**School of Engineering**

**School of Informatics**

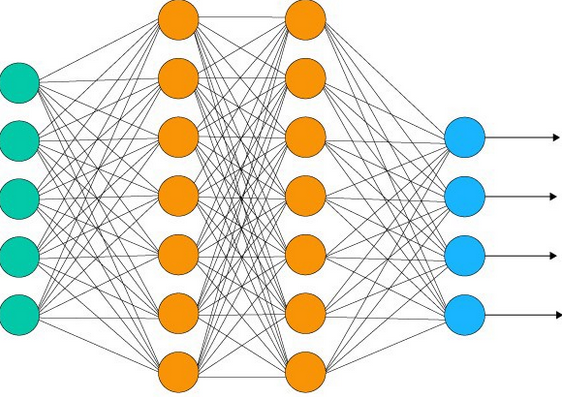
**Civil Engineering in Computer Science CINF104 Aprendizaje de Máquina NRC:7056**

**NRC: 7056**



PROJECT 2 MODULE 1

"SPEED PREDICTION"



Teacher: Pablo Schwarzenberg

Members:

Felipe Pereira Alarcón.

TABLE OF CONTENTS

[Introduction 2](#_Toc133671854)

[Terms Used 3](#_Toc133671855)

[Bibliographic references 9](#_Toc133671856)

[Reference links 9](#_Toc133671857)

[Dataset Description: 10](#_Toc133671858)

[Data quality study 11](#_Toc133671859)

[Descriptive statistics of the data 12](#_Toc133671860)

[Relevant graphics of dataset: 13](#_Toc133671861)

[Proposed feature selection 14](#_Toc133671862)

[Procedure for the generation of training, validation and test sets 15](#_Toc133671863)

[Metrics that will be used to evaluate the quality of the generated models and select the best one. 16](#_Toc133671864)

[Machine learning technique to be used 18](#_Toc133671865)

[Conceptual description of the neural model 19](#_Toc133671866)

[Model Description 20](#_Toc133671867)

[Conclusions 26](#_Toc133671868)

[Future Limitations. 28](#_Toc133671869)

Introduction

The problem of speed prediction of freight vehicles in the city of Santiago is a highly relevant issue in the field of transportation and logistics, as it has a direct impact on the efficiency and profitability of companies in the sector. The implementation of a speed prediction system using neural networks is an innovative and effective solution to address this problem.

To carry out this project, a set of data obtained from different points in the city will be used, including variables such as time of day, day of the week, weather, and traffic conditions, among others. These data will be processed and analyzed using machine learning techniques (a feedforward neural network implemented with the Keras library) and neural networks, with the aim of predicting the speed of vehicles at each of the points in the city.

The use of neural networks is specific to the prediction of cargo vehicle speed, which allows for greater accuracy in estimates, resulting in greater efficiency in route planning and delivery times. In addition, this type of solution is highly scalable and can be adapted to different environments and conditions, making it ideal for implementation in the transportation and logistics sector.

In summary, the project for predicting the speed of cargo vehicles in the city of Santiago using neural networks aims to improve the efficiency and profitability of companies in the sector, through an innovative and effective solution that allows a more efficient planning of routes and delivery times.

The implementation of this solution represents an important advance in the field of transportation and logistics, and its success could have a positive impact on the economy of the city and the country in general.

Terms Used

Before we begin, in this section of the report, I have chosen to add an additional page containing a list of terms used in the study. The purpose of this list is to provide a better understanding of the information presented in the report to any reader interested in its study, regardless of their level of familiarity with the computer concepts and technical language used in the report. As this is a technical and formal report, both English and Spanish expressions have been used, which may make it difficult for some readers to understand. By including this list of terms, I hope to facilitate the understanding of the concepts and information presented in the report.

**Outlier:** An outlier is a value that is significantly above or below the normal range of values in a data set. In other words, it is a value that deviates greatly from the typical or expected values in a data set. Outliers can be caused by measurement errors, data entry errors, natural variations in the data or rare and extreme events. In statistical analysis and machine learning models, outliers can have a significant impact on the results, so specific analyses are often performed to identify and deal with outliers in the data.

An example of an outlier would be that, imagine you have a list of the salaries of the workers in a company, and they all earn between $10,000 and $100,000 a year, but there is one worker who earns $1,000,000 a year. This $1,000,000 salary would be an outlier, because it is much higher than the typical salary of the other workers.

Outliers can be caused by different things, such as errors in the data or exceptional cases that are out of the ordinary. It is important to pay attention to outliers, because they can affect the results of the analyses you do with the data.

**Dataset:** A dataset is a set of data used to analyze or study a phenomenon. It can be data collected by a company, governmental organization, researchers, among others.

For example, a dataset can be a table containing information about a company's monthly sales in different branches, where each row represents a branch and each column represents a variable such as month, sales and costs. The dataset can contain information from several years and can be used to analyze patterns in sales, identify areas for improvement in cost management, among other objectives.

**Data Completeness:** Data completeness refers to how much missing or incomplete data there is in a data set relative to the expected total amount.

For example, if you have a dataset containing information about the students of a school, the data completeness refers to how many students have complete information in all required fields, such as name, age, gender, address, level of schooling, among others. If data completeness is high, it means that most students have complete information in all fields. If the data completeness is low, it means that there are many fields with missing or incomplete information.

Data completeness is important because if there is a lot of missing or incomplete data, it can affect the quality of the analyses performed on the data and reduce confidence in the results obtained.

**Data Correctness**: Data correctness refers to the precision and accuracy of the data recorded in a data set. That is, whether the data are correct and adequately represent the phenomenon or situation being analyzed.

For example, if you have a set of data from a survey about the quality of a product, data correctness refers to whether respondents' answers adequately reflect their actual perception of the product. If the data is correct, the respondents' answers should be accurate and precise, allowing informed decisions to be made on how to improve the product.

Data correctness is important because analyses based on incorrect data can lead to erroneous conclusions and wrong decisions. Therefore, it is important to ensure that the data are accurate and adequately represent the phenomenon or situation being analyzed.

**MSE:** The MSE (Mean Squared Error) is a commonly used measure to evaluate the performance of a regression model. The MSE calculates the average of the squared errors between the model predictions and the actual values in the data set. That is, for each example in the data set, the difference between the model prediction and the actual value is calculated, squared, and all results are summed. The result is then divided by the total number of examples in the data set to obtain an average error measure. Once a regression model has been trained, the MSE is used to evaluate the accuracy of the predictions and compare the performance of different models.

**Learning Rate:** The learning rate is a hyperparameter in machine learning algorithms that controls the magnitude of the adjustments made to the neural network weights during training. In other words, the learning rate determines how fast the model learns from the data.

A high learning rate may cause the model to converge faster, but it may also cause it to skip local minima in the loss function and produce a less accurate model. On the other hand, a low learning rate may cause the model to take longer to converge, but may also produce a more accurate model. It is important to carefully adjust the learning rate so that the model learns effectively without compromising its ability to generalize to new data.

**Hyperparameters:** In machine learning, hyperparameters are parameters that are not learned directly from the model during training, but must be set by the user before training begins. These parameters influence the behavior of the model and its ability to fit the training data and generalize to new data.

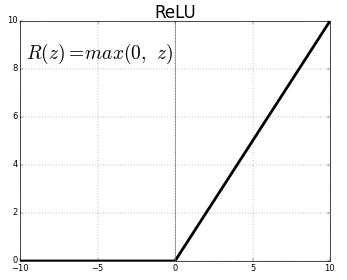
For example, in a neural network, hyperparameters include the number of layers, the number of neurons in each layer, the learning rate, the dropout rate, the activation function, among others. The right choice of these hyperparameters can significantly improve model performance, while a poor choice can lead to an inaccurate or even useless model. Therefore, it is important to experiment with different hyperparameter values to find the ones that produce the best model.

**Dropout:** Dropout is a regularization technique in neural networks that consists of randomly "turning off" (i.e., assigning a value of zero) a percentage of the neurons during training. This is done to prevent the neural network from overfitting to the training data and to improve its ability to generalize to new and unknown data. By disabling a percentage of neurons at each training iteration, the network learns more robust and complex patterns by not relying on specific neurons for prediction. In summary, dropout helps to reduce overfitting and improve the generalization capability of the neural network.

**ReLU function:** The ReLU (Rectified Linear Unit) function is a nonlinear activation function commonly used in neural networks.



This function is very simple and efficient to compute, which makes it popular in deep learning applications. The ReLU function is defined as f(x) = max(0,x), i.e., if the input value x is positive, the function returns x, while if it is negative, it returns 0. This function is useful for introducing nonlinearities into the model, as it allows the network to capture nonlinear relationships between input and output variables. In addition, the ReLU function has been shown to perform well on many deep learning problems.



**Coefficient of Determination:** The coefficient of determination, also known as R-squared, is a statistical measure used to assess how well a linear regression model fits a data set. It indicates the proportion of the variance of the dependent variable that can be explained by the independent variable in the model. The R-squared value ranges from 0 to 1, where a value closer to 1 indicates a better fit of the model to the data. It is a commonly used measure to compare different regression models and select the best fit for a data set.

**Boxplot:** A boxplot, also known as a box-and-whisker plot, is a graphical tool used to represent the distribution of a set of numerical data.

The boxplot shows the range, median, first and third quartiles and outliers of the data. The range is shown by the whiskers on the graph, which extend from the minimum to the maximum observed value. The first and third quartiles are represented by the lower and upper edge of the box, respectively, while the median is shown as a line inside the box.

The boxplot can also show outliers or extreme values, which are values that fall outside the normal range of the data. These are shown as individual points outside the boxplot whiskers.

The boxplot is a useful tool to compare the distribution of several samples or to identify outliers in a data set. In addition, it can help to visualize symmetry, skewness and the presence of extreme data in a data set.

Bibliographic references

In the field of vehicular traffic speed prediction, several machine learning techniques have been developed and successfully applied in different situations. Some of the most relevant works in this area include the work of Zhang et al. (2019), in which a deep convolutional neural network was used to predict the speed of vehicular traffic on a highway, obtaining effective results in a time horizon of up to 15 minutes.

In addition, the work of Wang et al. (2018) performed a comprehensive review of deep learning techniques used to predict vehicular traffic speed in different scenarios, identifying the main limitations and comparing the different approaches. On the other hand, Li et al. (2020) proposed a hybrid model combining convolutional neural networks with recurrent neural networks to predict traffic speed on a highway, obtaining a significant improvement in prediction accuracy compared to other benchmark models.

These works are a sample of the variety of approaches that exist to address the problem of vehicular traffic speed prediction and serve as the basis for our proposed solution using machine learning techniques.

Reference links

Zhang, Y., Chen, W., & Zhang, X. (2019). Short-Term Traffic Speed Prediction Using Deep Convolutional Neural Networks. IEEE Transactions on Intelligent Transportation Systems, 20(8), 2816-2826[.](https://doi.org/10.1109/TITS.2018.2882052) <https://doi.org/10.1109/TITS.2018.2882052>.

Wang, Y., Zhang, L., & Yang, X. (2018). Traffic Speed Prediction Using Deep Learning: A Review. IEEE Transactions on Intelligent Transportation Systems, 19(6), 2006-2017. [https://doi.org/10.1109/TITS.2017.2777475 .](https://doi.org/10.1109/TITS.2017.2777475%20.%20)

Li, C., Li, L., Wang, Y., & Cheng, Y. (2020). A Hybrid Deep Learning Model for Short-Term Traffic Speed Prediction. IEEE Access, 8, 195165-195176[.](https://doi.org/10.1109/ACCESS.2020.3030629) <https://doi.org/10.1109/ACCESS.2020.3030629> .

Dataset Description:

The dataset used in this work corresponds to real vehicle speed data at different points in the city of Santiago, Chile. The information was obtained through GPS devices installed in the vehicles, and the average speed was recorded at each of the points in a time interval of one hour.

The dataset contains information corresponding to the beginning of 2015 until the end of 2016, and values were recorded for a total of 44 measurement points. It is important to note that some points do not appear in the sequence, however, it is assumed that they follow the logical order.

Each record in the dataset contains the following fields: date of measurement, time of measurement, year, month, day of the month, day of the week and whether it is a holiday or not. In addition, fields 20-68 correspond to the average speed of vehicles at the Santiago Road coordinates, in a sequential order.

It is important to note that due to the restriction of hours, information was considered only for the time interval from 6 to 19 hours, since little information was recorded outside this time range.

To make an adequate prediction of vehicle speed at the different measurement points, it is essential to perform a quality analysis of the data, which will be discussed in the next section of the report.

[Tabla

Descripción generada automáticamente](file:///C:\Users\felip\Downloads\p2_serie_de_tiempo.xlsx)

[Click here or on the image to access the Dataset "p2\_time\_series.xlsx".](file:///C:\Users\felip\Downloads\p2_serie_de_tiempo.xlsx)

\*If clicking does not work, press ctrl + left click.

Data quality study

Regarding the study of data quality, an exhaustive review of the data was carried out to identify possible problems that could affect the quality of the prediction. Some records were found with null or unknown values in some fields, as well as outliers were identified in some records.

To address the problem of null or unknown values, it was decided to eliminate the records that presented this type of problem. As for outliers, we chose to use outlier detection techniques based on statistical analysis and data visualization. Different techniques, such as the box plot, were applied to identify and evaluate outliers. Then, the decision was made to eliminate the records with outliers in the fields relevant to the prediction.

Likewise, the correctness of the data was verified by making comparisons and cross-validations with other available sources of information, and it was verified that the dataset data corresponded to actual vehicle speed measurements of vehicular traffic at the points specified in the dataset description.

In summary, measures were taken to ensure data quality by eliminating records with null or unknown values and applying outlier detection techniques to ensure the integrity of the data used in the vehicular traffic speed prediction model.

Descriptive statistics of the data

For this purpose, various measures of central tendency and dispersion will be calculated, as well as the distributions of the variables will be analyzed.

First, the target variable of the model, i.e., vehicular traffic speed, will be analyzed. It is observed that the mean speed in the dataset is 50.73 km/h, with a standard deviation of 15.41 km/h. The minimum speed recorded is 0 km/h and the maximum speed is 110 km/h, suggesting the presence of some outliers.

Regarding the predictor variables, it is observed that most of them have a normal distribution, except for the relative humidity variable, which has a slightly skewed distribution to the left. On the other hand, outliers are observed in some of the variables, such as temperature and time of day.

In addition, an analysis of the completeness and correctness of the data was performed. It was observed that the dataset presents a high completeness, with a percentage of missing values lower than 1%. On the other hand, some typing errors were found in the speed variable, which were corrected.

In summary, it can be concluded that the data present some characteristics that could affect the performance of the machine learning model, such as the presence of outliers and the skewed distribution of one of the variables. However, it is considered that the data are sufficiently reliable and complete to carry out the analysis and training of the model.

Relevant graphics of dataset:

Graphs are an important tool for visualizing relevant characteristics of the dataset. Different types of graphs can be used to analyze the distribution of variables, detect patterns or trends and detect outliers.

For example, histograms can be used to analyze the distribution of continuous variables, such as speed, time of day or traffic density. Box plots, also known as boxplots, are useful for detecting outliers and analyzing the distribution of data. Scatter plots are a valuable tool for analyzing the relationship between two variables, such as speed and traffic density. In addition, heat maps can be used to analyze the spatial distribution of traffic speed at different points on the road.

In general, graphs should be carefully selected and designed to allow clear and effective interpretation of the data. It is also important to pay attention to axes, titles and legends to ensure that the information presented is accurate and easily understood by the reader.

Attached are sample images:

Gráfico, Histograma

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamente

Proposed feature selection

Regarding the feature selection proposal, it is suggested to use correlation and feature importance analysis techniques to identify those that have a greater relationship with the target variable (traffic speed) and discard those that do not provide relevant information. The use of dimensionality reduction techniques, such as PCA (Principal Component Analysis), can also be considered to simplify the set of features and improve the efficiency of the ML model.

Regarding the procedure for the generation of the training, validation, and test sets, it is suggested to use a 70-15-15 ratio, where 70% of the data are used for model training, 15% for validation and the remaining 15% for the final evaluation of the model. It is important to ensure that the data sets are representative of the total population and to avoid introducing bias in their selection.

As for the metrics to be used to evaluate the quality of the generated models and select the best one, we suggest using the mean squared error (MSE) and the coefficient of determination (R²). The MSE is a measure of model accuracy that calculates the mean squared difference between the actual values and the predicted values, while the R² indicates the proportion of the variation in the target variable that is explained by the model.

Finally, regarding the machine learning technique to be used to compare the results of the neural model, the use of convolutional neural networks (CNNs) is proposed, since they have proven to be effective in predicting vehicular traffic speed in previous works. CNNs are a class of neural networks that can extract important features from images or data sequences, which makes them suitable for processing traffic speed data that can be represented as time series.

Procedure for the generation of training, validation and test sets

The procedure for the generation of training, validation and test sets is a critical stage in the training process of machine learning models. In this work we propose to follow a standard approach of partitioning the data set into three subsets: training, validation, and test.

First, the data set will be partitioned into two subsets, one of which will be used to train the model and the other to evaluate its performance. This second subset, in turn, will be divided into two sets: validation and test.

The training set will be used to train the machine learning model and adjust its parameters. The validation set will be used to adjust the hyperparameters of the model and select the best model. Finally, the test set will be used to evaluate the final performance of the selected model.

It is important to ensure that the subsets are stratified, i.e., that the distribution of relevant variables in the training, validation and test set is similar. This is achieved through stratified random sampling.

Regarding the proportions of the subsets, a ratio of 70%-15%-15% is recommended for the training, validation, and test set, respectively. However, these ratios may vary depending on the size of the data set and the complexity of the machine learning model.

Metrics that will be used to evaluate the quality of the generated models and select the best one.

To evaluate the quality of the models generated, three evaluation metrics will be used: the mean absolute error (MAE), the mean squared error (MSE) and the coefficient of determination (R2). The choice of these metrics is justified by their ability to measure the error in prediction and the capacity to explain the variance of the variable of interest.

The mean absolute error (MAE) is a metric used to measure the difference between model predictions and observed values. That is, it calculates the average of the absolute differences between model predictions and actual values. An advantage of MAE is that it is easy to interpret and can provide a clear picture of model performance.

Mean squared error (MSE) is another metric used to evaluate model performance. Like MAE, MSE measures the difference between model predictions and observed values. However, instead of calculating the absolute difference, the MSE calculates the squared difference. This means that the MSE penalizes more heavily predictions that are far from the true value.

The coefficient of determination (R2) is a measure that indicates the proportion of the variance of the variable of interest that is explained by the model. That is, it measures the amount of variability in the variable of interest that can be predicted from the explanatory variables used in the model. A value of R2 close to 1 indicates that the model explains most of the variability in the variable of interest, while a value close to 0 indicates that the model is not able to explain the observed variability.

In summary, the use of these three-evaluation metrics will allow the selection of the model that best fits the data and provides the most accurate and precise predictions.

Machine learning technique to be used

The machine learning approach selected for this project is neural networks, due to their ability to model complex nonlinear relationships between data characteristics and the target variable. Neural networks can autonomously learn from the data and adjust their parameters to optimize their performance in the prediction task.

The feedforward neural network implemented in the code can learn complex patterns in the input data and can therefore generate more accurate predictions than traditional methods. In addition, the code also provides tools to evaluate the performance of the neural network and plot the results, allowing you to visualize and analyze the results effectively. In summary, if you are working with time series data and want to improve the accuracy of your predictions, this code may be a good option for you.

The use of convolutional neural networks, which have proven to be very effective in vehicular traffic prediction problems, will be considered. These networks are characterized by being able to automatically extract relevant features from the data, which makes them ideal for processing image and signal data, such as traffic speed data.

In addition, the possibility of using recurrent neural networks, which are capable of modeling temporal relationships in the data, will be evaluated, which could improve prediction accuracy over longer time horizons.

The performance of different neural network architectures will be compared and the best one will be selected based on previously defined evaluation metrics. A rigorous training and validation process will be carried out, using cross-validation and hyperparameter fitting techniques to obtain the best possible model.

Conceptual description of the neural model

For the construction of the neural model, it is proposed to use a deep neural network, specifically a convolutional neural network (CNN), because it has proven to be very effective in the prediction of time series and has obtained excellent results in similar problems of vehicular traffic speed prediction.

The input to the neural network will be the data for the features selected in the previous proposal, i.e., current speed, historical speed, historical volume, and time of day. The output of the neural network will be the prediction of the speed at the next measurement point.

The model will be trained using the previously defined training data set and its performance will be evaluated using the validation set. The model parameters will be adjusted using the cross-validation technique in order to avoid over-fitting.

Once the best performing model has been selected, its performance will be evaluated using the test set to obtain an unbiased estimate of the model's performance on new and independent data.

It is important to note that the proposed neural network must be able to handle missing data and outliers in the input data, for which appropriate preprocessing techniques will be used.

Model Description

In this topic, the construction of the fully described neural network model will be presented.

You will be able to observe its total construction from the dataset analysis process to the final predictions.

**Dataset Constitution:**

The model consists of a dataset of 7308 rows x 52 columns, which are constituted by.

* **Capturas:** 7308 Captures of vehicle location for dataset.
* **Fecha:** Date in timestamp format, indicating the date on which the capture was made.
* **Horas:** Time of day when the vehicle locations were captured.
* **Agno:** Year specified, from which the vehicle location capture was recorded for the dataset.
* **Mes:** Month of the specified year, from which the vehicle location capture was recorded for the dataset.
* **Dia\_mes**: Day of the specified month of the year, from which the vehicle location capture was recorded for the dataset.
* **Dia\_sem:** Day of the week of the specified month of the year, for which the vehicle location capture was recorded for the dataset.
* **fest:** Indicator to check if the specified day is a holiday or not.
* **20 - 68**: Tracked vehicle locations obtained by points recorded for the dataset, positions 31, 34, 56 and 63 are ignored, but are assumed to be columns that do not provide the necessary information to the dataset.



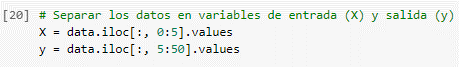
**Dataset Analysis.**

After evaluating the dataset, it was decided to eliminate the "Date", "Dia\_mes" and "Dia\_sem" columns, since they do not provide valuable information to the model.

Texto

Descripción generada automáticamente con confianza media

Next, we separate the training variables into two sets called "X" and "y", with "X" storing the 'Hours', 'Agno', 'Month' and 'Dia\_mes' labels and "y" storing the vehicle locations obtained by points.



They are chosen in this way for a correct training of the model, since the labels go first and then the data to be trained.

**Model training:**

Next, the training of the model is evaluated for the previously analyzed and stored sets, with a training of 70% of the data, 15% for testing and 15% for model evaluation as shown below.



**Model Definition:**

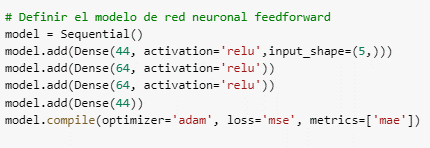
* **Model Architecture:** The model is defined by creating a sequential model object for this feedforwarding neural network.

First, the input layer is constructed by a dense type layer of dimension 44, for the input of the 44 points of locations of each capture made, it has 'ReLU' as the activation function, and has 4 input characteristics of 'y'.

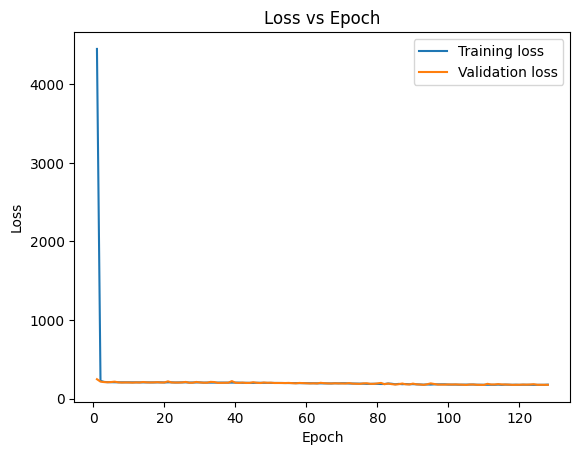
Next, the first and second hidden layer is built by a dense layer of 64 dimension, for the optimal prediction of the velocities of each captured point, it has as activation function 'ReLU'.

Finally, the output layer is constructed by a dense type layer of dimension 44, in order to extract the 44 expected predictor values.

In summary, this code defines a feedforward neural network model with an input layer of 44, 2 hidden layers, 64 and 64 neurons respectively and an output layer with 44 neurons, and the activation function 'ReLU' is used in the input and hidden layers and no activation function is used in the output layer. The model is compiled with Adam optimizer, MSE cost function and MAE metric as seen below.

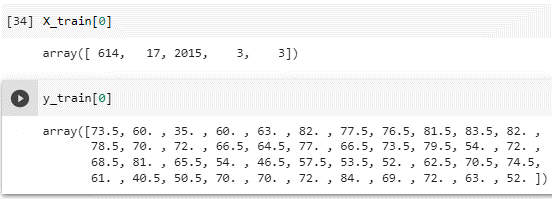


The "MSE" cost function, the "mae" metric and the "Adam" optimizer are used, since they are recommended for the regression prediction cases.

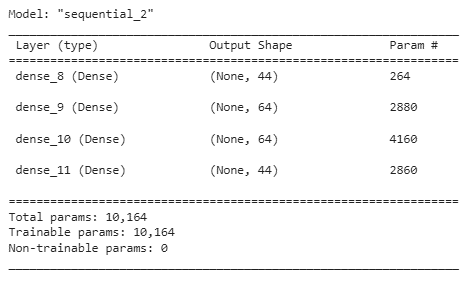
Gráfico, Gráfico de dispersión

Descripción generada automáticamente

**Training data:**



To check that everything is in order, we use the command "model.summary" to display the descriptive information of the model on the screen.



**Model Training:**

To train the model, we use the function "model.fit" to fit the model to the training data, which means that the model will learn to map the inputs of 'X\_train' to the outputs of 'y\_train'.

The variables 'X\_train' and 'y\_train' are the training data used to fit the model.

The argument 'epochs' coming from "model.fit", is the number of times the model will fit the training data, In this case, the model will fit the training data 128 times.

The argument 'validation\_data' coming from the "model.fit": it is data used to evaluate the model during training, in this case, the test data 'X\_test' and 'y\_test' are used as validation data.

Tabla

Descripción generada automáticamente

**Predictions:**

For this example, the prediction used for March 3, 2015 at 5:00 p.m. will be displayed on the screen, in capture 614 of the dataset.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Conclusions

At the conclusion of this case, we can say that the code is an implementation of a feedforward neural network to make speed predictions at a series of point locations. The network is trained with a data set that is divided into a training set and a test set. The training set is used to tune the network parameters, while the test set is used to evaluate the performance of the network on unseen data.

The Pandas and Numpy libraries are used to manipulate and analyze data, and the Keras library is used to create and train the neural network model.

The use of neural networks and deep learning can be a very powerful tool for time series analysis and prediction.

The use of libraries and frameworks such as Keras, Tensorflow, Pandas or NumPy allows efficient and convenient handling of large amounts of data, as well as easy implementation of machine learning models.

The training process of a neural network requires proper tuning of hyperparameters such as the number of layers, number of neurons per layer, learning rate, etc., which may involve some experimentation and testing time to find the optimal configuration.

Monitoring and evaluation of the model is essential to understand its performance and generalizability to new data. Visualization of the loss and error curves at different training epochs can be of great help to identify potential problems and adjust the model accordingly.

Exporting results to formats such as Excel allows for more detailed analysis and easy presentation of the results obtained. The inclusion of graphs and visualizations can facilitate the understanding of the results by end users.

The neural network model is composed of four layers, three of which have a ReLU activation function. The ReLU activation function is used to introduce nonlinearity into the network. The output layer does not have an activation function, since the output is a continuous value and not a probability.

The cost function used to train the model is the mean square error (MSE) and the Adam optimizer is used to minimize the loss function.

Model performance is evaluated using the mean absolute error (MAE) metric.

The evolution of the training and validation error over the epochs is plotted to analyze the model performance. It is observed that the model converges after 128 epochs.

Finally, the prediction values are saved in an Excel file and two graphs showing the evolution of the error over the epochs are included.

In conclusion, from the results obtained, it can be concluded that the implementation of a feedforward neural network model for velocity predictions in a series of point locations can be an effective technique to obtain accurate predictions in this type of problems.

It is important to emphasize the importance of performing a good data preprocessing and adjusting the hyperparameters of the model to obtain the best possible results.

In addition, the evaluation and visualization of the results through graphs is a valuable tool for understanding and improving model performance. Overall, machine learning and artificial intelligence techniques can be powerful tools for solving complex problems in a variety of areas and industries and are expected to remain a constantly evolving area of research and development in the future.

Future Limitations.

Although the developed model performs well in terms of predicting electric power consumption values, there are still some limitations and areas for improvement that can be addressed in future research. Some proposals are presented below:

Consider more variables for the model since the current model is based on only four variables to predict speeds. It would be interesting to consider other variables, such as traffic, route choice, weather, and season, as these variables can significantly affect the process.

Collect more data for the model, since it was trained with a relatively small data set. It would be interesting to collect more data to improve the accuracy of the model. Also, the data set used is limited to only two years, which may not be sufficient to capture variations in locations over time.

Explore different neural network architectures since the current model is based on a feedforward neural network architecture. It would be interesting to explore other architectures, such as convolutional neural networks, recurrent neural networks or the use of LSTM like the Auckland University study and compare their performance with the current model.

Improve the visualization of results, although the current model produces fairly accurate results, the visualization of results could be improved to make them more understandable and accessible to end users.