CCT College Dublin

Assessment Cover Page

To be provided separately as a word doc for students to include with every submission

|  |  |
| --- | --- |
| Module Title: | Big Data Process and Storage & Advanced Data Analytics |
| Assessment Title: | **Neural Networks for forecasting Big Data using Hadoop and PySpark technologies** |
| Lecturer Name: | David McQuaid  Muhammad Iqbal |
| Student Full Name: | Luis Joaquin Morelos Tovar |
| Student Number: | 2023221 |
| Assessment Due Date: | 07 April 2024 |
| Date of Submission: | 12 April 2024 |

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Neural Networks for forecasting Big Data using Hadoop and PySpark technologies

Luis Joaquin Morelos Tovar   
*Advanced Data Analytics & Big Data Storage & Processing CA*CCT College DublinDublin City, Ireland  
email address or ORCID

*Abstract*—In this analysis experimentation with Recurrent Neuronal Network (RNN) and Artificial Neural Networks (ANN) was made to find out which model would work better to predict the number of crashes in the studied area in New York City. The use of Artificial Neural Networks (ANN) is intended to identify the relation between each factor contained in our dataset. The problem consists in finding out how factors such as day of time, season or neighborhood could have an impact to predict the number of crashes and factors surrounding every event. For this analysis a dataset with records from December 2012 until March 2024 was used. This analysis intends to provide useful information to address the crash problematic in the studied New York area using Big Data Technologies such as Hadoop and Apache PySpark.

Keywords—RNN (Recurrent Neural Network), Artificial Neural Network (ANN), predict, crashes, factor, analysis, Hadoop, PySpark, Big Data

# Introduction

Car crashes are an important, not to mention major, problem in the US, just in 2021 car crashes caused about 46, 980 deaths in the U.S., (Ntional Safety Council, 2024) and according to the Forbes Analysis of 2020, a total of 5,250,837 collisions happened over the course of a single year (Christy Bieber, 2023). This has a great impact in US economy causing losses of billions of dollars per year, just to set an example, in 2010 it generated an overall cost to the American society of $277 billion USD (National Center for Statistics and Analysis, 2015).

Ney York city is one of the most important cities in the USA and even though is not consider the one with more crashes in the country, with over 100,000 crashes reported in 2022 (Jay S. Knispel, 2024), it certainly has a major problem of cashes, the reason behind it’s not listed as one of the cities with more car crashes is probably because New York City is the most populated city in the US (United States Census Bureau, 2024)

Vision Zero is a strategy created to eliminate all traffic fatalities and severe injuries (Vision Zero Network, 2024). This program was introduced in Sweden in the year 1997 where it led to a 67% drop in traffic-related deaths (Stephen Stock, 2024). In 2016 the government of the United States implemented the program Vision Zero looking to decrease the number of fatalities in crashes along the whole territory of the US. However, an analysis carried out by CBS News on traffic-related death data from 2016 to 2021 little to no difference in the number of deaths in most cities adopting the strategy was observed, furthermore deaths actually increased in 22 of the 27 cities with Vision Zero programs in place by 2019 (Stephen Stock, 2024)

This is the reason why been able to identify how the factors that causes the crashes are related is so important and can provide with information that can be helpful to develop a strategy that helps to decrease the number of incidents

The work done in this analysis intends to provide information on the relation between the different factors contained in the dataset and give answer to some questions that this problematic presents, question like… *Is there any time of the year where is more likely to be involved in a crash car accident? Is there any specific borough where drivers should be more careful? What’s is the location where more deceases due to car crashes occurs?*  By using ANNs, analysing factors such as street, time of day, season of the year, and some other will be used to train and test an Artificial Neural Network (ANN) to identify where the crashes is more likely to occur.

The results that this analysis provides can be the cape stone to start looking into new ways to approach this kid of problematics, problematics that are not only present in New York City, but in many other cities all over the world.

# Literature Review

A Neural network is a computational model that mimics the shape of a brain and the way it works by linking and nesting multiple algorithms, which are called neurons, in order to learn from the previously input data and afterwards predict a behaviour or find the relation between the input and output values that may not be easily identified by some other computational models or the human eye. In other words

NNs are composed by layers and there are mainly 3 different types of layers:

* Input Layer: This is where the number input data should be defined
* Hidden Layer: Sometimes also called Intermediate Layer, this is a layer where more neurons will be added and increasing the learning capacity
* Output Layer: This is the layer that provides a result, whatever is we are training the NN accomplish

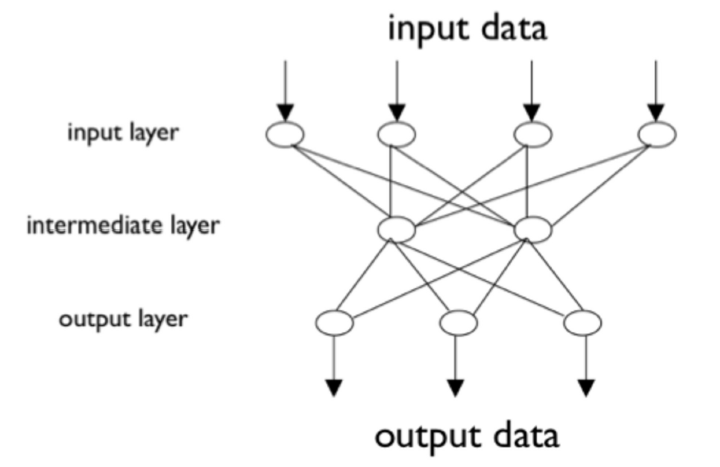


Fig. 1 Neural Network representation (Gallo, 2018)

The human brain, the original and natural Neuron Network is composed billions of nerve cells, called neurons, which are interconnected in a complex network (Gallo, 2018)and interact between each other to process information to help us make decision, such as if it’s safe to cross a street or not, classify the type of animal we have in front of us or simply measure the distance between us and an object so we can move or extend an arm or a leg just the exact distance to reach it.

ANNs, are a type of NN that provide the means for dealing with complex pattern-oriented problems of both categorization and time-series (trend analysis) types. (Steven Walczak, 2003)

On the other hand, a forecasting of the number of crashes using the provided data is important to develop a strategy based on the where is most likely to have an incident, this problem will be address by using RNNs, which is one of the most used Neural Networks to forecast time series.

[RNNs](https://www.sciencedirect.com/topics/computer-science/recurrent-neural-network) were designed to solve the problem of time series in sequential computer data. An RNN's input is made up of this and the preceding samples. (Anubhav Singh, 2022)

The data used in this analysis has been collected by the New York Police Department of Motor Vehicle Collisions, and by the time this analysis has been done it provides data from the year 2012 until the first trimester (March) of the year 2024, the website is been updated constantly so the data will probably change in a soon future.

The amount of traffic data available nowadays is gigantic for this reasons, Big Data Technologies can be used to ease the processing of the data and also because it allows us to retrieve live-time data.

Big Data is a term used to refer to datasets that are so large and are constantly growing that it becomes hard to deal with traditional data processing tools (Nada Elgendy, 2014), the characteristics of Big Data are:

* Volume
* Value
* Variety
* Velocity
* Veracity

In this analysis Hadoop Distributed File System was used as a way to demonstrate how data can be storage in a distributed file system and then access to it by a previously programmed connection

Hadoop is a framework for performing big data analytics which provides reliability, scalability, and manageability by providing an implementation for the MapReduce paradigm. Hadoop consists of two main components: the HDFS for the big data storage, and MapReduce for big data analytics (Nada Elgendy, 2014)

In the figure below a representation of how Hadoop works:

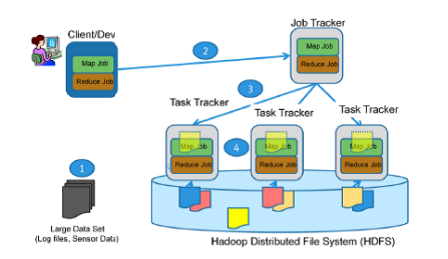


Fig. 2 MapReduce and HDFS

Figure 2 shows how MapReduce and HDFS works, in Step 1 there is a large dataset including log files, sensor data, or anything of the sorts, the HDFS stores replicas of the data, represented by the blue, pink, yellow icons, across the Data Nodes, In Step 2 the client defines and executes a map job and a reduce job on a particular dataset, and sends them both o the Job Tracker. The Job Tracker then distributes the jobs across the Task Trackers in Step 3 the Task Tracker runs the mapper, and the mapper produces output that is then stored in the HDFS file System. Finally in Step 4, the reduce job runs across the mapped data in order to produce the result. (Nada Elgendy, 2014)

Pyspark is the Python API for Apache Spark, and is an open source, distributed computing framework and set of libraries for real-time, large-scale data processing (Harshal S. Kudale, 2022)

Apache Spark is a powerful big data processing platform which adapts the hybrid framework. Even though Spark uses many similar principles to Hadoops MapReduce engine, Spark outperform the latter in terms of performance. For instance, given the same batch processing workload, Spark can be faster than MapReduce due to the ”full in-memory computation” feature used by Spark compared to the traditional read from and write to the disk used by MapReduce. Spark can run in standalone mode or it can be combined with Hadoop to replace MapReduce engine (Eman Shaikh, 2019)

# Problem Description

The problematic of crashes in New York city dates from years ago, and even though the government has implemented strategies loke Vision Zero and spends over 2 billion USD in programs aimed to reduce crashes alongside the country and prevent fatalities due to this factor, the problematic doesn’t seem to have a significative reduction (Stephen Stock, 2024)

This represents an opportunity to propose a model that helps to prevent and decrease the number of events per year

Based on the research done with similar problematics and the use of deep learning techniques, specifically NNs, ANN and RNN were selected to address this problematic due to the performance they have when it comes to forecasting time series and find relation between variables.

# Techniques and tools used

Big data storage and pre-processing

The dataset used was obtain from the New York Police Department of Motor Vehicles Collisions, and is a small representation of what Big Data really is, this data is constantly updated and it keeps growing and growing.

To work with this data a Virtual Machine running Ubuntu 22.04 Jammy JellyFish was used, once running the Ubuntu operating system, we install the open-source software Apache Hadoop to work with a Distributed FileSystem.

The Dataset was then put in the HDFS folder and then read using Pyspark, which is the Python interface from Apache Spark to process Big Data,

It is important to mention that the Hadoop framework was run only in 1 computer, a local node, which has a great impact when computing since all the processing lies on the same processor y and a limited size of RAM memory

The cleaning of the dataset was carryout using Apache PySpark connecting it to Hadoop so it was possible to read the dataset from the hdfs folder, the intend of this cleaning was to prepare the data so it could be used by the ANN and RNN models

During this process a major difference was observed between Python and PySpark, for starters, the command lines are quite different between one and the other, some of them would result easier to understand since they are very straight forward but it can take a bit of time to get used to the difference with this one and the ones in Python.

Another difference observed is that, technically it should be faster to process Big Data but unfortunately the resources and specifications of the hardware made it, not difficult, but slower than processing the data on the normal environment.

Fir this reason, only the cleaning of the data was made on PySpark, later the cleaned data was exported to the hdfs folder in csv format so it could be put on the desktop and then downloaded via flash device.

Even by just cleaning the data some valuable information can be acquire, information that may help to answer the research questions set at the beginning of this work, in the introduction, for example, if we have a look at the image below, we can see that most of the crashes occurs during the afternoon (green) and that morning and evenings have similar quantity of accidents (purple and red), on the other hand small hours represents a minority of accidents.

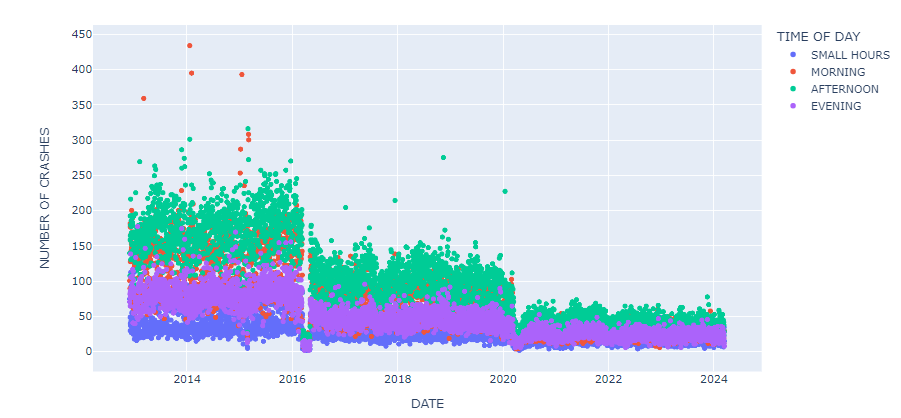


Fig. 3 Distribution of crashes along the day

In another interesting representation we can also see how seasons affect car accidents having a similar behaviour along the years but is also interesting to observe how the number of accidents decrees evenly by year.

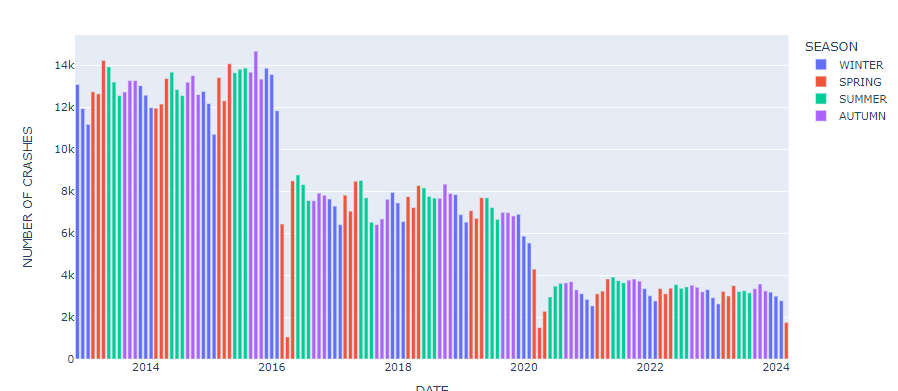


Fig. 4 Distribution by season by year

Prediction using ANN

The first Neural Network model chosen was ANN due to its capacity to predict multiple variables from an input. As discussed before, a neural network is composed by the 3 layers, input layer which is where we decide the size of input input neurons, the size of the input data, In this case the columns in the df that we consider our dependant variables and the output layer, the number of neurons in the output layer will depend on the result we want to get, multiple classification, binary classification, among others.

In this ANN several trails were made but in the end, the input layer consists on 500 neurons, an input x with shape equal to the columns in the df, the relu function activation, this function introduces the property of non-linearity to a deep learning model and solves the vanishing gradients issue. (Krishnamurthy, 2024) and a kernel initializer to prevent the vanishing.

The second layer was the first hidden layer with another 500 neurons to process the data, activation function and kernel initializer were kept the same.

A second hidden layer was added to the architecture, but it was removed since it didn’t prove any improvements in the results.

Last ,comes the output layer, in this experiment we tried to predit the borough where the crash will occur using all the other factors contained in the df, so the output layer has 4 neurons (1 for each borough) and the activation function used was softmax which performs well with multiclassification problems.

Once the ANN was architected, we compile the model and trained it, first we tried compiling the model using Stochastic Gradient Descentt (SGD) because according to the literature it performs well in large datasets because In stochastic gradient descent, instead of processing the entire dataset during each iteration, we randomly select batches of data. This implies that only a few samples from the dataset are considered at a time, allowing for more efficient and computationally feasible optimization in deep learning models. (Gupta, 2024), but optimizer adam prove better results.

The loss function used was Categorical CrossEntropy because it allows us to obtain a multi classification based on the original outputs.

Forecasting using RNN

Time series forecasting is important to machine learning applications in many domains, such as analysis of time series data for video, image, speech, finance, traffic, and disease. There are a variety of approaches to time series forecasting tasks and among them, recurrent neural networks (RNNs) are widely used models for time series forecasting problems. (Xueli Zhang, 2023)

For the forecasting an LSTM RNN was implemented, the reason being the size of the dataset an also because LSTM provides higher memory power to remember the outputs of each node for a more extended period to produce the outcome for the next node efficientl.

In this RNN and because a time series wereforecasted, the data was not splitted evenly like when running the ANN, instead, the forecast was made on the next month based on the last 3 months, and the parameter that was forecasted was the number of crashes that will occur.

Unlike ANNs, this type of Neural Network proved to be more demanding and more complex taking longer to process the input data, but also achieving better results in loss.

In the table below a description of the tool or technique or method is shown and the process carried out with it

| TOOL | Table Column Head |
| --- | --- |
| Hadoop | Big Data Storage (Read & Write data) |
| PySpark | Data Processing (Data Cleaning) |
| ANN | Predict outcome by linking the relation between the variables |
| RNN | Time series predictions |

# Results and Discussion

ANN Results analysis

The intend of this analysis is to forecast the number

of crashes for the next month based on the data provided, but also has the intention of analyze how much could the behavior in car crashes be predicted with all the factors obtained from the dataset, as mentioned before in the Techniques and Tools Used part.

ANN allows us to select from a range of loss functions and optimizers according to what we need to address the problem in front of us and based on the data we have, in this case, 2 optimizers were tested and since it is a multi-classification problem, only categorical\_crossentropy function was tested.

Different tests were done changing number of neurons, number of hidden layers, number of epochs, but the evidence shown is only for the trail that returned better results.

2 dependent variables were tested, ‘Borough’ and ‘Season’. The analysis using ‘Borough’ as the dependent variable is the one that achieve better results; however, this are no really promising.

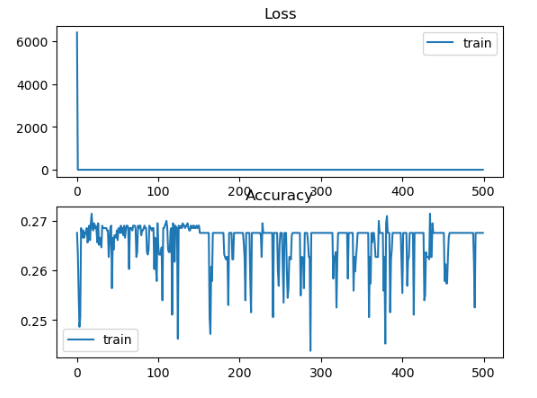


Fig. 5 (a) Loss & Accuracy Season SGD

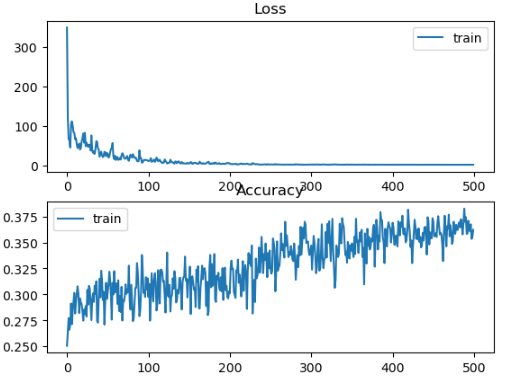


Fig. 5 (b) Loss & Accuracy Season ADAMS

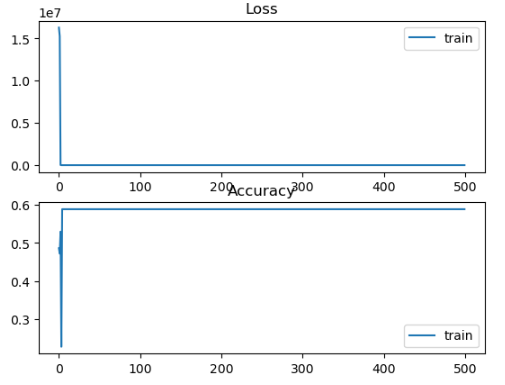


Fig. 5 (c) Loss & Accuracy Borough SGD

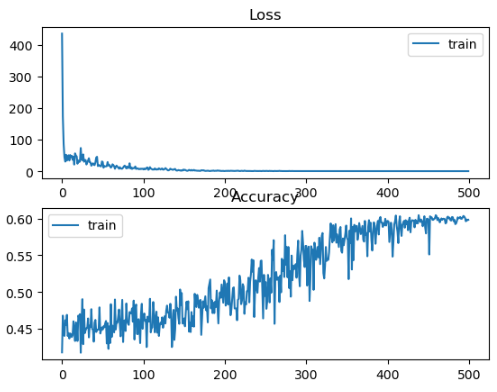


Fig. 5 (d) Loss & Accuracy Borough ADAMS

It can be observed in each graphic the accuracy and the loss for dependent variable depending on which optimizer is being used, and if truth is they look very much different, the reality is that the accuracy after fulfilling the training, is no that different, the values are shown in the table below

| OPTIMIZER | | BOROUGH | SEASON |
| --- | --- | --- | --- |
| SGD | LOSS | 0.9703 | 1.3841 |
| ACCURACY | 0.5894 | 0.2676 |
| ADAM | LOSS | 0.9191 | 1.2869 |
| ACCURACY | 0.6001 | 0.3713 |

Table 2 Comparison Loss & Accuracy

As stated before, there is no much of a difference between the optimizers, but what can be observed in the plots is that optimizer SGD has a more linear behavior, and can learn faster than using optimizer ADAM.

RNN Result Analysis

The use of the Recurrent Neural networks to analyze time series is widely accepted, and it represents a useful tool to forecast time series in a complex way.

The analysis carried out, is intended to provide useful information on future crash events based on the variables that are observed most commonly in every crash in New York City, parameters such as Borough, Season, Street, Contributing or direct factor, are just some of the variables that we can use to predict the number of events expected within a time period.

For this analysis a pre processing stage had to been followed, even with the cleaned dataset, there was still a bit more work to do before the data was actually ready to feed the RNN and expect the output prediction.

The timeframe selected was 1 month predicted using 3 moths before, the reason why 3 months were chosen is because it was intended to work with the seasons of the year, which last 3 months, and because the dataset last record was in march, so all the data was there to be used

Now that the reason behind the selection of the timeframe has been explained, let’s move forward to evaluate the performance of the RNN and the obtained results.

First of all, the training data needed to be created and for that a simple for loop was used fill the X data with 30 days and the Y data with only 1 day, the immediate next one, however, the values are too big and training the data with this values would require more hardware resources so the data was escalated to fit between the range of 0-1.

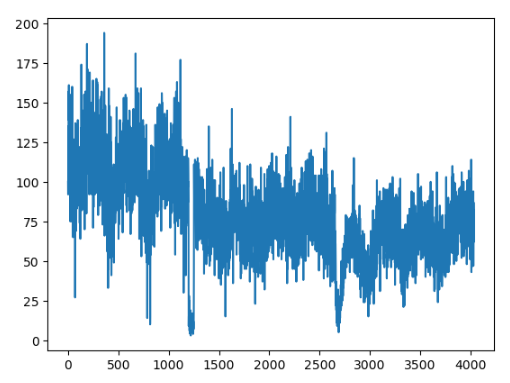


Fig.6 (a) Train data created (X=length df, Y=3 months + 1 month)

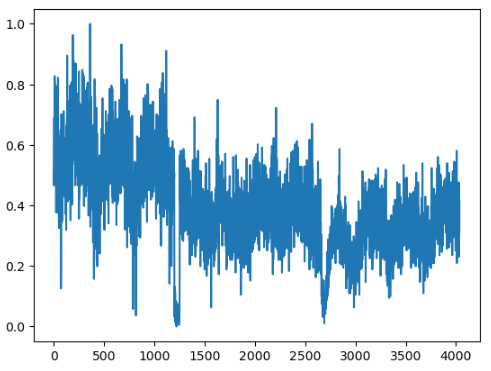


Fig. 6 (b) Trained data escalated 0-1

Once the training data was completed it was time to build the Neural Network. Recurrent Neural Networks are great to analyze time series, but when it comes to a large data series, the Long Term Short Memory (LSTM) RNN is a better fit to it, and therefore this was the type of RNN that was used, originally it was tested with 100 neurons in the input layer, 500 in the first hidden layer, 500 in the second hidden layer and 500 in the third hidden layer, 100 units in the output layer and it was compiled with a batch of 64 and 75 epochs, unfortunately the hardware available was not sufficient and the number of hidden layers was reduced to only 1 hidden layer, so the architecture was set in the following way:

* Input layer 100 neurons
* Hidden layer 100 neurons
* Output layer 100 neurons
* Batch 128
* Epochs 50

The optimizer used was ADAMS which is the most

commonly used and the loss function used was mean\_square\_error since our variables are continuous.

In the image below the behavior of the time series along the selected frame time can be observed, note how it would look like if there was actually a patron, going up and down all through the time period, this so called patron is the one we are trying to predict using the RNN

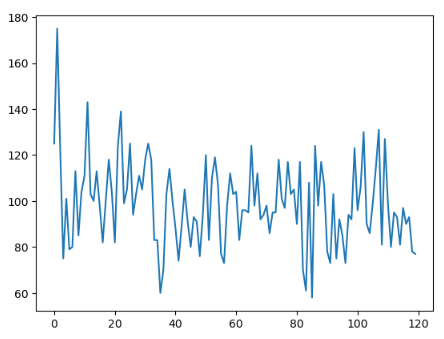


Fig. 7 Testing dataframe plot

Once the prediction has been done we can now compare how does the prediction fits the real data, and we do this by overlapping plots

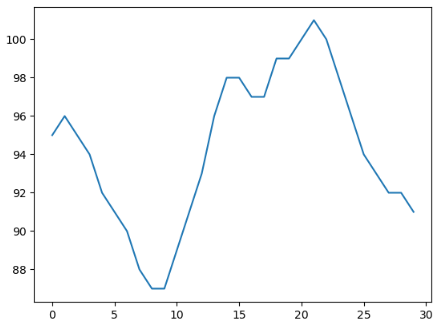


Fig. 8 (a) Predictions for the next 30 days

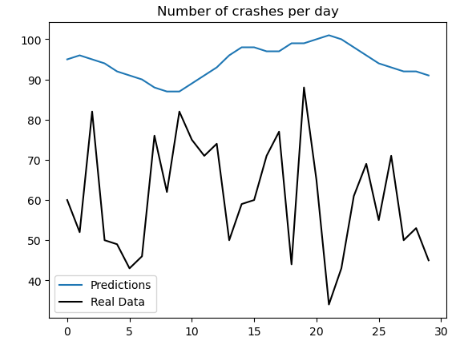


Fig 8 (b) Predictions vs real data

It can be observed in fig, 6 (a) that the predictions are actually following the up and down patron, but accuracy is not perfect. As mentioned before the limitations of the hardware made it difficult to run bigger test, with more neurons, more epochs or more hidden layers, nevertheless it is proved that time series can be analyzed and forecasted using neural networks.

# Conclusions

After the analysis carried out, it can be concluded

that car crashes is indeed a major problem in Ney York City, but it can also be concluded that is not as easy as it would have been thought to solve this problematic, there are many factors than have a direct impact in collisions, on this analysis only 2 where addressed, and even though the results are not as positives as it could have been expected.

On the other hand, a forecasting of the crashes can be partially done and it is expected that this somehow could help the local authorities to develop a new strategy that contributes to drop down the number of crashes in the city, furthermore, if there is a chance to combine this type of analysis with the already implemented strategy VISION ZERO, that could represent a great opportunity to join efforts and help prevent car accidents.

Are the initial questions answered trough this analysis? Probably not hundred percent, but definitely this analysis is on the track to provide an answer to the questions set and to all the questions that could pop out with further and more thorough analysis.

# References

* Anubhav Singh, K. S. V. N. V. A. M. S. S. P. P. P. R. L. C., 2022. Artificial intelligence in edge devices. In: ELSEVIER, ed. *Advances in Computers.* s.l.:s.n., pp. 437-484.
* Christy Bieber, J., 2023. *FORBES ADVISOR.* [Online]   
  Available at: https://www.forbes.com/advisor/legal/car-accident-statistics/  
  [Accessed 09 April 2024].
* Eman Shaikh, Y. A. I. M. I. N., 2019. *Apache Spark: A Big Data Processing Engine,* Khobar: IEEE Middle East and North Africa COMMunications Conference (MENACOMM.
* Gallo, C., 2018. *Artificial Neural Networks Tutorial.* First ed. July: Information Science Reference.
* Gupta, A., 2024. *A Comprehensive Guide on Optimizers in Deep Learning.* [Online]   
  Available at: https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-deep-learning-optimizers/  
  [Accessed 11 April 2024].
* Harshal S. Kudale, M. V. P. P. J. C., 2022. *A REVIEW OF DATA ANALYSIS AND VISUALIZATION OF OLYMPICS,* Pune, Maharashtra: International Research Journal of Modernization in Engineering Technology and Science.
* Jay S. Knispel, 2024. *Law offices of Jay S. Knispel Personal Injury Lawers.* [Online]   
  Available at: https://jknylaw.com/new-york-car-accident-lawyer/statistics/#:~:text=In%202022%2C%20there%20were%20100%2C508,most%20recent%20full%20data%20available.  
  [Accessed 09 April 2024].
* Krishnamurthy, B., 2024. *built in.* [Online]   
  Available at: https://builtin.com/machine-learning/relu-activation-function  
  [Accessed 10 April 2024].
* Nada Elgendy, A. E., 2014. *Big Data Analytics: A Literature Review Paper,* Cairo: Springer International Publishing Switzerland.
* National Center for Statistics and Analysis, 2015. *National Academies.* [Online]   
  Available at: https://trid.trb.org/View/1311862  
  [Accessed 09 April 2024].
* Ntional Safety Council, 2024. *Injury Facts.* [Online]   
  Available at: https://injuryfacts.nsc.org/motor-vehicle/overview/introduction/  
  [Accessed 09 April 2024].
* Stephen Stock, A. C. J. K. J. S., 2024. *Traffic deaths rise in U.S. cities despite billions spent to make streets safer.* [Online]   
  Available at: https://www.cbsnews.com/newyork/news/traffic-deaths-rise-despite-billions-spent-to-make-streets-safer/  
  [Accessed 10 April 2024].
* Steven Walczak, N. C., 2003. Introduction to Artificial Networks. In: *Encyclopedia of Physical Science Technology .* s.l.:s.n., p. 631.
* United States Census Bureau, 2024. *United States Census Bureau.* [Online]   
  Available at: https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork/PST045223  
  [Accessed 09 April 2024].
* Vision Zero Network, 2024. *Vision Zero Network.* [Online]   
  Available at: https://visionzeronetwork.org/about/what-is-vision-zero/  
  [Accessed 09 April 2024].
* Xueli Zhang, C. Z. J. Z. T. W. W. W. N., 2023. *Science Direct.* [Online]   
  Available at: https://www.sciencedirect.com/science/article/abs/pii/S0925231223000462  
  [Accessed 11 April 2024].