# Handling Missing Phenotype Data with Random Forests for Diabetes Risk Prognosis

**Beatriz López**[1], **Ramon Viñas**[2] and **Ferran Torrent-Fontbona**[3] and **José Manuel Fernández-Real**[4]

**Abstract.** Machine learning techniques are the cornerstone to handle the amounts of information available for building comprehensive models for decision support in medical practice. However, the datasets use to have a lot of missing information. In this work we analyse how the random forests technique could be used for dealing with missing phenotype values in order to prognosticate diabetes type 2.

## 1 INTRODUCTION

Diagnosis of type 2 diabetes is made typically using clinical criteria. However, some population studies, specially in which young people is involved, have provided evidence that the diagnosis should be supported by phenotype data [16]. This phenotype data is not just useful for handling inheritance factors, but also for understanding nutrition conditions in pre and post-natal stages (see [8] and [9] for a reviewed version). In fact, phenotype data could provide new possibilities for handling risk prognosis for both, type 1 and type 2 diabetes [17], and also find explanations for other combination processes known as undetermined diabetes or 1.5 diabetes [16].

Our work concerns on using phenotype data to building a clinical decision support system (CDSS) for diabetes 2 prognosis. To that end, we are provided with a huge dataset of patient samples, each one characterised by a considerable amount of phenotypes. Therefore, we require the application of a machine learning technique to obtain a prognosis model to be handled by the CDSS. In so doing, our challenge is to handle the considerable amount of missing information, a typical situation when dealing with phenotypes [14].

There are several methods to deal with missing data that can be organized in four categories [15]. First, methods that discard instances (i.e. samples) with missing information. Second, methods that acquire missing values to complete the information, which involves some additional costs. Third, imputation methods are the largest family, and can be in turn organized in three groups: predictive value computation methods (e.g. mean, mode, the most popular ones), distribution-based computation (which take into account the class or diagnose of the samples), and unique-value imputation (replacing the missing value by a given value that represents it). Finally, the fourth category of methods are the reduced-feature models which incorporate only the phenotypes known in a given query (test). These latter kind of methods have been shown to be the ones that most improve the prognosis accuracy [15]

[1] University of Girona, email: beatriz.lopez@udg.edu
[2] University of Girona, email: rvinast@gmail.com
[3] University of Girona, email: ferran.torrent@udg.edu
[4] Biomedical Research Institute of Girona, email: jmfreal@idibgi.org

Handling missing values by adding and removing features according to a given query as reduced model approaches do is quite similar to the random forests (RF) machine learning technique. RF is a method that combines several decision tree models to provide a classification outcome (i.e. prognosis) [5]. Each decision tree is learned by using a base learner method applied to a subset of features (phenotypes) that are randomly selected, as well as to a subset of samples that are also randomly chosen. In fact, the RF technique could be considered as a combination of discard instance methods and reduced-feature models for handling of missing values. However RF does not remove any information which could be useful towards a personalised prognosis. In this paper, we analyse such possibility by applying RF to prognosticate diabetes type 2 from a dataset of phenotypes with a considerable amount of missing values.

This paper is organized as follows. First, we describe in Section 2 some previous related work. Next, in Section 3 we explain our method. We continue in Section 4 by describing the experimentation carried out and providing the results obtained. We end the paper in Section 5 with some conclusions and discussion about future work.

## 2 RELATED WORK

The application of machine learning techniques to gene expression data is becoming a key issue for Biomedicine [3]. For example, [7] build a binary logistic regression model based on phenotypes and genotype data to risk prediction of inheritance diabetes. 5639 patients were considered in the study, from which samples with at most a 10% of missing features were considered. We are not provided with so many patient data, and we need to handle a higher number of missing information to keep enough samples for learning a model.

In [14] and approach for imputing missing phenotypes based on a method called co-trained is presented. Co-trained means that missing phenotypes are predicted (*in-silico* phenotypes) based on a second class of information (i.e. clinical data). The method is applied in phenotypes related to migraine. the use of in-silico phenotypes generation implies that two machine learning methods are combined (one for phenotype learning, the second one for disease prediction from the phenotypes), and transfer leaning complex issues should be taken into account. Our aim is to keep original data as much as possible, handling missing data in the machine learning technique itself.

Another interesting work is [11], which use self-organizer maps to look for associated diseases (kidney disease, retinopathy, hypertension). Self organized maps allows to obtain groups of biomakers than should next be interpreted by the clinicians. In our work, we are dealing with classification (i.e. prognosis), although [11] could be considered to extend the follow-up of diagnostic persons, in a hybrid methodology of [11] and ours.

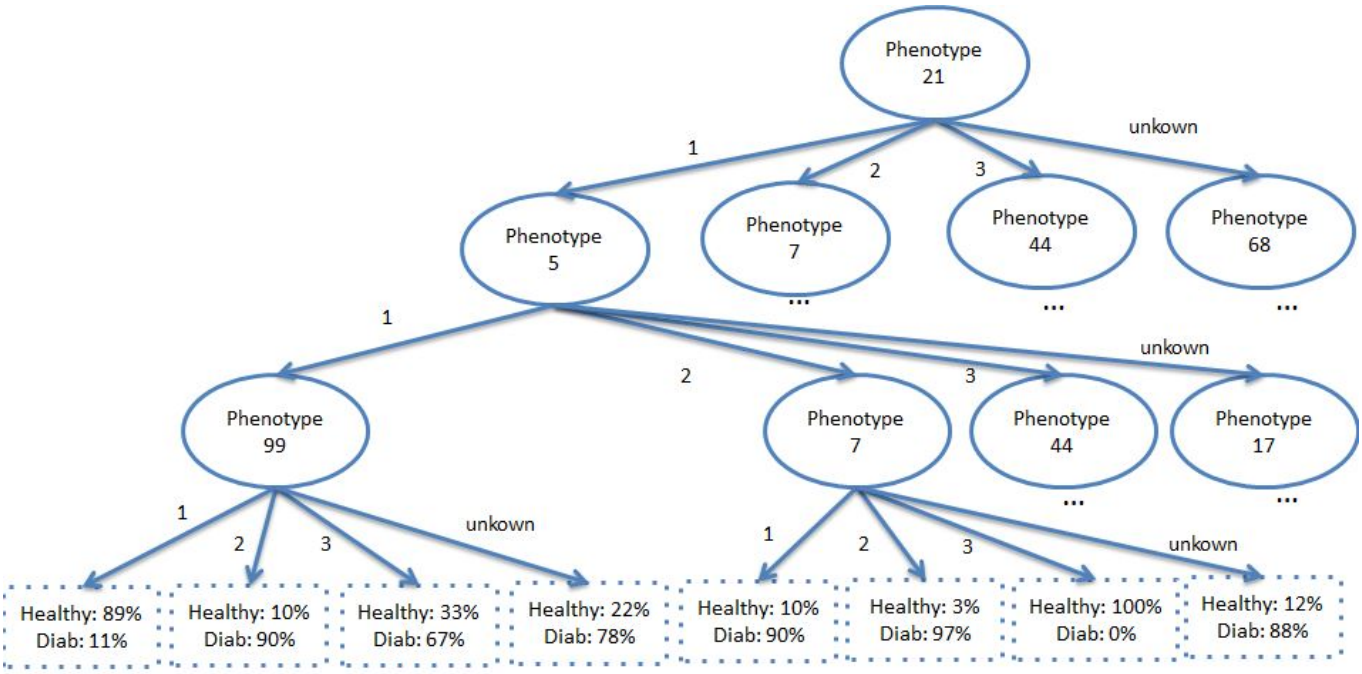


**Figure 1.** Example of Random Forests.

In [1] a comparison analysis among different imputation methods is performed, including instance deletion, mean imputation, median imputation, and k-nearest neighbour (knn) over a parametric and a non-parametric machine learning methods. The results highly depend on the characteristics of the data set, that is, the amount of missing features. Nevertheless, it seems that the case-deletion methods is the one that performs the worst, while the knn showed a higher robustness to missing data. The latter results agree with [2], where the authors analyse also several methods and demonstrate the out-performance of knn. The knn approach was analysed also in [15] as part of the reduced model approaches, and the results were slightly different, obtaining best performance with the authors approach called reduced-feature ensemble (RFE). RFE consists on generating several models, in which a feature is excluded in each of them. Given a query case, the outcomes of the different models are combined in a voting approach to obtain the final prediction value. This approach is also known as bagging ("bootstrap aggregating") [4]. However, bagging suffers from a higher correlation of the predictions [12]. The RF technique applied in our work decorrelate the base learners thanks to the random choice of features and samples.

## 3 METHODOLOGY

Our aim is to build a prediction model from phenotype data, which involves a considerable amount of missing values. The technique we are proposing is RF, because our hypothesis is that RF are able to handle missing information in a similar way than remove-feature and remove-instance missing information methods. However, RF does not discard any data a priori, which could provide nice properties regarding individualization (i.e. personalized prognosis).

RF is a supervised method, meaning that each instance or sample is labelled with the outcome (prognosis). Each instance is noted as $(x, y)$, where x is a list of attributes $a_1, a_2, \ldots, a_n$ and its values $v_1, v_2, \ldots, v_n$; and $y$ the class to which the patient belongs. In our particular case, $y \in C = \{healthy, diabetisType2\}$. Moreover, $a_i$ are the phenotypes, and we use $v_{ij}$ to denote the $j$ value of the $i$ phenotype. Each phenotype $i$ has $NVA_i$ values. In our particular case, $NVA_i = 4$ $(\forall i)$, 3 values, plus the *unknown* value. Therefore, we are considering phenotypes with missing information in our machine learning technique [5].

RF consists of an ensemble of $k$ classifiers $h_1(x), h_2(x), \ldots, h_k(x)$, being $h(x)$ the joint classifier [13, 5]. Each classifier $h_i(x)$ consists of a decision tree, in which nodes are attributes (see Figure 1). The selection of which attribute is collocated in a node is performed as follows: 1) by randomly selecting a subset of features, 2) an evaluation measure is applied to the selected attributes according to their capability to provide homogeneity partitions of the samples, and 3) the attribute with the highest score is chosen. In particular, we use the change of the Gini impurity function (GC) to compute the score, as described in Equation 1:

$$GC(a_i) = -\sum_{C_k \in C} p^2(C_k) + \sum_{j=1}^{NVA_i} p(v_{i,j}) \sum_{C_k \in C} p^2(C_k|v_{i,j}) \quad (1)$$

Once a node is set with an attribute $a_i$, the the data is split into as many sets as values the $a_i$ attribute has. Then, the tree is growth with new nodes in each branch that are obtained by repeating the attribute selection process. The stopping conditions is defined according to the number of instances remaining in a node: if this number is lower than a given threshold $\tau$, the algorithm stops. Samples used to build each tree are also selected randomly with replacement.

[5] In fact, this could be considered as a unique-value imputation method, as the unknown or missing value is treated as another attribute value.

Given a query case $q$, each decision trees provides an outcome, $h(q)$, and the final prediction is obtained by using a voting mechanism.

## 4 RESULTS AND DISCUSSION

In this section we describe our data, the experimental scenarios, and the results obtained.

### 4.1 Dataset description

The experimentation has been carried out with a dataset of 1074 patients, of whom we knew whether they had diabetes or they do not. For 196 patients, the diagnosis was unknown and therefore, have been removed from the dataset, remaining a total of 878 instances for experimentation. Each sample contains 101 phenotypes regarding diabetes type 2.

Regarding missing information, Figure 2 shows the distribution of missing data along the different samples. It is worthy to observe that some of the samples accumulates a huge percentage of missing information. On the other hand, Figure 3 shown the amount of missing values per phenotypes[6] (blue color). Phenotypes have been ordered in the x-axis according to their amount of missing values.

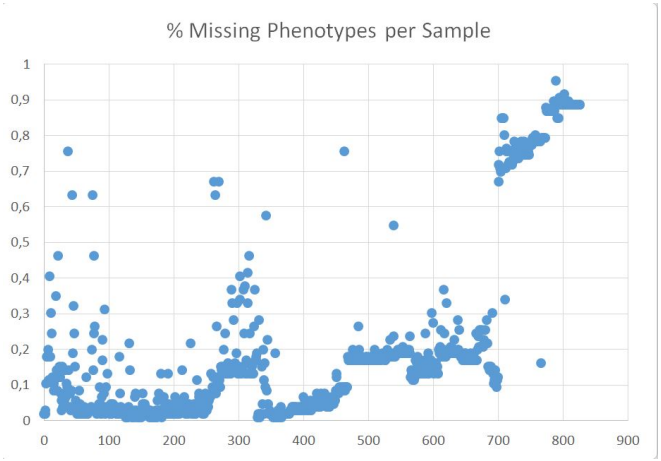


**Figure 2.** Percentage of missing phenotype values per sample. X-axis: cases.

### 4.2 Experimental set up

In order to analyse the implications of RF to handle missing data, the following experimental scenarios have been defined:

**Raw data** The dataset is used as provided.

**Reduced features** Features with the highest degree of missing information are removed. In particular, all features with more than 23% of missing values have been removed. This percentage has been set up according to the information visualized in Figure 3.

**Reduced samples** Samples with more than 25% of missing information has been removed. The percentage has been set up according to Figure 2.

[6] Phenotypes names are hidden for simplicity reasons and medical research confidentiality issues.

**Reduced features and samples** Both, the reduced features and samples criteria is applied to the dataset.

The number of decision tress has been set to k=1000. According to [5], as the number of trees increases, for almost surely the RF converges to the real predictor. The experimentation methodology used has been the stratified k-fold cross validation (we set 5 folds). Results are analysed in terms of accuracy.

### 4.3 Results

Table 1 shows the results obtained in the different scenarios. The highest accuracy is obtained when removing samples with a huge amount of missing values (in bold). On the other hand, it is interesting to observe that the results when removing features are very bad, even when the removed features contain a lot of missing values. This fact also impacts in the combination scenario. Therefore, RF is handling appropriately missing information. Internally, RF are building several trees in which the phenotypes with a high amount of missing features could be skipped, but the presence of all of the phenotypes are important for prognosis prediction. In that regard, individualization is keep in the model, favouring a personalized prognosis.

On the other hand, RF is not able to handle samples with a huge number of missing information (scenario raw data). Although internally samples are randomly selected for building the decision trees, RF require from some pre-processing that filter outs the data with a huge amount of missing information in order to provide good accuracy results. Therefore, a pre-processing step for performing such remove-instances method is still required.

**Table 1.** Accuracy results

| Scenario | Experiment | Accuracy |
|---|---|---|
| 1 | Raw data | 80.50% |
| 2 | Reduced features | 62.93% |
| 3 | Reduced samples | **86.91**% |
| 4 | Combine 2+3 | 62.67% |

## 5 CONCLUSION

The application of machine learning techniques to phenotype datasets for building models for disease prognosis need to deal with a huge amount of missing information. In this work we present an application of RF that shows how this technique could deal with missing information. Results show than RF can perform well with features with missing values. Keeping all phenotypes lead us to think that RF favours personalized prognosis, considering all the particularities of an individual. However, regarding samples, RF requires a minimum information in the samples to achieve good accuracy results.

As a future work, we need also to explore the combination of phenotype data with clinical information, as well as other environmental factors; diabetes type 2 is an heterogeneous disorder that require considering all these factors [10]. On the other hand, the use of RF causes a loss of the nice interpretation properties of a single decision tree. In that regard, the work of [6] could provide some insights.

## Acknowledgment

This project has received funding from the grant of the University of Girona 2016-2018 (MPCUdG2016) and the European Unions Hori-

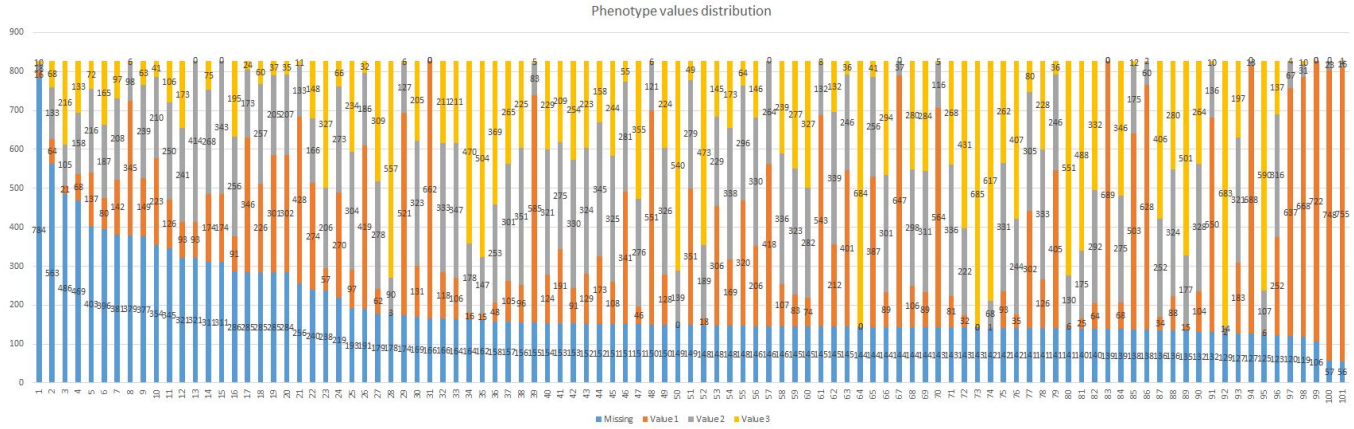


**Figure 3.** Distribution of phenotype values. Phenotypes are ordered according to the highest to lowest number of missing values (blue color).

zon 2020 research and innovation programme under grant agreement No 689810 (PEPPER). The work has been developed with the support of the research group SITES awarded with distinction by the Generalitat de Catalunya (SGR 2014-2016).

## REFERENCES

[1] Edgar Acuña and Caroline Rodriguez, 'The Treatment of Missing Values and its Effect on Classifier Accuracy', in *Classification, Clustering, and Data Mining Applications*, 639–647, Springer Berlin Heidelberg, Berlin, Heidelberg, (2004).

[2] Gustavo E A P A Batista and Maria Carolina Monard, 'An analysis of four missing data treatment methods for supervised learning', *Applied Artificial Intelligence*, **17**(5-6), 519–533, (2003).

[3] Riccardo Bellazzi and Blaz Zupan, 'Towards knowledge-based gene expression data mining.', *Journal of biomedical informatics*, **40**(6), 787–802, (dec 2007).

[4] Leo Breiman, 'Bagging Predictors', *Machine Learning*, **24**(2), 123–140, (1996).

[5] Leo Breiman, 'Random Forests', *Machine Learning*, **45**(1), 5–32, (2001).

[6] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch, 'BART: Bayesian additive regression trees', *The Annals of Applied Statistics*, **4**(1), 266–298, (mar 2010).

[7] Hüsamettin Gül, Yeim Aydin Son, and Cengizhan Açikel, 'Discovering missing heritability and early risk prediction for type 2 diabetes: a new perspective for genome-wide association study analysis with the Nurses' Health Study and the Health Professionals' Follow-Up Study.', *Turkish journal of medical sciences*, **44**(6), 946–54, (2014).

[8] C. N. Hales and D. J. P. Barker, 'Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis', *Diabetologia*, **35**(7), 595–601, (jul 1992).

[9] C. N. Hales and D. J P Barker, 'Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis', *International Journal of Epidemiology*, **42**(5), 1215–1222, (2013).

[10] H E Lebovitz, 'Type 2 diabetes: an overview.', *Clinical chemistry*, **45**(8 Pt 2), 1339–45, (aug 1999).

[11] Ville-Petteri Mäkinen, Carol Forsblom, Lena M Thorn, Johan Wadén, Daniel Gordin, Outi Heikkilä, Kustaa Hietala, Laura Kyllönen, Janne Kytö, Milla Rosengård-Bärlund, Markku Saraheimo, Nina Tolonen, Maija Parkkonen, Kimmo Kaski, Mika Ala-Korpela, Per-Henrik Groop, and FinnDiane Study Group, 'Metabolic phenotypes, vascular complications, and premature deaths in a population of 4,197 patients with type 1 diabetes.', *Diabetes*, **57**(9), 2480–7, (sep 2008).

[12] Kevin P. Murphy, *Machine Learning: A Probabilistic Perspective*, The MIT Press, 2012.

[13] Marko Robnik-Sikonja, 'Improving random forests', *Machine Learning: ECML 2004*, 12, (2004).

[14] Damian Roqueiro, Menno J Witteveen, Verneri Anttila, Gisela M Terwindt, Arn M J M van den Maagdenberg, and Karsten Borgwardt, 'In silico phenotyping via co-training for improved phenotype prediction from genotype.', *Bioinformatics (Oxford, England)*, **31**(12), i303–10, (jun 2015).

[15] Maytal Saar-Tsechansky and Foster Provost, 'Handling Missing Values when Applying Classification Models', *The Journal of Machine Learning Research*, **8**, 1623–1657, (2007).

[16] Hala Tfayli, Fida Bacha, Neslihan Gungor, and Silva Arslanian, 'Phenotypic type 2 diabetes in obese youth: insulin sensitivity and secretion in islet cell antibody-negative versus -positive patients.', *Diabetes*, **58**(3), 738–44, (mar 2009).

[17] Tiinamaija Tuomi, 'Type 1 and type 2 diabetes: what do they have in common?', *Diabetes*, **54 Suppl 2**(suppl 2), S40–5, (dec 2005).