Designing a system with machine learning to formative research projects management

C. Pineda¹, A. Morales¹, L. Lopez¹, C. Benitez¹ and S. Sanchez¹

¹Corporación Universitaria Antonio José de Sucre (Colombia).

Abstract. Artificial intelligence is marked among emerging technologies and is one of the most projected expectations for future dates. In the last decade this technology has been implemented in various sectors, without leaving the education and research sector, generating positive results. Among the most demanded algorithms are machine learning, because it gives the ability to "learn" to machines and information systems through experience and not rely only on data from system inputs in a given context. For this reason and also because of the difficulty that students and teachers express in the elaboration of training research documents, the objective of this research is to design a system that assists the user in the construction of the different sections of a formative research project. The research is descriptive and technological development. Among the results found is the behavior of the proposed system and identifies the different technologies to be used for future development. It is concluded that the system is structured by the MVC(Model View-Controller) architectural pattern which facilitates the scalability and maintenance of the system in the future, in addition to the implementation of the libraries of Python, NLTK(Natural Language Toolkit), Scikit-learn, Tensor Flow and Keras, allow the system to fully use machine learning for scientific and academic text prediction.

Keywords: supervised machine learning, unsupervised machine learning, artificial intelligence, intelligent tutoring systems and formative research.

1. Introduction
The Artificial Intelligence or AI, is considered to be the cognitive process simulation that influence the human decision making by computers or informatics systems [1]. For this these systems integrate mathematical modellings which through information banks detect patterns and this way these “understand” the behavior of context for which they were developed. In the last decades this technology has taken center stage in different sectors, generating expectations for the development and technology innovation of future years.

With regard to the field of education, AI(Artificial Intelligence) facilitates the automation of different processes supporting teaching and facilitating the learning process for students, ensuring that both the teacher and the student focus understand and generate new knowledges and they set aside operational and repetitive processes that do not contribute to vocational training. This aspect can benefit mainly to the field of formative and scientific research which are considered as one of the fundamental processes in the training of professional career and postgraduate students, because this wake up in the student curiosity and the searching and information classification skills for implement of innovation and development of new solutions for daily problems that allows the society progress. For this reason the Corporacion Universitaria Antonio José de Sucre, recognizes the importance of
scientific and formative research, relying on bodies that manage this activity within the programs it offers. Among these bodies are the research and innovation center that relies directly on the research committees of the faculties, which in turn relies on the training research activities developed by teachers, which carry out pedagogical strategies such as the creation of research hotbed in an extracurricular way and development of classroom projects within some subjects, with the intention of fostering an investigative culture within the academic community.

Based on the above, the Corporacion Universitaria Antonio José de Sucre seeks to innovate in the creation of an information system that allows not only the management of research and academic documents generated by teachers and students, but also that the system is supported by Artificial Intelligence to automate processes of elaboration of the sections of research documents, for this reason the central objective of this document is to analyze and investigate articles research results of different authors of the last 5 years, all these related to artificial intelligence, and propose the architecture of a system and identify the technologies with which it should be developed. The document is initially organized by setting out the theoretical framework, then describes the research methodology and finally disseminates the results, discussions and conclusions.

2. Theoretical Framework

In this section sets out concepts and theories according to authors related to artificial intelligence, machine learning supervised and unsupervised systems.

2.1. Artificial Intelligence

The Artificial Intelligence (AI) is understood as the ability of computers or informatics systems to learn to make decisions. This technology allows the automation of various process that are carried out in different sectors in which the human being performs, which has spread rapidly in recent years, because more technology devices are incorporated every day that simplify the tasks that human beings develop. [2]

The AI is supported in mathematical models which make use of databanks with similar characteristics, which this technology detects by means of different techniques of identification of patterns using linear or nonlinear multivariate functions. These characteristics can be exploited through techniques that make use of a regression model of the process, and then be contrasted with the real process and get a residue; and techniques that directly model data output for both normal and abnormal operating conditions. [3]

2.2 Machine Learning

The Machine Learning, is a discipline in the field of Artificial Intelligence, that it has as purpose, generate the ability of learn for machines and informatics systems to resolve daily problems. Learning involves obtaining information to improve performance skills and abilities in a context, by identifying patterns in multi-data analysis, allowing machines to perform operations autonomously as they gain experience in the process. In the context of machines, learning is a set of algorithms that allow you to make decisions based on accumulated experiences to solve problems [4].

That is, machine learning is a discipline that arises to analyze data in an automated way, using the technique of recognizing patterns [5], generating the ability for realize generating the ability to make predictive judgments and make the best decisions by analyzing a large amount of existing data, for this reason it is argued that the key to machines acquiring artificial intelligence is through machine learning, the use of this technique has been developed in various sectors, from aspects such as Internet search, tasks that are commonly performed, to political procedures such as the United States presidential elections, to being applied in military decision assistants of developed countries; in other words, always that exist the necessity for data analysis, the machine learning can be used. [6]

From another perspective it can be said that machine learning does not depend of a programming for respond in a certain way according to the inputs received, but rather that utilize techniques to extract same characteristics of behavior from data input, and based on that information, carry out the
evaluation of new entries. The internal algorithms that constitute the base of this machine learning have a strong statistic and algebraic component, with the consequent calculus ability [7].

2.3 Supervised and unsupervised machine learning algorithms
Among the machine learning algorithms the supervised and unsupervised algorithms were found. The unsupervised machine learning algorithms realize input-only processing. Unsupervised machine learning makes use of a dataset, which is not tagged or classified. The system is tasked with identifying patterns in pieces of data by recognizing similarities and grouping the data into categories. The unsupervised machine learning algorithms have the ability to update in relation to processing based only on inputs [7].

On the other hand, supervised machine learning algorithms have a manually classified dataset. On this particular set, the algorithm carries out two processes: find the best parameters for the algorithm, and evaluate the probability level with those parameters. Supervised learning is a technique for deducing a function from training or learning data [7].

2.4 Natural Language Processing
Natural Language Processing is one of many branches of knowledge from Artificial Intelligence, where it basically seeks to optimize and discover more thoroughly the way that the machines communicate with humans, using many different languages that exist in the world. For the elaboration of the Natural Language Processing, it is necessary to eliminate what is known as stop words to then lemmatize the remaining words in the dataset. The stop words makes references to words that have no meaning, for example: prepositions (to, during, according to) or pronouns (he, she, and we) etc. While lemmatize makes references to the process for remove words that are in plural, female or conjugated, the purpose of this is to look for a word that represents all these words in the same form, which are the ones that will appear in the corpus. Based on this a dictionary is developed, with a number of keywords that will serve for the development of natural language. [7]

3. Methodology
This research is experimental design, this type of research has absolute control of the study variables considering that the research topic is projected to design the system architecture with machine learning. As the type of research is concerned, it is descriptive and technological development. Descriptive because the features and functionalities of the different Python libraries are described to suit the study context. According to Mario Tamayo and Tamayo in his work "the process of scientific research", he defines this type of research as the registration, analysis and interpretation of the current nature and the composition or processes of phenomena. The approach is made on dominant conclusions or on how a person, group or thing is conducted or works in the present [8]. Also, the research is technological development type, according to MinCiencias this type of study implements research results, for the design or development of new technological products, as well innovation in technological products. This activity will include the materialization of research results in a blueprint, scheme or design, as well the creation of non-marketable prototypes [9]. The purpose of this research is to generate the design of a computer system that integrates machine learning to support the development of the sections of formative research documents.
To realize the research, two phases were developed:
In the initial phase, a documentary analysis of the information was carried out by implementing the analytical matrix instrument of content where academic research results of topics related to artificial intelligence and machine learning were grouped identifying the functional and non-functional requirements of the proposed system.
Among the selected technologies is the Python programming language, a language that has taken center stage in the development of intelligent systems or algorithms, because it is intuitive and compatible with multiple platforms. Python integrates the NLTK (Natural Language Toolkit), Scikit-
learn, Tensor Flow and Keras libraries, which have methods that allow to integrate machine learning activities.

In the second phase, the system architecture was designed by developing block diagrams, class diagrams, and flowcharts, which allow you to schematize the behavior that you want the proposed machine learning system to have, in addition to illustrating the relationship or integration of the different technologies that make up the system.

4. Results

Research has been conducted linking machine learning in artificial intelligence to education and analyzing the results of these researches, it has been proposed the design of a research project management system with artificial intelligence which includes machine learning. In Figure 1, is represented with a class diagram the structure in which the entire coding of the research project management system is composed, which it can also call Back End. In Figure 2, it was decided to represent using a block diagram, the architecture of the entire system, in this you will be able to observe the communication that exists between the different technologies used. Finally, it is shown in Figure 3, the behavior and processes that the artificial intelligence executes represented in a flowchart.

4.1 Defining how the Web Prototype Backend works

The following diagram illustrates the behavior of each of the Model-View-Controller sections that is implemented in the Web Prototype Backend (see Figure 1):

![Back End class diagram.](image)

The initial data flow when typing the URL (Uniform Resource Locator) in the browser, where the browser request runs to the web server, which immediately communicates with the view controller, which in turn is responsible for verifying that all data is properly hosted on the web server, redirecting to the page not found or the main interface of the website, from which the User Login can be accessed, at the time of login by user, the login controller communicates with AJAX (Asynchronous JavaScript
And XML), technology that is responsible for executing the actions of each of the main controllers, in this case it queries in the database and checks if the data is correct, it also checks the type of user that has been connected, and then redirects you to the home interface either the teacher or the student, if incorrect data is typed system would show the message "wrong user and/or password" and stay on the same interface.

AJAX (Asynchronous JavaScript And XML) is responsible for executing the tasks of each main controller, that is to say, AJAX has internal controllers for each of the controllers associated with this, and it also communicates directly with the database and with the client, AJAX is also responsible for running the interactive application, which will be the text writing section, this component allows interaction with the application and artificial intelligence asynchronously in the background, so the user would have no interruption when they are using the application.

Finally, there are multiple controllers for the different sections of the website, as they are, the controllers for teachers, students, programs, problematic cores, line of research, among other sections, so there is the possibility to find different views or interfaces depending on the section that the user wants to access.

4.2 Defining the Operation of the Web Prototype System

The overall operation of the system is as well as the task of each technology within the system below (see Figure 2):

![System block diagram.](image-url)
Server: it is responsible for providing website services to the client and responding to requests that the client demands.

Client: it is the user who normally executes requests and performs activities on the website.

HTML: It is a markup language that is used for the development of Internet pages. This is the acronym for Hyper Text Markup Language.

PHP: (recursive acronym for PHP: Hypertext Preprocessor) it is a very popular open source language especially suitable for web development and that can be embedded in HTML.

Web Server: It is responsible for containing all the information of the website from the server side, it also takes care of the transfer of information by implementing the SSL protocol (Secure Sockets Layer) for the protection of information.

SSL Protocol: These are cryptographic protocols that provide privacy and integrity in communication between two points in a communication network. This ensures that the information transmitted by that network cannot be intercepted or modified by unauthorized elements, so ensuring that only the senders and legitimate recipients are the ones who have access to the communication in full.

AJAX (Asynchronous JavaScript and XML): It is a web development technique for creating interactive applications or RIA (Rich Internet Applications). These applications run on the client, that is to say, in the browser for users while these maintain asynchronous communication with the server in the background. In this way it is possible to make changes to the pages without reloading them, improving the interactivity, speed and usability in the applications.

AI (Artificial Intelligence): It is artificial intelligence that will be responsible for providing the text prediction tool in the section of project writing by users, the artificial intelligence algorithm has a connection to Python libraries and dictionaries, this way real-time queries are performed to provide the text prediction service to these users, it has also machine learning, which gives artificial intelligence the ability to increase the content of word dictionaries and thus not just leave learning manually for developers through data updates.

SQL Server: It is the database manager that will contain all the information of the website referring to research documents, user and administrator data, among others.

4.3 Defining the functioning of artificial intelligence
The following is how Artificial Intelligence works from the point where the user within the repository decides to start writing their research document (see Figure 3):
Figure 3. Artificial Intelligence flowchart.

- At the beginning, artificial intelligence conducts a query to determine the research topic or core of the problematic that will be working with, in order to have a better reception and response of the data, that is to say, from the beginning of the process artificial intelligence will already be able to preselect words and phrases that go in context with the problem.
- Having completed the above process, artificial intelligence will begin to capture all the words that are being written within the application area, this process could be carried out in different ways either by means of a key listener or some tool that helps store words in data structures like vectors.
- Later, the Pre-processing stage of the data was initiated with the data that it has been received, this process is utilized to clean up data and that avoid all kinds of conflicts or error when running the processes of the text prediction model, for the pre-processing of the data is chosen to use the NLTK (Natural Language Toolkit) library for Natural Language Processing NLP and account is taken of 4 main tasks which are:
  o Deleting the stop words, this process is to remove words that could affect the results in this case the stop words would be connectors and helpers.
  o Tokenize words, Tokenize text is important because text cannot be processed without tokenizing it. The tokenization process means dividing larger parts into small parts. That is to say, convert paragraphs into sentences and make sentences into words according to your needs. NLTK contains a phrase tokenizer and a word tokenizer.
  o Get synonyms, WordNet is a database created for natural language processing. Includes synonym groups and a brief definition. It will be used to load synonyms of words for better search and response options.
  o Lemmatize Words means removing sharpening words and finding the word root. Search engines use this technique when indexing pages, since many people write different versions for the same word and all derive from the root word.
- After completing the Pre-Processing stage, the query is made in the dataset which is the set of words and sentences stored by artificial intelligence, so two decisions are made, in case
the words entered at that moment are related in the dataset, the process of loading the text prediction model immediately begins, otherwise it would initiate a machine learning process to quickly add the necessary words to provide a response to the user.

- If it starts the machine learning process, the first thing that is done is an internet search using the Web Scraping technique for which a tool called Beautiful Soup will be used that is responsible for executing this process, after searching on internet, it has to check if it belongs to the context and the theme that is being addressed in that moment, if it belongs to the theme the pre-processing stage is executed and then added to the dataset and that it is immediately available for use in the trained text prediction model, in case the query made on the internet does not match with the topic that is being worked on, the notification contains a request to the teacher which is that you must make the decision to add the words or phrases found on the internet or if you want to delete them that are not added.

- On the other hand, the process of text prediction and self-training of the model, consists of performing a post-processing of data using the Keras library, in this post processing stage the words are sorted in order of probability that they can be used, that is to say, these words are labeled with a probability obtained with a function called Softmax, the artificial intelligence immediately loads the dataset data into the state that is currently in and checks whether the words captured at the moment are related to some of the words in the dataset, in case of no coincidences this throws a message that says "no suggestions", otherwise if it finds matches, it loads the words in order of probability and they are displayed on the screen to the user, after this, a feedback stage is performed to the text prediction model with the results in order to generate a self-training of the model, That is to say, this text prediction model will be in constant training and updating your data to deliver the best possible results.

5. Conclusions

From the bibliographic review it is concluded that there are a number of techniques, technologies and algorithms that allow machine learning to be integrated into the software development process, allowing developers to select the most appropriate solution to solve the problem to be addressed.

Also, it is concluded that machine learning is one of the main areas of artificial intelligence because it allows the generation of knowledge autonomously in computer machines and informatics systems, by using pattern identification models that use mathematical functions that are supported in statistical and algebraic techniques for performing calculations and procedures.

It is concluded that the proposed system is a comprehensive system that integrates various technologies that achieve the proposed objective, this system is structured by the MVC architectural pattern which facilitates the scalability and maintenance of the system in the future, in addition to the implementation of Python's libraries, NLTK(Natural Language Toolkit), Scikit-Learn, Tensor Flow and Keras, allow the system to fully use machine learning for scientific and academic text prediction, finally the integration between the components and technologies of the system give reliability, creating positive expectations in the development of the system.
References

[1] «Inteligencia artificial en la docencia médica,» Educ Med Super, vol. 33, nº 3, 2019.
[2] «Tecnologías avanzadas e Inteligencia artificial: oportunidades para una Enfermería con competencias ampliadas,» Index Enferm, vol. 28, nº 3, 2020.
[3] L. Hurtado, E. Villarreal y L. Villarreal, «Detección y diagnóstico de fallas mediante técnicas de inteligencia artificial, un estado del arte,» Dyna rev.fac.nac.minas, vol. 83, nº 199, pp. 19-28, 2016.
[4] J. Ríos, G. Ulla y A. Gianni, «APRENDIZAJE AUTOMÁTICO EN UNA PLATAFORMA ONLINE CON FINES SOCIALES Universidad Tecnológica Nacional de Argentina,» Concurso de Trabajos Estudiantiles. pp. 30-39, 2019.
[5] A. Urbina y J. De la Calleja, «BREVE REVISIÓN DAPLACIONES EDUCATIVAS UTILIZANDO MINERÍA DE DATOS Y APRENDIZAJE AUTOMÁTICO,» REDIE, vol. 19, nº 04, pp. 84-96, 2017.
[6] W. Sun y X. Gao, «The Construction of Undergraduate Machine Learning Course in the Artificial Intelligence Era,» International Conference on Computer Science & Education (ICCSE), pp. 1-5, 2018.
[7] T. Baviera, «TÉCNICAS PARA EL ANÁLISIS DEL SENTIMIENTO EN TWITTER: APRENDIZAJE AUTOMÁTICO SUPERVISADO Y SENTISTRENGTH,» Revista De Comunicación Digital, pp. 1-18, 2017.
[8] Tamayo y T. Mario, El proceso de la investigación científica, Mexico: GRUPO NORIEGA EDITORES., 2003.
[9] MinCiencias, «Tipología de proyectos calificados como de carácter científico, tecnológico e innovación.,» MinCiencias, Bogota, 2016.
[10] I. Martín, RECONOCIMIENTO DE VOZ BASADO EN CARACTERÍSTICAS DNN BOTTLENECK, Madrid: Universidad Autonoma de Madrid, 2019.
[11] L. Ferreira, G. Spadon, C. Carvalho and F. Rodrigues, "UN ANÁLISIS COMPARATIVO DEL MODELADO AUTOMÁTICO DE ESTILOS DE APRENDIZAJE A TRAVÉS DE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO," IEEE Frontiers in Education Conference (FIE), 2018.
[12] H. Pardede, E. Suryawati, R. Sustika y V. Zilvan, "APRENDIZAJE DE FUNCIONES BASADO EN EL CODIFICADOR AUTOMÁTICO CONVOLUCIONAL NO SUPERVISADO PARA LA Detección AUTOMÁTICA DE ENFERMEDADES DE LAS PLANTAS.," International Conférence on Computer, Control, Informatics and its Applications (IC3INA), pp. 158-162, 2018.
[13] R. Tommy, G. Sundeep and H. Jose, "DETECCIÓN AUTOMÁTICA Y CORRECCIÓN DE VULNERABILIDADES MEDIANTE APRENDIZAJE AUTOMÁTICO.," Robin Tommy, Gullapudi Sundeep, Hima Jose, pp. 1062-1065, 2017.
[14] K. Bengoetxea, A. Atutxa and M. Iruskieta, "UN DETECTOR DE LA UNIDAD CENTRAL DE UN TEXTO BASADO EN TÉCNICAS DE APRENDIZAJE AUTOMÁTICO EN TEXTOS CIENTIFICOS PARA EL EUSKERA," Red de Revistas Científicas de América Latina, el Caribe, España y Portugal, pp. 37-44, 2017.
[15] O. Santa, L. Ramírez and F. Trujillo, "TÉCNICAS DE APRENDIZAJE AUTOMÁTICO APLICADAS A ELECTROENCEFALOGRAMAS.," Research in Computing Science 113, 2016.