UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

Analysis of the Error in RBF Networks with Missing Data

Project management (GEP)
Assignment 4: Final document

Xavier Martín Ballesteros

Bachelor Thesis
Specialization in Computing

Director: Luis Antonio Belanche Muñoz
GEP Tutor: Joan Sardà Ferrer

16th March, 2020

Contents

List of Figures

List of Tables

1 | Context and Scope

1.1 Context

This is a Bachelor Thesis of the Computer Engineering Degree, specialization in Computing, done in the Facultat d'Informàtica de Barcelona of the Universitat Politècnica de Catalunya directed by Luis Antonio Belanche Muñoz, doctorate in Computer Science.

1.1.1 Introduction

Most of the machine learning methods assume data completeness. Real-world data, however, is not usually perfect and contains missing data, which hinders obtaining good results. Learning from incomplete data has been recognized as one of the fundamental challenges in machine learning [6]. There is no general solution that performs the best in all problems that have to deal with missing data. Hence, the study for methods to deal with missing values is strongly justified.

Data missingness

Data missingness can be caused by numerous reasons. A new feature can be added during the research, therefore the collected instances so far will not have this feature value. Hardware equipment may not work as expected and does not collect some attributes in some instances. Merging databases can also lead to missing values in some rows. Responses are not collected in some cases (e.g. a person may decline to answer part of a questionnaire).

On the other hand, missingness can be directly related to the data. Thus, it is crucial to find the possible pattern(s) of missing values to determine the procedure to treat them. There are three different categories of missing data randomness [7]:

Let Y be a vector valued that is composed of two random variables Y_{obs} (variable observed in the data) and Y_{miss} (unobserved random variable). Parameter θ estimates Y. Let R be the pattern of missing responses, where $R_{ij} = 1$ if and only if Y_{ij} is missing (0 otherwise). R is conditioned by unknown parameters ψ.

- **Missing At Random (MAR):** The probability of an instance feature to be missing can depend on observed data but not on missing data.

$$P(R|Y_{obs}, Y_{mis}, \psi) = P(R|Y_{obs}, \psi) \tag{1.1}$$

- **Missing Completely At Random (MCAR):** Refers to data in which the probability of an instance feature to be missing depends neither on observed data, missing data nor itself.

$$P(R|Y_{obs}, Y_{mis}, \psi) = P(R|\psi) \tag{1.2}$$

- **Missing Not At Random (MNAR):** The probability of an instance feature to be missing is related to the values of the variable itself. The missingness is related to factors that are not measured by the researcher.

Missing data treatment

As Luengo, Garcıa, and Herrera [8] stated, "rates of missing data less than 1% are generally considered trivial, 1%-5% manageable. However, a rate of 5%-15% requires sophisticated methods to handle, and more than 15% may have a severe impact on any kind of interpretation and harm model's results". Consequently, there is a need for treating missing values. There are three ways to treat them:

- **Ignore missing values:** No method for imputation is used. In this case, all instances with at least one missing value (or more than a percentage) are deleted. They reduce the data set to a subset of complete data. However, this treatment of the data does not lead to good performances.
- **Parameter estimation:** This approach estimates the parameters of the selected distribution in the presence of missing data, for later impute the missing entries using the computed distribution.
- **Imputation methods:** Set of procedures that fill in the missing entries of the incomplete data sets based on the hidden patterns and relationships between attributes in the data set. Then, analyze the full data set as if the imputed values were actual observed values.
  - **Single imputation:** The missing data is filled in by a single value, computed only one time. However, it does not reflect the uncertainty about the prediction of the missing values. Single imputation works well when the amount of missing data is low. Nevertheless, with a large amount of missing values, it can cause a serious problem, as it treats the imputed value as an equal to the data that was not imputed.
  - **Multiple imputation:** Tries to solve the problem in single imputation methods. Instead of imputing the missing entry only one time, it repeats the process M times, obtaining M different complete data

sets. For each one, it is computed a parameter of interest that reflects the uncertainty due to missing values. Finally, the M results are combined into one getting a confidence interval of the variable imputed.

Radial Basis Function Networks

In this project, we will focus on the Radial Basis Function Networks (RBFNs) [3], a particular type of Artificial Neural Network (ANN). The difference with other types of ANNs is that they use radial basis functions as the activation function[1]. It has been used in a wide variety of fields including classification, interpolation, time-series analysis and image processing.

The architecture of the traditional RBFNs consists of three layers: one input layer, one hidden layer and one output layer. The input vector $\mathbf{x}$ is a n-dimensional array which is forwarded to each neuron in the hidden layer. Each neuron i in the hidden layer computes a RBF, typically the Gaussian[2]:

$$h_i(x) = e^{-\frac{(x-c_i)^2}{2\sigma_i^2}} \tag{1.3}$$

where u_i is the center of the neuron i and h_i, the output of that neuron. The outputs of the neurons are linearly combined with weights $\{w_i\}_{i=1}^m$ to produce the network output $f(x)$:

$$f(x) = \sum_{i=1}^{m} w_i h_i(x) \tag{1.4}$$

The network is trained as following. First, the centers are initialized using some unsupervised learning method such as the k-means clustering algorithm. The value of k is very important, as with too few centers the network will not make a good generalization (underfitting) and with too many centers the network will learn useless information from noisy data (overfitting)[3]. Once selected the centers, we have to compute σ_i (the width) in all neurons. One way to do it is by the mean distance between all points in the cluster i with respect to the cluster center c_i:

$$\sigma_i = \frac{1}{L} \sum_{k=1}^{L} d(x_k, c_i) \tag{1.5}$$

where L is the total number of points belonging to cluster i. Finally, we have to update the weights $\{w_i\}_{i=1}^m$ between the hidden and the output layer, which can be done using the gradient descent algorithm.

[1] A radial basis function (RBF) is a real-valued function in which their response decreases (or increases) monotonically with distance from a central point.

[2] Other types of RBF are multiquadric, inverse multiquadric and Cauchy functions.

[3] A good estimation of the k value comes from an empirical formula $k = 0.51 + \sqrt{0.43m_1m_2 + 0.12m_2^2 + 2.54m_1 + 0.77m_2 + 0.35}$, where m_1 is the number of input features (n) and m_2 is the number of output neurons [10].

RBFNs differ from Multilayer Perceptron (MLP) architecture in some aspects:

- RBFNs are faster to train than MLPs. However, classification will take less time in MLPs than in RBFNs.
- MLPs usually have more hidden layers.
- RBFNs need more neurons than in MLPs to obtain a similar accuracy.
- Interpretation of each node in the hidden layer is easier in RBFNs.

1.1.2 Problem to be resolved

The current problem is that we do not know which missing data treatment method performs better depending on the percentage of missingness in the data set nor the maximum error or the expected error that we can have when using them. Thus, in the data treatment process we have to experiment with multiple algorithms, resulting in a waste of time and resources. It would be better to know which method(s) to use depending on the characteristics of the data set we have (percentage of missing values, correlation between attributes...).

As stated above, it is crucial to find which missing data treatment methods lead to better results when we have to work with incomplete data sets. The aim of this project is to analyze the impact in the neurons of an RBFN due to missing data.

1.1.3 Stakeholders

The project has many involved parties, which can be grouped into two different groups depending on the interaction and benefits they have with the project itself.

The stakeholders that have direct interaction with the project are the **tutor** and the **researcher**. Luis Antonio Belanche Muñoz is the tutor of this project. Moreover, RBFNs is one of his areas of research and has wanted to analyze the error in these networks produced by the effect of missing values in the data set for a long time. Thus, he will lead and guide the researcher for the correct development of the project. The researcher, Xavier Martín Ballesteros, is responsible for planning, developing and documenting the project, as well as experimenting, analyzing and drawing conclusions.

Secondly, the stakeholders that do not interact with the project but receive direct benefits can be divided into two more groups: **companies**, that use machine learning methods (specifically, RBFNs) combined with data sets that have missing values to sell a product; and the **scientific community**, who get access to the study realized and can use the information and conclusions for further studies in this area.

1.2 Justification

1.2.1 Previous studies

In the past decades, there has been notably research in improving the performance of different imputation methods.

Dixon [5] found that ignoring missing values or imputing them by the 0 value lead to consistently poor performances, whereas other imputation methods (k-NN, normalize distances between two vectors and average of distances in the same attribute) were found to be generally good. Luengo, Garcıa, and Herrera [8] concluded that EventCovering method outperforms other methods almost all the time when working with RBFNs. Troyanskaya et al. [13] compared three different imputation methods (k-NN, mean imputation and singular-value decomposition imputation) for gene expression data.

Other researchers have focused on improving existing methods in different ways. Tutz and Ramzan [14] obtains better accuracy by improving versions of the nearest neighbor imputation method. They use a weighted nearest neighbor imputation method and combine it with the selection of the attributes depending on the correlation they have with the attribute of the missing value. Improvements in the computational cost of the methods have also been a research topic. A spanning-tree based algorithm was proposed in Delalleau, Courville, and Bengio [4] to obtain an efficient training of the mixture of Gaussians imputation method.

There has also been research in proposing more methods for processing missing data. Śmieja et al. [12] justify theoretically a mechanism that uses neural networks, replacing typical neuron's response in the first hidden layer by its expected value. Mesquita et al. [9] described a method to estimate the expected value of the Euclidean distance between two possibly incomplete feature vectors.

1.2.2 Justification

All the research mentioned above only present methods and compare the errors obtained with and without imputing values. However, any of them try to bound the error that can achieve the method nor give the expected error.

When dealing with missing data, it is very important to select which imputation method to use. This selection will directly affect the performance in the training part of the machine learning technique used. From our point of view, the information about the upper bound and the expected value of the error is needed to select which imputation method to use. It could be the case in which one imputation method usually performs better than another but has a greater upper bound for the error. With all this information, the company/researcher could decide which method to use depending on its specific requirements.

1.3 Scope

1.3.1 Objectives and sub-objectives

As presented in Section 1.1.2, the main objective in this project is to analyze the error in the neurons of the RBFN produced by the imputation of the missing values in incomplete data sets.

To accomplish this objective, the project has been subdivided in several sub-objectives:

Theoretical part

- Do research in RBFNs and in the methods that will be tackled in the project.
- For each method
  - Bound the error that can cause its use in the learning process of the machine learning algorithm. The lower bound will be 0, as it is the minimum possible error that we can get.
  - Give the expected value for the error.
  - Compute the space and time complexities.

Practical part

- Program the methods studied and the RBFN.
- For each method
  - Analyze its behavior for artificial data sets with missing data.
  - Analyze its behavior for real-world data sets with missing data.
- Compare the results obtained with the ones computed in the theoretical part.
- Draw conclusions about all the results obtained in the project.

1.3.2 Requirements

There are some requirements needed to ensure the quality of the final project.

- Errors must be bounded between 0 and 1. Otherwise, upper bound errors would be very large and it would be difficult to compare one method with the others.
- Select methods that belong to different ways for treating missing values (See 1.1.1).

- Select the optimal parameters of the methods using some model validation technique (e.g. cross-validation).
- Real-world data sets must have missing values. Otherwise, they would be treated the same as artificial data in the practical part.
- Optimize the code for all the imputation methods.
- Use good programming practices, with a readable style and least complexity possible.

1.3.3 Potential obstacles and risks

There can be some risks that prevent the correct functioning of the project. Besides, there can appear some obstacles during the execution of the project.

- **Deadline of the project.** There is a deadline for the delivery of the project that has to be accomplished. This forces to take drastic decisions during the development. Hence, there will have to be a good plan and meet the specified deadlines to be able to finish the project in time.
- **Impossibility of meeting with the tutor in person:** To be able to progress in the project, I have to do meetings with the tutor to see if the work done is correct, and to know which are the next steps to take. This task can be hampered due to a virus (Coronavirus) that can make the Spanish government decide to decree the state of the alarm. In such case, meetings would not be in person.
- **Bugs in some libraries.** Some libraries used could have bugs in some functions, which would make the code incorrect.
- **Inexperience in the programming language.** In case I choose to use a programming language I have never used (to be decided), I would have to spend time learning this new language, and the code may not be as clear as it should be.

1.4 Methodology and rigor

1.4.1 Methodology

The methodology that I will use for the project is the Kanban methodology, whose principal objective is to manage in a general way how the tasks are completed.

Kanban is a Japanese word that means "visual cards", where Kan means "visual" and Ban corresponds to "card". Hence, in this methodology it is used visual notes, where each one represent a task to do. The cards will be on a board with 4 different columns:

- **To do:** Composed by all tasks that have been defined but have not been started yet.
- **In progress:** Composed by all tasks that are being developed.
- **Testing:** Composed by all tasks that have been developed but still not tested for the correct functioning.
- **Completed:** Composed by all finished and tested tasks.

To control the work, we will use Asana, a web application designed to manage your work. It allows us to have virtual cards organized in several groups.

Note that as in some tasks we will not have to program nor test anything, some notes will skip the Testing step.

This methodology stands out for being very easy to use and update, as well as a very visual technique, which allows you to see really quick the status of all the tasks of the project.

1.4.2 Validation

We will use a GitHub repository as a tool for version control, which will allow us to facilitate the code availability (it is in the cloud) and the recovery from previous versions in case of critical failures. All code that has been developed and tested will be in the *master* branch. Code that is being developed will be placed in the *dev* branch. Finally, the *hotfix* branch will be used in case we have to solve a specific error in the code. In order to verify the implemented code, it will be passed through various tests to see whether the code works as expected or not.

In the practical part, the optimal parameters of the imputation methods will be selected using some model validation technique, such as the cross-validation. Each experiment will be done more than one time (number of times to be decided) and the average of the results will be the final result.

Lastly, face-to-face meetings will be scheduled once every two weeks with the tutor of the project. In these meetings it will be discussed the project status and the tasks to do during the following two weeks, before the next meeting. In case I have some problem in the project, extraordinary meetings will be arranged.

2 | Project Planning

This project lasts approximately 544 hours, distributed in 150 days starting from January 25th, 2020 until June 22nd, 2020. It has not been decided the date for the oral defense yet. Hence, the previous deadline is the earliest one we can have. It is planned to work 3,5 hours approximately every day even though exams can duly affect this time periodically.

2.1 Task definition

Following, it is presented all the tasks that will be carried out along the project. For each one, it is given a description, duration and dependencies with the other tasks. Table 2.1 summarizes all the information and Figure 2.1 illustrates the project schedule.

The project management is probably one of the most important group of tasks for the project. It defines the scope of it, the tasks and plans its distribution. Below are shown the multiple tasks for the project management.

- **ICT tools to support project and team management.** We need the latest technology, devices and concepts to support the development of a project of this kind. To do so, we have to research different types of software for different types of tasks (e.g. sharing documents and task planning).
- **Context and scope.** We have to indicate the general objective(s) of the project, contextualize it and justify the reason for selecting this subject area.
- **Project planning.** To achieve the project deadline we need a good planning for all the tasks. This will help us to know in which tasks we have to focus on more and which are the critical ones.
- **Budget and sustainability.** When doing a project it is very important to know what will be the total cost of it and the impact that will produce its development. Hence, this task focuses on making a budget and analyzing the sustainability of the project.

- **Final project definition**. We have to group the project done in the previous tasks, modifying the parts that were wrong.

- **Meetings.** Face-to-face meetings are scheduled once every two weeks with the tutor of the project. We will discuss the status and the following tasks to carry out. We have added extra time due to possible extraordinary meetings.

This project has a big part of research. Hence, before starting the experimental part it is mandatory to **do research** about previous studies to see the past and recent investigations in multiple missing data treatment methods. We will also have to document ourselves in the RBFNs area, as well as the algorithms and the statistical theory used in the studies.

In the theoretical part, we will focus specifically on the error generated in the neurons of the RBFN due to missing data in incomplete data sets, and what is the biggest and the expected effect it can have. This part is divided in three tasks.

- **Bound the error** between 0 and 1 (included) for every method applied for the missing values' treatment.

- **Compute the expected value for the error** for every method.

- **Compute the time and space complexities of the methods.** It involves the total understanding of each step for all the algorithms.

Before starting with the experimentation it is needed to program the network and methods to impute and estimate the missing data. Each function will have to be tested for the correct functioning. It has been divided also in three tasks.

- **Program a Radial Basis Function Network.** Even though there are some libraries which includes these type of Artificial Neural Networks, we will have to do several modifications in the code during the experimentation. Thus, it is better to program one RBFN from scratch.

- **Program the multiple missing data treatment methods for incomplete data sets.** As well as the previous task, we will have to modify parts of the code for experimental purposes, therefore it is better to program them also from scratch.

- **Test the correct functioning** of the programmed network and methods. This implies to create multiple test sets for each function to figure out if they work as expected even in the worst case scenarios.

The experimental and analysis part is the most important one, as this project needs proof that the results computed in the theoretical part agree with the results that will be obtained here. It has three different tasks.

- **Obtain the data sets** to experiment with. It can be divided in two subtasks. On the one hand, we have to create artificial data sets and delete some values to induce data missingness. On the other hand, we have to obtain multiple real-world data sets.
- **Experiment** with the programmed methods and the created and collected data sets. Some methods may have tuning parameters. In those cases, we will first have to select the optimal parameters using some model validation technique (e.g. cross-validation).
- **Analyze the results** obtained in the experiments and **draw conclusions**.

Once we have finished will all the previous tasks, we will have to document everything. Firstly, we will have to **collect all the information** obtained in the experimental and analysis part. Afterwards, we can start **writing the documentation** of the project.

Finally, we will have to **prepare for the oral defense** for the presentation of the project. To do so, we will think about possible questions that may come to mind to the senior FIB TFG tribunal members.

2.2 Resources

Every project needs resources to be able to organize it properly and carry out its correct development. These resources have been divided in 4 different groups: human, hardware, software and material resources.

2.2.1 Human resources

In this project we find three human resources. Firstly, the researcher is the responsible for the development of the project. He will have to plan, experiment, analyze and document the project. On the other hand, the tutor of the project is responsible for leading and guiding the researcher for the correct development of the project. Finally, the GEP tutor is in charge of helping the researcher to do the project management correctly during the first month of the project.

2.2.2 Hardware resources

One of the essential resources needed is a computer. In this project it will be used two different types of personal computers:

- ASUS desktop computer: 8GB of RAM and Intel(R) Core(TM) i5-4460 CPU @ 3.20GHz.
- Lenovo laptop computer: 8GB of RAM and Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz.

Moreover, we also have to take into account all the resources for the connection to the network (e.g. the router). Finally, we may also use a USB memory in case it is needed to pass information from one computer to another one.

2.2.3 Software resources

We need multiple software resources. Each one will help us in a specific part of the project. To be able to manage the meetings we will use Google Calendar and in case of not being able to meet the tutor in person we will use Skype or Google Hangouts Meet. All the information for the project will be saved in a GitHub repository. The Gantt chart will be created using Ganttproject. We will need some programming language(s) to code the RBFN and the missing data treatment methods. Besides, we will use overleaf or TeXmaker as our text editor. We will use one or the other depending on the size of the document we are writing.

2.2.4 Material resources

In research projects there is always the need to get knowledge from previous studies and the area in question. In order to obtain this knowledge we will have to read books and papers.

2.3 Risk management: alternative plans

During the course of the project there may appear obstacles that hinder its progress. The potential risks and obstacles have been introduced in Section 1.3.3. In this section, we will present how can they be solved by introducing new tasks and rearranging the planning. Moreover, we say for each one which is the level of risk it has.

- **Deadline of the project [High risk].** It can be caused by a bad estimation of the tasks and its duration. As we are doing the planning before having started anything, it is very probable that it will happen. This can be solved by doing a second planning in a more advanced point of the project. We would need to reuse the PC, Ganttproject software, the calculator, the GEP tutor and the researcher. In case we are running out of time and a new planning does not help, we can still solve this problem by increasing the number of working hours per day.
- **Impossibility of meeting with the tutor in person [Extreme risk].** In such case, the tutor and me would have to find alternative ways to communicate with each other. For the basic communication, we would continue using the email. For the meetings, we would have to find some software (e.g. Skype and Google Hangouts Meet) to be able to accomplish the meeting task. The task to find a new way for the meeting is estimated to last between 2 and 3 hours, because we first have to communicate through the email and set the new meeting rules. This new software, fortunately, is free to use.
- **Bugs in some libraries [Medium risk].** In these type of projects it may arise technical difficulties we were not prepared for. It is very

common to use third-party libraries when programming. Some functions from a library may have a bad functioning due to some mysterious bug. One possible solution for this problem is to wait until there is an update of the library that solves the bug. However, we cannot do this due to the strong deadline we have. Consequently, we would have to program that function from scratch and test its correct functioning, which would increase our total duration of the project. The new task would have the PC, the programming language and the researcher as resources.

- **Inexperience in the programming language [Medium risk].** In case we decide to use a programming language the researcher has never used, we will have to create a new task of a duration of 25-30 hours that will go before the programming part. Thus, we will create a new dependency: the tasks in the programming part cannot begin before the end of the "learning the programming language" task. The new task would have the PC, the programming language and the researcher as resources.

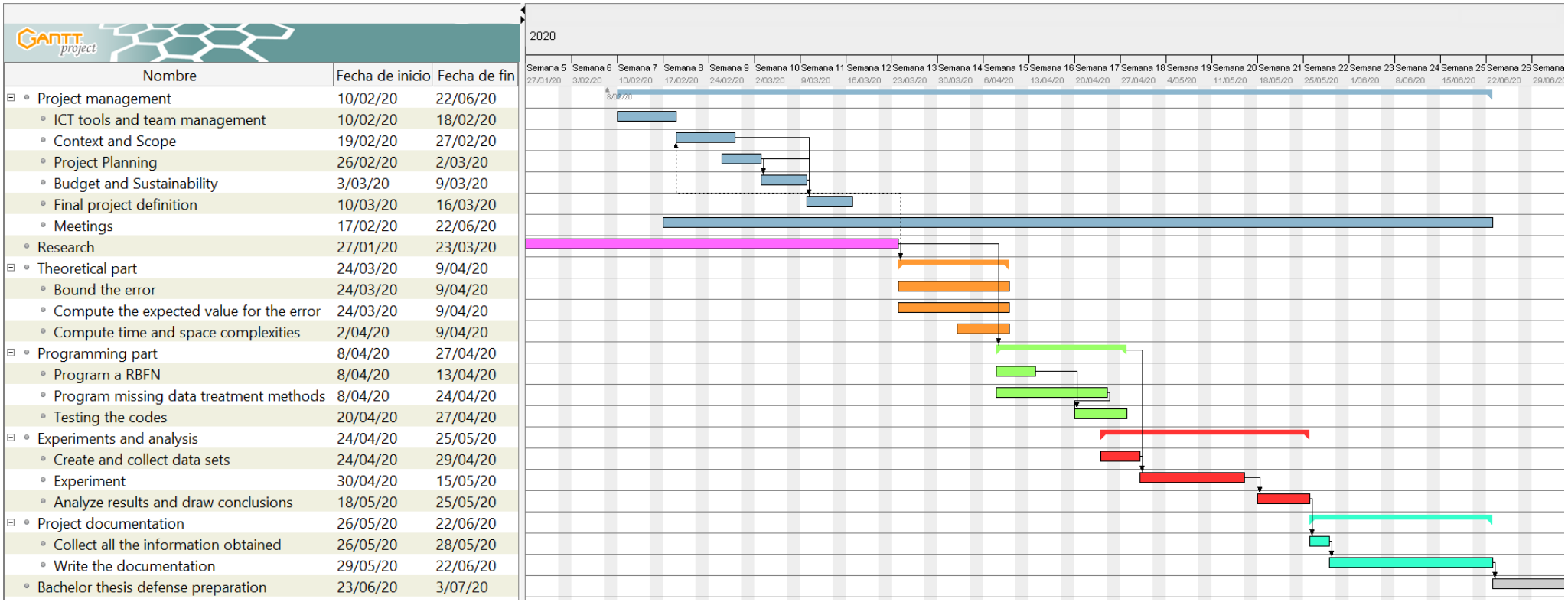
Due to all these risks and obstacles we have overestimated the time for the Meeting task. In case some obstacle appears, we can have an extraordinary meeting to solve it.

ID	Name	Time(h)	Dependencies	Resources
T1	Project management	**94**		
T1.1	ICT tools and team management	4		PC, R
T1.2	Context and Scope	25	T2*	PC, overleaf, GEPT, R
T1.3	Project Planning	10		PC, overleaf, Ganttproject, calculator, GEPT, R
T1.4	Budget and Sustainability	15	T1.3	PC, overleaf, calculator, GEPT, R
T1.5	Final project definition	20	T1.2, T1.3, T1.4	PC, overleaf, GEPT, R
T1.6	Meetings	20		T, R
T2	Research	**85**		PC, papers, books, R
T3	Theoretical part	**105**		
T3.1	Bound the error	40	T2	PC, papers, books, R
T3.2	Compute the expected value for the error	40	T2	PC, papers, books, R
T3.3	Compute time and space complexities	25	T2	PC, papers, books, R
T4	Programming part	**50**		
T4.1	Program a RBFN	10	T2	PC, programming language, papers, books, git, R
T4.2	Program missing data treatment methods	25	T2	PC, programming language, papers, books, git, R
T4.3	Testing	15	T4.1, T4.2	PC, programming language, papers, books, git, R
T5	Experiments and analysis	**110**		
T5.1	Create and collect data sets	10		PC, R
T5.2	Experiment	60	T4, T5.1	PC, data sets created/collected, R
T5.3	Analyze results and draw conclusions	40	T5.2	PC, results obtained, R
T6	Project documentation	**80**		
T6.1	Collect all the information obtained	10	T5.3	PC, results obtained, R
T6.2	Write the documentation	70	T6.1	PC, TeXmaker[1], project resources, R
T7	Bachelor thesis defense preparation	**20**	T6.2	PC, project resources, results obtained, R
Total		**544**		

Table 2.1: Summary of the information of the tasks. [Own creation]
*: The task that creates the dependency does not have to end to be able to begin with the task.
GEPT: GEP tutor.
T: Tutor.
R: Researcher.

[1]When the document is very large, it is recommended to use some offline text editor rather than overleaf, as it will compile much faster.

2.4 Gantt chart



Nombre	Fecha de inicio	Fecha de fin
Project management	10/02/20	22/06/20
ICT tools and team management	10/02/20	18/02/20
Context and Scope	19/02/20	27/02/20
Project Planning	26/02/20	2/03/20
Budget and Sustainability	3/03/20	9/03/20
Final project definition	10/03/20	16/03/20
Meetings	17/02/20	22/06/20
Research	27/01/20	23/03/20
Theoretical part	24/03/20	9/04/20
Bound the error	24/03/20	9/04/20
Compute the expected value for the error	24/03/20	9/04/20
Compute time and space complexities	2/04/20	9/04/20
Programming part	8/04/20	27/04/20
Program a RBFN	8/04/20	13/04/20
Program missing data treatment methods	8/04/20	24/04/20
Testing the codes	20/04/20	27/04/20
Experiments and analysis	24/04/20	25/05/20
Create and collect data sets	24/04/20	29/04/20
Experiment	30/04/20	15/05/20
Analyze results and draw conclusions	18/05/20	25/05/20
Project documentation	26/05/20	22/06/20
Collect all the information obtained	26/05/20	28/05/20
Write the documentation	29/05/20	22/06/20
Bachelor thesis defense preparation	23/06/20	3/07/20

Figure 2.1: Gantt chart. [Own creation]

3 | Budget and Sustainability

It will be described the elements to consider when doing the budget estimation, which includes personnel costs per task, generic costs and other costs. Moreover, it will be defined management control mechanisms to control the deviations that can appear in the project due to unforeseen obstacles. Finally, it will be answered some questions regarding the sustainability aspect in the project.

3.1 Budget

3.1.1 Personnel costs per activity

In this section it is computed the total cost for each task defined in Section 2.4. The cost for one task will be calculated summing the cost of the work of the personnel. The cost for each worker will be computed multiplying his cost per hour by the amount of time that they will be involved in the activity.

In this project there are 5 types of personnel, each one with a different cost per hour. Firstly, the **project manager** is responsible for the planning and correct development of the project. The GEP tutor, the tutor and me will play this role. Secondly, the **programmer** and the **tester** have to program the code and verify its correct functioning. This two roles will be played only by me. On the other hand, the **researcher** has to experiment, analyze the results and draw conclusions. I will play this role. Finally, there is the **technical writer** who has to document everything that involves the development and results of the project. This role will also be played by me. Following, it is shown the annual salary of the different project roles.

Now, we can compute the total cost for each task. Table 3.3 shows the distribution of time for the personnel for each task, and its total cost. This is known as CPA.

3.1.2 Generic costs

Amortization

One aspect to take into account is the amortization of the material resources used in the project. It is considered an average of 3,5 working hours per day

Role	Annual Salary (€)	Total including SS (€)	Price per hour (€)	Role played by
Project manager	39.004	50.705,2	28,97	GEPT, T, R
Programmer	26.198	34.057,4	19,46	R
Tester	20.592	26.769,6	15,29	R
Researcher	35.259	45.836,7	26,19	R
Technical writer	26.263	34.141,9	19,50	R

Table 3.1: Annual salary of the different project roles. It has been estimated the total number of hours worked per year, which is 1750. Information obtained from [1].
GEPT: GEP Tutor.
T: Tutor.
R: Researcher.

during a total of 150 days. It is estimated that an 80% of the project has been carried out using the desktop computer whereas the other 20% has been done using the laptop computer. The equation to compute the amortization for each resource is the following:

$$\text{Amortization (€)} = \text{Resource price} \times \frac{1}{4 \text{ years}} \times \frac{1}{150 \text{ days}} \times \frac{1}{3.5 \text{ hours}} \times \text{hours used} \tag{3.1}$$

The amortizations in this project are only of hardware because all the software used is free to use. The resources will be used a total of 544 hours. The amortization costs are shown below.

Hardware	Price (€)	Time used (h)	Amortization (€)
ASUS desktop computer	1.200	435,2	248,68
Lenovo laptop computer	858,99	108,8	44,50
Total			**293,18**

Table 3.2: Amortization costs for the hardware resources used in the project. [Own calculations]

ID	Name	Total hours	Hours					Cost (€)
			Project manager	Programmer	Tester	Researcher	Technical writer	
T1	Project management	**94**	**94**	**20**	**20**	**20**	**20**	**4.331,98**
T1.1	ICT tools and team management	4	4	0	0	0	0	115,88
T1.2	Context and Scope	25	25	0	0	0	0	724,25
T1.3	Project Planning	10	10	0	0	0	0	289,7
T1.4	Budget and Sustainability	15	15	0	0	0	0	434,55
T1.5	Final project definition	20	20	0	0	0	0	579,4
T1.6	Meetings	20	20	20	20	20	20	2.188,2
T2	Research	**85**	**0**	**0**	**0**	**85**	**0**	**2.226,15**
T3	Theoretical part	**105**	**0**	**0**	**0**	**105**	**0**	**2.749,95**
T3.1	Bound the error	40	0	0	0	40	0	1.047,6
T3.2	Compute the expected value for the error	40	0	0	0	40	0	1.047,6
T3.3	Compute time and space complexities	25	0	0	0	25	0	654,75
T4	Programming part	**50**	**0**	**35**	**15**	**0**	**0**	**910,45**
T4.1	Program a RBFN	10	0	10	0	0	0	194,6
T4.2	Program missing data treatment methods	25	0	25	0	0	0	486,5
T4.3	Testing	15	0	0	15	0	0	229,35
T5	Experiments and analysis	**110**	**0**	**0**	**0**	**110**	**0**	**2.880,9**
T5.1	Create and collect data sets	10	0	0	0	10	0	261,9
T5.2	Experiment	60	0	0	0	60	0	1.571,4
T5.3	Analyze results and draw conclusions	40	0	0	0	40	0	1.047,6
T6	Project documentation	**80**	**0**	**0**	**0**	**0**	**80**	**1.560**
T6.1	Collect all the information obtained	10	0	0	0	0	10	195
T6.2	Write the documentation	70	0	0	0	0	70	1.365
T7	Bachelor thesis defense preparation	**20**	**20**	**0**	**0**	**0**	**0**	**579,4**
Total		**544**	**114**	**55**	**35**	**320**	**100**	**15.238,83**

Table 3.3: Personnel cost for each task defined. [Own calculations]

Electric cost

Regarding the electric cost, the actual cost of the kWh is 0,1198€[2]. We only count the expenses of the hardware when they are turned on. I should remark that for the desktop computer, we have to add also the monitor expenses. Table 3.4 shows the individual and total costs of energy consumption.

Hardware	**Power (W)**	**Time used (h)**	**Consumption (kWh)**	**Cost (€)**
ASUS desktop computer	280	435,2	121,856	14,6
Lenovo laptop computer	65	108,8	7,072	0,85
Total				**15,45**

Table 3.4: Electric cost of the hardware resources used in the project. [Own calculations]

Internet cost

The internet rate costs 100€ per month. Taking into account that the project lasts 5 months and that the working hours per day are 3,5 the internet cost is 5 months * (100€/month) * (3,5h/24h) = 72,92€.

Water cost

The water cost in my zone area costs around 32,5€ per person. Using the equation in the previous section and using the same number of working hours per day, the water cost is the following: 5 months * (32,5€/month) * (3,5h/24h) = 23,7€.

Travel cost

I have to use the public transport to meet with the tutor once every two weeks. The expected number of travels is twice the number of meetings, therefore I have to do 40 travels during the project. To do so, I have to use the T-Casual[1] of only 1 zone, which costs 10 journeys/11,35€. Consequently, the total cost due to travel is (11,35€/10 journeys) * 40 journeys = 454€.

Work space

This project will mainly be developed in my house, located in Badalona. The rental cost of this house is 1.200€/month. I will not use all the space, thus I find convenient divide the rent by 3, as it could be two more students living in here. Hence, the cost of the space per month is 400€ and the project duration is 5 months, therefore the total cost of the work space is 2.000€.

[1] In case the meetings were more often, I could have used the T-mes (monthly ticket) and less T-Casuals, which would have been less expensive.

Generic cost of the project

Table 3.5 summarizes all the generic costs of the project introduced in the previous sections. The total cost is computed summing the CPA cost with the CG cost (generic cost).

Concept	**Cost (€)**
Amortization	293,18
Electric cost	15,45
Internet cost	72,92
Water cost	23,7
Travel cost	454
Work space	2.000
CG cost	**2.859,25**

Table 3.5: Generic cost of the project. [Own calculations]

3.1.3 Other costs

Contingencies

During the development of the project it can appear unforeseen events, which take part of our budget. For this reason, it is always necessary to prepare a fund of contingency to be prepared to face these events. To the total cost (CPA + CG = 18.098,08), we have to add a 15% of contingency margin. With this, the computed contingency cost is 2.714,71.

Incidental costs

We have also to take into account the cost of applying alternative plans in case unexpected events occur during the course of the project. The alternative plans can be found in Section 2.3. Table 3.6 show the total cost to solve these events. The cost for each incident is computed multiplying the price it would cost by the risk probability that the event occurs.

Incident	**Estimated cost (€)**	**Risk (%)**	**Cost (€)**
Deadline of the project (20 hours)	547,05	30	164,12
Impossibility of meeting with the tutor in person (2 - 3 hours)	0	100	0
Bugs in some libraries (5 hours)	117,3	15	17,6
Inexperience in the programming language (25 - 30 hours)	583,8	15	87,57
Total			**269,29**

Table 3.6: Incidental costs of the project. [Own calculations]

3.1.4 Total cost

The total cost expected for the project is found in Table 3.7, computed using all the justified costs calculated in the previous sections.

Activity	**Cost (€)**
CPA cost	15.238,83
CG cost	2.859,25
Contingency	2.714,71
Incidental cost	269,29
Total	**21.082,08**

Table 3.7: Total cost of the project. [Own calculations]

3.1.5 Management control

In big projects, it is very probable that the budget and time estimations will not be 100% fulfilled due to obstacles (expected and unforeseen). Consequently, we need to define a model for controlling the potential budget deviations.

Every time we finish a task, we have to compute the deviation of all the involved costs in it (CPA, CG, contingency and incidents). Following are listed the different indicator formulas for the different deviations we can have:

- **Human resources deviation:** It is caused when the personnel do less or more hours than the expected. We compute this deviation as shown below.

$$\text{Human resources deviation} = \sum_{i \in pit} (estimated\_cost\_per\_hour_i - real\_cost\_per\_hour_i) \times total\_real\_hours_i \tag{3.2}$$

  where pit refers to all personnel involved in that task.

- **Amortization deviation:** In case we use the hardware resource less or more time than the expected the amortization cost will vary.

$$\text{Amortization Deviation} = \sum_{i \in hr} (estimated\_usage\_hours_i - real\_usage\_hours_i) \times price\_per\_hour_i \tag{3.3}$$

  where hr refers to all hardware resources.

- **Travel cost deviation:** Probably the number of meetings will vary from the expected due to unexpected obstacles in the project.

$$\text{Travel Costs Deviation} = (estimated\_number\_journeys - real\_number\_journeys) \times 11,35 \tag{3.4}$$

- **Total cost deviation**: Groups all deviations in the different tasks. It does not take into account contingencies nor incidents.

$$\text{Total Costs Deviation} = estimated\_general\_costs - real\_general\_costs \tag{3.5}$$

Doing this, we can visualize and comprehend easily where and why has been a deviation and how much is the deviation cost. In case the total cost deviation is negative, we will have to use the budget reserved for contingencies.

Finally, we will update a list of incidental costs in case any unforeseen event occurs. In case any of these events happen, we will have to use the budget part reserved for incidents.

3.2 Sustainability

3.2.1 Self-assessment

Every time I heard the word sustainability or sustainable I was thinking of something (e.g. a product) which helps the development of the society trying not to harm the environment nor the society and that supports long-term ecological balance. I thought there was only two dimensions involved: the environmental and the social. However, I realized that there is also an economic dimension.

The poll has made me think about the causes, consequences and solutions about social, economic and environmental problems that exist today. Moreover, it has made me to analyze the relation of these problems with the Information and Communications Technology (ICT).

On the other hand, I must remark that I have been surprised with the quantity of indicators that there are to evaluate the different dimensions. Besides, it is very important to measure the different impacts in each dimension using these indicators before starting the project. Doing this, we are able to detect where do we have problems regarding the sustainability and how can we solve them using the tools we have.

To sum up, after doing the poll and thinking about the sustainability aspect in a project, I have come to the conclusion that I did not have much idea about it. I was only looking "on the surface" of the topic but I had never deepened in it. Finally, it has made me think about the importance to take into account the sustainability when planning and developing a project from the point of view of the three different dimensions.

3.2.2 Economic dimension

Regarding PPP: Reflection on the cost you have estimated for the completion of the project

The reflection on the estimated cost for the project, as well as the management control, can be found in Section 3.1 of the document. It has taken into account the human, hardware, software and material resources. The salary for the

personnel has been obtained searching through the internet [1]. Moreover, we have also measured other costs (contingencies and incidental costs) and their effects in the budget.

From my point of view, the budget is reasonable and could be carried out in a real life project of this area. However, we could decrease the work space by developing the project in a coworking space.

Regarding Useful Life: How are currently solved economic issues (costs...) related to the problem that you want to address (state of the art)?

Nowadays, the economic issue related to the problem is very high, because there is no research that says which imputation model is better than the rest, or when should we apply each method with real data sets depending on the percentage of missing values. Consequently, the studies have to program and test the different algorithms and experiment with each one, which costs a lot.

One way to reduce the economic cost in our project would be by reusing some software resource such as the code that have been previously programmed. Doing this, we could reduce the total time for the programming and testing tasks (T4) which would reduce the total economic cost for the project.

How will your solution improve economic issues (costs ...) with respect other existing solutions?

As explained in the previous question, in the recent studies and in real life cases, they have to experiment with the different methods to see which one(s) works better for their own data sets. This has a very high economic cost. In case we succeed in this project, we should be able to explain and justify which method should be used depending on the multiple cases. This would decrease a lot the economic cost, as they would no longer need to experiment with all the methods and choose the one that gives the best results.

3.2.3 Environmental dimension

Regarding PPP: Have you estimated the environmental impact of the project?

I have not estimated the environmental impact of the project. Nevertheless, this project does not waste material but rather has a high electricity consumption for the experiments. Each experiment will be done several times and the average of the results will be the final result. Hence, there will be a high waste of electricity.

Regarding PPP: Did you plan to minimize its impact, for example, by reusing resources?

As said before, the only resource that could be reused is the programmed code by other researchers. With this, we would not need to develop the program and testing tasks (T4) and we could decrease the total electricity consumption.

Note that we cannot use results obtained in other researches because we have to experiment with our own methodology and data sets. Thus, we cannot reuse resources regarding the experimental part.

Regarding Useful Life: How is currently solved the problem that you want to address (state of the art)?

As there is no way to know which method works the best for a particular data set with a specific percentage of missing values, people have to experiment with all the algorithms and then choose the best one.

How will your solution improve the environment with respect other existing solutions?

Our solution will advise which method should be used depending on the percentage of missingness in the data set. Thanks to that, the electricity consumption can be decreased, as they should no longer experiment with all the existing algorithms.

3.2.4 Social dimension

Regarding PPP: What do you think you will achieve -in terms of personal growth- from doing this project?

First, this project will help me to introduce in the research world and learn which are the steps needed to do for research of this kind. Secondly, be the principal responsible for a big project will help me to organize myself better and plan the projects correctly. Finally, it will also help me to measure the sustainability aspects in a project.

Regarding Useful Life: How is currently solved the problem that you want to address (state of the art)?

Currently, the problem is being solved by comparing the performances of the different missing data treatment methods depending on the data set, and selecting which outperforms the rest.

How will your solution improve the quality of life (social dimension) with respect other existing solutions?

The aim of this project is to reduce the time that people need to compare the performances of the methods by advising which ones should give the best results depending on the percentage of missing data in the data set. Hence, people would have more time to focus on other important tasks rather than choosing the best method.

Regarding Useful Life: Is there a real need for the project?

There is a real need for the project. Probably, it will help more people that do not have high power computers that can only execute few algorithms in the same computer. People who have high power computers can execute multiple methods at the same time and therefore can evaluate the results very quickly.

On the other hand, however, those who do not have high power computers will be able to select which method(s) is the best for the specific case they have and, if they want, compare the result with few other algorithms.

Nevertheless, the results and conclusions will serve to guide people to which methods to choose/experiment. In this way, they will be able to turn away some algorithms and will decrease the power consumption, the economic cost and the total time spent on it.

Bibliography

[1] `https://glassdoor.com`. [Online; Accessed March 5th, 2020].

[2] `https://tarifasgasluz.com/comercializadoras/endesa/precio-kwh`. [Online; Accessed March 6th, 2020].

[3] David S Broomhead and David Lowe. *Radial basis functions, multi-variable functional interpolation and adaptive networks*. Tech. rep. Royal Signals and Radar Establishment Malvern (United Kingdom), 1988.

[4] Olivier Delalleau, Aaron Courville, and Yoshua Bengio. "Efficient EM training of Gaussian mixtures with missing data". In: *arXiv preprint arXiv:1209.0521* (2012).

[5] John K Dixon. "Pattern recognition with partly missing data". In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.10 (1979), pp. 617–621.

[6] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[7] JA Little Roderick and B Rubin Donald. "Statistical analysis with missing data". In: *Hoboken, NJ: Wiley* (1987).

[8] Julián Luengo, Salvador Garcıa, and Francisco Herrera. "A study on the use of imputation methods for experimentation with Radial Basis Function Network classifiers handling missing attribute values: The good synergy between RBFNs and EventCovering method". In: *Neural Networks* 23.3 (2010), pp. 406–418.

[9] Diego PP Mesquita et al. "Euclidean distance estimation in incomplete datasets". In: *Neurocomputing* 248 (2017), pp. 11–18.

[10] GAO Da-Qi. "On structures of supervised linear basis function feedforward three-layered neural networks [J]". In: *Chinese Journal of Computers* 1 (1998).

[11] Friedhelm Schwenker, Hans A Kestler, and Günther Palm. "Three learning phases for radial-basis-function networks". In: *Neural networks* 14.4-5 (2001), pp. 439–458.

[12] Marek Śmieja et al. "Processing of missing data by neural networks". In: *Advances in Neural Information Processing Systems*. 2018, pp. 2719–2729.

[13] Olga Troyanskaya et al. “Missing value estimation methods for DNA microarrays”. In: *Bioinformatics* 17.6 (2001), pp. 520–525.

[14] Gerhard Tutz and Shahla Ramzan. “Improved methods for the imputation of missing data by nearest neighbor methods”. In: *Computational Statistics & Data Analysis* 90 (2015), pp. 84–99.