z \sim \mathcal{N}(0,1)$ and projects it to a vector $\hat{\mathbf{x}}$; and (ii) a discriminator D that seeks to discriminate between real and synthetic data. The goal of $G(\cdot)$ is to generate synthetic data with characteristics that are as close to real data as feasibly possible. $G(\cdot)$ and $D(\cdot)$ aim to optimize a zero-sum min-max game, with the value function $V(G, D)$ as follows:

$$\min_G \max_D V(G, D) \; = \; \mathbb{E}_{x \sim \rho_{data}(x)}[logD(\mathbf{x})] + \mathbb{E}_{z \sim \rho_z(z)}[log(1 - D(G(\mathbf{z}))),$$

where $\rho_{data}(x)$ and $\rho_z(z)$ are the distribution of the real data and that of the samples generated by G, respectively. Symbols $\mathbf{x}$ and $\mathbf{z}$ represent the samples from the input and the latent space, and E_x and E_z are the expected log-likelihood from the different outputs of both real and generated samples.

GANs have been used in a variety of applications because of their great performance in generating synthetic data, particularly receiving a lot of interest in the computer vision field due to their ability to generate synthetic images [19]. Although GANs have been extensively applied for generating new images, only a few research studies have proposed these models when dealing with structured data. Among them, CTGAN proposed by [21] has been proposed for generating tabular mixed data (handling both continuous and categorical features). As stated, categorical features present a challenge for GANs since both the generator and the discriminator need to be differentiable. To solve this, the Wasserstein divergence and the weight-clipping with a gradient penalty are used by the CTGAN [21].

2.2.4. Quality Metrics for Synthetic Data

Let be a dataset $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ represented by a $N \times d$ matrix, consisting of $i = \{1, \ldots, N\}$ samples and $j = \{1, \ldots, d\}$ features, with the i-th sample represented by a vector $\mathbf{x}_i \; = \; [x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(d)}]$. $\mathcal{X}$ was partitioned into training subset $\mathcal{X}_{train}$ and test subset $\mathcal{X}_{test}$, with 80% and 20% of the samples, respectively. From $\mathcal{X}_{train}$, we selected the samples associated with the minority class $\mathcal{X}_R$ (real data) and oversampling techniques were used for creating the synthetic dataset $\mathcal{X}_S$. The number of synthetic samples generated directly depends on the Imbalance Ratio (IR), defined as $IR = N_{min}/N_{maj}$, with N_{min} and N_{maj} indicating the number of samples of the minority and the majority class, respectively.

The main goal of oversampling techniques is to generate synthetic samples from original data that capture underlying data structure, including the distribution of features and correlations between them. To examine whether the features of the synthetic dataset truly mimic the real features, we used several data quality metrics. To quantify the similarity between a pair of real and synthetic PMFs of $\mathcal{X}_R$ and $\mathcal{X}_S$, and for measuring univariate attribute fidelity, we used the symmetric Kullback–Leibler Divergence (KLD) [55] and the Hellinger Distance (HD) [56].

- KLD [55] measures the similarity between two PMFs. Given two PMFs, P_v and Q_v, for a given feature v, the KLD is defined as:

$$KLD(P_v||Q_v) \; = \; \frac{1}{2}KLD_{ns}(P_v||Q_v) + \frac{1}{2}KLD_{ns}(Q_v||P_v)$$

$$KLD_{ns}(P_v||Q_v) = \sum_i^{c_f} P_v(i) log \frac{P_v(i)}{Q_v(i)}$$

$$KLD_{ns}(Q_v||P_v) = \sum_i^{c_f} Q_v(i) log \frac{Q_v(i)}{P_v(i)}.$$

  Note that c_f is the number of categories for a given feature v, and ns denotes the non-symmetric KLD. When both distributions are identical the symmetric KLD is zero, while larger values indicate a larger discrepancy between the two PMFs.
- HD [56] quantifies the similarity between two PMFs, P_v and Q_v, for a specific feature v as follows:

$$HD(P_v, Q_v) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{c_f}\left(\sqrt{P_v(i)} - \sqrt{Q_v(i)}\right)^2}$$

where c_f is the number of categories for a given feature v. HD is ranged between 0 and 1, with 0 indicating that the two distributions are practically identical and 1 showing that they are the furthest apart.

To complement the information provided by the aforementioned measures, we additionally proposed in this work two metrics, named Mean Absolute Error Probability (MAEP) and Repeated Sample Vector Rate (RSVR):

- MAEP measures the absolute difference between two PMFs. Given PMFs P_v and Q_v for a given feature v, MAEP is defined as follows:

$$MAEP(P_v, Q_v) = \sum_{i=1}^{c_f} |P_v(i) - Q_v(i)|,$$

  where c_f is the number of categories for a given feature. KLD, HD and MAEP allow us to assess univariate attribute fidelity by determining the similarity between marginal distributions of $\mathcal{X}_R$ and $\mathcal{X}_S$.
- RSVR indicates the rate of repeated sample vectors in the synthetic data $\mathcal{X}_S$, denoting how well the oversampling methods create unique vectors. It is worth noting that this metric is directly related to the number of samples generated. As IR increases, the probability of appearance of repeated vectors may augment, and consequently the RSVR value.

It is important to remark that KLD, HD and MAEP are defined here at the feature level, i.e., they are computed for a single feature. To calculate an overall measurement, the average value across all features is computed by aggregating the contribution of each feature.

Furthermore, to evaluate if the oversampling methods capture adequately the relationships between features, the Pairwise Correlation Difference (PCD) [57] and the Log-Cluster Metric (LCM) [58] were considered.

- PCD quantifies the difference between the correlation matrices of $\mathcal{X}_R$ and $\mathcal{X}_S$ [57], thus measuring the correlation among features that the different methods of oversampling can capture. It is defined as:

$$PCD(X_R, X_S) = \|Corr(X_R) - Corr(X_S)\|.$$

  $Corr(\cdot)$ represents the correlation among features for $\mathcal{X}_R$ and $\mathcal{X}_S$, and $\|\cdot\|$ is a matrix norm that measures the similarity between two matrices (a scalar). Since we handled categorical features, Cramér's V correlation [59] was used. Small PCD values indicate that relationships between features are preserved, meaning that $\mathcal{X}_R$ is more similar to $\mathcal{X}_S$ in terms of linear correlations across the features.
- LCM [58] measures the similarity of the underlying latent structure of $\mathcal{X}_R$ and $\mathcal{X}_S$ in terms of clustering. This metric relies on two steps. Firstly, $\mathcal{X}_R$ and $\mathcal{X}_S$ were merged into one single dataset, and then a cluster analysis using the k-means algorithm [60] was performed on this dataset with a fixed number of clusters k [57]. LCM is defined as follows:

$$LCM(X_R, X_S) = log\left(\frac{1}{k}\sum_{i=1}^{k}\left[\frac{n_j^R}{n_j} - c\right]^2\right),$$

  n_i being the number of samples in the i-th cluster, n_i^R the number of samples from the real dataset in the i-th cluster, and $c = n^R/(n^R + n^S)$ (where n^R and n^S are the numbers of real and synthetic samples, respectively) [57]. High values of LCM denote disparities in the cluster memberships, indicating high differences in the distribution of $\mathcal{X}_R$ and $\mathcal{X}_S$ [57].

2.2.5. ML Classifiers and Figures of Merit

The ML classifiers used in this work were the following: LASSO [32], SVM [33], DT [34] and KNN [35]. To select the best hyperparameters for each classifier, we considered a k-fold Cross Validation (CV) [33] strategy with the training subset. We followed the 3-fold CV approach, and the sensitivity and the Area Under the Curve (AUC) of the receiver operating characteristic were considered as figures of merit to evaluate the predictive performance of the classifiers. Although binary classification has been studied extensively and standard evaluation measures are used to characterize classifier performance, many are unsuitable in imbalanced scenarios since performance in the majority class will be overrepresented. In imbalanced learning, the main objective is to improve the classification of minority class samples while maintaining reasonable performance concerning the majority ones. For this reason, we decided to use sensitivity and AUC metrics. Sensitivity measures the impact on the predictive performance of the minority class, while AUC provided us with a compromise metric between sensitivity and specificity. The following hyperparameters were tuned: λ for the LASSO model, C for the SVM model, the maximum tree depth and the minimum number of samples to split a node for DT, and the number of neighbors for KNN. The performance and post-hoc interpretability of the classifiers (LASSO, SVM, DT, KNN) were obtained when considering different resampling techniques (RUS and oversampling techniques (SMOTEN, TVAE, and CTGANs) and different class balancing strategies (only undersampling strategy, only oversampling strategy, and hybrid strategy) using different IRs.

3. Results

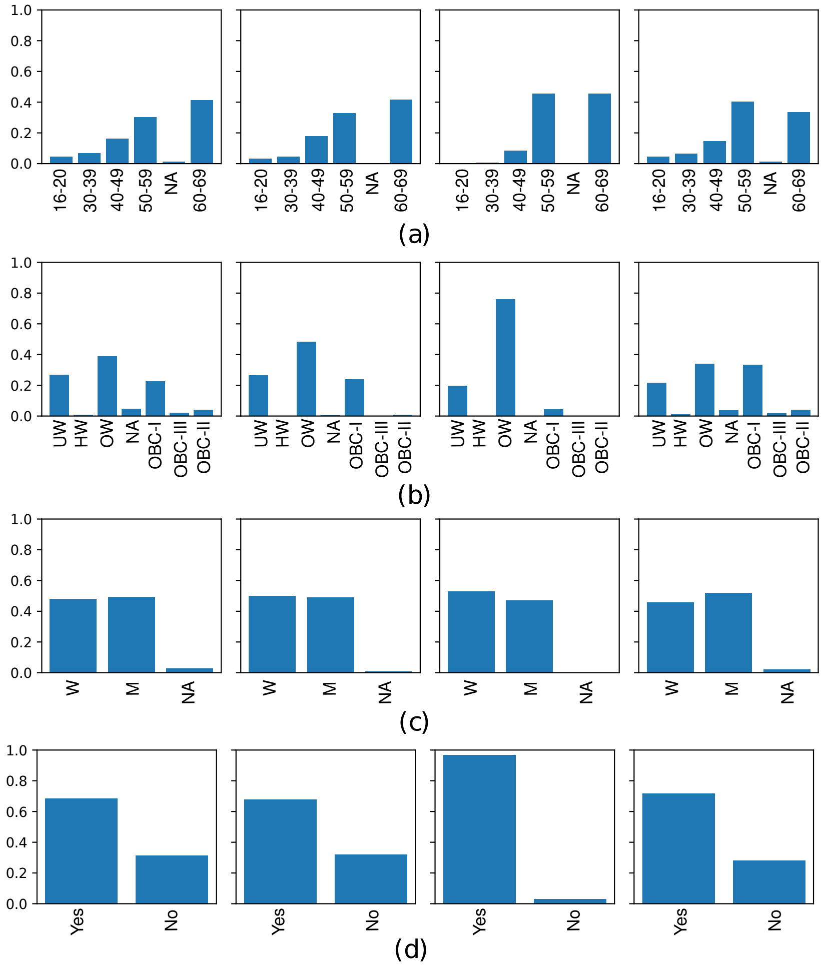
In this section, we analyzed the impact of combining one oversampling technique and one ML approach in a binary classification scenario for identifying between healthy individuals and CVD subjects. The experimental setup and then the figures of merit obtained are presented, including a comparison of classification performance by using all features and those selected by FS. Finally, post-hoc model interpretability based on the importance of the features was conducted.

3.1. Experimental Setup

The dataset was randomly split into training (80%) and test (20%) subsets (see Figure 2), and five independent training and test partitions were considered to further evaluate the performance of the model. The training subset was used for the model design, while the test subset was used to evaluate its performance (i.e., to evaluate its generalization ability). Bootstrap resampling was used to remove those features that were non-relevant and uninformative for predicting the target variable. Using this method, $\hat{d} = 14$ features were selected from the $d = 26$ initial features. Before training the classifiers, different values of the smoothing parameter w (between $[0.0, 1.0]$) were investigated to ensure proper use of the target encoding technique selected, addressing the over-fitting issue raised by target-agnostic approaches. The AUC values were used to select the best w for our dataset. Experimental results showed that lower values of w offer better AUC values, and consequently were more suitable for the binary classification scenario. In this paper, a regularization smoothing parameter of $w = 0.1$ was chosen for subsequent analyses.

3.2. Quality Evaluation of Synthetic Clinical Data

Several data quality metrics were considered to evaluate the similarity between synthetic and real data according to different IRs. As stated in Section 2.2.4, the PMFs associated with the different features were estimated as the previous step for computing these metrics. In Figure 3, the estimated PMFs associated with age, BMI, sex, and high cholesterol are depicted. The panels in the first column of the figure show the estimated PMF for those features when considering real data, and the remaining represent the PMFs estimated from synthetic data when using SMOTEN, TVAE and CTGAN. We aim to measure the similarity between the PMF of real data and that estimated with synthetic data, comparing how well the different oversampling methods can reproduce the PMF estimated using real data.



(a)

(b)

(c)

(d)

Figure 3. Estimations of the PMFs associated with: (**a**) age; (**b**) BMI; (**c**) sex; and (**d**) high cholesterol. Estimated PMF obtained using real data (first column); SMOTEN (second column); TVAE (third column); and CTGAN (fourth column).

Remarkably, the PMFs estimated using data generated with TVAE do not follow the probability distribution linked to the real data, showing a lack of probability values in certain categories. For instance, if we look at the age feature (see Figure 3a), in the PMF of synthetic data obtained with TVAE (third column), there are categories without samples, specifically for 16–20, 30–39 and NA. In the same manner, for the BMI feature (see Figure 3b), the categories HW, NA, OBC-II and OBC-III do not have values. Additionally, in some categories, the probability is much higher than that obtained when using real data (see the category OW in the BMI feature). We conclude that TVAE is the method that worst mimics the distribution for the four features considered. This further points out the low performance of TVAE for replicating samples of categorical data. By analyzing SMOTEN, we find similar insights that in TVAE. There are categories in the estimated PMF with zero values (see panels (a) and (b) associated with age and BMI). Regarding CTGAN, note that the PMFs estimated using the generated data are quite similar to those estimated using real data. All categories have probability values different from zero and they are similar to those estimated when considering real data. The insights drawn from these figures allow us to understand how similar the synthetic and real data are to one another in terms of univariate attribute fidelity, CTGAN being the oversampling approach that more precisely emulates the real distributions.

The next step was to analyze the data quality metrics considering different IR values (see Figure 4). We observed that, in terms of KLD, HD, MAEP, and RSVR (panels (a), (b), (c) and (d), respectively), CTGAN showed the best results. Note that lower values in these metrics indicate more similar distribution probabilities associated with features. Regarding PCD (Figure 4e), which measures whether the correlations between features are preserved, CTGAN also reached the lowest values compared to SMOTEN and TVAE. For LCM (Figure 4f), where the lower values indicate fewer differences in the distribution between $\mathcal{X}_R$ and $\mathcal{X}_S$, it was SMOTEN that reached the lowest values, followed by CTGAN. Regarding RSVR (Figure 4d), CTGAN showed the lowest values, indicating that it generates fewer repeated vectors.

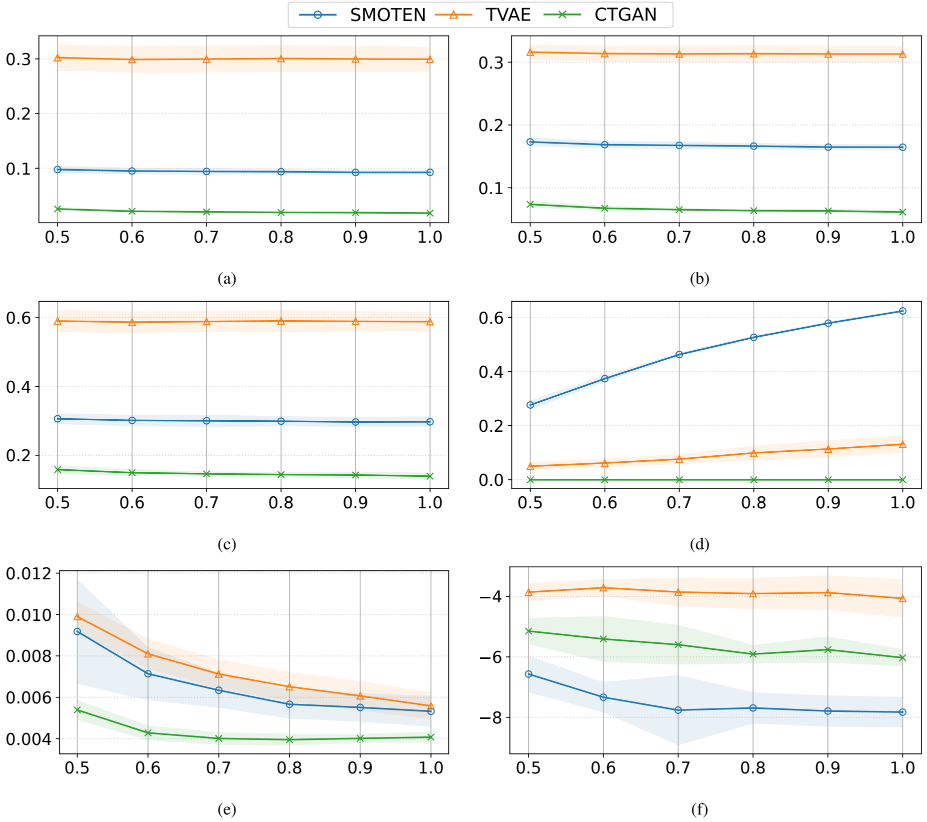


Figure 4. Mean ± std for the synthetic data quality metrics when considering a range of IR values (from 0.5 to 1) and different oversampling techniques (SMOTEN, TVAE and CTGAN). (**a**) KLD; (**b**) HD; (**c**) MAEP; (**d**) RSVR; (**e**) PCD; (**f**) LCM.

Table 2 presents a summary of the quality metrics that analyzes: (*i*) the similarity between the features of $\mathcal{X}_R$ and $\mathcal{X}_S$ (KLD, HD, and MAEP) and (*ii*) the relationships between features captured (PCD and LCM). The results of RSVR are shown to identify the percentage of repeated vectors generated by the oversampling methods. Note that these metrics were obtained considering IR = 1.0 and 5 different training partitions, and we present the mean and the standard deviation (std). As previously mentioned, CTGAN more accurately simulated true distributions, achieving the best KLD, HD, and MAEP values. Regarding PCD, which measures how well the correlations between features are captured, CTGAN also reached the best performance (the lowest value). CTGAN also obtained the lowest RSVR value, generating less repeated vectors associated with subjects. It is in LCM where SMOTEN reached a better performance (lowest values). Therefore, we conclude the potential of CTGAN for generating categorical synthetic data against other oversampling methods.

Table 2. Mean ± std (evaluated on 5 training partitions) of the data quality metrics for different oversampling techniques when considering IR = 1.0. The best values are shown in bold.

Method	KLD	HD	MAEP	RSVR	PCD	LCM
SMOTEN	0.092 ± 0.004	0.168 ± 0.006	0.298 ± 0.014	0.609 ± 0.011	0.005 ± 0.001	**−8.116 ± 0.407**
TVAE	0.299 ± 0.022	0.313 ± 0.013	0.590 ± 0.032	0.177 ± 0.020	0.006 ± 0.001	−3.926 ± 0.517
CTGAN	**0.017 ± 0.002**	**0.061 ± 0.003**	**0.145 ± 0.006**	**0.001 ± 0.001**	**0.003 ± 0.001**	−6.110 ± 0.235

3.3. Classification Performance

This section presents the classification results provided by linear (LASSO and SVM) and nonlinear (DT and KNN) classifiers using SMOTEN, TVAE, and CTGAN and considering different resampling strategies.

In Figure 5, we show the mean and std of sensitivity and AUC values on 5 test subset partitions, considering different classifiers (LASSO, SVM, DT and KNN) and oversampling strategies. It can be seen that there is a direct relationship between the IR and the model performance in terms of sensitivity and AUC, indicating that the higher the number of synthetic samples generated, the better the performance of the ML models. By analyzing Figure 5a,c,e, it can be observed that the sensitivity values present high variability for the DT model and all oversampling methods. On the contrary, the linear models LASSO and SVM present lower variability (low std), being the most robust classifiers. For small IR values (0.4), the AUC is approximately 0.6 for all models (see Figure 5b,d,f), achieving better performance when the IR increases. The highest AUC values are obtained when CTGAN linear models are considered (see Figure 5f).

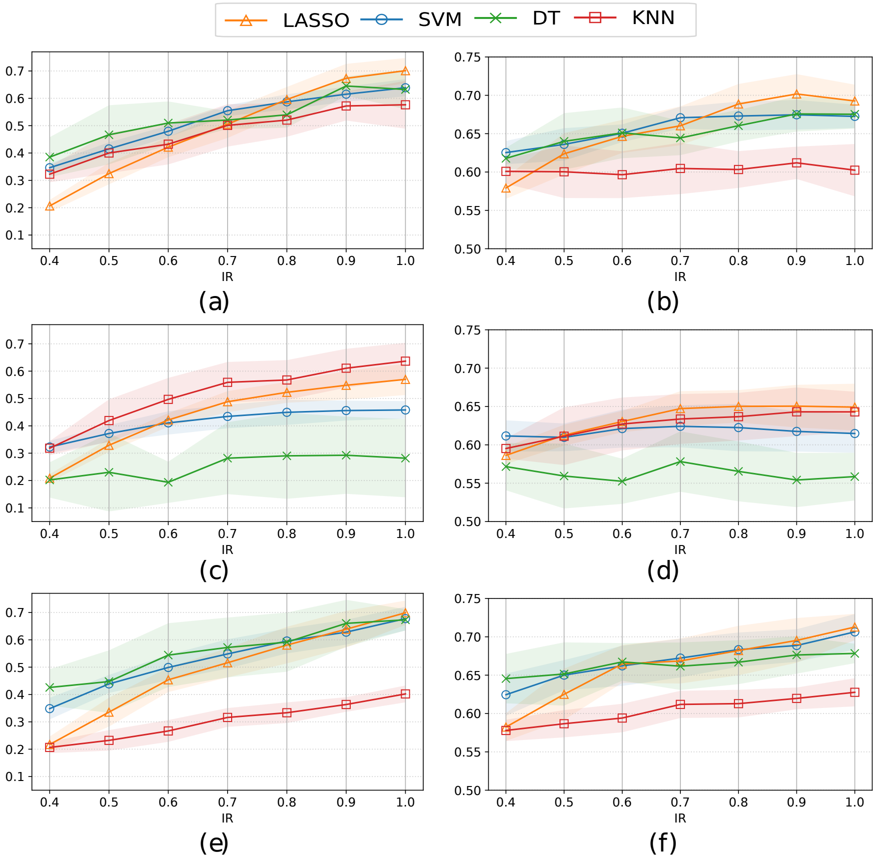


Figure 5. Mean ± std of the sensitivity (left panels) and AUC (right panels) considering 5 test subset partitions and different IR values, classifiers (LASSO, SVM, DT and KNN) and an oversampling approach with: (**a**,**b**) SMOTEN; (**c**,**d**) TVAE; and (**e**,**f**) CTGAN.

In general, the best figures of merit are obtained when applying linear models and considering SMOTEN and CTGAN techniques. The results obtained for TVAE are slightly different, where higher values are obtained with the nonlinear KNN model than with the linear models, specifically when IR increases. We can conclude that the best performance is obtained when considering CTGAN and IR = 1.0, i.e., when the number of samples of the minority and the majority classes is the same.

Next, we show in Figure 6, the sensitivity and AUC values when using different classifiers and the hybrid resampling approach. Two main insights can be drawn. Firstly, in general terms and referring to the six panels, it can be observed that the increase in the IR values (from 0.5 to 1.0), which refers to the increase in the generation of the number of samples of the minority class, does not have a positive impact on the classification measures (neither for sensitivity nor for AUC). For this approach, generating a high number of synthetic samples from the minority class and training the ML classifiers with a more substantial number of synthetic data does not imply obtaining a better predictive performance. This indicates that, after a given IR value, no matter how many additional samples are added, the performance remains unchanged. In terms of sensitivity and AUC, it is interesting to note that linear models (LASSO and SVM) using SMOTEN and CTGAN provide best performance than when using nonlinear models such as DT and KNN.

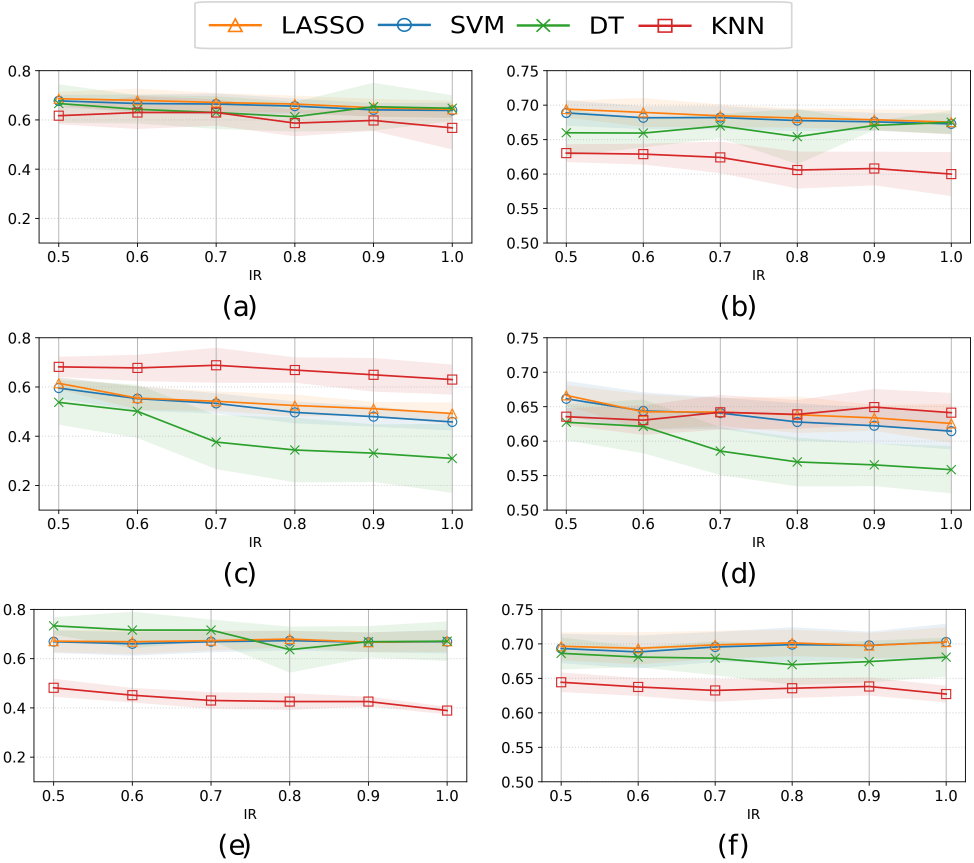


Figure 6. Mean ± std of the sensitivity (left panels) and AUC (right panels) considering 5 test subset partitions and different IR values, classifiers, and a hybrid approach (by combining undersampling and oversampling strategies) with: (**a**,**b**) SMOTEN; (**c**,**d**) TVAE; and (**e**,**f**) CTGAN.

Table 3 summarizes the best sensitivity and AUC values using different resampling methods and class balancing strategies. These results are presented considering all variables ($d = 26$) and only the selected variables ($\hat{d} = 14$) when using the bootstrap resampling method. Comparing the figures of merit obtained using 26 and 14 features, it is observed that models using 14 features show a slight improvement in the classification performance for most of the models. Mainly, this improvement can be seen when considering TVAE and the hybrid technique (average AUC value of 0.66 with all variables versus 0.70 with FS). However, the best performance is obtained when CTGAN and the hybrid strategy are considered, also outperforming the results provided by the undersampling approach.

Table 3. Sensitivity and AUC values (mean ± std) on 5 test subsets when training ML models using different resampling strategies with all features and FS. The highest average performance for each figure of merit is marked in bold.

Method	Balancing Strategy	Sensitivity (All)	Sensitivity (FS)	AUC (All)	AUC (FS)
RUS	Under	0.711 ± 0.059	0.707 ± 0.041	0.697 ± 0.025	0.706 ± 0.021
SMOTEN	Over	0.701 ± 0.046	0.708 ± 0.054	0.701 ± 0.025	0.700 ± 0.012
	Hybrid	0.686 ± 0.028	0.707 ± 0.014	0.694 ± 0.012	0.709 ± 0.020
TVAE	Over	0.636 ± 0.067	0.640 ± 0.013	0.650 ± 0.027	0.668 ± 0.024
	Hybrid	0.615 ± 0.023	0.694 ± 0.041	0.661 ± 0.026	0.704 ± 0.016
CTGAN	Over	0.699 ± 0.044	0.695 ± 0.021	0.702 ± 0.026	0.707 ± 0.017
	Hybrid	**0.716** ± **0.043**	**0.709** ± **0.036**	**0.712** ± **0.017**	**0.711** ± **0.021**

3.4. Analyzing Risk Factors Using Interpretability Methods

The average values of the coefficients in the five partitions assigned to each feature when training the classifiers using different oversampling methods (SMOTEN, TVAE and CTGAN), with an IR ranging from 0.4 to 1.0, are shown in Figure 7. Note that the values of the coefficients associated with each feature do not change substantially as IR increases. As a result, we conclude that the increasing fraction of synthetic samples generated has not much impact on the fact that some features are more relevant to predicting CVD. Regarding the LASSO model and the DT model for each specific oversampling approach, the coefficient values are similar, with many of them being zero in both models, nullifying the impact of that variable. However, the value of the coefficients fluctuates between positive and negative values for the SVM model (middle panels), with positive values indicating a greater influence of the variable on the prediction of subjects with CVD and negative values indicating a greater influence of the variable on the prediction of healthy individuals.

Note also that the weight (coefficient) assigned to each variable varies depending on the oversampling strategy used. This means that the method considered for generating synthetic samples could influence the features that are deemed most important to decide the final classification. In this sense, when applying SMOTEN and CTGAN, and especially when comparing LASSO and DT, the pattern of the coefficient values is comparable, with age, BMI, high cholesterol, and sex being the most significant factors in predicting CVD. TVAE, on the other hand, exhibits a somewhat different pattern; in this case, while age remains the most important predictor in predicting CVD, it is followed by preprocessed meat, strenuous activity, and BMI.

We focus now on SHAP [36], a post-hoc interpretability approach used for noninterpretable ML models. Figure 8 shows the SHAP summary plot when training the nonlinear model KNN with the CTGAN oversampling method, the oversampling class balancing strategy, and IR = 0.6. The summary plot provides different information concerning the interpretability of the model. It presents the importance of the 20 most relevant features as well as their impact on the prediction. The remaining features did not provide relevant information, and it was decided to make the cut-off on these 20 features. Each point represented in the summary plot is a Shapley value for a specific feature and a specific sample. The position on the x-axis shows the Shapley value assigned to each sample, whereas the y-axis shows the relevance of the features in descending order. For the KNN classifier, the three most relevant features in the prediction are the presence of high cholesterol, age, and sex, all coinciding with the results obtained for the linear models. As for the colors, the SHAP summary plot represents the value of the feature from lowest (blue) to highest (red). Thus, observing the most relevant variables, we can conclude that higher values of high cholesterol, age and sex are mostly associated with the prediction of subjects with CVD, and lower values of these variables are associated with the prediction of healthy individuals. On the contrary, for example, in the case of the alcohol variable, higher values of this variable are relevant in predicting healthy individuals, and lower values are related to predicting subjects with CVD.

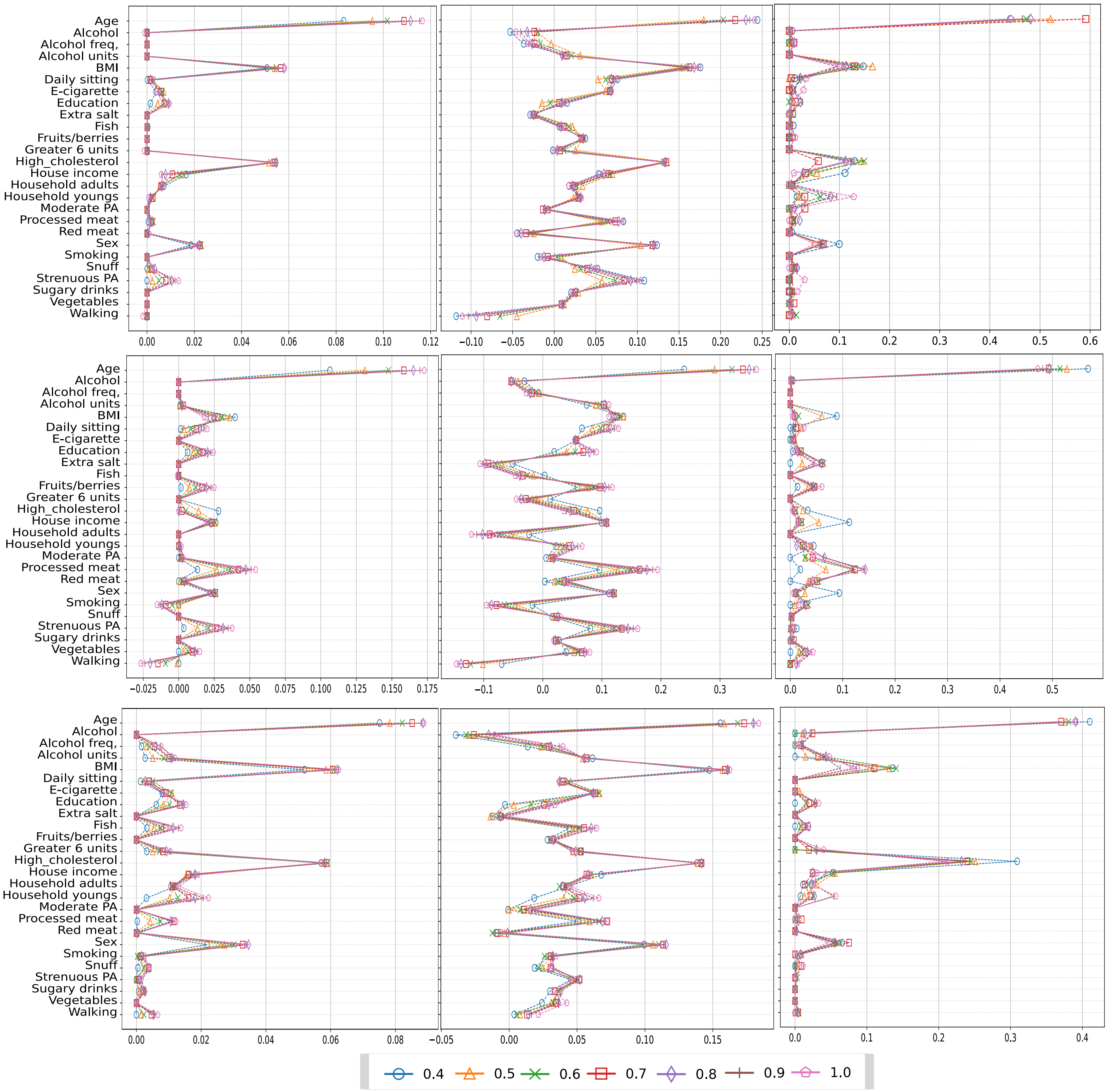


Figure 7. Coefficient values for different IR when using LASSO (**left panels**), SVM (**middle panels**), and DT (**right panels**) and the oversampling methods with all features: SMOTEN (**first row**); TVAE (**second row**); and CTGAN (**third row**).

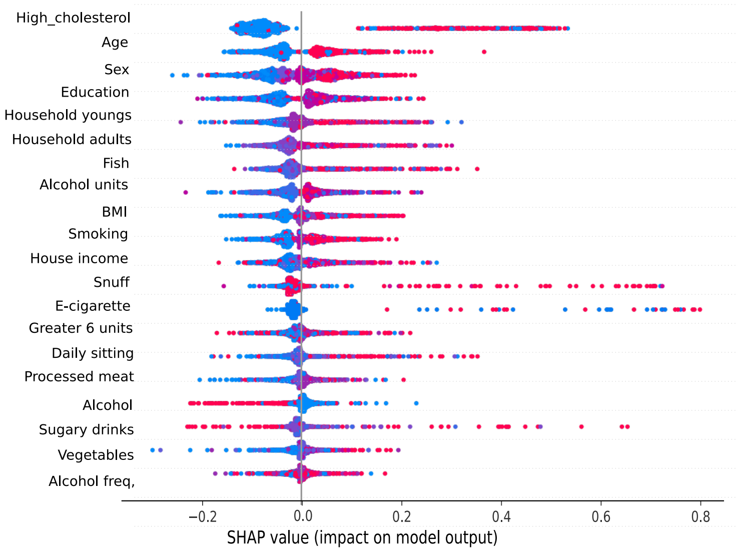


Figure 8. SHAP summary plot of KNN model using the CTGAN oversampling technique and all subjects as well as their impact on the prediction. Only the 20 most relevant features are shown.

4. Discussion

In this work, different resampling methods were used, highlighting the use of CTGANs as an oversampling technique to generate synthetic data for achieving data balance among different classes in a classification problem. In particular, a dataset with real-world data from healthy individuals and individuals with CVD was employed to carry out the synthetic data generation. Different metrics for quality assessment of the generated synthetic datasets were analyzed and discussed.

Our results demonstrated the high potential of CTGANs for generating categorical data, keeping relevant information, and improving classification performance. The following five findings are particularly important to this study. Firstly, the GAN-based model generates high-quality synthetic data, yielding a PMF that is very close to that estimated using real data. This makes a significant contribution to the literature, showing the potential of GANs in the clinical setting [21]. Secondly, it has been demonstrated that linear classifiers outperform nonlinear ones when using target encoding. Thirdly, the combination of the GAN and LASSO-based models yields an AUC value of 71%. Finally, because of the interpretive capabilities of our models, the findings of this study may help to improve both the prediction and prevention of CVD and the knowledge of the risk factors related to CVD. The most important risk factors for CVD prediction identified in all models were the presence of high cholesterol, age, BMI, and sex.

The previous findings are consistent with the primary CVD risk factors identified in the literature [61–63]. According to [61], BMI is one of the most critical criteria to consider since excessive adiposity is a significant risk factor for morbidity and death from type 2 diabetes, CVD, and different types of cancer. On the one hand, individuals from high-income countries are more likely to consume healthy foods, according to [64]. On the other hand, low-income people tend to consume more fat and less fiber, which explains the average importance of characteristics such as fish or meat consumption in the prediction of CVD. Furthermore, individuals in low- and middle-income countries are more likely to drink alcohol and smoke than socioeconomic groups in high-income countries, which explains the importance of these features for predicting CVD, according to [65].

Despite the potential benefits of using clinical data for research, these data are highly sensitive, and their use is restricted by privacy legislation and organizational guidelines [66]. Additionally, patient data are regulated by laws protecting patients' privacy such as the *Health Insurance Portability and Accountability Act* in the United States and the *General Data Protection Regulation* in the European Union [67]. The sharing of public health data has always been hampered by privacy concerns. Furthermore, in the clinical setting, most of the populations studied are commonly unbalanced, with the class of patients with a certain disease typically being smaller than the class of healthy individuals. In this way, synthetic data could allow researchers to delve deeper into complex medical issues, eliminating challenges such as a lack of access to protected data and addressing the issue of class imbalance [49].

As discussed throughout the previous sections, one of the main challenges when working with categorical data is to conduct an appropriate encoding process. In a prior author's work [68], one-hot encoding was evaluated, while in the current paper, target-encoding was considered. Target-encoding with regularization was shown to be an efficient technique for encoding categorical features, which do not increase the dimensionality in the analysis and allow us to achieve better results in predictive classification metrics. Despite its utility for handling the cardinality of categorical features, as argued, target-encoding presents certain limitations. It does not perform well when categories have few training samples because mean target values for these categories may be not representative (assigning extreme values), thus changing the original data distribution and deteriorating the subsequent performance of predictive algorithms. As future work, other types of techniques, such as integer, frequency, indicator, hash, leaf, and impact encoding, as well as a generalized linear mixed model encoder (detailed in [25]), can be explored and validated using data of CVD individuals in predictive tasks. In line with this, in real-world applications, the presence of categorical and numerical data is common. Our results showed that CTGAN outperforms other oversampling techniques, working reasonably well with one type of data and in one scenario of binary classification. An extended analysis can be conducted using heterogeneous clinical datasets that contain mixed-type data and consider multi-class classification scenarios. This will provide us with insights into the robustness and effectiveness of GAN-based models in more complex applications from a data perspective.

Further works will assess the findings achieved in this work by employing several different real-world clinical and nonclinical datasets. In particular, Framingham [69] and Steno [70] datasets can be used for evaluating resampling techniques. Additionally, an extended analysis of the use of CTGAN, which outperforms other oversampling techniques for identifying CVD disease in this paper, using these clinical datasets, can reveal the robustness of GAN-based models in scenarios with heterogeneous data. Furthermore, the quality of the synthetic data could be assessed by designing specific classifiers able to provide discrimination between real and synthetic data, evaluating in this way the performance of the oversampling techniques. In this sense, ensemble classification techniques have been demonstrated to improve performance in predictive applications [71–74] , which is a line of a research that can be further explored. Finally, other filter, wrapper and embedded FS techniques (such as mRmR, RELIEF, LASSO, Random Forest) could be studied in order to confirm the findings of this work related to the most relevant features selected to improve the performance and the interpretability of the ML models.

5. Conclusions

ML methods have become increasingly important for improving the performance of prediction models that could support decision-making. However, although these approaches have been applied in real-world scenarios, the class imbalance problem is a significant drawback in the development and performance of ML models. The main challenge lies in the fact that skewed class distributions hinder the proper learning process. In the clinical setting, most of the populations being studied are undersampled compared to healthy individuals, which limits the application of ML models. The use of synthetic data in conjunction with real data allows us to address these issues, improving the performance of ML models and achieving a better NCD prediction. In this paper, we evaluated the effectiveness of oversampling techniques (especially GAN-based approaches) in a scenario of binary classification to identify patients with CVD. Experimental results showed that the combination of a GAN-based model (CTGAN) and LASSO achieved an AUC value of 71%, outperforming the results of other oversampling strategies and the match-point method of the undersampling strategy. These advances in health informatics could help with clinical decisions, potentially changing the course of a chronic disease or health condition. Finally, the favorable impact of these decisions on cost savings and patient satisfaction would be significant from both clinical and socioeconomic standpoints.

Author Contributions: Conceptualization, I.M.-J. and C.S.-R.; methodology, I.M.-J. and C.S.-R.; software, C.G.-V. and D.C.-M.; validation, I.M.-J., H.F. and C.S.-R.; formal analysis, I.M.-J. and C.S.-R.; investigation, C.G.-V. and D.C.-M.; resources, I.T.G., C.G. and C.S.-R.; data curation, I.T.G., M.-L.L.; writing—original draft preparation, C.G.-V., D.C.-M., H.F. and C.S.-R.; writing—review and editing, I.M.-J., I.T.G., M.-L.L., C.G. and C.S.-R.; visualization, C.G.-V., and D.C.-M.; supervision, I.M.-J. and C.S.-R.; project administration, C.G. and C.S.-R.; funding acquisition, C.S.-R. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partly supported by European Commission through the H2020-EU.3.1.4.2., European Project WARIFA (Watching the risk factors: Artificial intelligence and the prevention of chronic conditions) under Grant Agreement 101017385; and by the Spanish Government by the Spanish Grants BigTheory (PID2019-106623RB-C41/AEI/ 10.13039/501100011033), and AAVis-BMR (PID2019-107768RA-I00/AEI/10.13039/501100011033); the Community of Madrid (YEI grant TIC-11649); Project Ref. 2020-661, financed by Rey Juan Carlos University and Community of Madrid; and by the Research Council of Norway (HELSE-EU-project 269882). This work was completed while Himar Fabelo was beneficiary of the FJC2020-043474-I funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR".

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data can be provided under a reasonable official request.

Conflicts of Interest: All authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CTGAN	Conditional Tabular Generative Adversarial Network
CVD	Cardiovascular Disease
DT	Decision Tree
FS	Feature Selection
GAN	Generative Adversarial Network
HD	Hellinger Distance
IR	Imbalance Ratio
KLD	Kullback-Leibler Divergence
KNN	K-Nearest Neighbors
LASSO	Least Absolute Shrinkage and Selection Operator
LCM	Log-Cluster Metric
MAEP	Mean Absolute Error Probability
ML	Machine Learning
PA	Physical Activity
PCD	Pairwise Correlation Difference
RSVR	Repeated Sample Vector Rate
RUS	Random Under Sampling
SHAP	Shapley Additive Explanations
SMOTE	Synthetic Minority Oversampling Technique
SMOTEN	Synthetic Minority Oversampling Technique Nominal
SVM	Support Vector Machine
TVAE	Tabular Variational Autoencoder

References

1. Raghupathi, W.; Raghupathi, V. Big data analytics in healthcare: Promise and potential. *Health Inf. Sci. Syst.* **2014**, *2*, 3. [CrossRef] [PubMed]
2. Bengio, Y.; Lecun, Y.; Hinton, G. Deep learning for AI. *Commun. ACM* **2021**, *64*, 58–65. [CrossRef]
3. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
4. Chen, Z.; Lin, T.; Xia, X.; Xu, H.; Ding, S. A synthetic neighborhood generation based ensemble learning for the imbalanced data classification. *Appl. Intell.* **2018**, *48*, 2441–2457. [CrossRef]
5. Ladeira Marques, M.; Moraes Villela, S.; Hasenclever Borges, C.C. Large margin classifiers to generate synthetic data for imbalanced datasets. *Appl. Intell.* **2020**, *50*, 3678–3694. [CrossRef]
6. Liu, R. A novel synthetic minority oversampling technique based on relative and absolute densities for imbalanced classification. *Appl. Intell.* **2023**, *53*, 786–803. [CrossRef]
7. Pérez, J.; Arroba, P.; Moya, J.M. Data augmentation through multivariate scenario forecasting in Data Centers using Generative Adversarial Networks. *Appl. Intell.* **2023**, *53*, 1469–1486. [CrossRef]
8. Zhu, T.; Luo, C.; Zhang, Z.; Li, J.; Ren, S.; Zeng, Y. Minority oversampling for imbalanced time series classification. *Knowl.-Based Syst.* **2022**, *247*, 108764. [CrossRef]
9. Malhotra, R.; Kamal, S. An empirical study to investigate oversampling methods for improving software defect prediction using imbalanced data. *Neurocomputing* **2019**, *343*, 120–140. [CrossRef]
10. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
11. Fernández, A.; Garcia, S.; Herrera, F.; Chawla, N.V. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *J. Artif. Intell. Res.* **2018**, *61*, 863–905. [CrossRef]
12. Liang, X.; Jiang, A.; Li, T.; Xue, Y.; Wang, G. LR-SMOTE–An improved unbalanced data set oversampling based on K-means and SVM. *Knowl.-Based Syst.* **2020**, *196*, 105845. [CrossRef]
13. Taft, L.; Evans, R.S.; Shyu, C.; Egger, M.; Chawla, N.; Mitchell, J.; Thornton, S.N.; Bray, B.; Varner, M. Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery. *J. Biomed. Inform.* **2009**, *42*, 356–364. [CrossRef] [PubMed]
14. Ijaz, M.F.; Attique, M.; Son, Y. Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. *Sensors* **2020**, *20*, 2809. [CrossRef]
15. Goh, K.H.; Wang, L.; Yeow, A.Y.K.; Poh, H.; Li, K.; Yeow, J.J.L.; Tan, G.Y.H. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nat. Commun.* **2021**, *12*, 711. [CrossRef]
16. Pereira, R.M.; Bertolini, D.; Teixeira, L.O.; Silla, C.N., Jr.; Costa, Y.M. COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios. *Comput. Methods Programs Biomed.* **2020**, *194*, 105532. [CrossRef]
17. Pahar, M.; Klopper, M.; Warren, R.; Niesler, T. COVID-19 cough classification using machine learning and global smartphone recordings. *Comput. Biol. Med.* **2021**, *135*, 104572. [CrossRef]
18. Tan, L.; Yu, K.; Bashir, A.K.; Cheng, X.; Ming, F.; Zhao, L.; Zhou, X. Toward real-time and efficient cardiovascular monitoring for COVID-19 patients by 5G-enabled wearable medical devices: A deep learning approach. *Neural Comput. Appl.* **2021**, 1–14. [CrossRef]
19. Gui, J.; Sun, Z.; Wen, Y.; Tao, D.; Ye, J. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 3313–3332. [CrossRef]
20. Jurado-Camino, M.T.; Chushig-Muzo, D.; Soguero-Ruiz, C.; de Miguel Bohoyo, P.; Mora-Jiménez, I. On the Use of Generative Adversarial Networks to Predict Health Status Among Chronic Patients. In *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2023, Lisbon, Portugal, 16–18 February 2023*; ScitePress: Setubal, Portugal, 2023; pp. 167–178.
21. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling tabular data using conditional gan. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, Canada, 8–14 December 2019; pp. 7335–7345.
22. Choi, E.; Biswal, S.; Malin, B.; Duke, J.; Stewart, W.F.; Sun, J. Generating multi-label discrete patient records using generative adversarial networks. In Proceedings of the Machine Learning for Healthcare Conference, Boston, MA, USA, 18–19 August 2017; pp. 286–305.
23. Meijers, W.C.; Maglione, M.; Bakker, S.J.; Oberhuber, R.; Kieneker, L.M.; de Jong, S.; Haubner, B.J.; Nagengast, W.B.; Lyon, A.R.; van der Vegt, B.; et al. Heart failure stimulates tumor growth by circulating factors. *Circulation* **2018**, *138*, 678–691. [CrossRef]
24. Gram, I.T.; Skeie, G.; Oyeyemi, S.O.; Borch, K.B.; Hopstock, L.A.; Løchen, M.L. A Smartphone-Based Information Communication Technology Solution for Primary Modifiable Risk Factors for Noncommunicable Diseases: Pilot and Feasibility Study in Norway. *JMIR Form. Res.* **2022**, *6*, e33636. [CrossRef]
25. Pargent, F.; Pfisterer, F.; Thomas, J.; Bischl, B. Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Comput. Stat.* **2022**, *37*, 2671–2692. [CrossRef]
26. Berisha, V.; Krantsevich, C.; Hahn, P.R.; Hahn, S.; Dasarathy, G.; Turaga, P.; Liss, J. Digital medicine and the curse of dimensionality. *NPJ Digit. Med.* **2021**, *4*, 153. [CrossRef] [PubMed]
27. Montavon, G.; Samek, W.; Müller, K.R. Methods for interpreting and understanding deep neural networks. *Digital Signal Process.* **2018**, *73*, 1–15. [CrossRef]

28. Chushig-Muzo, D.; Soguero-Ruiz, C.; de Miguel-Bohoyo, P.; Mora-Jiménez, I. Interpreting clinical latent representations using autoencoders and probabilistic models. *Artif. Intell. Med.* **2021**, *122*, 102211. [CrossRef] [PubMed]
29. Stiglic, G.; Kocbek, P.; Fijacko, N.; Zitnik, M.; Verbert, K.; Cilar, L. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1379. [CrossRef]
30. Marchese Robinson, R.L.; Palczewska, A.; Palczewski, J.; Kidley, N. Comparison of the predictive performance and interpretability of random forest and linear models on benchmark data sets. *J. Chem. Inf. Model.* **2017**, *57*, 1773–1792. [CrossRef]
31. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine learning interpretability: A survey on methods and metrics. *Electronics* **2019**, *8*, 832. [CrossRef]
32. Rao, N.; Nowak, R.; Cox, C.; Rogers, T. Classification with the sparse group lasso. *IEEE Trans. Signal Process.* **2015**, *64*, 448–463. [CrossRef]
33. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006.
34. Safavian, S.R.; Landgrebe, D. A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man, Cybern.* **1991**, *21*, 660–674. [CrossRef]
35. Zhang, M.L.; Zhou, Z.H. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognit.* **2007**, *40*, 2038–2048. [CrossRef]
36. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4768–4777.
37. Bush, K.; Kivlahan, D.R.; McDonell, M.B.; Fihn, S.D.; Bradley, K.A.; ACQUIP. The AUDIT alcohol consumption questions (AUDIT-C): an effective brief screening test for problem drinking. *Arch. Intern. Med.* **1998**, *158*, 1789–1795. [CrossRef] [PubMed]
38. Hagströmer, M.; Oja, P.; Sjöström, M. The International Physical Activity Questionnaire (IPAQ): A study of concurrent and construct validity. *Public Health Nutr.* **2006**, *9*, 755–762. [CrossRef] [PubMed]
39. Chushig-Muzo, D.; Soguero-Ruiz, C.; Engelbrecht, A.P.; Bohoyo, P.D.M.; Mora-Jiménez, I. Data-driven visual characterization of patient health-status using electronic health records and self-organizing maps. *IEEE Access* **2020**, *8*, 137019–137031. [CrossRef]
40. Cerda, P.; Varoquaux, G.; Kégl, B. Similarity encoding for learning with dirty categorical variables. *Mach. Learn.* **2018**, *107*, 1477–1494. [CrossRef]
41. Rodríguez, P.; Bautista, M.A.; Gonzalez, J.; Escalera, S. Beyond one-hot encoding: Lower dimensional target embedding. *Image Vis. Comput.* **2018**, *75*, 21–31. [CrossRef]
42. Sachan, S.; Almaghrabi, F.; Yang, J.B.; Xu, D.L. Evidential reasoning for preprocessing uncertain categorical data for trustworthy decisions: An application on healthcare and finance. *Expert Syst. Appl.* **2021**, *185*, 115597. [CrossRef]
43. Carrizosa, E.; Restrepo, M.G.; Morales, D.R. On clustering categories of categorical predictors in generalized linear models. *Expert Syst. Appl.* **2021**, *182*, 115245. [CrossRef]
44. Mumtaz, S.; Giese, M. Hierarchy-based semantic embeddings for single-valued & multi-valued categorical variables. *J. Intell. Inf. Syst.* **2022**, *58*, 613–640.
45. Micci-Barreca, D. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explor. Newsl.* **2001**, *3*, 27–32. [CrossRef]
46. Cai, J.; Luo, J.; Wang, S.; Yang, S. Feature selection in machine learning: A new perspective. *Neurocomputing* **2018**, *300*, 70–79. [CrossRef]
47. Mora-Jiménez, I.; Tarancón-Rey, J.; Álvarez-Rodríguez, J.; Soguero-Ruiz, C. Artificial Intelligence to Get Insights of Multi-Drug Resistance Risk Factors during the First 48 Hours from ICU Admission. *Antibiotics* **2021**, *10*, 239. [CrossRef] [PubMed]
48. Martínez-Agüero, S.; Soguero-Ruiz, C.; Alonso-Moral, J.M.; Mora-Jiménez, I.; Álvarez Rodríguez, J.; Marques, A.G. Interpretable clinical time-series modeling with intelligent feature selection for early prediction of antimicrobial multidrug resistance. *Future Gener. Comput. Syst.* **2022**, *133*, 68–83. [CrossRef]
49. Haixiang, G.; Yijing, L.; Shang, J.; Mingyun, G.; Yuanyue, H.; Bing, G. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.* **2017**, *73*, 220–239. [CrossRef]
50. Liu, X.Y.; Wu, J.; Zhou, Z.H. Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **2008**, *39*, 539–550.
51. Elreedy, D.; Atiya, A.F. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Inf. Sci.* **2019**, *505*, 32–64. [CrossRef]
52. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.
53. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Volume 27, pp. 2672–2680.
54. Zavrak, S.; İskefiyeli, M. Anomaly-based intrusion detection from network flow features using variational autoencoder. *IEEE Access* **2020**, *8*, 108346–108358. [CrossRef]
55. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
56. Le Cam, L.; LeCam, L.M.; Yang, G.L. *Asymptotics in Statistics: Some Basic Concepts*; Springer Science & Business Media: New York, NY, USA, 2000.

57. Goncalves, A.; Ray, P.; Soper, B.; Stevens, J.; Coyle, L.; Sales, A.P. Generation and evaluation of synthetic patient data. *BMC Med. Res. Methodol.* **2020**, *20*, 108. [CrossRef]
58. Woo, M.J.; Reiter, J.P.; Oganian, A.; Karr, A.F. Global measures of data utility for microdata masked for disclosure limitation. *J. Priv. Confidentiality* **2009**, *1*, 111–224. [CrossRef]
59. Rayward-Smith, V.J. Statistics to measure correlation for data mining applications. *Comput. Stat. Data Anal.* **2007**, *51*, 3968–3982. [CrossRef]
60. MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability, Oakland, CA, USA, 27 December 1966–7 January 1967; Volume 1, pp. 281–297.
61. Malik, V.S.; Willett, W.C.; Hu, F.B. Global obesity: Trends, risk factors and policy implications. *Nat. Rev. Endocrinol.* **2013**, *9*, 13–27. [CrossRef]
62. Dahlöf, B. Cardiovascular disease risk factors: Epidemiology and risk assessment. *Am. J. Cardiol.* **2010**, *105*, 3A–9A. [CrossRef] [PubMed]
63. Wagner, K.H.; Brath, H. A global view on the development of non communicable diseases. *Prev. Med.* **2012**, *54*, S38–S41. [CrossRef] [PubMed]
64. Mayen, A.L.; Marques-Vidal, P.; Paccaud, F.; Bovet, P.; Stringhini, S. Socioeconomic determinants of dietary patterns in low-and middle-income countries: A systematic review. *Am. J. Clin. Nutr.* **2014**, *100*, 1520–1531. [CrossRef] [PubMed]
65. Marmot, M.; Bell, R. Social determinants and non-communicable diseases: Time for integrated action. *Bmj* **2019**, *364*. [CrossRef] [PubMed]
66. Benaim, A.R.; Almog, R.; Gorelik, Y.; Hochberg, I.; Nassar, L.; Mashiach, T.; Khamaisi, M.; Lurie, Y.; Azzam, Z.S.; Khoury, J.; et al. Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies. *JMIR Med. Inform.* **2020**, *8*, e16492. [CrossRef] [PubMed]
67. Yale, A.; Dash, S.; Dutta, R.; Guyon, I.; Pavao, A.; Bennett, K.P. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing* **2020**, *416*, 244–255. [CrossRef]
68. García-Vicente, C.; Chushig-Muzo, D.; Mora-Jiménez, I.; Fabelo, H.; Gram, I.T.; Løchen, M.L.; Granja, C.; Soguero-Ruiz, C. Clinical Synthetic Data Generation to Predict and Identify Risk Factors for Cardiovascular Diseases. In *Proceedings of the Heterogeneous Data Management, Polystores, and Analytics for Healthcare: VLDB Workshops, Poly 2022 and DMAH 2022, Virtual Event, 9 September 2022*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 75–91.
69. Mahmood, S.S.; Levy, D.; Vasan, R.S.; Wang, T.J. The Framingham Heart Study and the epidemiology of cardiovascular disease: A historical perspective. *Lancet* **2014**, *383*, 999–1008. [CrossRef]
70. Vistisen, D.; Andersen, G.S.; Hansen, C.S.; Hulman, A.; Henriksen, J.E.; Bech-Nielsen, H.; Jørgensen, M.E. Prediction of first cardiovascular disease event in type 1 diabetes mellitus: The Steno Type 1 Risk Engine. *Circulation* **2016**, *133*, 1058–1066. [CrossRef]
71. Abdar, M.; Zomorodi-Moghadam, M.; Zhou, X.; Gururajan, R.; Tao, X.; Barua, P.D.; Gururajan, R. A new nested ensemble technique for automated diagnosis of breast cancer. *Pattern Recognit. Lett.* **2020**, *132*, 123–131. [CrossRef]
72. Latha, C.B.C.; Jeeva, S.C. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Inform. Med. Unlocked* **2019**, *16*, 100203. [CrossRef]
73. Xiao, Y.; Wu, J.; Lin, Z.; Zhao, X. A deep learning-based multi-model ensemble method for cancer prediction. *Comput. Methods Programs Biomed.* **2018**, *153*, 1–9. [CrossRef] [PubMed]
74. Kazemi, Y.; Mirroshandel, S.A. A novel method for predicting kidney stone type using ensemble learning. *Artif. Intell. Med.* **2018**, *84*, 117–126. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.