XLIII
CILAMCE
21-25, November, 2022
FOZ DO IGUAÇU - PR - BRAZIL

Application of Artificial Intelligence in Jurisprudence Search Engine

Edyene Cely Amaro Oliveira[1], Ingrid Haas[1], André Isaac Ferreira, Ana Rúbia Martins Ferreira Vieira, Caio Adriano Rodrigues dos Santos, Camila Chaves Mariano, Camilo Leal Ferreira, Daniel Teófilo da Silva, Eduarda Isabelle Correia Schneider, Filipe André Marcelino E Oliveira, Gabriel Martins Ferreira, Ivan Pedro, Jeniffer Lorrane Costa Sousa Silva, Jonathan Enrique Silva Moreira, Lana de Souza Medeiros, Luciano França da Silveira Júnior, Maria Eduarda da Silva Viana, Matheus Henrique Marcelino e Oliveira, Nathan Siman Teixeira, Patricia Gomes Martins, Thamirys de Jesus Campos, Viviane Santana

[1]*Dept. Centro Universitário Una campus Aimorés*
Rua dos Aimorés, 1451 - Lourdes, Belo Horizonte - MG, 30140-071
edyene.oliveira@prof.una.br

Abstract. Search for jurisprudence is an arduous task that requires hours of commitment on the part of the legal professional. Currently, there are several software that help in the search for the ideal document that the lawyer needs to corroborate his processes. However, such systems use search for key words, that is, from a phrase or set of words typed by the user, the software performs research in each document of the databases of jurisprudence. This is a slow process and, by using keywords, the documents resulting from the search are mostly irrelevant with the intention of the lawyer. Thus, this project aims to implement a web software to perform search in databases of jurisprudence using artificial intelligence in its search engine. The methodology applied consists of obtaining the database with all published jurisprudence and training an artificial intelligence algorithm with all the content of the documents. After training the model, it will be inserted into a web software that will receive the lawyer's research. Currently, an artificial intelligence model was generated with 420 documents and the results were satisfactory. In the tests, the model reached 95% accuracy.

Keywords: Artificial Intelligence, Jurisprudence, Law Research, Search Engine

1. Introduction

The total amount of data created, captured, copied and consumed worldwide reached 64.2 *zettabytes in* 2020 [1]. Such data comes from various actions, such as decisions in legal courts around the world. In Brazil there are more than 90 courts (Superior Courts, Courts of Justice, Regional Labor Courts, Military Courts of Justice, among others).

The courts are present in several states and make many decisions daily. In 2020, for example, 25 million judgments and terse decisions were handed down [2]. However, such decisions sometimes need to be interpreted and law uses the Greek word hermeneutics which means the art or technique of interpreting and explaining a text or discourse.

Legal hermeneutics is a part of law where it aims at the systematic study of the processes, principles, and rules to be used by law enforcement operators for better interpretation of laws. However, interpreting a law depends on it not being clear, otherwise it does not need interpretation in *claris cessat interpretatio*. Some authors diverge from this information, because laws can have their dark side, confusing too difficult to interpret, so it can confuse the operator of law [3].

To guide in the interpretation of the law, there are formal and material sources. Material source is the real and political source of law, while the formal sources are: Federal Constitution of 1988, complementary laws, ordinary laws, amendments, delegated laws, provisional measure, delegated laws, decrees, legislative decrees, regulatory decrees, resolutions, international treaties, infralegal norms such as ordinances, circulars, work orders, and previous jurisprudence and judicial law [4].

Thus, a formal source widely used in law are jurisprudence. The term "jurisprudence" *comes from latin jus* ("fair") *and prudentia* ("prudence" or "wisdom") *or jurisprudentia*, means a set of decisions of a particular court to a certain subject. Case-law provides legal certainty within a decision [5].

Judicial decisions after being made are stored in databases and, with this, there are currently millions of judicial decisions (jurisprudence) in various databases in Brazil. The search for jurisprudence by law-a-seators is an arduous task since search systems are not optimized. It is common to search for a theme on websites of the Supreme Court (Superior Federal Court) and STJ (Superior Court of Justice) and receive many different results, and the only way to achieve the goal is by reading document by document. It is a time-consuming task where law operators work several hours to find a result that will assist them in the proposed theme.

Therefore, the objective of this work is to design and implement software that uses artificial intelligence to assist in the search for jurisprudence facilitating the lawyer's day-to-day life. He will have more agility and assertiveness in the search for these newsletters.

In the literature review were found projects such as: JUSBRASIL: Site that allows research of jurisprudence, being able to seek the Supreme Court (STF), Supreme Court of Justice (STJ), Superior Electoral Court (TSE), Superior Labor Court (TST), Supreme Military Court (STM), National Standard Class (TNU), in addition to the Federal Regional Courts (TRF), Regional Electoral Courts (TRE), Regional Labor Courts (TRT), Courts of Justice (TJ) and Courts of State Accounts (ECA), just selecting which courts you want to seek. It is a site widely used by lawyers because it covers the whole of Brazil, besides having instructions for use, making the use of the user easier. JUSBRASIL does not have the efficiency of precisely finding a case law. He performs the search for similar words and returns all available jurisprudence, so that if not quite specific, the search will be quite wide.

- DIGESTO: It is possible to do research, filter the courts, in the same way as JUSBRASIL, acquiring all kinds of jurisprudence in Brazil. It is not as used as JUSBRASIL but has its value among law students. DIGESTO is also unable to accurately search for the researched document.

- LEGJUR: A research site less comprehensive than JUSBRASIL and DIGESTO but has its differential that allows searches in certain areas of law separately. In addition, it has a legal forum, with relevant discussions of law, causing this point a great differential for the user who needs a different opinion on a specific topic

This project aims to speed up the search for these newsletters more easily for all who are going to use this source of law. In financial terms, using this software the lawyer will save time, because the proposal of the system is to deliver only relevant documents through the researched theme. For large offices that usually use interns to perform the search, they will be able to allocate such professionals in other tasks. In short, regardless of whether you are just a legal professional or a large group, the software will provide agility in the search for jurisprudence. With this, lawyers will be able to focus on interpreting the law and providing quality services to their clients.

The overall objective of the project is to implement a software with search engine capable of performing intelligent searches in large databases of jurisprudence and return only the relevant documents to the lawyer.

As secondary objectives can be highlighted so far:

1. Generation of webcrawler software to search for jurisprudence documents in public databases, such as TJMG. In the first phase, the search focused only on TJMG in the civil area.
2. Design and implementation of preprocessing algorithms for transformation of the document (.pdf) into text (.txt), and cleaning of the data for classification of the Machine learning algorithm.
3. Implementation and identification of a PLN neural network algorithm (Natural Language Processing) suitable for classification of texts.
4. Implementation of a software with attractive and user-friendly interface to receive the phrase and / or key words typed by the user and perform the search of jurisprudence using artificial intelligence.

2. Methodology

The process began with the generation of the database of case-law documents. For this work, data were obtained from the Court of Minas Gerais (TJMG) of the Civil area. Thus, documents were extracted from the period between February 6, 2019, and December 11, 2019. The total amount of files totaled 420 documents of the *Portable Document Format type*, that is, with extension .pdf.

A *webcrawler software was implemented to* search for jurisprudence documents on the TJMG (https://www5.tjmg.jus.br/jurisprudencia/formEspelhoAcordao.do) website. Next, functionality was implemented to transform the files (.pdf) into text documents (.txt) and perform cuts in parts that have no relevance to the algorithm, such as headers. All documents in this court have the same layout.

After obtaining the files, functionality was implemented to perform data cleansing. Thus, they were removed from the text: punctuation, articles, words smaller than 2 characters, numbers, words that are not composed purely by alphanumeric and connective characters.

No documents have been discarded, only transformed, and cleaned. This process is important because the adopted criteria extract words that do not have relevance in the search. In addition, documents have not gone through the anonymization process since the data is public. Any citizen has access to any case law document.

After this step, an algorithm capable of generating grouping of documents according to similarities between them was implemented. Several tests were generated for analysis and training of the classification algorithm, such as 3, 8, 20 classes,

Another algorithm for classifying documents has also been implemented. This algorithm used artificial neural networks MLP (*MultiLayer Perceptron) to* classify the documents.

Finally, a software with user interface was implemented to receive a sentence and return documents that have information relevant to the search used as a filter.

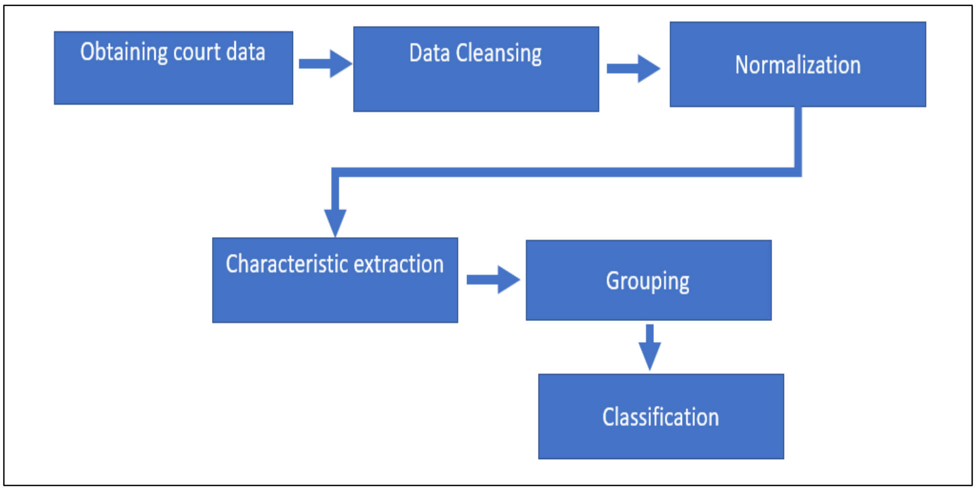


Figure 1 - Flow of the process of classification of jurisprudence
Source: the authors

After training the artificial intelligence algorithm, the model will be included in the web system and made available for users to search for jurisprudence. Figure 2 can be seen using the software process.

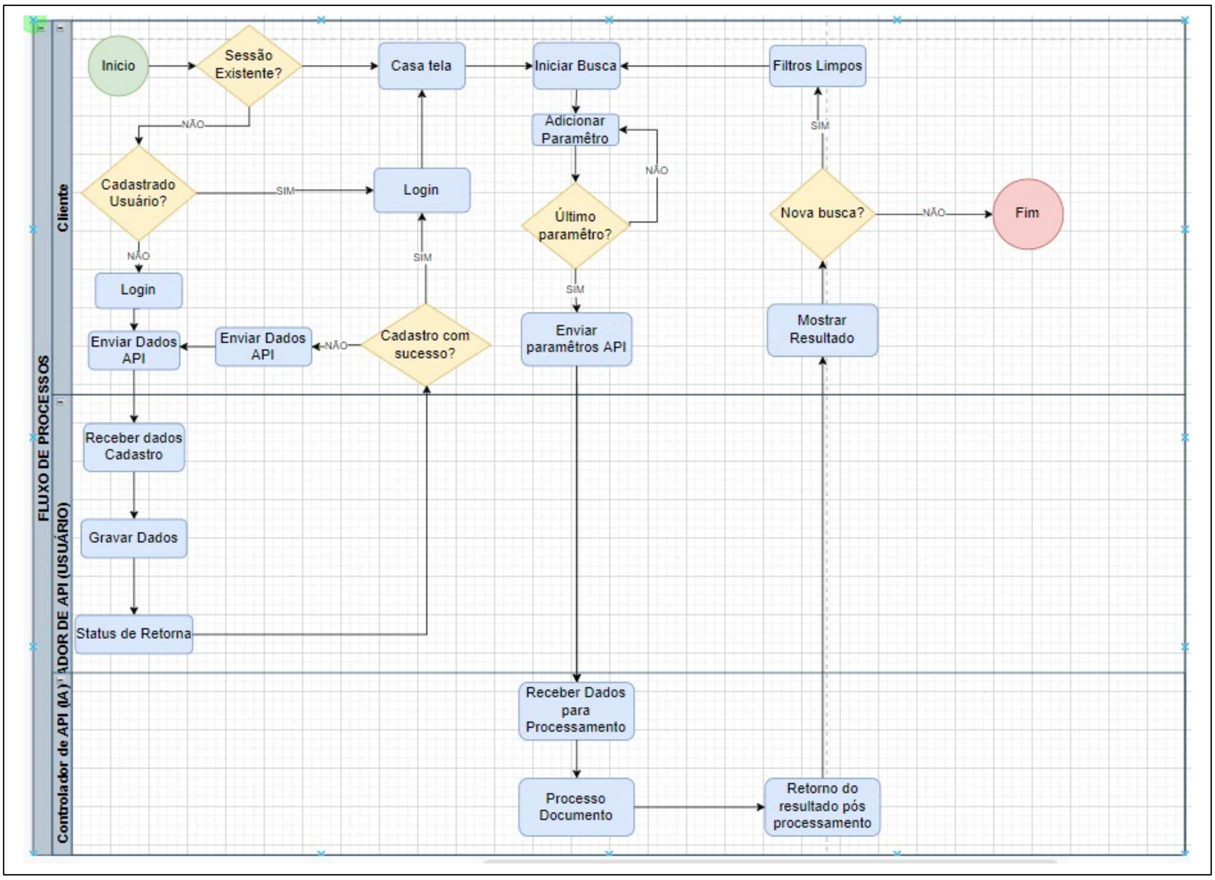


Figure 2 - Flow of the activity of search for jurisprudence by the user
Source: the authors

3. Results

The results obtained so far consist of training with MLP neural networks. Another detail is that currently in the courts throughout Brazil it is presumed that there are millions of jurisprudence documents available for consultation. Therefore, in this work, an extremely low number of documents was used to perform the tests in relation to the total, since the amount totaled 420 files. Table 1 shows data used for grouping and sorting.

Table 1 - Data used for grouping and sorting

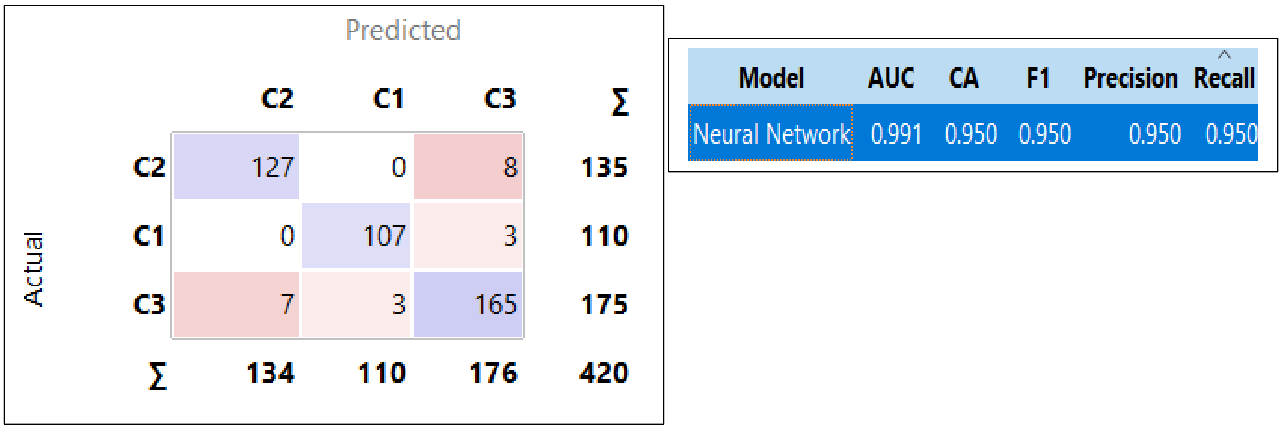
Data - area	Number of files (pdf documents)	Number of words
Civil - TJMG	420	1.024.411

The measure of similarities used for grouping was the Cosine Distance. The configuration of the hyperparameters of the MLP neural network were Neurons: 100 in the first layer, 50 in the second and third; learning rate: 0.001; Activation functions: Tangent on the first layer, Relu on the second and third layers; Optimization algorithm: Adam and 200 times.

Validation methods with manual data separation and cross-validation were used. In the manual separation, the 90% cut was used for training and 10% for tests. And, in cross-validation, tests were performed with 5 and 10 *folds*.

Results and conclusions of The MLP RNA considering the metrics and confusion matrix using:

a) 90% data for training and 10% for testing.



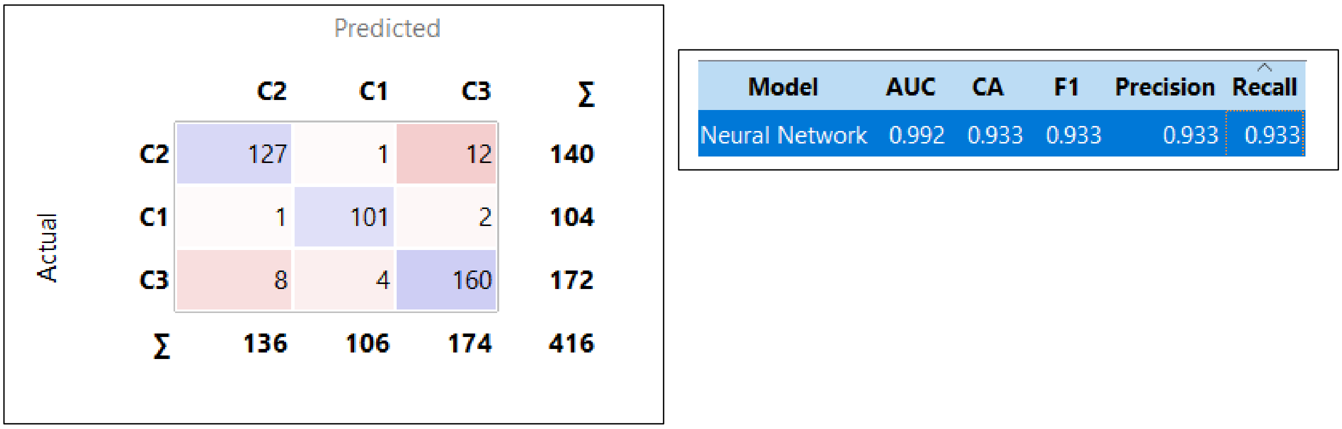
Predicted

Actual	C2	C1	C3	∑
C2	127	0	8	135
C1	0	107	3	110
C3	7	3	165	175
∑	134	110	176	420

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.991	0.950	0.950	0.950	0.950

Figure 3 - Confusion matrix and results of metrics of: AUC, accuracy, F1, Precision and Recall.

b) 5 Folds



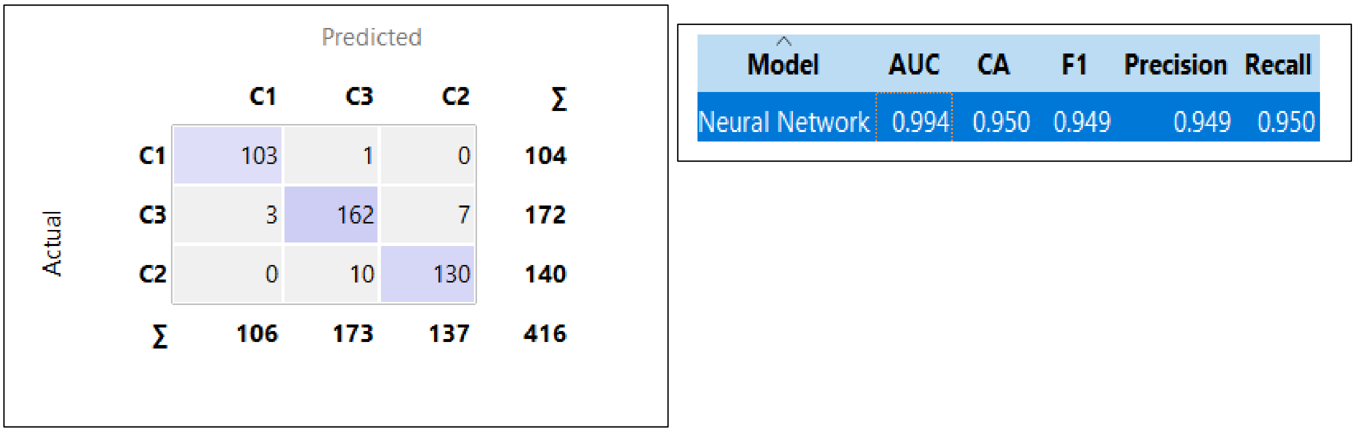
Predicted

Actual	C2	C1	C3	∑
C2	127	1	12	140
C1	1	101	2	104
C3	8	4	160	172
∑	136	106	174	416

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.992	0.933	0.933	0.933	0.933

Figure 4 - Confusion matrix and results of the metrics of: AUC, Accuracy, F1, Precision and Recall. Cross-validation.

c) 10 Folds



Predicted

Actual	C1	C3	C2	∑
C1	103	1	0	104
C3	3	162	7	172
C2	0	10	130	140
∑	106	173	137	416

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.994	0.950	0.949	0.949	0.950

Figure 5 – Confusion matrix and results of metrics of: AUC, Accuracy, F1, Precision and Recall. Cross-validation.

The results showed that the model presented high accuracy. Figure 3 shows that the model correctly classified 399 of the 420 documents. Of class C1 of the 110 belonging to it, the model classified 107 correctly. In class 2 the model ranked 127 of the 135 documents and finally class 3 the model ranked 165 correctly and missed 10.

In the first case in which 5 *folds were used* in cross-validation, the model correctly classified 388 of the 420 documents evaluated. The class with the highest error was C2 with 13 documents incorrectly classified by C3 with 12.

In the classification using 10 *folds* the model ranked 395 documents correctly out of a total of 420. Classes C2 and C3 showed a higher number of errors, adding 10 documents classified incorrectly. Whereas class C1 presented only 1 incorrectly classified document.

Figure 6 shows the cost and performance graph of the model.

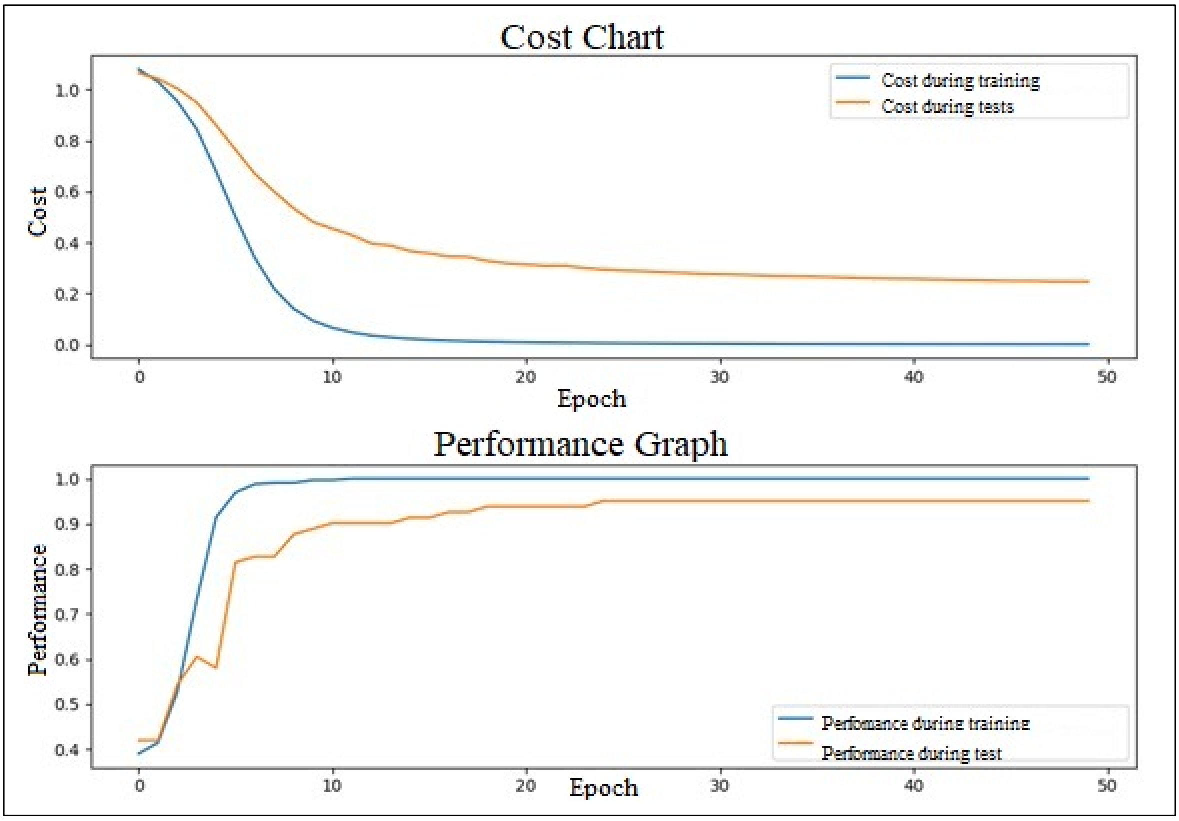


Figure 6 - Cost and performance graph of the MLP model.

In the cost chart it is noticed that in training the model starts with error above 1 and falls quickly. Around season 20 the model stabilizes with values very close to zero. In the case of the tests the fall was not as fast as in training but fell and stabilized around the 45th season.

In the performance graph, the training showed accuracy close to 95% stabilizing around the season 30. And in the case of the tests there were oscillations in some points as close to season 5 and 9 and was stabilized with learning close to 92% around the 40th season.

4. Conclusions

In this project, tests were carried out using a very small amount of jurisprudence documents extracted from the TJMG database. The objective was to identify the feasibility of implementing the project.

Tests were performed using grouping of the obtained documents and an artificial neural network model was trained to know such documents. Then a prototype user interface was implemented to receive the research and use the model to search for the appropriate jurisprudence.

The tests performed in this prototype indicated that the methodology works, because the model obtained accuracy above 95%. In addition, in tests performed using the user interface, AI only needed 3 seconds to indicate where the search-related documents were. Speed occurs since AI works differently from a keyword search system. She knows the documents, so when she receives the keywords, she points out where the documents are. Unlike traditional systems that when receiving words, the system needs to read document by document in search of similar words.

Therefore, this software has great potential to be innovative in law. The project is less than 6 months old and there is already a prototype. Next semester it is expected to have obtained the full case law base or at least 1 million files. In addition, other algorithms will be implemented, such as the *Transformer Neural Network*. This model is the most current technology for NLP jobs.

References

[1] Statista, “Volume de dados/informações criados, capturados, copiados e consumidos em todo o mundo de 2010 a 2025”, 2022. https://www.statista.com/statistics/871513/worldwide-data-created/.

[2] C. N. de Justiça, “Justiça em Números”, 2021.

[3] R. L. França, “Hermenêutica jurídica”, São Paulo, p. 190, 2009.

[4] E. C. B. Bittar, “Introdução ao estudo do direito: humanismo, democracia e justiça”, *2. ed. Saraiva*, São Paulo, p. 616, 2019.

[5] L. M. Pereira, “Intepretação das normas jurídicas”, *Damásio*, São Paulo, 2021.