computers

MDPI

*Article*

# Trading-Off Machine Learning Algorithms towards Data-Driven Administrative-Socio-Economic Population Health Management

**Silvia Panicacci [1,*], Massimiliano Donati [1], Francesco Profili [2], Paolo Francesconi [2] and Luca Fanucci [1]**

[1] Department of Information Engineering, University of Pisa, Via G. Caruso 16, 56122 Pisa, Italy; massimiliano.donati@unipi.it (M.D.); luca.fanucci@unipi.it (L.F.)
[2] Tuscan Agenzia Regionale Sanità, 50141 Florence, Italy; francesco.profili@ars.toscana.it (F.P.); paolo.francesconi@ars.toscana.it (P.F.)
* Correspondence: silvia.panicacci@phd.unipi.it

**Abstract:** Together with population ageing, the number of people suffering from multimorbidity is increasing, up to more than half of the population by 2035. This part of the population is composed by the highest-risk patients, who are, at the same time, the major users of the healthcare systems. The early identification of this sub-population can really help to improve people's quality of life and reduce healthcare costs. In this paper, we describe a population health management tool based on state-of-the-art intelligent algorithms, starting from administrative and socio-economic data, for the early identification of high-risk patients. The study refers to the population of the Local Health Unit of Central Tuscany in 2015, which amounts to 1,670,129 residents. After a trade-off on machine learning models and on input data, Random Forest applied to 1-year of historical data achieves the best results, outperforming state-of-the-art models. The most important variables for this model, in terms of mean minimal depth, accuracy decrease and Gini decrease, result to be age and some group of drugs, such as high-ceiling diuretics. Thanks to the low inference time and reduced memory usage, the resulting model allows for real-time risk prediction updates whenever new data become available, giving General Practitioners the possibility to early adopt personalised medicine.

**Keywords:** decision support system; population health management; explainable artificial intelligence; machine learning; big data

check for updates

**Citation:** Panicacci, S.; Donati, M.; Profili, F.; Francesconi, P.; Fanucci, L. Trading-Off Machine Learning Algorithms towards Data-Driven Administrative-Socio-Economic Population Health Management. *Computers* **2021**, *10*, 4. https://dx.doi.org/10.3390/computers10010004

Received: 30 November 2020
Accepted: 19 December 2020
Published: 25 December 2020

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Copyright:** © 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Thanks to advances in therapies and treatments, to declining of fertility and to immigration, the population is getting older and older [1,2]. It is expected that one fourth of the U.S. population will be over 65 years old by 2060 [3]. This situation will not be different in the other countries. In addition, the population aged 85 years and over is expected to almost double in the next 25 years [4].

Population ageing influences the cost of health [5]. In fact, nowadays more than half of the elderly population is affected by more than 2 co-existing chronic diseases (multimorbidity), with increasing prevalence in very old people. Over the next 15 years, the number of old people affected by chronic diseases will increase by more than 50% [6]. These "complex" patients are the major users of the healthcare systems, because of their higher risk of hospitalisation and death, leading to higher healthcare expenditures than people with no chronic conditions or with a single chronic condition [7,8].

It is well known in the literature that the implementation of the Chronic Care Model (CCM) have reduced the healthcare costs and have improved the outcomes, especially in the management of single chronic conditions [9]. Although CCM has attempted to lead to proactive care of patients [10], the current care model is still based on a reactive approach, where the General Practitioners (GPs) usually react to patients' symptoms.

This poor tendency to prevent and to act before the symptoms often leads to higher risk of hospitalisation, hospital readmission and mortality, adversely affecting quality of life (QoL) and economic healthcare resources.

Identifying relevant sub-populations for proactive management, i.e., the highest-risk (complex) patients, becomes the key for early diagnoses and therapies, preventing or postponing some adverse events, delivering them the most appropriate care program according to the individual needs and thus improving QoL and the allocation of the available resources.

However, the correct identification of a cohort of high-risk patients to be monitored is a hard work for the doctors [11]. In fact, the patient medical record is not enough to define the health status, but also biology/genetics, socio-economic factors, culture, environment and behaviour should be considered [12]. In addition, in the big data era, a huge amount of data is available, coming from different sources (e.g., administrative flows, wearable and IoT devices and telemedicine platforms) with frequent updates [13–15]. If treated and analysed with proper methods, these data can give some useful information, decrease the hospital readmission rate and reduce the healthcare costs of more than 25% [16].

Population Health Management (PHM) is the automated process of using big data for the definition of patients cohorts and for the stratification of the groups by the risk of hospitalisations. Its final aim is improving clinical outcomes while lowering costs [17]. Thanks to the advances in machine learning (ML) algorithms and to the digitalization of the informative flows, PHM can really help in the identification of the target patients and in the implementation of the proactive approach. However, it is important that the selected ML model is explainable and understandable by the GPs in order that they can trust them and they are fully committed to follow their outcomes.

This paper presents the trade-off of machine learning algorithms, taking as input datasets composed by administrative and socio-economic data deriving from periods of study of different lengths, to develop an explainable PHM first level screening tool. Such a tool is for the early identification of high-risk patients (from the clinical point of view, and so complex patients), who will be re-analysed by the GPs during the second level screening phase, to obtain a final group of patients to monitor with specific plans of care. Its final aim is multiple: supporting GPs in early identification of high-risk patients, improving patients' QoL by decreasing hospitalisations, readmissions and mortality and reducing healthcare costs.

A preliminary version of this work has been presented at the conference IEEE CBMS 2018 [18] by the same authors. This article provides an extended state-of-the-art analysis, additional details regarding the extraction of the input features, new models trained with different datasets and, consequently, new results and a more detailed discussion of these results, including the explanation of the best model through the analysis of the most important variables.

After this introduction, Section 2 reports the analysis of the state-of-the-art. Section 3 shows the phases of data pre-processing, considering the available data in Italy, the extraction of the input features and the final datasets and modelling (how to deal with imbalanced data, implementation of the models, parameters tuning and feature selection). Section 4 presents the results, which are discussed in Section 5, together with a detailed explanation of the best model. At last, conclusions are drawn in Section 6.

## 2. Related Works

With advances in artificial intelligence and new technologies for data collection and storage, the healthcare industry can exploit big data analytics to improve the outcomes [19].

Multiple supervised ML algorithms have been implemented and validated for the prediction of single chronic diseases, starting from different data sources. Swain [20] built Logistic Regression (LR) and Decision Tree (DT) models to evaluate the risk of being obese in the U.S. population, using a database constructed with interviews. Chen et al. [21] proposed a Convolutional Neural Network (CNN) for the prediction of cerebral infarction, us-

ing both structured and unstructured hospital data. Meng et al. [22] compared LR, DT and Artificial Neural Network (ANN) performance in the prediction of diabetes, using data extracted from questionnaires, while Worachartcheewan et al. [23] proposed Random Forest (RF) for the same aim. Regarding the prediction of heart diseases, Latha et al. [24] used ensemble methods (boosting, bagging and stacking), other than the classical Naive Bayes (NB), RF, DT, ANN and Projective Adaptive Resonance Theory (PART), starting from the Cleveland heart dataset. NB, RF, LR, Support Vector Machine (SVM) and ensemble methods were also used by Dinesh et al. [25] for the prediction of cardiovascular disease. In addition, Panicacci et al. [26] implemented and validated RF and Least Absolute Shrinkage and Selection Operator (LASSO) for the identification of heart failure patients, using Italian administrative data. Tengnah et al. [27] tried to diagnose hypertension with DT and ANN. Yang et al. [28] started from a survey database (ELSA, English Longitudinal Study of Ageing) to predict dementia, comparing Gradient Boosting Machine (GBM), CNN, RF, Regularized Greedy Forest (RGF) and LR, while Cattelani et al. [29] developed a risk prediction model for depression, using three different European databases.

In addition, Electronic Health Records (EHRs) [30] are largely employed as a data source for chronic diseases prediction, even if they are not structured and collected for research projects [31,32]. Some examples of their use are described in [33] for the prediction of childhood obesity, in [34] for diabetes predictions, in [35] to predict heart failure, in [36] for cardiovascular disease, in [37] for hypertension and in [38] for dementia.

However, predicting a single chronic disease does not completely overlap with the identification of high-risk patients, since the highest-risk patients are the complex patients, usually affected by multimorbidity, but the onset of multiple chronic conditions is not taken into account in these studies.

Anyway, EHRs are also used, especially in the USA, for the identification of high-cost patients, considering then the occurrence of multimorbidity, and for predicting hospitalisations [39–44]. These predictive models are usually exploited by the insurance companies to adjust the insurance premium of the patients.

Unfortunately, EHRs are not available at the same level of maturity in all the regions in Italy and they are not usable for prediction analysis. Here, only administrative data (hospitalisations, procedures, outpatient services, drugs, exemptions and emergency room visits) are collected by public authorities to be used for healthcare predictive models. In three Italian regions, Tuscany, Emilia-Romagna and Lazio, Bellini et al. [45], Louis et al. [46] and Balzi et al. [47] tried to identify high-risk patients using the available administrative data, but they employed statistical methods. Statistical methods are more explainable than ML algorithms, but at the same time they have limitations in the amount of data to process. For this reason, the authors made some critical a-priori decisions, regarding input variables and involved patients. However, to the best of our knowledge, administrative data have already been treated only with statistical methods in the literature.

The idea of this work is to implement and test state-of-the-art ML algorithms, already employed in the prediction of a single chronic disease and healthcare costs, for the identification of high-risk patients. These models were adapted to be used with administrative and socio-economic data.

## 3. Materials and Methods

In this section, we describe the phases of data pre-processing and modelling. In [18], multiple ML algorithms were implemented and validated using as input a dataset with 5-years of historical and socio-economic data to develop a PHM tool for the identification of high-risk patients. Starting from those results, we decided to trade-off not only ML models but also input datasets with different populations and periods of historical data, to find the best matching model/dataset to identify the target population.

The models tried to solve a binary classification problem, since the output was high-risk/low-risk for each patient. The algorithms were evaluated with three golden metrics: Positive Predictive Ratio (PPR) [48], which represents the capability of discriminating

the positive class from the negative one; F1-Score, the harmonic mean between Positive Predictive Value (PPV) and Sensitivity (SE); F2-Score, which weighs SE higher than PPV by placing more emphasis in false negatives [49]. They are defined as follows:

$$PPR = \frac{PPV}{1 - NPV} \tag{1}$$

$$F1Score = 2 \times \frac{PPV \times SE}{PPV + SE} \tag{2}$$

$$F2Score = 5 \times \frac{PPV \times SE}{4 \times PPV + SE} \tag{3}$$

where SE, PPV, Negative Predictive Ratio (NPV) and Specificity (SP) are defined as:

$$SE = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{4}$$

$$PPV = \frac{TruePositives}{TruePositives + FalsePositives} \tag{5}$$

$$NPV = \frac{TrueNegatives}{TrueNegatives + FalseNegatives} \tag{6}$$

$$SP = \frac{TrueNegatives}{TrueNegatives + FalsePositives} \tag{7}$$

With respect to [18], in this work we included also the F2-Score on top of the PPR and F1-score, in order to minimise the false negatives more than the false positives. This is very important for any first level screening tool.

### *3.1. Data Pre-Processing*

In Italy, every time each resident benefits from a health service, he/she produces some digital traces. These data are first collected by the healthcare facilities where the service is provided (i.e., hospitals, clinics, pharmacies and Local Health Units (LHUs)) and then they are transmitted to the regions. Regions send anonymized data (fiscal codes are encrypted according to Italian Law, no. 675/1996 [50]) to the Agencies of Regional Health Services (ARHSs), the only institutions in charge of comparative-effectiveness analysis [51]. Before any processing, unstructured data are extracted, transformed and loaded (ETL process) on a relational database (DB). In addition to these administrative data, ARHSs store also socio-economic data collected by ISTAT [52] with the national census. Socio-economic data are in the order of census sections. In the case of the Tuscany Region, mARSupio is the DB where administrative and socio-economic data are stored [53]. Figure 1 shows the architecture of the system for the collection of data.

More in detail, we could rebuild the medical history of each patient using administrative data. The interesting information for this work were diagnoses and procedures done during hospital admissions, assistive, diagnostic and rehabilitation outpatient services, prescribed drugs and exemptions for any reason. These administrative flows are usually mainly complete, because a specific reimbursement is provided for each record.

The problem addressing this paper is a binary classification one (high-risk/non-high-risk). Moreover, the analysis was shifted in the past to implement supervised algorithms with the possibility of evaluating relevant performance (retrospective study): the goal was to identify clinically high-risk patients in 2015. With the support of a group of medical experts, we defined high-risk patients as the ones who will have an avoidable hospitalisation or will die the following year with respect to the period of study (2015 in this case). According to the Agency for Healthcare Research and Quality (AHRQ), avoidable hospitalisations are inpatients that could be prevented with early intervention and good outpatient care [54]. They can be identified in hospital admissions for angina without procedure, congestive heart failure, hypertension, chronic obstructive pulmonary disease, diabetes

short-term complication, diabetes chronic complication (renal, ocular, neurological, circulatory, etc.), uncontrolled diabetes and lower-extremity amputation among patients with diabetes. People of the initial population who had avoidable hospitalisations or died in 2015 were then identified using data in mARSupio. They represented the positive class 'B' of this supervised binary classification problem. All the others in the initial population were part of the negative class 'G'.

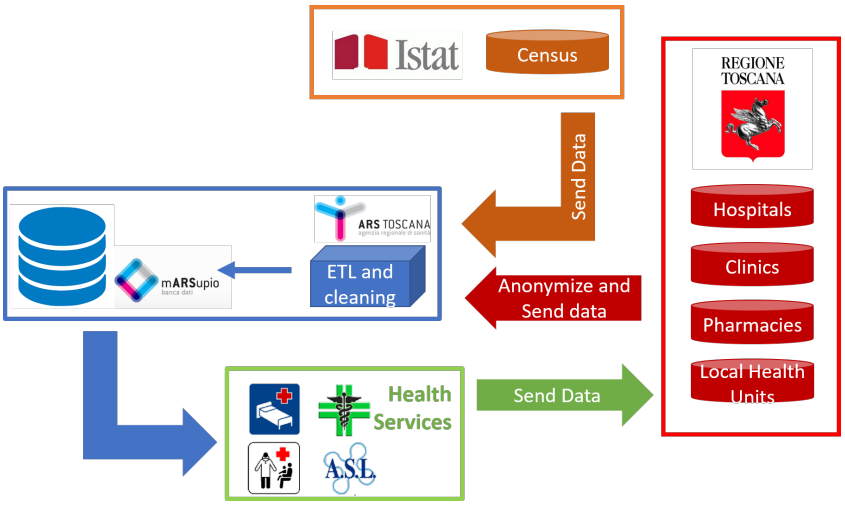



**Figure 1.** Architecture of the system for the collection of data.

We considered multiple initial populations, according to the reference period of study: people were observed for 5-, 4-, 3-, 2-years and 1-year periods. Therefore, since the output was calculated for 2015, the periods of study were 2010–2014, 2011–2014, 2012–2014, 2013–2014 and only 2014, respectively, as shown in Figure 2. The five populations involved in the study were composed by all the residents in the LHU of Central Tuscany, alive on 1st January 2015, who have lived in Tuscany at least the 80% of the days of the entire period of study. This restriction was due to the fact that mARSupio contains medical information only for the residents in Tuscany and so some crucial events could be lost for people who have left Tuscany for a significant period of time. As a result of this limitation, the initial population grew when considering a shorter period of study: 1,529,714, 1,580,899, 1,605,627, 1,629,651 and 1,648,897 for 5-, 4-, 3-, 2-years and 1-year periods of study, respectively.

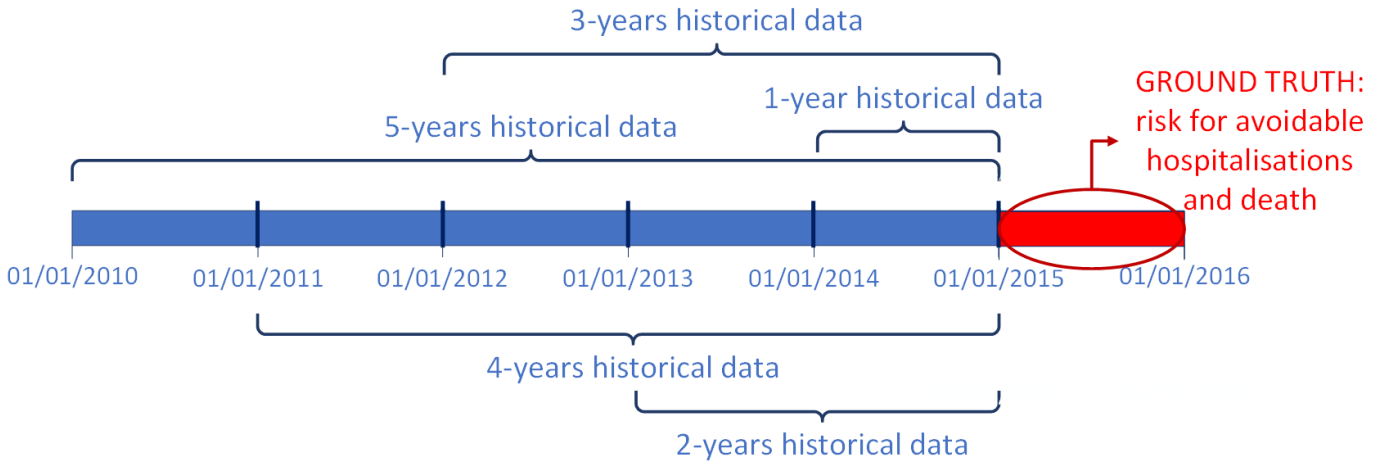



**Figure 2.** Building input datasets.

All the interactions with the healthcare system occurred in the period of data collection were considered with the same weight for every population. Of course, it was infeasible

considering each diagnosis, each procedure, each outpatient service, each drug and each exemption separately as input for the models, since they are identified by tens of thousands of different codes. Therefore, for each class, the codes were grouped according to the type of data and the input variables for every dataset were generated as described in the following:

- Diagnoses were grouped by single-level Clinical Classification Software (CCS) [55]. They are 283 clinically homogeneous mutually exclusive categories, which cluster all the 14,000 diagnosis codes, according to International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM, Fifth Edition) [56]. We counted the number of hospital admissions and summed the number of days of hospitalisation for each CCS for each patient for different dates. For example, if patient A has been hospitalised on 1st January 2014 for 3 days with ICD-9-CM codes 250.60 (diabetes with neurological manifestations, type II or unspecified type, not stated as uncontrolled), 250.70 (diabetes with peripheral circulatory disorders, type II or unspecified type, not stated as uncontrolled) and 278.00 (obesity), since the two codes 250.60 and 250.70 belong to CCS 50 (diabetes mellitus with complications) and the code 278.00 belongs to CCS 58 (other nutritional endocrine and metabolic disorders), this event counts as 1 hospital admission with 3 days of hospitalisation for CSS 50 and 1 hospital admission with 3 days of hospitalisation for CCS 58. Another admission of 5 days with code 250.60 on 30th March 2014 for the same patient will be summed to the previous one for CCS 50, for a total of 2 hospital admissions and 8 days of hospitalisation for CCS 50 for patient A. Considering admissions and days as two different variables for each CCS, the total number of input attributes for diagnoses was 566.
- Procedures were grouped by single-level CCS. For procedure classification, the scheme contains 231 mutually exclusive categories, grouping 3900 ICD-9-CM procedure codes [57]. The ICD-9-CM codes of each category usually refer to the same system/organ. The number of procedures done for each CCS for each patient on different dates was counted. The total number of input features for procedures was then 231.
- Outpatient services were divided into 76 ad-hoc groups. The outpatient services codes are about 2000, but a unique regrouping method is not defined yet. Therefore, they were divided first in visits, diagnostics, laboratory, therapeutics and rehabilitation and then we went into more detail, considering criteria like methods and purpose, defining 76 mutually exclusive and homogeneous ad-hoc groups. The services made on different dates for every group and every patient were counted. Each group represented an input variable, for a total of 76 variables.
- For drugs, Anatomical Therapeutic Chemical classification system (ATC) [58] was chosen. The third level (ATC3) was selected for classification. It is the therapeutic/pharmacological subgroup of the drug itself and it is usually used to identify 265 therapeutic subgroups of about 3350 chemical substances. The number of drugs taken in the period for different ATC3 classes and on different dates was counted, for a total of 265 attributes for drugs.
- Exemptions were partitioned into 28 ad-hoc mutually exclusive groups. Starting from almost 800 codes, the classification was done considering the motivation of the exemption, according to chronic diseases, specific services (e.g., emergency room or sports medicine) and income. Every class was a categorical variable (for a total of 28 input features) with three values: "0" for no exemptions for the group, "1" meaning at least an active exemption for the group at 1st January 2015, and "2" if the exemption has expired during the period of study.

The process of defining and generating the input variables for each dataset from the medical codes is shown in Table 1.

**Table 1.** Process of defining input variables from diagnoses, procedures, outpatient services, drugs and exemptions codes.

| Class | Codes | Grouping Method | Groups | Input Variables |
|---|---|---|---|---|
| Diagnoses | ∼14,000 | CCS | 283 | 566 |
| Procedures | ∼3900 | CCS | 231 | 231 |
| Outpatient Services | ∼2000 | Ad-hoc | 76 | 76 |
| Drugs | ∼3350 | ATC3 | 265 | 265 |
| Exemptions | ∼800 | Ad-hoc | 28 | 28 |

Since 80% of what affects the health outcomes and the clinical phenotype is associated to health behaviours, social and economic factors and physical environment [59], we included in the input datasets also socio-economic data. Unfortunately, this kind of information is not available for the single person, but derives from the last ISTAT census (2011), where data are aggregated for census section. For each census section, we calculated the dependency index (the ratio between the number of people in non-working age and the number of people in working age [60]), the median level of education (a categorical variable with values 4, 3, 2, 1 for degree, high school, secondary school, primary school and nothing, respectively), the median marital status (from 0 to 3, meaning single, married, divorced and widowed), the percentage of the working population, the percentage of strangers, the median number of family members and the percentage of rented houses. In addition, three variables related to the environment of living were considered: the density of the municipality of residence, the characteristic of inner area (with possible values centre, middle, belt, outlying and outermost) and the classification as fragile area (fragile/non-fragile). Socio-economic variables were then 10.

Finally, age on 1st January 2015 and gender were included.

Therefore, the complete set of input features was composed by 1178 variables, including both administrative and socio-economic data (Table 2).

**Table 2.** Summary of the input variables of each final dataset.

| Class | | Input Variables |
|---|---|---|
| **Medical Variables** | Diagnoses | 566 |
| | Procedures | 231 |
| | Outpatient Services | 76 |
| | Drugs | 265 |
| | Exemptions | 28 |
| **Socio-Economic Variables** | Personal | 2 |
| | Municipality of residence | 3 |
| | Census Section of residence | 7 |
| **Total** | | 1178 |

The final datasets had every person in a row and 1179 columns (1178 input attributes plus the binary output variable). Their dimensions were then 1,529,714 × 1179, 1,580,899 × 1179, 1,605,627 × 1179, 1,629,651 × 1179 and 1,648,897 × 1179 for the five periods of study, from the longer to the shorter.

Since in the DB not all the people were associated to a census section, all the datasets presented some missing values, about the 10% of the values of the seven columns related

to the census section of residence. They were replaced with the mean value in continuous variables and with the median value in categorical variables.

All the datasets were very unbalanced towards the negative output class. The prevalence of the positive class, in fact, varied from 1.42% to 1.34%, from the 5-years population to the 1-year population.

### *3.2. Modelling and Feature Selection*

The entire modelling phase for each dataset is described in Figure 3. It was executed on a Linux server with 64 GB of RAM and 64 CPU cores, using a program written in R language.

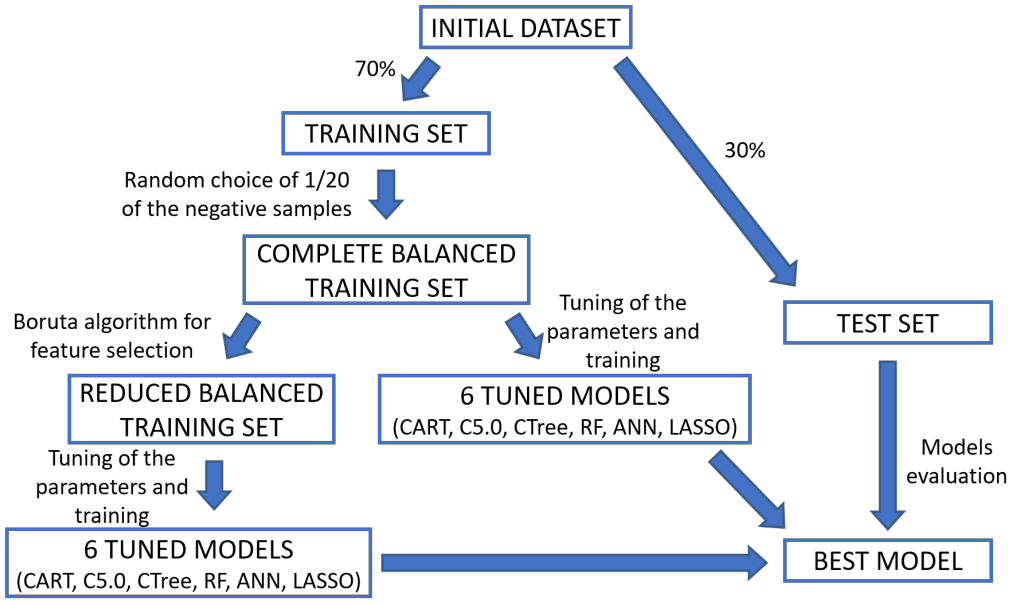



**Figure 3.** Modelling phase for each input dataset.

Every dataset was split into two partitions, training and test sets, the 70% and the 30% of the original dataset, respectively. During the split, the proportion of the output classes was maintained: all training and test sets had the same prevalence of the original datasets. In order to achieve better performance, handling the problem of imbalanced data, the five training sets were under-sampled [61], taking one random sample every 20 samples belonging to the negative class. The positive samples, on the contrary, were all kept. Since the original prevalence was slightly different among the initial training sets, the obtained prevalence of the balanced training sets varied from 22.36% to 21.38%, as shown in Table 3. This ratio was considered a good trade-off between the total balancing (50% of positives and 50% of negatives) and the deletion of too many samples. On the contrary, the test sets were not balanced, because performance must be evaluated on a real sample of the Tuscan population, with the original prevalence.

The balanced training sets were then used to train six ML algorithms, the ones used in [18], except NB.

In fact, NB assumes that input features are independent, but in this problem dependency exists among attributes: this explains its poor performance. Therefore, the algorithms implemented in this study were Classification and Regression Tree (CART), C5.0, Conditional Inference Tree (CTree) (three types of DTs differing for the splitting criterion), RF, ANN and LASSO. The output of every algorithm was the probability to belong to the positive class 'B'. The threshold for discriminating the two classes was set to 0.5 in this phase.

We applied 10-fold cross validation with grid search to tune the parameters of each algorithm for each balanced training set in order to find the best combination of parameters to maximise PPR.

**Table 3.** Positive 'B' and negative 'G' classes distribution for complete datasets, initial and balanced training sets and tests sets.

| Class | Complete Dataset | | Initial Training Set | | Balanced Training Set | | Test Set | |
|---|---|---|---|---|---|---|---|---|
| | **Samples** | **%** | **Samples** | **%** | **Samples** | **%** | **Samples** | **%** |
| | | | | **5-years** | | | | |
| B | 21,711 | 1.42 | 15,198 | 1.42 | 15,198 | 22.36 | 6513 | 1.42 |
| G | 1,508,003 | 98.58 | 1,055,603 | 98.58 | 52,780 | 77.64 | 452,400 | 98.58 |
| | | | | **4-years** | | | | |
| B | 21,963 | 1.39 | 15,375 | 1.39 | 15,375 | 21.98 | 6588 | 1.39 |
| G | 1,558,936 | 98.61 | 1,091,256 | 98.61 | 54,562 | 78.02 | 467,680 | 98.61 |
| | | | | **3-years** | | | | |
| B | 22,033 | 1.37 | 15,424 | 1.37 | 15,424 | 21.77 | 6609 | 1.37 |
| G | 1,583,594 | 98.63 | 1,108,516 | 98.63 | 55,425 | 78.23 | 475,078 | 98.63 |
| | | | | **2-years** | | | | |
| B | 22,083 | 1.36 | 15,459 | 1.36 | 15,459 | 21.55 | 6624 | 1.36 |
| G | 1,607,568 | 98.64 | 1,125,298 | 98.64 | 56,264 | 78.45 | 482,270 | 98.64 |
| | | | | **1-year** | | | | |
| B | 22,119 | 1.34 | 15,484 | 1.34 | 15,484 | 21.38 | 6635 | 1.34 |
| G | 1,626,778 | 98.66 | 1,138,745 | 98.66 | 56,937 | 78.62 | 488,033 | 98.66 |

For CART [62,63], the DT that uses Gini index as splitting criterion, the parameters to tune were those used for pruning: the complexity parameter (*cp*) (the minimum improvement to do to attempt a split) and the minimum number of observations that must exist in a node to perform a split (*minSplit*). On the contrary, the maximum depth of a tree was not tuned. While *minSplit* was different for the various datasets, the best *cp* was always the same (0.0001).

Additionally for C5.0 [64,65], a DT with Information Gain used as splitting criterion, the parameter chosen for tuning was linked to the pruning phase: the minimum number of samples that must be put in at least two of the splits (*minCases*), meaning the degree of fitting the initial tree (higher values cause a more approximate fit, making a sort of pre-pruning). The confidence factor (*CF*), which regards the severity of pruning, was set to the default value 0.25. However, also optimum *minCases* resulted to be the same for the five training sets (50).

CTree [66,67] uses statistical tests to evaluate the association of the input features and the target variable and to perform splits. In this study, the maximum *p*-value in order to implement a split was set to 0.05, while *minSplit* (with the same meaning as in CART) was tuned. Of course, the optimal values were different from CART.

For RF [68,69], an ensemble of trees, we tuned *ntree* (the number of trees in the forest) and *mtry* (the number of randomly selected variables chosen at each node). The best value of this last parameter was always 34, which corresponds to the default value in classification problems (the square root of the number of input features). We used the Gini index as splitting criterion for each tree.

ANNs [70,71] were built with one hidden layer for interpretability motivations. The number of hidden neurons (*HN*) was tuned. Each ANN run for at most 1000 iterations. The resulting *HNs* were different for all the datasets.

At last, for LASSO [72,73], the $\lambda$ parameter (i.e., the shrinkage penalty) was tuned. The optimal values were all in the order of $10^{-3}$.

The summary of the tuning phase with parameters and results for each algorithm and each training set is shown in Table 4.

The created models were really big and difficult to be interpreted because of the great number of input features. Moreover, variables' redundancy can lead to overfitting and can add noise, decreasing performance, and some of the 1178 input attributes could be irrelevant, increasing only computational time and memory required to store the model itself. Therefore, a feature selection (FS) process was executed on the five complete balanced training sets separately. We chose the Boruta algorithm to select the most important variables to predict the outcome [74], since it is both computationally efficient and simple and it does not require user defined parameters to tune. Boruta is a wrapper method built around RF. More in detail, it performs a top-down search for the most predictive attributes and iteratively deletes the less relevant variables: (1) at each run, an RF is built with an extended dataset, composed by the original dataset *w/o* unimportant attributes plus shuffled confirmed and tentative input features; (2) for every variable a statistical test is executed; (3) each attribute is confirmed or eliminated if the importance of the feature is significantly higher or significantly lower than the maximum importance among the shuffled attributes; (4) otherwise, if statistical significance is not found, the attribute is left tentative and the algorithm restarts from step 1. In this work, Boruta was run on the five complete balanced datasets separately, with $p$-value set to 0.01. After 100 runs for each dataset, it produced five reduced training sets (Figure 3) with different input features.

The reduced training sets were composed as follows:

- for the 5-years dataset, 280 variables were confirmed as important (99 diagnoses, 58 procedures, 35 outpatient services, 80 drugs, 6 exemptions, age and gender);
- for the 4-years dataset, 225 features were selected as important (65 diagnoses, 38 procedures, 38 outpatient services, 76 drugs, 6 exemptions, age and gender);
- for the 3-years dataset, 244 variables were chosen (58 diagnoses, 62 procedures, 39 outpatient services, 74 drugs, 9 exemptions, age and gender);
- for the 2-years dataset, 199 attributes were confirmed as important (47 diagnoses, 39 procedures, 29 outpatient services, 71 drugs, 11 exemptions, age and gender);
- for the 1-year dataset, 153 features were selected as important (44 diagnoses, 23 procedures, 24 outpatient services, 54 drugs, 6 exemptions, age and gender).

The six ML models were then re-built and re-tuned in the same way as before, but with the reduced balanced training sets. The best combinations of parameters before and after the feature selection process are shown in Table 4.

For CART, if *cp* was always 0.0001 (for all the datasets, both before and after FS), *minSplit* became smaller or remained equal after the FS process. C5.0's best *minCases* was 50 in all the cases, while *minSplit* for CTree was usually different w.r.t. its value before FS (except for 5- and 3-years datasets). In the case of RF, *ntree* did not vary with the reduction of input features, while *mtry* usually decreased. As regards ANN, the number of hidden neurons always increased after FS. At last, for LASSO, $\lambda$ went from order of $10^{-3}$ to order of $10^{-4}$.

**Table 4.** Tuning of the parameters for each algorithm vs. complete and reduced balanced training sets.

| Algorithm | Parameters and Ranges | Training Set | Best Values before FS | Best Values after FS |
|---|---|---|---|---|
| CART | *cp* in [$10^{-1}$, $10^{-2}$, $10^{-3}$, $10^{-4}$, $10^{-5}$] *minSplit* in [50, 70, 100, 150, 200] | 5-years | ($10^{-4}$, 200) | ($10^{-4}$, 100) |
| | | 4-years | ($10^{-4}$, 200) | ($10^{-4}$, 150) |
| | | 3-years | ($10^{-4}$, 100) | ($10^{-4}$, 100) |
| | | 2-years | ($10^{-4}$, 100) | ($10^{-4}$, 100) |
| | | 1-year | ($10^{-4}$, 70) | ($10^{-4}$, 70) |
| C5.0 | *minCases* in [50, 70, 100, 150, 200] | 5-years | 50 | 50 |
| | | 4-years | 50 | 50 |
| | | 3-years | 50 | 50 |
| | | 2-years | 50 | 50 |
| | | 1-year | 50 | 50 |
| CTree | *minSplit* in [50, 70, 100, 150, 200] | 5-years | 200 | 200 |
| | | 4-years | 50 | 70 |
| | | 3-years | 50 | 50 |
| | | 2-years | 70 | 50 |
| | | 1-year | 70 | 50 |
| RF | *ntree* in [300, 500, 700, 1000, 1500, 2000] *mtry* in [10, 16, 24, 34] | 5-years | (1000, 34) | (1000, 34) |
| | | 4-years | (500, 34) | (500, 24) |
| | | 3-years | (500, 34) | (500, 24) |
| | | 2-years | (800, 34) | (800, 34) |
| | | 1-year | (2000, 34) | (2000, 16) |
| ANN | *HN* in [2, 3, 4, ..., 13, 14, 15] | 5-years | 10 | 13 |
| | | 4-years | 11 | 14 |
| | | 3-years | 8 | 12 |
| | | 2-years | 5 | 6 |
| | | 1-year | 9 | 13 |
| LASSO | $\lambda$ in [$4.6\times10^{-3}$, $2.6\times10^{-3}$, $1.5\times10^{-3}$, $1.4\times10^{-3}$, $1.2\times10^{-3}$, $7.8\times10^{-4}$, $5.3\times10^{-4}$, $4.9\times10^{-4}$, $3.3\times10^{-4}$] | 5-years | $1.5\times10^{-3}$ | $4.9\times10^{-4}$ |
| | | 4-years | $4.6\times10^{-3}$ | $3.3\times10^{-4}$ |
| | | 3-years | $1.2\times10^{-3}$ | $7.8\times10^{-4}$ |
| | | 2-years | $1.4\times10^{-3}$ | $3.3\times10^{-4}$ |
| | | 1-year | $2.6\times10^{-3}$ | $5.3\times10^{-4}$ |

## 4. Results

The results were evaluated on the test set, to find the best model in terms of PPR, F1-Score and F2-Score. Figure 4 shows the results of all the models for these golden metrics.

Considering each model trained with the five complete datasets, PPR, F1-Score and F2-Score usually increase reducing the length of the period of study. There are only two exceptions: C5.0, where PPR with 3-years dataset is slightly worse than PPR with 4-years dataset, and CTree, where F1- and F2-Score obtained with 3-years dataset are almost the same of the results reached with 4-years dataset. However, for all the models, the best length of the period of historical data is 1 year. In addition, comparing the performance

achieved by the different models with the same input dataset, for every length of the period of study, the best model is RF, while the worst model is CTree. As a result, in the first phase where complete datasets were used, RF with 1-year of historical data is the best model.

These considerations are valid also for the second phase (after FS), considering the models obtained from the reduced training sets: better results are achieved when decreasing the length of the period of study and, with the same input dataset, RF behaves better than the other models.

From Table 5, which summarises the performance achieved by the models using 1-year datasets (the best length of period of study) before and after FS, we can note that results are usually slightly better after FS. This is valid also for the other datasets, as highlighted in Figure 4, where dotted lines (results obtained with reduced datasets) overlap or exceed continuous lines (performance achieved with complete datasets). There are only some exceptions: in terms of F1- and F2-Score, CTree 1-year, LASSO 1-year, RF 5-years and ANN 3-years behave better with the complete datasets; considering PPR, ANN 1-year achieves higher outcomes with the complete dataset; ANN 5-years reduces its performance for all three golden metrics after FS.

Therefore, among the 60 models (=6 algorithms × 10 datasets, where the 10 datasets are given by 5 periods × 2 phases, i.e., before and after FS), RF taking as input the reduced 1-year dataset results in the best one.

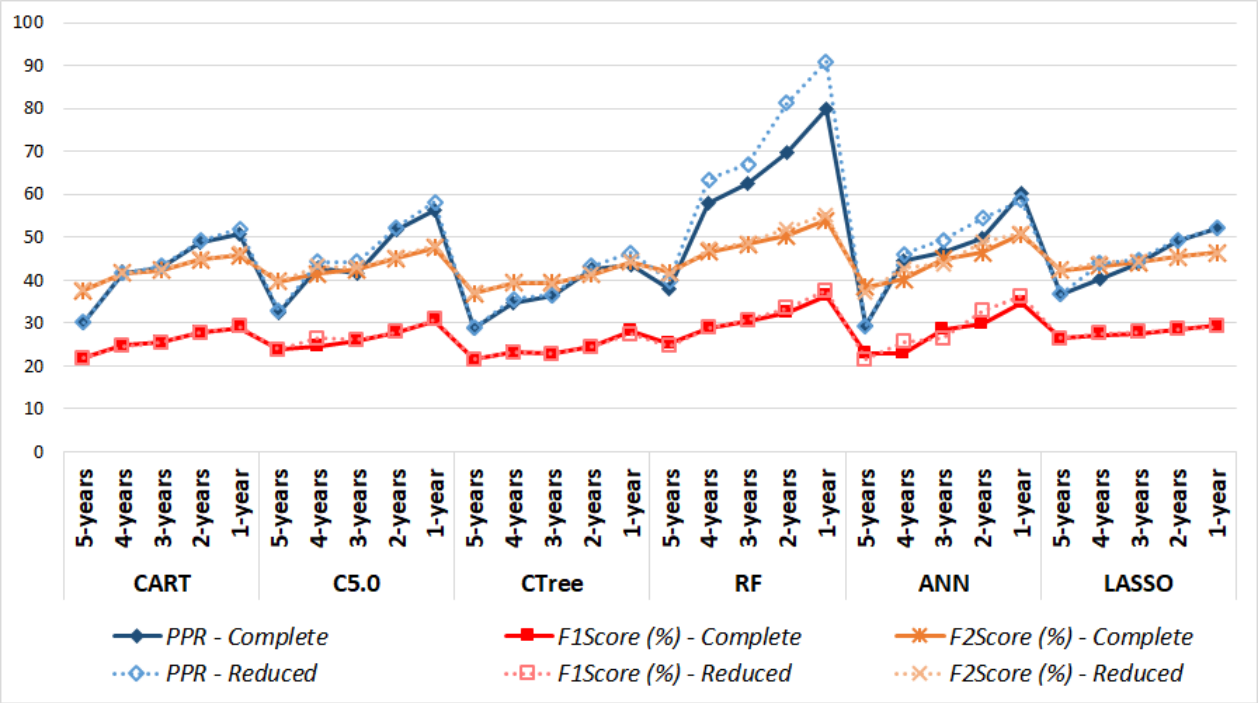



**Figure 4.** Comparison of the performance in terms of golden metrics of all the models trained with both complete balanced datasets and reduced balanced datasets, using the best parameters.

**Table 5.** Performance achieved by all the models using complete and reduced 1-year datasets.

| Model | Dataset | PPR | F1Score (%) | F2Score (%) | SE (%) | SP (%) | PPV (%) | NPV (%) |
|---|---|---|---|---|---|---|---|---|
| CART | Complete | 50.6 | 28.72 | 45.69 | 75.34 | 95.25 | 17.75 | 99.65 |
| | Reduced | 51.92 | 29.22 | 46.21 | 75.43 | 95.37 | 18.12 | 99.65 |
| C5.0 | Complete | 56.31 | 30.31 | 47.48 | 76.32 | 95.55 | 18.91 | 99.66 |
| | Reduced | 58.12 | 30.84 | 48.05 | 76.55 | 95.65 | 19.31 | 99.67 |
| CTree | Complete | 43.68 | 28.12 | 44.28 | 71.8 | 95.39 | 17.48 | 99.6 |
| | Reduced | 46.33 | 27.45 | 44.2 | 74.54 | 94.99 | 16.82 | 99.64 |
| RF | Complete | 79.99 | 36.48 | 53.85 | 78.87 | 96.55 | 23.73 | 99.7 |
| | Reduced | 91 | 37.52 | 55.31 | 80.87 | 96.6 | 24.43 | 99.73 |
| ANN | Complete | 60.09 | 34.7 | 50.63 | 72.98 | 96.63 | 22.76 | 99.62 |
| | Reduced | 58.72 | 36.21 | 51.05 | 70.22 | 97.04 | 24.4 | 99.58 |
| LASSO | Complete | 51.96 | 29.5 | 46.4 | 75.12 | 95.46 | 18.35 | 99.65 |
| | Reduced | 52.18 | 29.26 | 46.26 | 75.52 | 95.37 | 18.14 | 99.65 |

## 5. Discussion

The Boruta algorithm was used to select the most predictive variables for the identification of high-risk patients in each dataset. The confirmed features were different in the five cases and their number decreased considering the datasets in decreasing order of length of periods of study. The only exception was the 3-years dataset, where the number of selected features (244) was higher than the number of selected features in the 4-years dataset (225).

In all the cases, the most predictive variables are consistent with the identification of the target population. They are mostly related to malignant tumours, to cardiovascular and respiratory system, to kidney and bowel and to other chronic diseases, such as diabetes.

We can observe that decreasing the length of the period of study, the importance of diagnoses in the dataset usually decreases (99 of 280 for the 5-years dataset vs. 44 of 153 for the 1-year dataset), while the impact of drugs and outpatient services increases (80 groups of drugs and 35 groups of outpatients of 280 for the 5-years dataset vs. 54 groups of drugs and 24 groups of outpatients of 153 for the 1-year dataset). This is probably due to the fact that drugs and outpatient services have more incidence in a near future. On the contrary, some specific diagnoses, for example some chronic diseases, can become fatal over the years. Finally, the weight of procedures and exemptions is almost stable over the datasets and it is also comparable with the weight of these classes on the complete datasets.

In all the reduced sets, both the number of hospital admissions and the number of days of hospitalisation are usually confirmed together for the same CCS. This means that these variables carry different information. When only one attribute of the couple is selected, the number of days of hospitalisation is confirmed, underlying its more relevance w.r.t. the number of admissions.

Another characteristic valid for all reduced datasets is that age and gender are always included in the reduced set of input features, while all the other socio-economic attributes are always rejected. In fact, these features refer to a non-homogeneous group of people, i.e., the residents in the same municipality or in the same census section. Unfortunately, they do not approximate well the individual. However, the exemptions for socio-economic conditions are always included in the reduced datasets, highlighting the importance of the socio-economic status in this problem.

Obviously, reducing the number of input features, the models have become faster (both for training and inference) and lighter in terms of memory usage. Furthermore, the performance achieved with the reduced datasets is usually slightly better than that obtained with the complete datasets, as shown in Figure 4. Therefore, features deleted with the Boruta algorithm were probably useless and added just noise. Since all the selected algorithms perform a sort of feature selection while building, with pruning or shrinking coefficient and weights, the results for the various metrics are not significantly better after FS. However, the models built with the reduced training sets are the best candidates for the first level screening tool, to extract in advance patients requiring specific programs of care, for several reasons. Firstly, they lower complexity during the collection of data, thanks to the reduction of the number of input features. Secondly, they can be compliant for the integration with mobile devices, such as tablets and smartphones, and, consequently, with telemedicine platforms [75], since they reduce both computational time and required memory. The inference time, lower than 1 ms per patient (about 108 s for RF for the entire 1-year reduced test set, composed by 494,668 rows), in fact, could allow to re-classify the patients every time a significant event occurs. At last, the results of the models built with the reduced training sets are comparable or even better than the ones achieved by the models built with the complete training sets.

Considering each model built with datasets with different historical data, it figures out that there is an improvement in performance decreasing the length of the period of study. The highest risk of hospitalisation or death is then defined by recent medical events. Moreover, considering old events with the same weight of newer ones can cause the misclassification of some samples, as demonstrated by worse performance with longer periods of study. Thus, from the trade-off on datasets deriving from periods of study of different lengths, it turns out that the best method is to consider only one year of medical history to identify high-risk patients next year. This result simplifies even more the collection of data, because it requires to go back just one year. Moreover, the input features are only 153 in this case.

Random Forest with reduced 1-year dataset results to be the best model among all. In particular, it achieves very high PPR (91), as shown in Table 5: people identified as high-risk have the risk of avoidable hospitalisation or death about 91 times greater than the others. Moreover, as highlighted in Table 5, it reaches both high sensitivity and high specificity (almost 81% and almost 97%, respectively), classifying the most part of the samples correctly. If featuring high specificity is quite a simple task, because of the great imbalance of the classes, high sensitivity is a very appreciable result for the same reason. Thus, most high-risk patients will be recognised in advance and monitored more carefully by their GPs; instead, for the majority part of lower-risk people, the healthcare approach will not be modified, because of their small needs. Conversely, RF features a high number of false positives (people identified as high-risk by the model, but really low-risk). In addition, this result is due to the very low outcome prevalence [76]. However, reducing the length of the period of study and passing from 5-years to 1-year of historical data, PPV increases of about 10%, from 14.64% [18] to 24.43% (Table 5), because higher-risk patients are identified considering only newer events. In addition, since the model will be used for the first level of screening, at this step it is acceptable to include more people than necessary in the positive class. Some of them will be excluded during the second level of screening made by the GPs, who can consider also behavioural, social and environmental factors. In this first phase, it is much more important to reduce the number of false negatives (people classified as negatives, but really belonging to the positive class), to exclude very few patients having really need of specific treatments. This aim is achieved and confirmed by high SE and high NPV (Table 5).

During the testing phase, the probability threshold was set to 0.5. The problem of false positives can be handled changing this probability threshold. Increasing the threshold, in fact, the model becomes more and more selective, predicting very high-risk patients. In this way, false positives decrease, but also true positives decrease, causing the increase

of false negatives. This scenario can be taken into consideration and really implemented, since in this case a limited list of high-risk patients is produced for the GPs. On the other hand, decreasing the probability threshold, more patients are identified as high-risk, increasing true positives and reducing false negatives, but in this case the single GP could have problems in enrolling such a large number of patients in a specific healthcare model. The Receiver Operating Characteristic (ROC) curve of RF with reduced 1-year dataset (Figure 5) shows sensitivity and false positive rate ($1 - SP$) at different thresholds. Figure 5 highlights the threshold used during the testing phase (0.5) and the threshold that best trades-off minimisation of false positives and false negatives (0.28), i.e., the closest point to (0, 1). It is not obvious that the best threshold in terms of metrics is also the best one in the practical scenario according to the current healthcare system. The Area Under the Curve (AUC) reaches a very high value (0.967) and points out the excellent performance of the selected model.

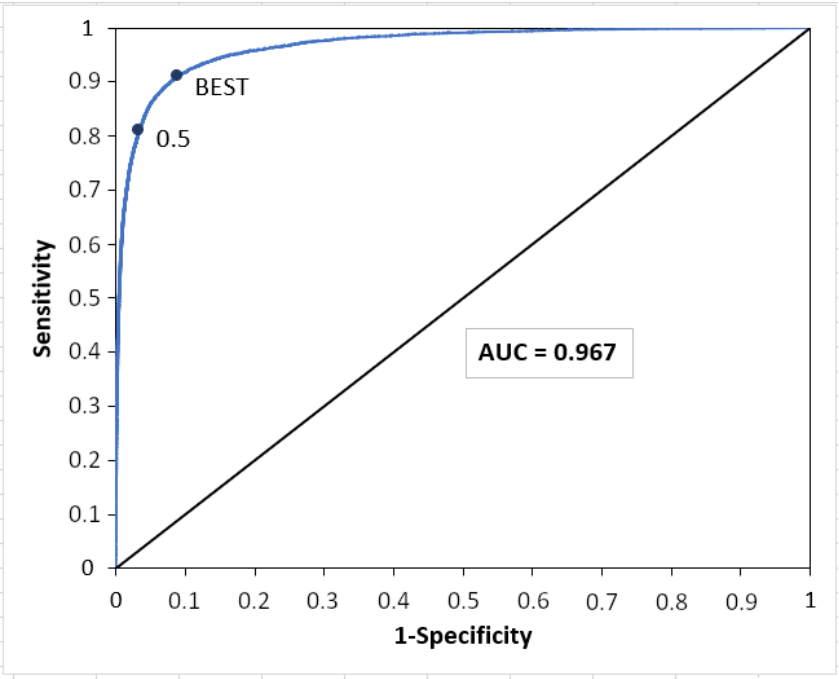



**Figure 5.** Receiver Operating Characteristic (ROC) curve of Random Forest (RF) which takes in input the reduced 1-year dataset.

RF can also be explained to the GPs, giving them some information regarding the importance and the weight that each input feature has in the final predictions. In mathematical terms, the importance of an input variable $A_k$ in the forest is the average of the importance of $A_k$ in every tree. For each tree, the importance of $A_k$ is given by the difference between the error rate of predictions obtained with the original dataset and the error rate of predictions obtained with a dataset which coincides with the original one, except for $A_k$, which is randomly permuted. Moreover, the importance of each variable can be expressed using other metrics, such as mean minimal depth, *times_a_root*, accuracy decrease and Gini decrease [77]:

- The mean minimal depth of $A_k$ is the average of the minimal depth of $A_k$ in each tree, where the minimal depth of $A_k$ in each tree represents the length of the path from the root to the node where the variable $A_k$ is used for splitting. The more important the feature is, the smaller mean minimal depth is, since it means that the variable discriminates well the two classes, maximising the reduction of impurity in the set.
- *Times_a_root* of $A_k$ is the number of trees where the variable $A_k$ is at the root. Obviously, it is strongly related inversely to the mean minimal depth: with the growth of *times_a_root*, mean minimal depth decreases.

- Accuracy decrease for variable $A_k$ is the mean decrease of accuracy in predictions if $A_k$ is randomly permuted. If it is larger, the variable is more important and affects more the final predictions.
- Gini decrease of $A_k$ is the mean decrease of Gini index when splitting on $A_k$ itself. As for accuracy decrease, the larger it is, the greater is the weight of the attribute in the predictions.

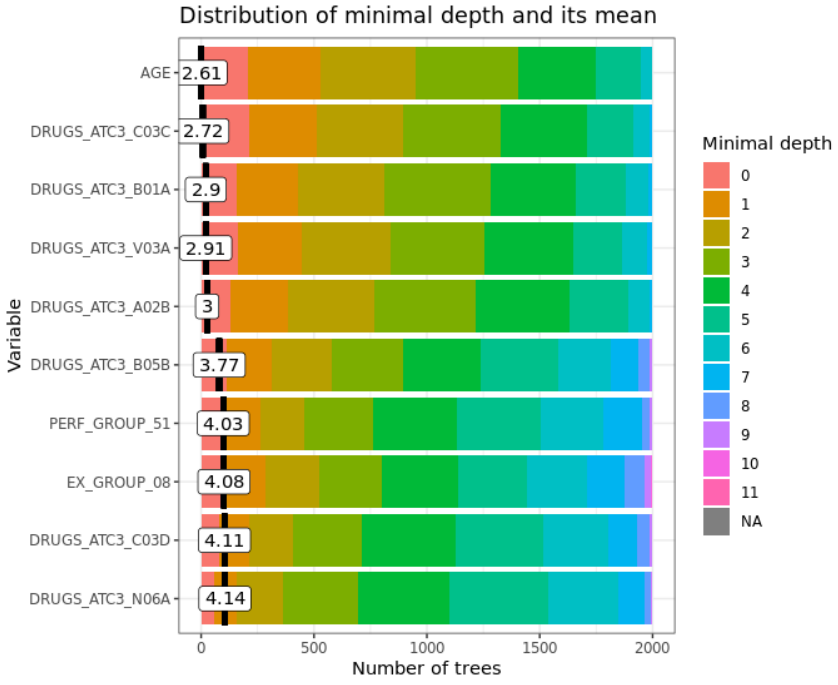
From the two plots of Figure 6, it is easy to note that age is the variable which most affects the output, since it has the maximum decrease both in accuracy (∼7%) and in Gini index (∼3761). This is the reason why it is usually at the root (in 209 trees) and it has the minimum mean minimal depth (2.61). This is not surprising, as residents of all ages are included in the study. Even if the variable DRUGS_ATC3_C03C (number of high-ceiling diuretics) has similar mean minimal depth (2.72), it has very lower accuracy decrease (∼1.5%) and Gini decrease (∼2058). The other top variables are mostly drugs, and there are only one group of outpatient services and one group of exemptions in this set. They are mostly linked to chronic diseases.

Their meaning is explained in Table 6. No hospital admissions nor days of hospitalisation for specific diagnoses nor procedures are included in the set of the most important features for the model.

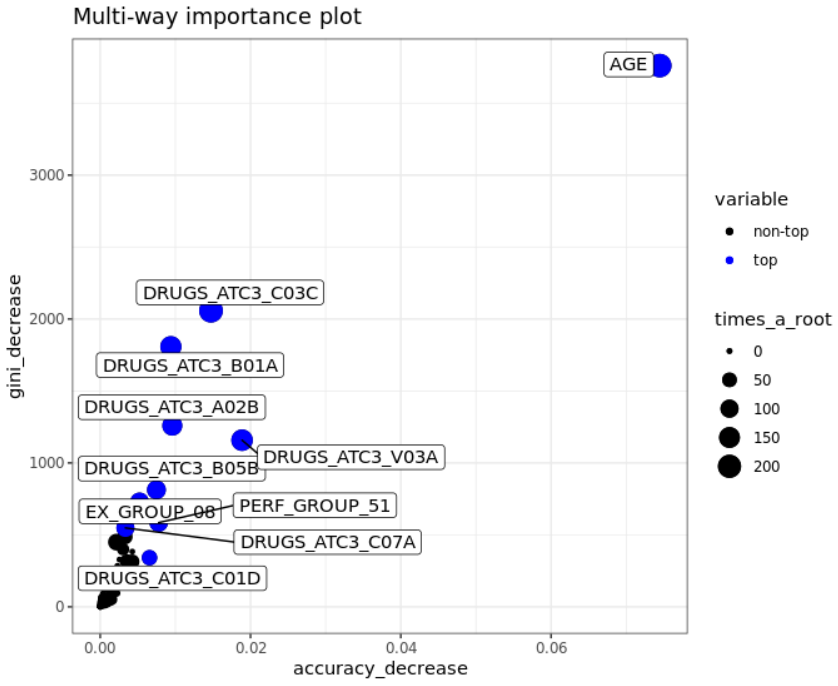
The inclusion of almost all the residents in an area (the LHU of Central Tuscany) and the consideration of features that are usually excluded can be a motivation of the good performance achieved by our best model, if compared with state-of-the-art studies. For example, the algorithm currently used in Tuscany for the identification of high-risk patients [45] excludes a-priori the patients without a hospitalisation in the previous three years. They achieve almost 6 for PPR and almost 72.8% as SE. With RF, PPR reaches 91 and SE increases to 80.87%.

**Table 6.** Description of the most important input features for RF trained with reduced 1-year dataset.

| Input Features | Description |
|---|---|
| AGE | Age on 1st January |
| DRUGS_ATC3_A02B | Number of drugs for peptic ulcer and gastro-oesophageal reflux disease |
| DRUGS_ATC3_B01A | Number of antithrombotic agents |
| DRUGS_ATC3_B05B | Number of intravenous solutions |
| DRUGS_ATC3_C01D | Number of vasodilators used in cardiac diseases |
| DRUGS_ATC3_C03C | Number of high-ceiling diuretics |
| DRUGS_ATC3_C03D | Number of potassium-sparing agents |
| DRUGS_ATC3_C07A | Number of beta-blocking agents |
| DRUGS_ATC3_N06A | Number of antidepressant |
| DRUGS_ATC3_V03A | Number of other therapeutic products |
| EX_GROUP_08 | Exemption for disability |
| PERF_GROUP_51 | Number of immunohematology laboratory exams |




(**a**)




(**b**)

**Figure 6.** Most important input features for RF trained with the reduced 1-year dataset: (**a**) distribution of minimal depth and its mean (marked by a vertical bar); (**b**) multi-way importance plot vs. *accuracy_decrease*, *gini_decrease* and *times_a_root*.

In Tuscany region, statistical methods are currently used to stratify the population by risk and for the identification of high-risk patients, to produce lists for the GPs to implement a proactive healthcare approach. These lists are extracted once per year. The proposed method fits into this context, replacing statistical methods. In fact, in this way, the selection of high-risk patients could be refined, thanks to better performance. Moreover, because

of the very low inference time and memory footprint, the risk of hospitalisation or death could be recalculated every month, when mARSupio is updated with new medical records, in a synchronous way for all the residents. The granularity of this process could also be increased. In fact, GPs could interact directly with the system by means of an easy graphical user interface. More in detail, patients could be registered in the platform at the beginning (when their GP approves the use of this system). At this time, historical data of the last year are uploaded and the initial risk index is computed. Then, the GP could access to a web platform, a mobile application or a computer program and, every time a medical or economic event occurs for each patient, update their level of risk. In this way, the risk index will be computed in real-time, asynchronously for each patient, and GPs could rapidly react to a patient predicted risk by adjusting relevant therapies or any other required action. This system could also be adopted in other Italian regions, where a structured administrative database is present and available and data are detailed as in mARSupio.

## 6. Conclusions

Population affected by multiple chronic conditions is increasing with population ageing and soaks up the most part of the healthcare resources. The early identification of complex patients becomes then crucial, but state-of-the-art approaches are used for the identification of sub-populations with a single chronic condition or use clinical data, which are not available in Italy. The idea of this work is to develop a data-driven administrative-socio-economic population health management tool based on machine learning algorithms for the selection of high-risk (complex) patients.

This paper presents a trade-off of several machine learning algorithms and multiple input datasets with the aim of identifying high-risk patients in the Tuscan population, to select the best model with the best data for a population health management tool. Datasets differ among themselves in the length of the period of study and are composed by administrative and socio-economic data. The best model in terms of three golden metrics, i.e., PPR, F1-Score and F2-Score, results to be Random Forest with historical data collected in 1-year period. The final input attributes of the dataset are only 153 of the initial 1178. Among them, age and some groups of drugs are the ones that most affect the output. This novel approach for feature selection leads to better performance of selected Random Forest model w.r.t. state-of-the-art algorithms. This is demonstrated by the fact that Tuscany region is considering the adoption of the proposed method for the identification of high-risk patients in the coming years.

Moreover, this approach allows for real-time risk prediction for any given patient as soon as new medical or economic data are gathered by the system. In this way, the General Practitioner can rapidly react defining a proper personalised medicine for the patient in order to reduce the predicted risk. The predictive model could be even improved by exploiting patient vital parameters in case the patient is enrolled in a telemedicine program.

**Author Contributions:** Conceptualization, S.P., F.P. and M.D.; Methodology, S.P. and F.P.; software, S.P.; validation, F.P. and M.D.; formal analysis, S.P.; investigation, S.P.; resources, F.P. and P.F.; data curation, S.P. and M.D.; writing–original draft preparation, S.P. and M.D.; writing–review and editing, S.P., M.D., F.P. and L.F.; visualization, F.P.; supervision, P.F. and L.F. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References


1. Mitchell, E.; Walker, R. Global ageing: Successes, challenges and opportunities. *Br. J. Hosp. Med.* **2020**, *81*. [CrossRef] [PubMed]
2. Anderson, G.F.; Hussey, P.S. Population Aging: A Comparison Among Industrialized Countries. *Health Aff.* **2000**, *19*. [CrossRef] [PubMed]
3. Colby, S.L.; Ortman, J.M. *Projections of the Size and Composition of the U.S. Population: 2014 to 2060. Population Estimates and Projections*; Current Population Reports; Census Bureau: Washington, DC, USA, 2015.
4. Nash, A. National Population Projections: 2018-Based. 2019. Available online: www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationprojections/bulletins/nationalpopulationprojections2018based (accessed on 29 November 2020).
5. Légaré, J. Population Aging: Economic and Social Consequences. In *International Encyclopedia of the Social & Behavioral Sciences*, 2nd ed.; Elsevier: Oxford, UK, 2015; pp. 540–544. [CrossRef]
6. Kingston, A.; Robinson, L.; Booth, H.; Knapp, M.; Jagger, C. Projections of multi-morbidity in the older population in England to 2035: Estimates from the Population Ageing and Care Simulation (PACSim) model. *Age Ageing* **2018**, *47*, 374–380. [CrossRef] [PubMed]
7. Marengoni, A.; Angleman, S.; Melis, R.; Mangialasche, F.; Karp, A.; Garmen, A.; Meinow, B.; Fratiglioni, L. Aging with multimorbidity: A systematic review of the literature. *Ageing Res. Rev.* **2011**, *10*, 430–439. [CrossRef] [PubMed]
8. Thavorn, K.; Maxwell, C.J.; Gruneir, A.; Bronskill, S.E.; Bai, Y.; Koné Pefoyo, A.J.; Petrosyan, Y.; Wodchis, W.P. Effect of socio-demographic factors on the association between multimorbidity and healthcare costs: A population-based, retrospective cohort study. *BMJ Open* **2017**, *7*. [CrossRef]
9. Bodenheimer, T.; Wagner, E.H.; Grumbach, K. Improving Primary Care for Patients with Chronic Illness. *JAMA* **2002**, *288*, 1775–1779. [CrossRef]
10. Boehmer, K.R.; Dabrh, A.M.A.; Gionfriddo, M.R.; Erwin, P.J.; Montori, V.M. Does the chronic care model meet the emerging needs of people living with multimorbidity? A systematic review and thematic synthesis. *PLoS ONE* **2018**, *13*, e0190852. [CrossRef]
11. Shadmi, E.; Freund, T. Targeting patients for multimorbid care management interventions: The case of equity in high-risk patient identification. *Int. J. Equity Health* **2013**, *12*, [CrossRef]
12. Safford, M.M.; Allison, J.J.; Kiefe, C.I. Patient Complexity: More Than Comorbidity. The Vector Model of Complexity. *J. Gen. Intern. Med.* **2007**, *22*, 382–390. [CrossRef]
13. Andreu-Perez, J.; Poon, C.C.Y.; Merrifield, R.D.; Wong, S.T.C.; Yang, G.Z. Big Data for Health. *IEEE J. Biomed. Health Inform.* **2015**, *19*, 1193–1208, [CrossRef]
14. Bates, D.W.; Saria, S.; Ohno-Machado, L.; Shah, A.; Escobar, G. Big Data in Health Care: Using Analytics to Identify And Manage High-Risk and High-Cost Patients. *Health Aff.* **2014**, *33*, 1123–1131, [CrossRef] [PubMed]
15. Raghupathi, W.; Raghupathi, V. Big Data in Healthcare: Promise and Potential. *Health Inf. Sci. Syst.* **2014**, *2*, [CrossRef] [PubMed]
16. Dash, S.; Shakyawar, S.K.; Sharma, M.; Kaushik, S. Big data in healthcare: Management, analysis and future prospects. *J. Big Data* **2019**, *6*, 1–25. [CrossRef]
17. Bresnick, J. How to Get Started with a Population Health Management Program. 2016. Available online: healthitanalytics.com/features/how-to-get-started-with-a-population-health-management-program (accessed on 29 November 2020).
18. Panicacci, S.; Donati, M.; Fanucci, L.; Bellini, I.; Profili, F.; Francesconi, P. Population Health Management Exploiting Machine Learning Algorithms to Identify High-Risk Patients. In Proceedings of the 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), Karlstad, Sweden, 18–21 June 2018; pp. 298–303. [CrossRef]
19. Mehta, N.; Pandit, A. Concurrence of big data analytics and healthcare: A systematic review. *Int. J. Med. Inform.* **2018**, *114*, 57–65. [CrossRef] [PubMed]
20. Swain, A.K. Mining big data to support decision making in healthcare. *J. Inf. Technol. Case Appl. Res.* **2016**, *18*, 141–154. [CrossRef]
21. Chen, M.; Hao, Y.; Hwang, K.; Wang, L.; Wang, L. Disease Prediction by Machine Learning over Big Data from Healthcare Communities. *IEEE Access* **2017**, *5*, 8869–8879. [CrossRef]
22. Meng, X.H.; Huang, Y.X.; Rao, D.P.; Zhang, Q.; Liu, Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *Kaohsiung J. Med. Sci.* **2013**, *29*, 93–99. [CrossRef]
23. Worachartcheewan, A.; Shoombuatong, W.; Pidetcha, P.; Nopnithipat, W.; Prachayasittikul, V.; Nantasenamat, C. Predicting Metabolic Syndrome Using the Random Forest Method. *Sci. World J.* **2015**. [CrossRef]
24. 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*. [CrossRef]
25. Dinesh, K.G.; Arumugaraj, K.; Santhosh, K.D.; Mareeswari, V. Prediction of Cardiovascular Disease Using Machine Learning Algorithms. In Proceedings of the 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, India, 1–3 March 2018; pp. 1–7. [CrossRef]
26. Panicacci, S.; Donati, M.; Fanucci, L.; Bellini, I.; Profili, F.; Francesconi, P. Exploring Machine Learning Algorithms to Identify Heart Failure Patients: The Tuscany Region Case Study. In Proceedings of the 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), Cordoba, Spain, 5–7 June 2019; pp. 417–422. [CrossRef]
27. Tengnah, M.A.J.; Sooklall, R.; Nagowah, S.D. Chapter 9-A Predictive Model for Hypertension Diagnosis Using Machine Learning Techniques. In *Telemedicine Technologies*; Academic Press: New York, NY, USA, 2019; pp. 139–152. [CrossRef]

28. Yang, H.; Bath, P.A. The Use of Data Mining Methods for the Prediction of Dementia: Evidence from the English Longitudinal Study of Aging. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 345–353. [CrossRef]
29. Cattelani, L.; Murri, M.B.; Chesani, F.; Chiari, L.; Bandinelli, S.; Palumbo, P. Risk Prediction Model for Late Life Depression: Development and Validation on Three Large European Datasets. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 2196–2204. [CrossRef] [PubMed]
30. What Is Electronic Health Record (EHR)? 2019. Available online: https://www.healthit.gov/faq/what-electronic-health-record-ehr (accessed on 29 November 2020).
31. Myers, L.; Stevens, J. Using EHR to Conduct Outcome and Health Services Research. In *Secondary Analysis of Electronic Health Records*; Springer International Publishing: Cham, Switzerland, 2016; pp. 61–70. [CrossRef]
32. Alamri, A. Ontology Middleware for Integration of IoT Healthcare Information Systems in EHR Systems. *Computers* **2018**, *7*, 51. [CrossRef]
33. Hammond, R.; Athanasiadou, R.; Curado, S.; Aphinyanaphongs, Y.; Abrams, C.; Messito, M.; Gross, R.; Katzow, M.; Jay, M.; Razavian, N.; et al. Predicting childhood obesity using electronic health records and publicly available data. *PLoS ONE* **2019**, *14*. [CrossRef] [PubMed]
34. Anderson, J.P.; Parikh, J.R.; Shenfeld, D.K.; Ivanov, V.; Marks, C.; Church, B.W.; Laramie, J.M.; Mardekian, J.; Piper, B.A.; Willke, R.J.; et al. Reverse Engineering and Evaluation of Prediction Models for Progression to Type 2 Diabetes: An Application of Machine Learning Using Electronic Health Records. *J. Diabetes Sci. Technol.* **2016**, *10*, 6–18. [CrossRef]
35. Panahiazar, M.; Taslimitehrani, V.; Pereira, N.; Pathak, J. Using EHRs and Machine Learning for Heart Failure Survival Analysis. *Stud. Health Technol. Inform.* **2015**, *216*, 40–44.
36. Pike, M.M.; Decker, P.A.; Larson, N.B.; Sauver, J.L.S.; Takahashi, P.Y.; Roger, V.L.; Rocca, W.A.; Miller, V.M.; Olson, J.E.; Pathak, J.; et al. Improvement in Cardiovascular Risk Prediction with Electronic Health Records. *J. Cardiovasc. Transl. Res.* **2016**, *9*, 214–222. [CrossRef]
37. Sun, J.; McNaughton, C.D.; Zhang, P.; Perer, A.; Gkoulalas-Divanis, A.; Denny, J.C.; Kirby, J.; Lasko, T.; Saip, A.; Malin, B.A. Predicting changes in hypertension control using electronic health records from a chronic disease management program. *J. Am. Med. Inform. Assoc.* **2013**, *21*, 337–344. [CrossRef]
38. Barnes, D.E.; Zhou, J.; Walker, R.L.; Larson, E.B.; Lee, S.J.; Boscardin, W.J.; Marcum, Z.A.; Dublin, S. Development and Validation of eRADAR: A Tool Using EHR Data to Detect Unrecognized Dementia. *J. Am. Geriatr. Soc.* **2020**, *68*, 103–111. [CrossRef]
39. Jin, Z.; Cui, S.; Guo, S.; Gotz, D.; Sun, J.; Cao, N. CarePre: An Intelligent Clinical Decision Assistance System. *ACM Trans. Comput. Healthc.* **2020**, *1*. [CrossRef]
40. Morawski, K.; Dvorkis, Y.; Monsen, C.B. Predicting Hospitalizations from Electronic Health Record Data. *Am. J. Manag. Care* **2020**. [CrossRef]
41. Miotto, R.; Li, L.; Dudley, J.T. Deep Learning to Predict Patient Future Diseases from the Electronic Health Records. In *Advances in Information Retrieval*; Springer International Publishing: Cham, Switzerland, 2016; pp. 768–774.
42. Kim, Y.J.; Park, H. Improving Prediction of High-Cost Health Care Users with Medical Check-Up Data. *Big Data* **2019**, *7*. [CrossRef] [PubMed]
43. Shenas, S.A.I.; Raahemi, B.; Tekieh, M.H.; Kuziemsky, C. Identifying high-cost patients using data mining techniques and a small set of non-trivial attributes. *Comput. Biol. Med.* **2014**, *53*, 9–18. [CrossRef] [PubMed]
44. Morid, M.A.; Kawamoto, K.; Ault, T.; Dorius, J.; Abdelrahman, S. Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation. *Amia Annu. Symp. Proc.* **2017**, *2017*, 1312–1321. [PubMed]
45. Bellini, I.; Barletta, V.R.; Profili, F.; Bussotti, A.; Severi, I.; Isoldi, M.; Bimbi, M.V.F.; Francesconi, P. Identifying High-Cost, High-Risk Patients Using Administrative Databases in Tuscany, Italy. *BioMed Res. Int.* **2017**. [CrossRef]
46. Louis, D.Z.; Robeson, M.; McAna, J.; Maio, V.; Keith, S.W.; Liu, M.; Gonnella, J.S.; Grilli, R. Predicting risk of hospitalisation or death: A retrospective population-based analysis. *BMJ Open* **2014**, *4*. [CrossRef] [PubMed]
47. Balzi, D.; Carreras, G.; Tonarelli, F.; Degli Esposti, L.; Michelozzi, P.; Ungar, A.; Gabbani, L.; Benvenuti, E.; Landini, G.; Bernabei, R.; et al. Real-time utilisation of administrative data in the ED to identify older patients at risk: Development and validation of the Dynamic Silver Code. *BMJ Open* **2019**, *9*. [CrossRef]
48. Linn, S.; Grunau, P.D. New patient-oriented summary measure of net total gain in certainty for dichotomous diagnostic tests. *Epidemiol. Perspect. Innvoation* **2006**, *3*. [CrossRef]
49. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
50. Tutela Delle Persone e di Altri Soggetti Rispetto al Trattamento dei dati Personali [Protection of Persons and Other Subjects with Regard to Personal Data Processing]. 1996. Available online: http://www.garanteprivacy.it/web/guest/home/docweb/-/docwebdisplay/docweb/28335 (accessed on 29 November 2020).
51. Donatini, A. The Italian Health Care System. 2020. Available online: https://international.commonwealthfund.org/countries/italy/ (accessed on 29 November 2020).
52. ISTAT. Available online: https://www.istat.it/ (accessed on 29 November 2020).
53. Toscana, A.R.S. MARSupio Database. 2018. Available online: https://www.ars.toscana.it/marsupio/database/ (accessed on 29 November 2020).

54. AHRQ. Potentially Avoidable Hospitalizations. 2018. Available online: www.ahrq.gov/research/findings/nhqrdr/chartbooks/carecoordination/measure3.html (accessed on 29 November 2020).
55. Elixhauser, A.; Steiner, C.; Palmer, L. Clinical Classification Software (CCS). 2016. Available online: http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp (accessed on 29 November 2020).
56. ICD-9-CM Diagnosis Codes. Available online: www.icd9data.com/2012/Volume1/default.htm (accessed on 29 November 2020).
57. ICD-9-CM Procedure Codes. Available online: www.icd9data.com/2012/Volume3/default.htm (accessed on 29 November 2020).
58. WHOCC. Anatomical Therapeutic Chemical Classification System (ATC). 2018. Available online: www.whocc.no/atc/structure_and_principles/ (accessed on 29 November 2020).
59. HealthCatalyst. Population Health Management: Systems and Success. 2017. Available online: https://www.healthcatalyst.com/population-health/ (accessed on 29 November 2020).
60. Eurostat. Projected Old-Age Dependency Ratio. 2019. Available online: https://ec.europa.eu/eurostat/web/products-datasets/-/tps00200 (accessed on 29 November 2020).
61. Brownlee, J. 8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset. 2015. Available online: machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/ (accessed on 29 November 2020).
62. Breiman, L.; Friedman, J.; Olsen, R.A.; Stone, C.J. *Classification and Regression Trees*; Wadsworth International Group: Belmont, CA, USA, 1984.
63. Therneau, T.; Atkinson, B.; Ripley, B. Package 'Rpart'. 2019. Available online: https://cran.r-project.org/web/packages/rpart/rpart.pdf (accessed on 29 November 2020).
64. Pandya, R.; Pandya, J. C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. *Int. J. Comput. Appl.* **2015**, *117*, 18–21. [CrossRef]
65. Kuhn, M.; Weston, S.; Culp, M.; Coulter, N.; Quinlan, R. Package 'C50'. 2020. Available online: https://cran.r-project.org/web/packages/C50/C50.pdf (accessed on 29 November 2020).
66. Hothorn, T.; Hornik, K.; Zeileis, A. Unbiased Recursive Partitioning: A Conditional Inference Framework. *J. Comput. Graph. Stat.* **2006**, *15*, 651–674. [CrossRef]
67. Hothorn, T.; Hornik, K.; Strobl, C.; Zeileis, A. Package 'Party'. 2020. Available online: https://cran.r-project.org/web/packages/party/party.pdf (accessed on 29 November 2020).
68. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
69. Wright, M.N.; Wager, S.; Probst, P. Package 'Ranger'. 2020. Available online: https://cran.r-project.org/web/packages/ranger/ranger.pdf (accessed on 29 November 2020).
70. Maind, S.B.; Wankar, P. Research Paper on Basic of Artificial Neural Network. *Int. J. Recent Innov. Trends Comput. Commun.* **2014**, *2*, 96–100.
71. Ripley, B.; Venables, W. Package 'nnet'. 2020. Available online: https://cran.r-project.org/web/packages/nnet/nnet.pdf (accessed on 29 November 2020).
72. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. [CrossRef]
73. Friedman, J.; Hastie, T.; Tibshirani, R.; Narasimhan, B.; Simon, N.; Qian, J. Package 'glmnet'. 2020. Available online: https://cran.r-project.org/web/packages/glmnet/glmnet.pdf (accessed on 29 November 2020).
74. Kursa, M.B.; Jankowski, A.; Rudnicki, W.R. Boruta—A System for Feature Selection. *Fundam. Inform.* **2010**, *101*. [CrossRef]
75. Donati, M.; Celli, A.; Ruiu, A.; Saponara, S.; Fanucci, L. A Telemedicine Service System Exploiting BT/BLE Wireless Sensors for Remote Management of Chronic Patients. *Technologies* **2019**, *7*, 13. [CrossRef]
76. Altman, D.G.; Bland, J.M. Statistics Notes: Diagnostic tests 2: Predictive values. *BMJ* **1994**, *309*, 102. [CrossRef]
77. Paluszynska, A.; Biecek, P.; Jiang, Y. Package 'RandomForestExplainer'. 2020. Available online: https://cran.r-project.org/web/packages/randomForestExplainer/randomForestExplainer.pdf (accessed on 29 November 2020).