ISSN 1870-4069

Predictive Model as a Tool for Acquiring a Certification for Client Companies and Certifying Entities with Machine Learning

Edgar Gonzalo Cossio Franco[1], Jorge Alberto Delgado Cazarez[2], Daniel Noel Torres Godoy[3]

[1] Instituto de Información Estadística y Geográfica de Jalisco, Mexico
[2] Universidad de Guadalajara, Mexico
[3] Universidad Enrique Díaz de León, Mexico
kofrran@gmail.com, guero10delgado@gmail.com, dt_godoy@hotmail.com

Abstract. Companies that seek to certify their processes do so with the aim of guaranteeing quality, although few seek it and least of all do so. CMMI is a model that certifies the maturity of the development of products and services. The present work has two proposals: the first is a tool in Java that determines when a client company is apt or not to a CMMI certification and the second is an intelligent analysis model based on machine learning that determines, from predictions, scenarios for decision making.

Keywords: CMMI, machine learning, predictive model, java.

1 Introduction

The quality of a company can be measured by the maturity of its processes [1]; In the case of software development companies, certification of capability maturity model integration (CMMI) is sought in order to guarantee, in addition to quality, productivity, customer satisfaction, performance and business model [2].

In Mexico, for the year 2016, there was a registry of 781 certified development centers with some quality model, as shown in Figure 1, 62% have MOPROSOFT certification while 35% have CMMI certification and 3% PACE.

In Figure 2 they have the certifications that have the development results that have 2 and 3, and those that have CMMI, that have more than one constellation. Thus, a representative of the Development Centers that have verification and certification in two quality models with 31, b Represents Development Centers that have verification, certification and evaluation in three quality models with 2 and c Represents Development Centers that have certification in more than one CMMI constellation with 11.

The importance of companies starting to see in certifications a door of opportunity to ensure maturity in construction processes that translate into quality is increasing as without certification they will be building software of poor quality.

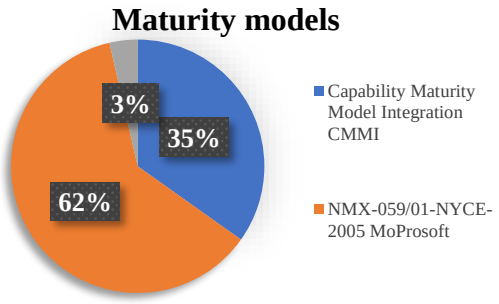

Fig. 1. Maturity models [3].

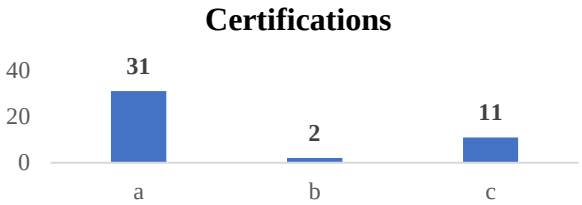

Fig. 2. Certifications center [3].

A. CMMI

The Capability Maturity Model Integrated (CMMI) is a reference framework for improving the processes of developing products and services of companies. It consists of five levels, three constellations and nineteen process areas [4]. Figure 3 shows CMMI maturity levels.

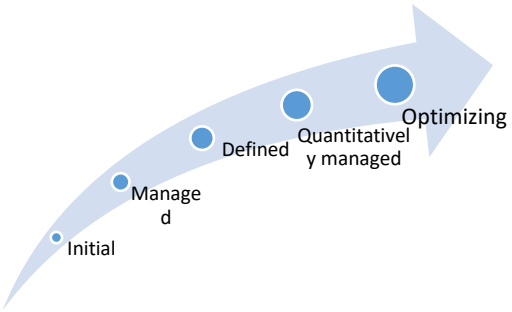

Fig. 3. CMMI Process [5].

The initial level shows a lack of control and chaos in the process; there are no maturity, control or documentation mechanisms. Success depends on superhuman effort.

In level two, the improvement actions are after something happened. No improvement scenarios are anticipated. There is knowledge of project management. At level three there are defined processes and there are metrics.

At level four, the collected data is used to manage and improve processes.

At level five, the processes continuously improve in a natural, sustained and constant manner.

B. Machine Learning

Machine learning is defined as the process by which computers learn automatically [6]. It is based on algorithms such as Bayesian classifiers for probabilities, in nearest neighbor classifiers for similarities, in artificial neural networks, in decision trees, in genetic algorithms, clustering, fuzzy logic and bioinspired algorithms to build global and local searches [7,8].

In 1943, Warren McCullock and Walter Pitts, in an attempt to explain the biological brain and apply it to the design of artificial intelligence, introduced the first concept of a simplified brain cell [9].

Later, in 1957, "Frank Rosenbelt proposed an algorithm that would automatically learn the optimal weight coefficients that are then multiplied with the input features in order to make the decision of whether a neuron fires or not" [9].

The classification is a supervised task belonging to Machine Learning; this identification the properties of a batch of data and, taking into account parameters or labels, previously configured, assign this data to one of the labels; this refers to a supervised action.

One of the algorithms dedicated to the classification is j48; its function is based on generating a decision tree with the variables it has, through partitions made recursively according to the search strategy first in depth [10]. The grouping is an unsupervised task of Machine Learning, which, automatically, divides the data into groups to help us obtain the groups [11]. This refers to an unsupervised task.

The k-means algorithm achieved the grouping principle in a fast and efficient way. Make a random selection of observations, such as clustered groups, then, later, assigned to the nearest point. This division of patched space is known as a Voronoi tasseling [12].

Some of the strengths of the k-means algorithm:

- It uses simple principles to identify clusters that can be expanded in non-statistical terms.
- It is highly flexible and can be adapted to address almost all its deficiencies with simple adjustments.
- It is quite efficient and performs well by dividing the data into useful groups. [11]. Unlike classification algorithms, association rules are used for the discovery of unsupervised knowledge in large databases. In the case of association rules, it is not necessary for the data to be tagged ahead of time, the program simply triggers a set of data, which are expected to find interesting.
- [11] The analysis of association rules is used to find a good connection between a large number of variables. The most common example is the analysis of the market basket, however, it can also be used for topics such as:
    - Search for interesting and frequent patterns in DNA and protein sequences in a cancer data analysis.
    - Find patterns of purchases or medical claims that occur in combination with fraudulent credit cards or insurance use.

- Identify combinations of behavior that cause customers to abandon their cell phone service or update their cable television package.

According to Lantz Brett, you can say that this type of task, if it can be done by a person, you need to have a person who has an expert level in the area in which you work and that, based on your great experience, can define some pattern or algorithm. However, it is the case in which the database is very large, it is an impossible job for a single person; look for a needle in a haystack [11].

2 Problem

Currently, companies that seek to certify their maturity processes in the construction of products and services do not succeed because, more than beyond the methodologies, standards, documentation and good practices, it has to do with the change in the way of being of capital intellectual, that is, of the people who are involved in the areas. According to [13] the problem lies in fears, ignorance, resistance to change, not wanting to invest money, little training and lack of time, as shown in Figure 4.

Fig. 4. Issues implementing CMMI [9].

3 Proposal

In the present work a predictive model is presented to implement CMMI by means of machine learning which has two objectives: the first objective is to support the client companies (CliC) to know their company or is candidate to aspire to certification through a developed program in JAVA that determines the score based on the answers provided by the personnel involved in the project. The second objective is to support the certification company (CertC) in the analysis of data for decision making and thus identify the client companies (CliC) that are candidates for certification and those that do not, by filtering, classifying, grouping and association.

4 Methodology

This section shows the proposed methodology which is distinguished in the sections, the first explains the tool and the second the intelligent model. In Figure 5 the proposed methodology is shown.

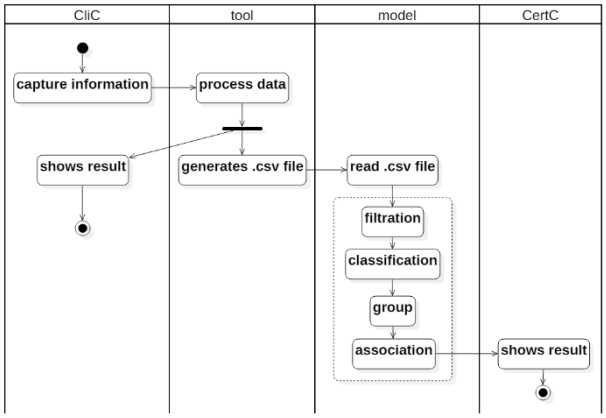

Fig. 5. Methodology.

On CliC Street the client company captures the information in the system. This is where the CMMI questionnaire is answered.

The data is sent to the street tool for processing in the JAVA software. In parallel, the results are sent to CliC and a .csv file is created with the captured information.

The .csv file is read by model and starts the analysis process with machine learning which consists of filtering, classifying, grouping and associating the information of the file.

Finally, the results are shown in CertC.

CliC

It answers a questionnaire of one hundred and twenty questions in order to know its status regarding the good practices of the CMMI level two and three independently of the constellation. The answers given to the questionnaire can be given by more than one person.

The dimensions considered for the questionnaire are: institution: thirteen questions, project management: fifty-four questions, organization: twenty-two questions, engineering: thirty-one questions to give a total of one hundred and twenty questions.

It is expected that the answers will be provided with ample and solid knowledge of what there is and what the company does. The answers must be dichotomous, that is, yes or no. Each response has a same weighting that will determine the level of maturity in each dimension and will be compared with the maximum of each dimension. Table 1 shows the process areas with the weights.

Tool

The proposed tool is based on the JAVA programming language. In this tool the survey is developed. In the Figure 6 the process is shown.

The tool processes the data provided by the CliC represented by the hundred and twenty dichotomous questions. A weight is assigned to each item of the survey. There is a question to request more answers; if there are more, the data is received and reprocessed until there are no more answers.

Table 1. Process areas.

Process area	Top Score
PP	16
PMC	6
MA	4
SAM	7
CM	6
REQM	6
PPQA	3
RD	9
TS	6
GG	13
PI	7
VER	6
VAL	5
OPD	7
OPF	3
OT	6
RSKM	5
DAR	1
IPM	6
TOTAL	122 points

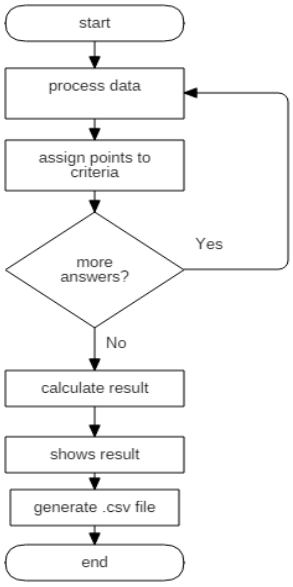

Fig. 6. Process of the tool.

When there is no more, the result is calculated and displayed at the same time that a .csv file is generated that will go to the model process.

The tool has 4 sections, one for each dimension. Figure 7 shows the prototype where it is possible to respond to the questions of each dimension by means of radio button controls to make the response to the reagents more usable and simple.

When the capture of a section is finished, the next button is clicked so that by means of this element it is moved to another section of the form. When you get to the engineering section there is a button to calculate yourself that will perform two actions; the first is to take all the answers and match them with the weighting criteria for each one of them; the second action is to create a .csv file with the results of the survey to send it to the model.

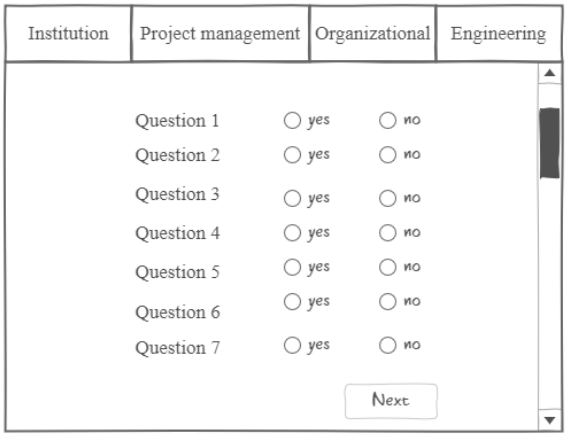

Fig. 7. Prototype.

The general process consists of five methods: Init, GetData, ShowQuestion, CreateRadioButton, CreateQuestions; and two buttons: Next and Back. The Init method positions the window and reads the file where the questions are, to do it in a dynamic way and to be able to read other files with more questions if required.

```
private void Init(){
    setLocationRelativeTo(null);
     this.setResizable(false);

    ReadCsv read = new ReadCsv();

    try{
        List = read.GetData();

        ShowQuestion(From,To);
        BtnBack.setVisible(false);

    }catch(IOException ex){}
}
```

The GetData method reads the file in the ISO-8859-1 format (coding of the Latin alphabet, as accented letters). This already read file is saved in an ArrayList to be able to handle it.

```
public ArrayList GetData() throws IOException{

    BufferedReader br = null;
    ArrayList list = new ArrayList();

    try {

       br = new BufferedReader(new InputStreamReader(new FileInputStream(Read), "ISO-8859-1"));

       //br =new BufferedReader(new FileReader(Read));

       String line = br.readLine();

       while (null != line) {
          line = br.readLine();

          String Num = line.split(",")[0];
          line = line.replace(Num + ",", "");
          line = Num + "&" + line;

          if(!line.equals(""))
            list.add(line);
       }

    } catch (Exception e) {

    } finally {
       if (null!=br) {
          br.close();
       }
    }

    return list;
}
```

The ShowQuestion method shows and generates the questions already read from the file with their respective selectable responses in a dynamic way.

```
private void ShowQuestion(int From, int TO){
    int y = 30;

    for(int x = From; x < TO; x++){
        String Sep[] = List.get(x).toString().replaceAll("\"", "").split("&");

        CreateQuestion("<html>" + Sep[0] + ".- " + FirstMayus(Sep[1]) + "</html>", y);
        y = y + 40;
        CreateRadioButton(y);
        y = y + 30;
    }
}
```

Model

The objective of the model is to build a predictive scheme, based on machine learning, for which the process is shown in Figure 8 where it is possible to identify the sections through which the information captured in the tool is transformed.

The process consists of providing the .csv file generated by the tool to the WEKA software in which it is necessary to create a .arff file from the .csv where the structure of attributes and data must be specified in the order of numeric, real, categorical, dichotomous in such a way that if the arrangement of options is shown inside the file.

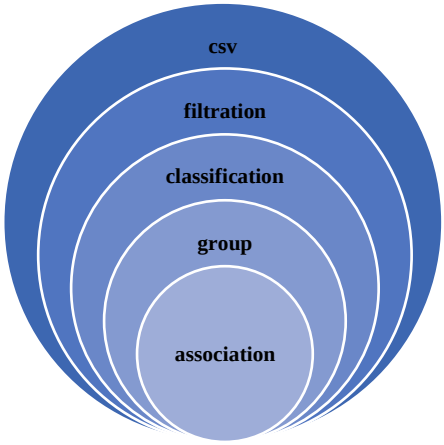

Fig. 8. Model process.

To continue with the filter, WEKA is informed of the location of the .csv file and it is opened. To work in the proper format you need to convert the .csv to .arff. The format that is used to work with the reagents is to assign them a name with q1, q2, q3 up to q120 for the reagents of the CMMI diagnosis for level two and three all as numeric types. The .arff file must contain, in addition to the .csv information, a special structure; @relation, @attribute and @data, as shown below.

```
@relation results

@attribute question {r1,r2,r3,r4,r5}

@attribute q1 numeric
@attribute q2 numeric
@attribute q3 numeric
@attribute q4 numeric
@attribute q5 numeric
@attribute q6 numeric
@attribute q7 numeric
@attribute q8 numeric
@attribute q9 numeric
@attribute q10 numeric

@datar1,0,0,1,0,1,1,0,1,1,1
```

@relation establishes the name of the file, @attribute shows each one of the questions of the questionnaire (for illustrative purposes, only the first ten records of the hundred and twenty are shown), @data contains the information that accompanies each attribute, that is, they are the data that make up the file result .csv.

In WEKA to perform the transformation from .csv to .arff the Arffviewer tool must be used, there the .csv is opened and saved as .arff.

The .arff is opened by the Explorer application and the file is indicated. Once opened, it proceeds with classification, grouping and association.

CertC

The certifying company observes the results of WEKA processing to determine or predict future scenarios. The results obtained are shown in the following section.

5 Results

The results obtained are based on the choice of one of the one hundred and twenty questions of the questionnaire; by convention, the one question that was chosen by the team is the most representative for the purpose of the certification; the question is: 1. Do you have an organizational policy that dictates the discipline of process monitoring?

The classification showed that only 5 of the 5 studied did not have it. It is presumed at this point that more than half do have an organizational policy. The grouping built two groups. The association shows five rules:

1. q73=1 ==> q16=1
2. q86=0 ==> q16=1
3. q96=1 ==> q16=1
4. q112=0 ==> q16=1
5. q86=0 ==> q73=1

Derived from the rules, the following findings are obtained:

1: Who do I establish operational scenarios that help me describe the product specification (flowcharts, presentations, prototypes, demos, sequence diagrams, etc.) ALSO document the life cycle that describes the phases through which these projects go unwinding
2: Those who do NOT document the architecture of our solutions and their design DO document the life cycle that describes the phases through which these projects are evolving.
3: Who do I run tests in customer environments to ensure that I meet their expectations. This may also involve UAT tests or guarantee periods of my solution ALSO we document the life cycle that describes the phases through which these projects are developing.
4: Those who do NOT document the standard work environment for our projects (materials, tools, licenses, corporate software, etc.) DO document the life cycle that describes the phases through which these projects are evolving.
5: Those of us who do NOT check that the technical description of interfaces is complete and well defined IF I establish operational scenarios that help me describe the product specification (flowcharts, presentations, prototypes, demos, sequence diagrams, etc.).

Therefore:
Those who document the life cycle that describes the phases through which these projects are evolving
As well
* I do establish operational scenarios that help me describe the product specification (flowcharts, presentations, prototypes, demos, sequence diagrams, etc.)

But no
* We check that the technical description of interfaces is complete and well defined
* Yes, I run tests in client environments to ensure that I meet your expectations. This may also involve UAT tests or guarantee periods of my solution
But no
* We document the architecture of our solutions and their design
* We document the standard work environment for our projects (materials, tools, licenses, corporate software, etc.)

6 Conclusions

Based on the results, the conclusion of this paper is that:

It is possible that companies can aspire to a certification under a trust scheme because they count on the present proposal with a tool that allows them to know their status.

That certification companies can, based on an analysis scheme, predict the behavior of client companies seeking certification.

Save resources (time, money and effort) by using this proposal.

7 Future Work

As part of the future work it is expected to build the tool an Android so that in this way it is more practical to apply the surveys as well as a possibility to store the historical data in a MySQL database.

It is also expected to have the application in a web environment in the same way with connection to MySQL in order to have the records of all the surveyed companies stored and be able to apply these results in another Big Data scheme.

References

1. Noyel, M., Thomas, P., Charpentier, P., Thomas, A., Brault, T.: Implantation of an on-line quality process monitoring. In Proceedings of 2013 International Conference on Industrial Engineering and Systems Management (IESM), pp. 1–6 (2013)
2. CMMI Institute - Maturity Profile ending 31 December 2017. (s. F.). Retrieved June 2, 2018, from https://cmmiinstitute.com/resource-files/public/quality/maturity-profiles/maturity-profile-ending-31-december-2017
3. PADRON_CENTRO DEVELOPMENT CURRENT_2016_abr-18.pdf. (s. f.) Retrieved June 3, 2018, from https://prosoft.economia.gob.mx/doc/PADRON_CENTRO%20DE%20DESARROLLO%20VIGENTE_2016_abr-18.pdf
4. Chrissis, M., Konrad, M., Shrum, S.: CMMI for Development: Guidelines for Process Integration and Product Improvement, Third Edition. Addison-Wesley Professional (2011)

5. Menezes, W.: Capability Maturity Model Integrated. Encyclopedia of Software Engineering, Volume 1, 1112–120 (2002)
6. Huang, K., Yang, h., King, I., Lyu, M.: Machine Learning. Modeling Data Locally and Globally. Springer (2008)
7. Kubat, M.: An Introduction to Machine Learning. Springer (2015)
8. Mohamed, K.: Machine Learning for Model Order Reduction. Springer (2018)
9. RaschKa, S.: Python Machine Learning. UK, Birmingham: Packt (2015)
10. Eckert, K.B., Suénaga, R.: Analysis of Attrition-Retention of College Students Using Classification Technique in Data Minig. Form. Univ. 8(5), (2015)
11. Lantz, B.: Machine Learning with R. UK, Birmingham: Packt (2013)
12. Rohwer, R., Wynne-Jones, M., Wysotzki, F.: Neural Networks. In: Michie, D., Spiegelhalter, D.J., Taylor, C.C. (eds): Machine Learning, Neural and Statistical Classification, pp. 84–106. Inglaterra, Birmingham (1994)
13. Palacios, H., Porcell, N.: Obstacles when implanting the CMMI model. Bogotá (2008)