The application of artificial intelligence languages in business intelligence for the prediction of COVID-19 data.

Sara EL HABBARI, M'hamed-amine SOUMIAA and Mohamed MANSOURI

LAMSAD, Hassan first university-Settat, National School of Applied Sciences- Berrechid, Morocco
elhabbarisara@gmail.com, medamine17@hotmail.com, mansouri1969@yahoo.fr

Abstract: Since the discovery of artificial intelligence, information systems are constantly evolving, especially decision-making ones. Thanks to AI, today's decision-making tools treat in a more effective manner massive data, real-time analysis, and optimize the decision-making process of companies in the short and medium-term. To further optimize the profits of these companies, it is necessary to enrich their decision-making platforms with predictive visions using some advanced AI technologies. This is what we are going to discuss in detail throughout this article

Keywords: Covid-19, Business Intelligence, Predictive Analysis, Power BI, Artificial Intelligence, Machine Learning, Deep Learning, LSTM, ARIMA, TBATS, ANN, Python, R, RMSE.

1 INTRODUCTION

Business Intelligence (BI) tools have seen real progress over the last decade. Several companies have adopted BI tools to help them better understand their environment and better drive their strategic decision-making.

According to Dresner Advisory Services survey results reported in the American economic magazine Forbes [1], the BI is used at all decision-making levels of companies, especially executive management, finance, sales, and operations. Also, each decision-making level uses its appropriate BI technologies [2].

Analyzing the survey results by technologies [3], we can see that the most frequently used analyses are descriptives or diagnostics. This means that BI tools are mainly used to load, integrate, prepare, and present historical data on dashboards to better understand past events. We can also see that companies do not place predictive, descriptive, and diagnostic analyses at the same level of importance while they are correlated. This observation is confirmed by the results of the IDC France study [4], which reveals a rate of only 9% of French enterprises that have used predictive solutions, even though they are a priority for 26% of business departments. According to the same study, this low rate can be explained by organizational, technical, and human obstacles, especially in large structures.

Despite its current low utilization rate, predictive analysis remains an important part of BI. Thanks to this type of analysis, we will be able to:

- Process large data volumes in real-time.
- Have a vision of what is likely to happen in the future.
- Anticipate the appropriate actions to take to increase profits and avoid crises.

The importance of this type of analysis also lies in its use of advanced artificial intelligence technologies [5], allowing efficient data processing.

Currently, several scientific articles talk about predictive analysis but very few link it to BI. The research studies found along those lines (Ex: [6] - [7]) uses BI tools to mainly

understand the data (correlations, frequencies, etc.) and then exploit the results on predictive analysis independently of the BI tool.

The purpose of this article is to explain how to have these analyses on BI tools in addition to descriptives and diagnostics ones. More precisely, to prove the efficiency of BI tools in terms of data prediction, to explain the approach of integrating this type of analysis on these tools, to test different predictive models including machine learning and deep learning ones, and to qualify the precision of these different models. The application of these points will be done by working on a sample of data of a current important topic which is COVID-19.

2   PREDICTIVE ANALYSIS

Predictive analysis uses techniques such as machine learning and deep learning, which are artificial intelligence (AI) technologies, to predict what is likely to happen. It will never be able to predict the future, but it can examine existing data and establish a probable outcome.Among its application examples, we can mention:

- In the health field: Predict the probability that a patient with specific symptoms will have a heart attack.
- In the sales field: Predict what a customer expects in a given month or quarter, based on his purchase history.
- In the marketing field: Predict the appropriate marketing actions to be launched after collecting and analyzing the data from digital marketing, social media, call centers, mobile applications, etc.

These predictions are made by applying advanced mathematical algorithms that we will see in a few examples in the next chapters. In this article, **we mainly focused on time series prediction problematics.**

3   METHODOLOGY

3.1 Dataset

For this article, we chose to work on **COVID-19** data. The information was extracted from the database of the Center for Systems Science and Engineering (**CSSE**), Johns Hopkins University (**JHU**) [8]. The choice of this dataset was made according to several criteria:

- Use a reliable data source.
- Use a complete global dataset.
- Have diversified data (confirmed, deceased, and recovered cases).
- Have data broken down by several analysis axes specifically the temporal one essential for the type of predictions we want to implement.
- Have data with a daily refresh to stay up to date.
- Have the ability to extract data in a format that can be easily interpreted by a BI tool.

3.2   BI tool

The solution we chose to implement predictive analysis is **Power BI**: a collection of software services, applications, and connectors developed by Microsoft. With Power BI, we

can connect to multiple different sources of data, combine them into a data model, and use this model to build visuals and dashboards. Power BI differs from other tools by its graphical richness, ease of use, and powerful features using advanced programming languages such as Python and R. For all these reasons and more, Power BI is positioned for the past few years as the Leader of BI platforms according to Gartner Magic Quadrant [9]. For this work, we used the free version: **Power BI desktop**.

3.3   Machine learning environment

Before starting the implementation of predictive analysis, it is useful to prepare a machine learning environment by making a set of installations:
- Anaconda.
- R packages.
- Python packages.
- Specific libraries such as NumPy, Pandas, Matplotlib, Keras, Tensorflow, Sklearn, Statsmodels.

These installations are done independently of Power BI, but once connected, the tool automatically detects their locations.

3.4 Data Integration

Once the installations completed, we created a Power BI application and we loaded the COVID-19 data using the following steps:
- Establish connections to data sources: In our case, the data is accessible through a URL. So we used the "Web" functionality of Power BI [10] to get the global statistics of confirmed, recovered, and deceased cases from the CSSE database.
- Prepare data: To facilitate the use of the collected data, some transformations have been done such as formatting dates, transposing data to have them in rows, and not in columns, and renaming fields.
- Create a data model: This means creating links between the three data sources (recovered, confirmed, and deceased cases) to be able to cross them on a single graph or to analyze them using the same analysis axis such as dates, countries, etc.

3.5   Visualization of historical data

Once the data model was established, we created some dashboards, starting with global distributions and then zooming in on Moroccoas shown in Figures 1 and 2.
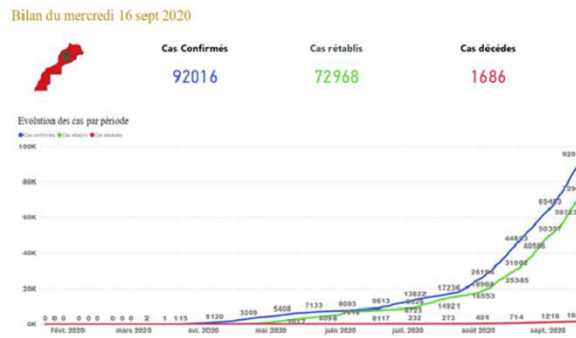

Figure 1: Worldwide distribution of confirmed cases.

Figure 2: Evolution of confirmed, recovered, and deceased cases in Morocco.

At this stage, we have only restituted the historical data to get a clearer idea of the evolution of COVID-19. After analyzing the visuals, a question came to our mind: **What will be the situation in the comingdays, especially in Morocco?** To answer this question, it was necessary to do predictive analyses and this is what we will explain in detail in the next chapter.

4   RESULTS INTERPRETATION

In this section, we will present the results of several predictive analyses realized in different ways on Power BI. For these analyses, **we focused on the confirmed cases in Morocco and we made predictions over 10 days from 09/16/2020**.
We will begin by presenting the results of each analysis. Then, we will qualify the exactitude of each predictive model to find the best one in terms of precision. Finally, we will explain the difference in the results of some models.

4.1   Results of forecast option

On some Power BI graphics, especially line charts, we can add easily a forecast to historical data by using the analytics pane [11]. The Figure 3 presentsa view of the results returned after applying the necessary parameters and the Table 1 illustrates the results of the predicted confirmed cases:

Figure 3: Prediction results of confirmed cases in Morocco using the forecast option of Power BI.

Table 1: Prediction results of confirmed cases in Morocco using the forecast option of Power BI.

Date	Predicted confirmed cases
09/17/2020	93862
09/18/2020	95706
09/19/2020	97551
09/20/2020	99396
09/21/2020	101241
09/22/2020	103086

09/23/2020	104931
09/24/2020	106776
09/25/2020	108621
09/26/2020	110466

The setting of this option is the easier and the fastest one compared to the models that we will going to present in the following sections.

4.2   Results of the prediction models using R

In addition to the graphics available on Power BI, it is possible to import other ones with advanced functionalities from the Microsoft AppSource [12]. The graphics imported and used in this work are:

- Forecasting with ARIMA: This type of graphic applies one of the most commonly used methods for time series forecasting, ARIMA (Auto-Regressive Integrated Moving Average).
- Forecasting TBATS: (T: Trigonometric seasonality; B: Box-Cox transformation; A: ARIMA errors; T: Trend; S: Seasonal components) one of the models commonly used in time series forecasting and more specifically for series with complex seasonal models.
- Forecast using Neural Network by MAQ Software: Neural networks are based on advanced data training and learning algorithms. They are generally recognized by their performance and their ability to return results that are more or less close to reality.

To use these visuals, we have to import them [13] and apply the necessary settings according to our analysis needs. Once the settings are applied, the graphics automatically run their appropriate algorithm on R and return the results asshown in Figures 4 ,5,6, and Table 2.

Figure 4: Prediction results of confirmed cases in Morocco using the ARIMA model.
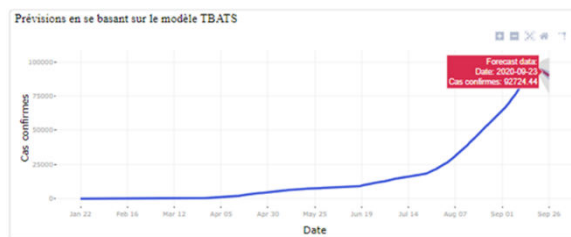

Figure 5: Prediction results of confirmed cases in Morocco using the TBATS model.
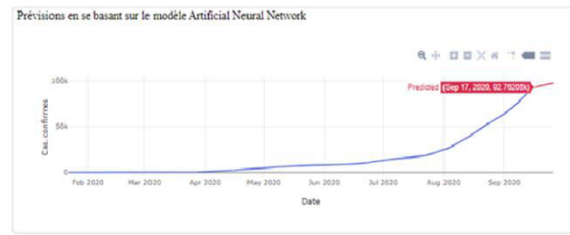
Figure 6: Prediction results of confirmed cases in Morocco using the ANN model.

Table2: Prediction results of confirmed cases in Morocco using the models: ARIMA, TBATS and ANN.

Date	ARIMA	TBATS	ANN
09/17/2020	92016	93246	92762
09/18/2020	93895	94087	93459
09/19/2020	95774	94520	94108
09/20/2020	99532	94547	94712
09/21/2020	101412	94207	95274
09/22/2020	103291	93570	95795
09/23/2020	105170	92724	96278
09/24/2020	107049	91767	96725
09/25/2020	108928	90792	97139
09/26/2020	110807	89877	97521

4.3 Results of the prediction models using Python

While analyzing the results returned by the ARIMA, TBATS, and ANN models using R, a question came to our mind: **If we redo one of these models on Python, will we get the same results?** To answer this question, we chose to reimplement the ARIMA model and the neural network on Python by following the steps below:
- Realize the Python code of each model on the Jupyter Notebook and test the results referring to the tutorials [14] [15].
- Integrate the codes on Power BI [16] and link the Python prediction data to the existing one.
- Create new line charts on Power BI to visualize predicted data.

Note: The neural network we have chosen to implement is called LSTM (Long Short Term Memory). This prediction model is distinguished by its intelligent and efficient use of memory during calculations, which allows it to return results closer to reality.

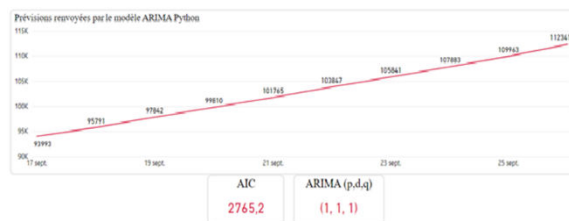The Figure 7,8, and the table 3 presents a view of the results:

Figure 7: Prediction results of confirmed cases in Morocco using the ARIMA model implemented on Python.
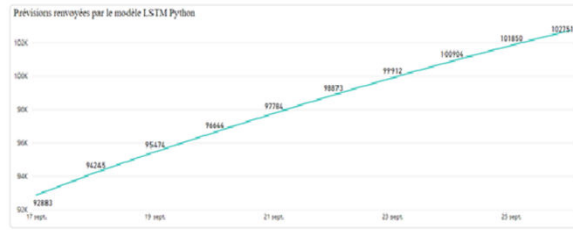
Figure 8: Prediction results of confirmed cases in Morocco using the LSTM model implemented on Python.

Table3: Prediction results of confirmed cases in Morocco using the LSTM and ARIMA models implemented on Python.

Date	ARIMA	LSTM
09/17/2020	93993	92883
09/18/2020	95791	94245
09/19/2020	97842	95474
09/20/2020	99810	96644
09/21/2020	101765	97784
09/22/2020	103847	98873
09/23/2020	105841	99912
09/24/2020	107883	100904
09/25/2020	109963	101850
09/26/2020	112341	102751

5   QUALIFICATION OF THE PREDICTIVE MODELS

Given the difference between the results obtained in Figure 9, it was necessary to qualify the precision of each model to identify the most reliable.
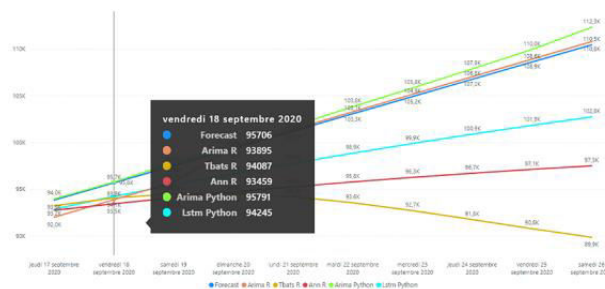

Figure 9: Results of the different predictive models tested on Power BI.

The evaluation of predictive models is done in several ways, such as calculating evaluation indices like the root mean square error (RMSE). The lower the RMSE value, the better is the model evaluated in terms of precision. The formula of the RMSE is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i)^2}$$

Avec :

-RMSE, la racine de l'erreur moyenne quadratique (dimension de X),
-n, le nombre d'observations (sans dimension),
-ϵ_i, l'écart entre chaque observation *i* et mesure *i* (dimensions de X),

The application of the RMSE requires actual and predicted data for the same period. So we waited a few days to get the real statistics, then we used Excel to calculate the RMSE. The Table 4 illustrates the results by model:

Table4: RMSE results.

Prediction method	RMSE
"Forecast" option	2812
ARIMA R	3073
TBATS R	14095
ANN R	10551
ARIMA Python	1940
LSTM Python	7367

The lowest RMSE corresponds to the **ARIMA Python**. We can then consider that the predictive analyses of this model are the most reliable.

6   DIFFERENCE BETWEEN THE RESULTS OF THE SAME MODEL

Among the points that drew our attention when implementing predictive models, the difference between results generated by Python and R for the same model. This is the case of the ARIMA model for example. This difference leads us to examine in more detail the R source code of the imported graphic "Forecasting with ARIMA" [17] and to compare it to the one we have implemented on Python.

After examining the codes, it turned out that we do not use the same functions or the same approach to build the model. On the R script, for example, the identification of the optimal parameters to fit the ARIMA model was done using the "auto.arima" function. The identification is managed automatically by this function, which is not the case on the Python script we developed. On our code, we generated all the possible combinations of the parameters, and we chose the one that returns the lowest AIC index using dedicated Python syntax and not a predefined automatic function.

We can then conclude that the implementation of prediction models on python certainly requires more effort than the use of the ready-to-use models proposed by Power BI, but allows us to have better predictions and this is what the result of the RMSE also confirms.

7   CONCLUSION

Through this article, we have proven that with the same BI tool we can have both historical and predictive analyses. By combining these two types of analyses, we were able to have an idea not only on the evolution of covid-19 in Morocco in the previous months but also on the risks we can face in the coming days. It has also been shown that with predictive models using artificial intelligence technologies, we can have better results in terms of data prediction.

What has been applied on the COVID-19 dataset is of course valid for any type of data that can be represented in a time series. Furthermore, the setting used for the days is also valid for the months and the years if we wish to extend the prediction area further.

Finally, it is important to talk about the added value of this work in terms of cost: using a single BI tool for historical and future analysis costs less than using several tools, each one of them specialized in a particular type of analysis. Costs can also be better managed by applying optimization actions deduced from the predicted data.

REFERENCES

[1] Louis Columbus (2020), What You Need To Know About BI In 2020, Forbes magazine, Chart "FunctionsDriving Business Intelligence".

[2] Louis Columbus (2020), What You Need To Know About BI In 2020, Forbes magazine, Chart "Technologies and Initiatives Strategic to Business Intelligence Objetives by Industry".

[3] Louis Columbus (2020), What You Need To Know About BI In 2020, Forbes magazine, Chart "Technologies and Initiatives Strategic to Business Intelligence".

[4] Thierry Hamelin (2019), Du Big Data aux modèles prédictifs : où en sont les entreprises françaises.

[5] Calum McClelland (2017), The Difference Between Artificial Intelligence, Machine Learning, and Deep Learning, Leverege.

[6] M.-L. Ivan and M. R ˘aducu TRIFU (2016), "Using Business Intelligence Tools for Predictive Analytics in Healthcare System," IJACSA) Int. J.Adv. Comput. Sci. Appl., vol. 7, no. 5, 2016.

[7]M.RAJARAJESWARI, NIRANJANA.SV (2019) ,"CUSTOMER BEHAVIOR ANALYSIS AND PREDICTION",SSN: 0374-8588Volume 21 Issue 14, December 2019.

[8] JHU CSSE (2019), COVID19/csse_covid_19_data/csse_covid_19_time_series at master · CSSEGISandData/COVID-19 · GitHub

[9] Gartner Magic Quadrant for Analytics and Business Intelligence Platforms (2020), 2020 Gartner Magic Quadrant | Power BI (microsoft.com)

[10]Microsoft documentation (2019), Connect to a webpage from Power BI Desktop - Power BI | Microsoft Docs

[11] Microsoft documentation (2020), Use the Analytics pane in Power BI Desktop - Power BI | Microsoft Docs

[12] Microsoft AppSource (2020), Microsoft AppSource – l'emplacement pour les applications métier

[13] Microsoft documentation (2018), Utiliser des visuels Power BI basés sur R dans Power BI - Power BI | Microsoft Docs

[14] Thomas Vincent (2017), A Guide to Time Series Forecasting with ARIMA in Python 3, ARIMA Time Series Data Forecasting and Visualization in Python | DigitalOcean

[15] Ian Felton (2019), A Quick Example of Time-Series Prediction Using Long Short-Term Memory (LSTM) Networks, A Quick Example of Time-Series Prediction Using Long Short-Term Memory (LSTM) Networks | by Ian Felton | The Startup | Medium

[16] Microsoft documentation (2020), Run Python Scripts in Power BI Desktop - Power BI | Microsoft Docs

[17]Boris Efraty (2018), powerbi-visuals-forcastingarima/script.r at master · microsoft/powerbi-visuals-forcastingarima · GitHub