Business Intelligence & Data Visualization for the Assessment of the Economic Situation Based on SVM & K-NN

Mohamed-Ali Ksouri and Sameh Souli

September 8, 2023

Business Intelligence & Data Visualization for the assessment of the economic situation based on SVM & K-NN

Mohamed-Ali Ksouri[1], and Sameh Souli [2-3]

[1] Esprit School of Business, Industrial Zone Chotrana II, B.P. 160 Technological Pole
El Ghazela, Ariana 2083, Tunisia.
[2] Faculty of Sciences of Tunis, University Tunis El-Manar, 2092 Tunis, Tunisia
[3] Esprit School of Business, Industrial Zone Chotrana II, B.P. 160 Technological Pole
El Ghazela, Ariana 2083, Tunisia.

Abstract. In this paper, you will find explanations, described steps, and definitions of key concepts, which have contributed to the final results of this Business Intelligence project and Data science research.

This work has required techniques, methods, technologies and disciplines of computer science, including Business Intelligence (B.I), which has the role of converting raw data into meaningful decision-making information through data collection, integration and cleaning operations, unique data storage and organization techniques outside the traditional and real-time data analysis and visualization.

In sync with B.I, our second major discipline of computer science, Data Science, is more focused on extracting knowledge from data, using statistical techniques and advanced algorithms to solve complex problems and discover hidden patterns and trends.

Keywords: Business Intelligence, Data Science, Data Mining, Dashboard, Extract-Transform-Load, KPI, Consulting, Datawarehouse, linear regression, ARIMA, SVM, KNN, K-means, XGBOOST, Power BI, MSBI, SSIS, f-1 score, Data-Viz, Data Analytics, Staging Area, Operational Data Store.

1 Introduction

This Scientific paper hovers over a project that consists of implementing different types of tasks, all operation around data, its integration, cleansing, querying and analyzing, in order to prepare visual presentations, aggregations, decision model and detailed analysis of our attributes. There will be carefully tailored machine learning algorithms to look for the fits, parameters and combinations in order to; either predict future behavior of the intended financial variables, or to look for deep patterns that further explain the fluctuations, or to carry out the performance results of different classification models by comparing between their different metrics such as precision, and indexes. **[1] [2]**

Our full data is going to consist of financial ratios and indicators used in the analysis of the economic climate of Tunisia and for this we used the Richard Bull “Financial Ratios” Book to choose our ratios **[7]**

This paper is organized as follows: Section 2 reviews previous or related works.

Section 3 is going to be about Business Intelligence, 4 Data visualization, 5 Data Science operations and lastly, we are going to evaluate the results in addition to some relevant discussions.

2 Review of related works

Prior to the analysis phase, it is very common to run ETL (Extract, transform and load) operations in order to clean and organize our data sources according to the needs of the endeavor.

Speaking in terms of finance now the collection of the indicators has been sought to fit the description of the financial indicators of Bull Richards in his book Financial Ratios **[7]**, so we need to know that according to his words; A ratio that is too meaningful may be too volatile to be useful. A ratio that is too stable may not provide enough information about the company's financial health.

Many use data warehousing techniques in order to load them with cleansed data so that it encloses two business goals; a Business Intelligence datawarehouse for OLAP (On-line Analytical Processing) and the main machine learning and data mining goals that takes data fragments from the datawarehouse as input source.

In addition, Collaboration between domain experts, data scientists, and finance professionals is essential for successful implementation and deployment of financial machine learning projects, which are similar to the ones used in the Financial Whirlpool, A System Story of the Global Recession of KAREN L. HIGGINS.**[8]**

3 Business Intelligence

3.1 Data Integration

This part of the process is covering ETL operations and analysis-oriented data storage management.

It has been done with the help of the data integrator tool SSIS (SQL Server Integration Services) of the Microsoft BI suite.

The aim of this part is to start from the dirty data and overlapped database structure to reach a predefined standard formatting across all fields of the same types, these types that are also predefined.

3.2 Data modeling

In this part, we are going to implement the decisional
database model by doing the necessary merges and joins, the establishment of the primary and foreign keys by setting the cardinalities for the dimensions, that are arranged

in a star schema mount type around the Fact table. It was a bottom-up Kimball approach that was opted for the creation of the dimensions. **[3]**

The final product of this step of the process enlarges the possibilities of the data querying and adds more possibilities for conditioning and filtering thanks to the star schema that makes all of the tables inter-connected. **[5]**

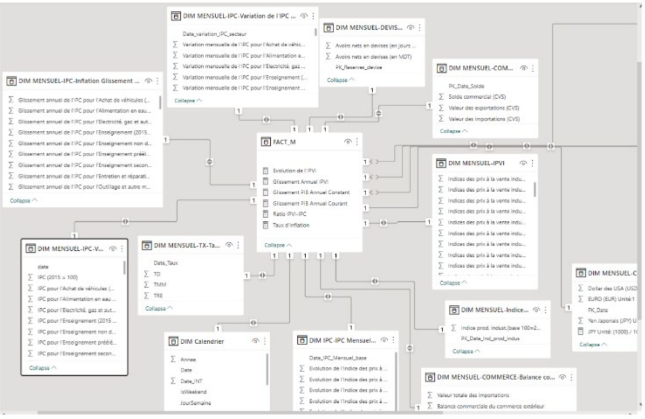

DIM MENSUEL-IPC-Variation de l'IPC ...
Date_variation_IPC_secteur
Variation mensuelle de l'IPC pour l'Achat de véhic...
Variation mensuelle de l'IPC pour l'Alimentation e...
Variation mensuelle de l'IPC pour l'Electricité, gaz ...
Variation mensuelle de l'IPC pour l'Enseignement (...
Variation mensuelle de l'IPC pour l'Enseignement ...
DIM MENSUEL-DEVIS...
Avoirs nets en devises (en jours ...
Avoirs nets en devises (en MDT)
PK_Reserves_devise
DIM MENSUEL-COM...
PK_Date_Solde
Solde commercial (CVS)
Valeur des exportations (CVS)
Valeur des importations (CVS)
DIM MENSUEL-IPC-Inflation Glissement ...
Glissement annuel de l'IPC pour l'Achat de véhicules (...
Glissement annuel de l'IPC pour l'Alimentation en eau...
Glissement annuel de l'IPC pour l'Electricité, gaz et aut...
Glissement annuel de l'IPC pour l'Enseignement (2015...
Glissement annuel de l'IPC pour l'Enseignement non d...
Glissement annuel de l'IPC pour l'Enseignement préél...
Glissement annuel de l'IPC pour l'Enseignement secon...
Glissement annuel de l'IPC pour l'Entretien et réparati...
Glissement annuel de l'IPC pour l'Outillage et autre m...
FACT_M
Evolution de l'IPVI
Glissement Annuel IPVI
Glissement PIB Annuel Constant
Glissement PIB Annuel Courant
Ratio IPVI-IPC
Taux d'inflation
DIM MENSUEL-IPVI
Indices des prix à la vente indu...
DIM MENSUEL-IPC-V...
date
IPC (2015 = 100)
IPC pour l'Achat de véhicules (...
IPC pour l'Alimentation en eau ...
IPC pour l'Electricité, gaz et aut...
IPC pour l'Enseignement (2015 ...
IPC pour l'Enseignement non d...
IPC pour l'Enseignement préél...
IPC pour l'Enseignement secon...
DIM MENSUEL-TX-Ta...
Date_Taux
TD
TMM
TRE
DIM MENSUEL-C
Dollar des USA (USD
EURO (EUR) Unité 1
PK_Date
Yen Japonais (JPY) U
JPY Unité: (1000) / 1
DIM MENSUEL-Indice...
Indice prod. indust.(base 100=2...
PK_Date_Ind_prod_indus
DIM Calendrier
Annee
Date
Date_INT
IsWeekend
JourSemaine
DIM IPC-IPC Mensuel...
Date_IPC_Mensuel_base
Evolution de l'Indice des prix à ...
DIM MENSUEL-COMMERCE-Balance co...
Valeur totale des importations
Balance commerciale du commerce extérieur
Collapse
1

Fig. 1: Star Schema

3.3 Data analysis

This is the fundamental analysis part that serves as a first insight of the collected data. It allows the visual interpretation of the attributes in relation to one another, and permits their aggregation through what we call a Key Performance Indicator, that basically is an illustration of an arithmetical formula that references values that indicates about a certain key goal **[9].**

Data Analysis Expressions language, known as DAX is a data-oriented language that serves as one of our analysis tools, it contains many predefined functions that helps in shaping more the insights.

4 Data Visualization

In this part, we proceed, keeping in mind the specifications of the project, by finalizing the import of the datawarehouse database in the visual interpreter of the data that is Power BI.

Then pages of dashboards according to each subject is made following a significant color choice and template to compare financial indicators such as the GDP and its different structures, inflation rate, the index of the industrial production, the consumer price index, industrial selling price index, the total debt ratio, currency exchange rates, policy rate, average currency market rate, savings yield rate.

Best practices for these visualizations have been inspired from the 2010 paper by Heer & Bostock **[13]** and also some previous works have been considered in order to opt for the most convenient graphics and visuals **[14].**

5 Data science

The next topic is going to be about the data science operations and in this part, we will be discussing the specifications for these two algorithms and the difference between both in the process of classifying a dataset containing the evolution of these four attributes: PIB (GDP), TMM (currency exchange rates), TD (policy rate), TRE (the total debt ratio).

5.1 K-Nearest Neighbors

According to the book: Introduction to Machine Learning with Python by Andreas C. Müller and Sarah Guido [15], this method was first developed by Evelyn Fix and Joseph Hodges, this supervised learning algorithm offers the options of a linear regression kernel or a classification kernel, in this project, we are going to turn our attention to the mathematical function classification kernel.
First the model takes in the number of neighbors which is 4 in our case. Then, it calculates the distance between x and xi, “x” being the data point in question and “xi” a data point in the training dataset **[15]**.

Secondly, sort the distances in ascending order and select the k data points with the smallest distances and lastly Predict the target value for the new data point x as the average (or weighted average) of the target values of the k-nearest neighbors.
After the standard scaling and classification process with the k-means algorithm we proceeded to a KNN model for the PIB and TMM data, and with a 4 number of neighbors and the employment of the Euclydian distance calculations just like did Trevor Hastie, Robert Tibshirani, and Jerome Friedman in their "A Tutorial on k-Nearest Neighbors Classification". **[16]**

$$d(x,\ y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

Fig. 2: Euclydian distance **[16]**

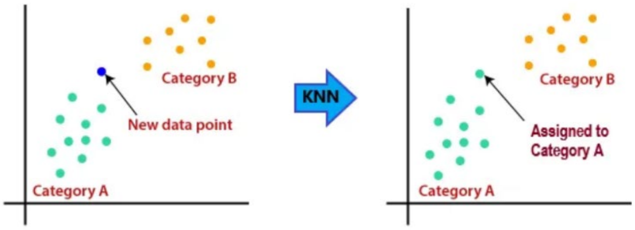


Fig. 3: KNN **[12]**

5.2 Support Vector Machine

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for both classification and regression tasks. Its main objective is to find the optimal hyperplane that best separates the data points of different classes in a high-dimensional feature space.

Same variables for this model, this algorithm starts by a linear separation by finding hyperplane that maximizes the margin between the two classes the nearest data points to the hyperplane form the support vector. In conclusion, SVM aims to maximize the margin while minimizing the classification error and once the SVM is trained, it can be used to classify new unseen data points.

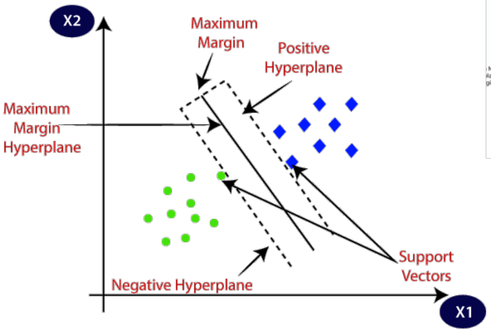


Fig. 4: SVM **[11]**

5.3 Comparison between K-NN and SVM

In reference to previous implementation by data scientists it is clear that SVM has a strong kernel and is a more complex classifier, especially when it comes to non-linear data, in addition, its robust kernel offers less execution time in comparison to its contestant the K-NN algorithm that when the volume of the dataset relatively goes up, its accuracy rate goes down.

In terms of memory usage and system build time, both models perform equally at a first glance of a small dataset, then there is a noticeable difference that starts to develop by the growth of the dataset volume, for instance: for a 2 attributes 1000 records length dataset, build times are 3.273 and 9.155 respectively for SVM and KNN **[10].**

In nature these two algorithms differ according to different criterias as mentioned in this table below:

Tab. 1: SVM & KSS Comparison

Criteria	SVM	KNN
Memory & build time	0.0642s, requires less memory, as only support vectors are stored	0.527s Requires more memory, as it stores all training data points
dataset size	Efficient with large and small dataset volumes	Suitable for small to medium-sized datasets
Hyperparameters	Tuning hyperparameters is required for optimal performance	Only k (the number of neighbors) needs tuning
Overfitting	Less prone to overfitting due to the margin maximization	More prone to overfitting, especially with small k

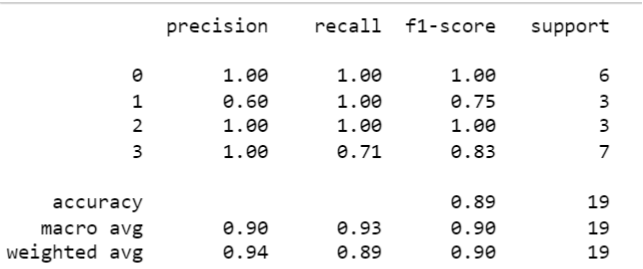
6 Classification results & discussion

6.1 Dataset

As mentioned earlier, the dataset is a composition of the three main ratios of the Tunisian Central Bank and we are going to be interested in the TMM and the PIB variables which are two labelled attributes consisting of real numbers

6.2 Experimental results

For the K-NN the results were the following: only 2 confusions in the class predictions out of 19 and an f-1 score of 0.89.



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         6
           1       0.60      1.00      0.75         3
           2       1.00      1.00      1.00         3
           3       1.00      0.71      0.83         7

    accuracy                           0.89        19
   macro avg       0.90      0.93      0.90        19
weighted avg       0.94      0.89      0.90        19
```

Fig. 5: f-1 score K-NN

Underneath is shown a figure of the confusion matrix as results of the execution of the k-NN algorithm.

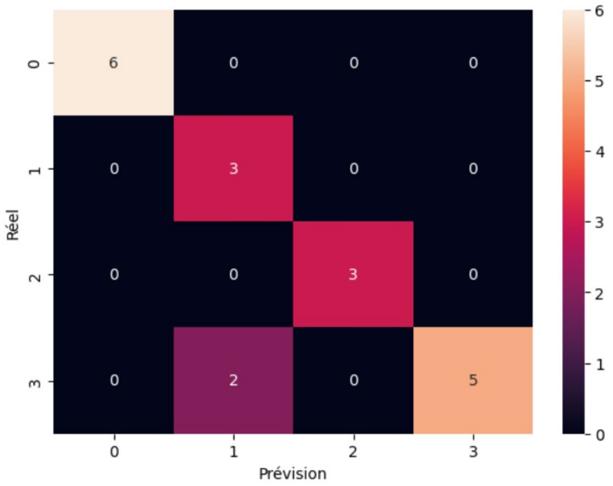
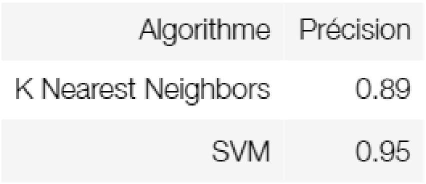


Fig. 6: K-NN Confusion Matrix

For the SVM, our parameters were the following; kernel=poly, polynomial degree=3, regulation rate=100, that resulted in a precision score = 0.95.

In conclusion, these are the results



Algorithme	Précision
K Nearest Neighbors	0.89
SVM	0.95

Fig. 7: Final results

6.3 Discussion

From a financial perspective, these two ratios seem to be fluctuating together every certain period of time, so to surround the mean periods of times, we did the clustering of the dataset depending on k-means generated classes according to differences between numbers in order to get the classification's experimental results mentioned above.

Our data source has been the website of the
Tunisian Central Bank:
(https://www.bct.gov.tn/bct/siteprod/stat_index.jsp)

From a technical perspective now, we can notice the clear outperformance of the SVM model in terms of precision and results due to first its parameters options and second to its powerful kernel with complex mathematic formulas.

Conclusion

In conclusion the results of the classifications have good scores but we can also optimize by switching over to advanced techniques such as the ones used in Nikolaos Passalis and Avraam Tsantekidis paper on Price Trailing for Financial Trading Using Deep Learning **[17]**.

With deep learning process, neural networks offer more flexibility in the selection of different parameters and their different weights. It also offers more advantages that make the data and model fit more appropriate especially in the case of the introduction of more attributes that opens the way for a new type of operations.
These operations could be fraud detection, stock price evaluation or return on investment predictions, since we already have the ratios that take part in all these aspects.

References

1. Data Warehousing: Design, Development, and Management by Ralph Kimball and Margy Ross (2nd edition, 1998)
2. Vapnik, V., Chapelle, O.: Bounds on Error Expectation for Support Vector Machines.
Journal Neural Computation 12, 2013–2036 (2000)
3. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling by Ralph Kimball, et al. (3rd edition, 2008)
4. Hsu, C.-W., Lin, C.-J.: A comparison of methods for multi-class support vector machines.
J. IEEE Transactions on Neural Networks 13, 415–425 (2002)
5. Data Preparation for Analytics: A Practical Guide by Michael Wittmann (2016)
Data Cleaning: A Practical Guide by Cathy O'Neil (2016)
6. Panos Vassiliadis, University of Ioannina, Greece on A Survey of Extract–Transform–Load Technology "International Journal of Data Warehousing & Mining, 5(3), 1-27, July-September 2009"
books
7. Richard Bull "Financial Ratios" Book, 1st Edition - October 26, 2007
8. Karen L. Higgins "Financial Whirlpools, A Systems Story of The Great Global Recession", 1st Edition - March 26, 2013
9. James F. Shortle and Jeffrey M. Theiler, The Effects of Data Aggregation in Statistical Analysis, 1976 Edition
10. J. S. Raikwal and Kanak Saxena "Performance Evaluation of Svm And K-Nearest Neighbor Algorithm Over Medical Data Set, International Journal of Computer Applications (0975 – 8887) Volume 50 – No.14, July 2012
11. Https://Www.Javatpoint.Com/Machine-Learning-Support-Vector-Machine-Algorithm, 2023.
12. Https://Www.Javatpoint.Com/K-Nearest-Neighbor-Algorithm-For-Machine-Learning, 2023.

13. Heer, J.; Bostock, M.; Ogievetsky, V. “Ieee Transactions on Visualization and Computer Graphics” 2010

14. Abdulalem Ali 1, Shukor Abd Razak 1,2, Siti Hajar Othman 1, Taiseer Abdalla Elfadil Eisa 3, Arafat Al-Dhaqm 1, Maged Nasser 4, Tusneem Elhassan 1, Hashim El-shafie 5 And Abdu Saif “Financial Fraud Detection Based 0n Machine Learning: A Systematic Literature Review” Published: 26 September 2022

15. Introduction To Machine Learning with Python by Andreas C. Müller And Sarah Guido, 1st Edition October 2016

16. Trevor Hastie, Robert Tibshirani, And Jerome Friedman, "A Tutorial On K-Nearest Neighbors Classification" Published November 2018

17. Avraam Tsantekidis, Nikolaos Passalis,” Price Trailing for Financial Trading Using Deep Learning” Published in July 2020

18. Slamet Wiyono, Dega Surono Wibowo, M. Fikri Hidayatullah and Dairoh, “Comparative Study of KNN, SVM and Decision Tree Algorithm for Student’s Performance Prediction”, International Journal of Computing Science and Applied Mathematics, Vol. 6, No. 2, August 2020.