International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-10, Issue-2, March 2022
https://doi.org/10.55524/ijircst.2022.10.2.4
Article ID IRPV1025, Pages 19-28
www.ijircst.org

A Framework for Voting Behavior Prediction Using Spatial Data

Shobhit Kumar Ravi[1], Shivam Chaturvedi[2] Dr. Neeta Rastogi[3], Dr. Nikhat Akhtar[4], and Dr. Yusuf Perwej[5]

[1] B.Tech Scholar, Department of Computer Science & Engineering, Ambalika Institute of Management & Technology, Lucknow, India
[2] Assistant Professor, Department of Computer Science & Engineering, Ambalika Institute of Management & Technology, Lucknow, India
[3, 5] Professor, Department of Computer Science & Engineering, Ambalika Institute of Management and Technology, Lucknow, India
[4] Associate Professor, Department of Computer Science & Engineering, Ambalika Institute of Management & Technology, Lucknow, India

Correspondence should be addressed to Shobhit Kumar Ravi; yusufperwej@gmail.com

ABSTRACT- The greatest method to anticipate the future is to look at what has happened in the past. We shall present important election behavioral predictions in this paper. This study article will focus on the data offered by Present age-wise voting statistics, voter demographics, votes cast, and spatial correlation among surrounding states in order to validate that a place's exit poll data. The major goals of our paper are to first encourage voting among different age groups based on projected circumstances, and then to understand the influence of a state's neighbours. Conclusively studying the entire voting scenario of previous years, which will aid in the forecast of citizens' voting behavior in the approaching years, as well as recognizing the root cause of the weaker portions and improving upon the flaws for a better future. Our main goal is to use some current voting data from a region to train and determine the major voting population in the various states of the United States based on their geographical influence. This will aid in the analysis of the current situation as well as assisting the government in creating awareness in places where it is missing.

KEYWORDS- Spatial Analysis, Voter Turnout, Data Analytics, ARIMA Model, Recurrent Neural Network (RNN), Time Series Forecasting.

I. INTRODUCTION

New mathematical, graphical, and computational models with the ability to evaluate data and obtain relevant information are urgently required. Computer scientists and political analysts can learn from one other if they grasp the value of data and data analysis. To be useful, data gathered from voters and residents must be examined. In a democratic society [1], one of the common decision-making methods that have become the topic of many studies is "voting." Voting is a method of determining how a group of people's choices influence collective decision-making. Representative selection is a highly unpredictable process. In political science, the representative opinion survey has long been the apex of empirical research [2]. Researchers may now investigate human behavior on a whole new scale using traces left behind by our digital interactions, thanks to the recent massive development in digital platforms [3]. Political studies involving millions of individuals have evolved within the subject of computational social science, replacing surveys with a few thousand respondents and yielding crucial new knowledge about our digital and analogue lives. Scholars have demonstrated the possibility for predicting election outcomes based on digital data from a variety of platforms, including YouTube [4], Google [5], Twitter [6], and Facebook [7] in the subfield of election forecasting. The revealed preference theory of non-market interactions includes a section on voting decisions. The party with the most votes in the public vote and the party with the most parliamentary seats or delegates are the two ways to look at who wins an election. It is desired that the party that receives the most votes also receives the most seats; however this is not always the case [8]. Elections in which one party gets a majority of the popular vote but loses a majority of legislative seats or delegates are typically considered as undemocratic, and are thought to reflect unjust election laws that favor one party or kind of voter systematically. The literature on electoral projections is fundamental because studies should feed it, as in the case of statistical analysis, which is in short supply [9]. The prediction of the 2009 German elections, as shown in [10]

why the party won the German election of 2009 or the trouble with predictions: was done by taking into account the frequency of mentions and obtaining the total mentions, replication of mentions, and percentages of mentions. The sample is less than a month long and is taken on random days. It also takes into account the followers' progression. The results are analyzed quantitatively. Voter turnout and voting behavior are inextricably linked with the public's attitude toward voting [11]. It is therefore critical to comprehend how the general public reacts to voting, as this can serve as an important indicator of how democracy will evolve in the future.

II. RELATED WORK

Understanding where and why political change occurs in a country is central to political geography. While the party for which to vote, or even abstain from voting, is a personal decision, it is influenced by a variety of sociological, cultural, and geographical factors. People living in big [12] cities, for example, are more likely to vote differently than those living in the countryside, owing to different sociocultural backgrounds and political issues at stake. First as well as foremost, spatial data analysis using spatial analysis methods. To that end, it is critical to discuss the key characteristics of spatial data that distinguish spatial analysis from traditional statistical analysis. Spatial data, in particular, exhibit two characteristics: spatial heterogeneity and spatial auto- correlation [13]. The uneven distribution of a trait, event, or relationship across space is referred to as spatial heterogeneity [14]. Spatial autocorrelation refers to the fact that data from locations in space close to one another are more likely to be similar than data from locations further apart. As a result, attributing correct empirical predictions to one model or another is frequently difficult [15]. One reason for this is that, despite differences in theoretical approach, the models frequently have identical voting behavior consequences. This phenomenon [16] was introduced by the author as the first Law of Geography, “All things are related, but nearby things are more related than distant things”. While the majority of studies in this emerging field have concentrated on predicting aggregated electoral results [17], a smaller group of studies has concentrated on the challenge of predicting individual political orientation. Notably, Ceron et al. [18] demonstrated how political orientation can be determined by comparing individuals' writing style with the writings on politicians' public Facebook profiles, while David et al. [19] demonstrated how political orientation can be determined by comparing individuals' writing style with the writings on politicians' public Facebook profiles. Furthermore, the Uniform Random Sampling Procedure has been widely used in public opinion polls to assess electoral perspectives. The sample size is primarily determined by the variance of the population: because the population of events to be estimated is more diverse, the sample size will increase regardless of population size [20]. While many of these studies achieve high prediction accuracies, this accuracy is frequently achieved by restricting the study to the most active users [21]. Using Twitter data, the author presented a mood analysis methodology for predicting social sentiment of political events. It was discovered that the proposed method is useful in observing online user behavior toward political issues during elections by pre-classifying tweets with positive and negative labels using a Naive Bayes classifier [22].

III. SPATIAL DATA

The availability of location data for both individuals and businesses has exploded in recent years. Spatial data adds a new dimension to data and reveals patterns that would otherwise go undetected. Spatial data is also referred to as geospatial data [23], spatial information, or geographic data. Spatial data is made up of points, lines, polygons, and other geographic and geometric data primitives that can be mapped by location, stored with an object as metadata, or used to locate end user devices by a communication system. Spatial data can be classified as either scalar or vector [24]. Each provides unique information about geographical or spatial locations. Users can save spatial data in a variety of formats because it can contain more than just location-specific data. This data analysis [25] provides a better understanding of how each variable affects individuals, communities, populations, and so on. Raster data is made up of grid cells that are identified by row and column. The entire geographic area is divided into groups of individual cells, each of which represents a different image [26]. Raster data includes satellite images, photographs, scanned images, and so on. Points, polylines, and polygons make up vector data. Points represent wells, houses, and so on. Polylines are used to represent roads, rivers, and streams, among other things. Polygons represent villages and towns.

IV. PROPOSED METHODOLOGY

In our project, we used a spatial data mining approach to develop a voting behaviour prediction model. Our main goal is to train some existing set of voting data from an area and determine the major voting population in the various states of a country based on their geographical influence on one another.

A. Dataset

In order to analyses the voting scenario, we used census data from previous years. The information was obtained from (the official US Government Census Data). The following is some background information on census data. A census is the process of collecting and [27] recording information about members of a given population in a systematic manner. The term is most commonly associated with national population and housing censuses, though agriculture, business, and traffic censuses are also common. The United Nations defines the essential characteristics of population and housing censuses as "individual enumeration, universality within a defined territory,

simultaneity, and defined periodicity, and recommends that population censuses be conducted at least once every ten years. We analyzed data from 2002 to 2016, because data prior to 2002 was not properly prepared and difficult to use.

B. Model

After analysing the various Machine Learning [28] models, we concluded that Time Series Models are the best fit for predicting future data, and we chose the Spatial Autocorrelation Distance algorithm to take the spatial attribute into account. Time series analysis [29] refers to methods for analysing time series data to extract meaningful statistics and other data characteristics. The use of a model to predict future values based on previously observed values is known as time series forecasting [30]. Among the various time series models, the ARIMA model [31] provided the best fit because it is one of the models that predicts data in a pure linear way and works on stationary data, which aids in better prediction. We used the Spatial Autocorrelation Distance Algorithm to determine the similarity between neighbouring states for each state in the United States, so that we can account for the influence of neighbouring states when using the ARIMA model. To compute the correlation, we used the SCIPY library, which uses the following formula to compute the spatial distance between two 1-Dimensional arrays.

$$1 - \frac{(u - \bar{u}) \cdot (v - \bar{v})}{\|(u - \bar{u})\|_2 \|(v - \bar{v})\|_2}$$

Where x.y is the dot product of x and y, and is the mean of the elements of u.

V. IMPLEMENTATION

The first step was to convert the existing dataset into a format that could be used. We preprocessed data from 2002 to 2016 and combined all altered entries into a single table. The Excel file in CSV format shown in figures 1 and 2 was read and converted using pandas.

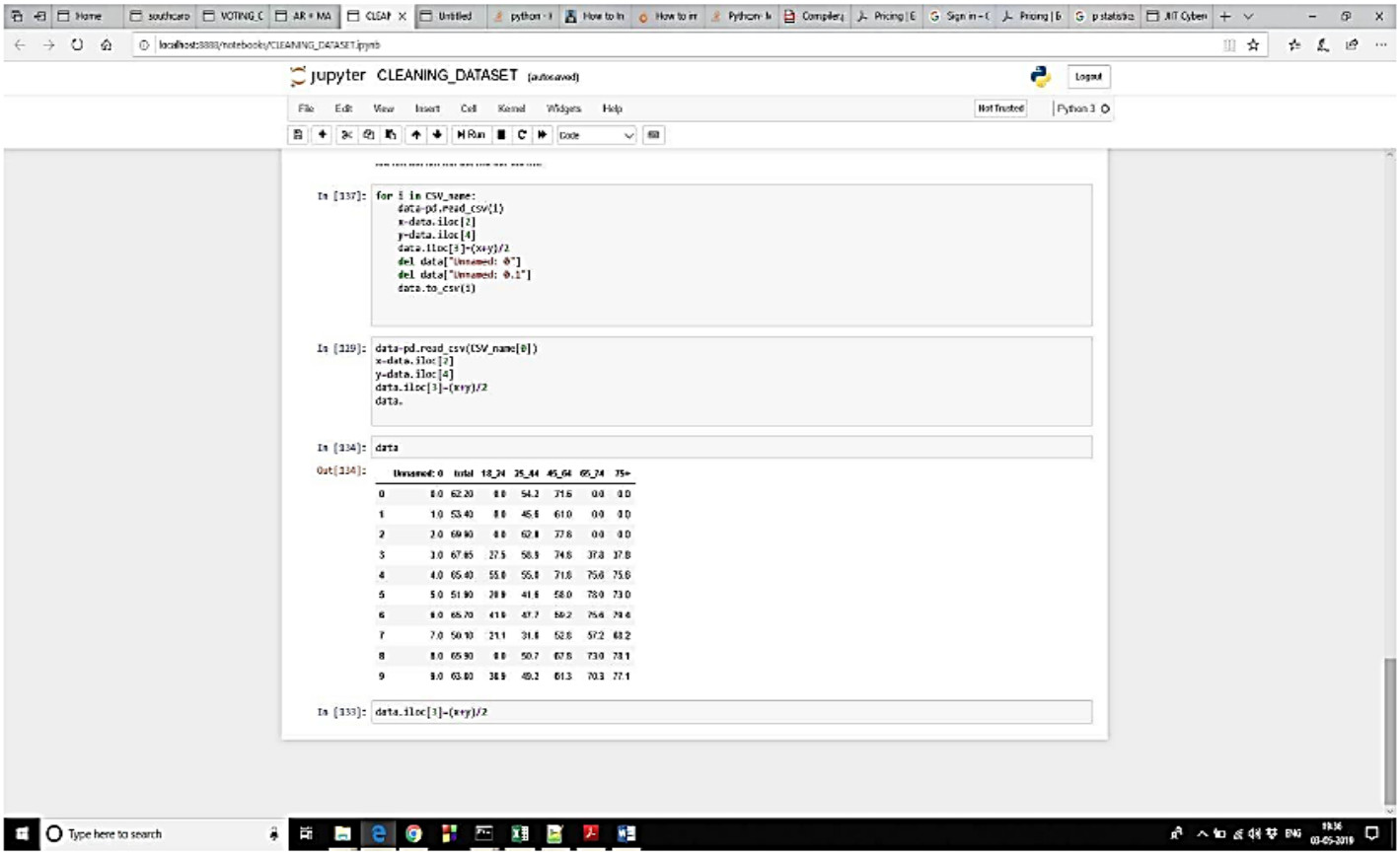

Figure 1: The Data Cleaning I

We used the ARIMA model to forecast the distribution of the following year based on previous year's data. Our data needs to be steady in order to use the ARIMA model. Autoregressive Integrated Moving Average is the abbreviation for Autoregressive Integrated Moving Average. ARIMA models [32] try to characterise the

correlations in the data with each other, whereas exponential smoothing models were built on a description of trend and seasonality in the data. Seasonal ARIMA, as seen in figure 3, is an improvement over ARIMA. We were able to attain stationary in moving average while adopting ARIMA, as illustrated in Figure 4. So, in order to get better results, we set the auto regression parameter to zero. We anticipated the influence of neighbouring states in order to add a geographical component to our model. To do so, we used the method described above to determine the degree of similarity between the state and its neighbours, and then proposed a value based on the data's similarity. The weighted average method produced more accurate findings than the ARIMA model [33]. The results were then plotted on a map of the United States to show the voting pattern in a visual way.Recurrent Neural Network (RNN) [34] is a type of Neural Network where the output from previous step is fed as input to the current step. All of the inputs and outputs in typical neural networks are independent of one another, however in some circumstances, such as when predicting the next word of a phrase, the prior words are necessary, and so the previous words must be remembered (see Figure 5). As a result, RNN was created, which used a Hidden Layer to tackle this problem [35]. The Hidden state, which remembers certain information about a sequence, is the most essential element of RNN. RNNs have a "memory" that stores all information about the calculations. It employs the same settings for each input since it produces the same outcome by performing the same task on all inputs or hidden layers. Unlike other neural networks, this decreases the complexity of the parameters.

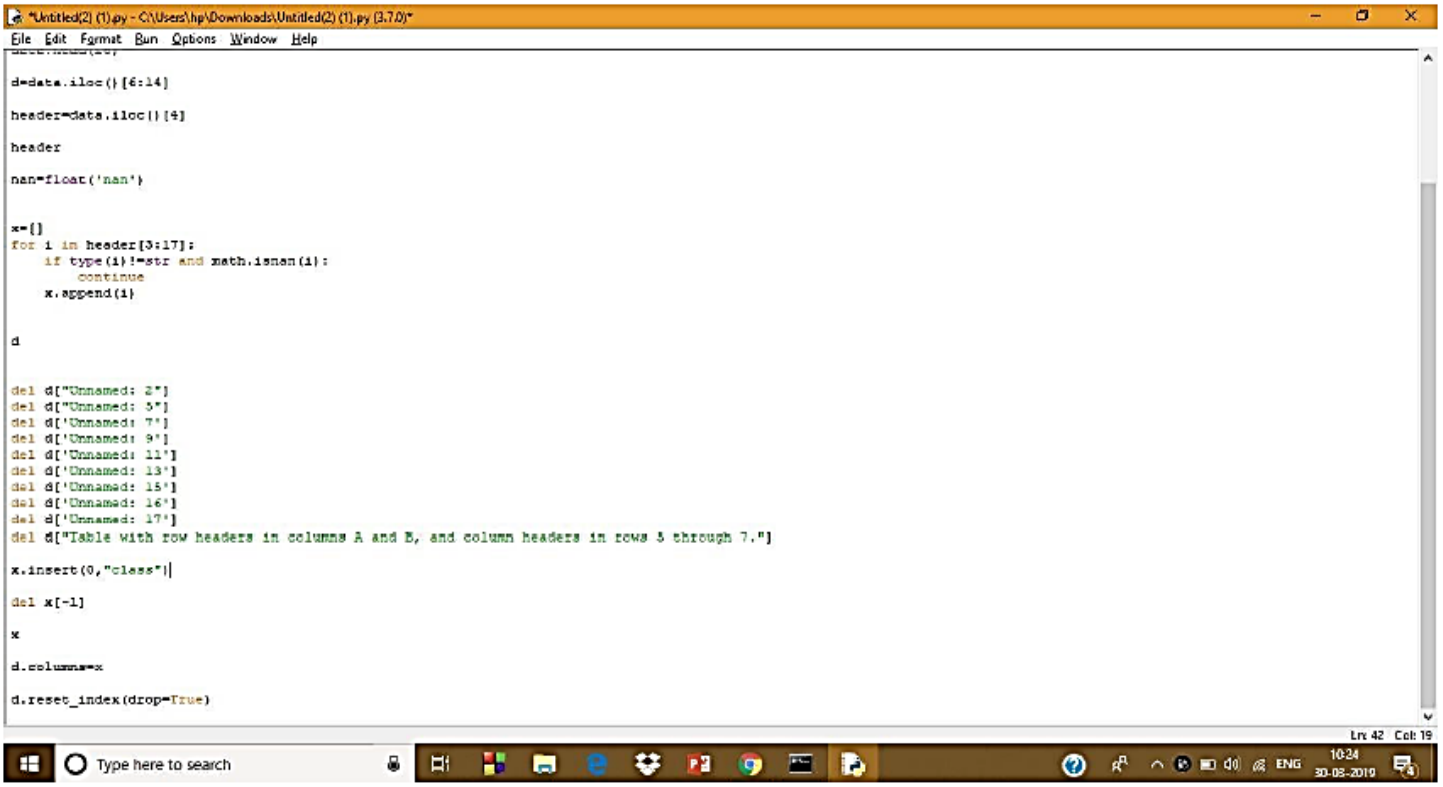

Figure 2: The Data Cleaning II

A stationary time series has statistical features such as mean, variance, autocorrelation, and others that remain constant over time, as shown in figure 6. Most statistical forecasting methods [36] are founded on the notion that through mathematical modifications, the time series may be rendered approximately stationary (unchanged). We merely anticipate that the statistical features of an unchanged series will remain the same in the future as they have been in the past. Rolling means (or moving averages) are commonly employed in time series data to smooth out short-term variations and highlight long-term patterns. If a statistical model predicts future values based on previous values, it is called autoregressive [37]. An autoregressive model, for example, might try to forecast a stock's future prices based on its historical performance.

VI. RESULTS AND DISCUSSION

This model will present us with key election behavioural predictions. This study project will focus on the following data points in order to authenticate a place's exit poll data current age-wise voting statistics [38], voter demographics,

and votes cast, and Spatial Correlation illustrated in figure 6 across neighbouring states.

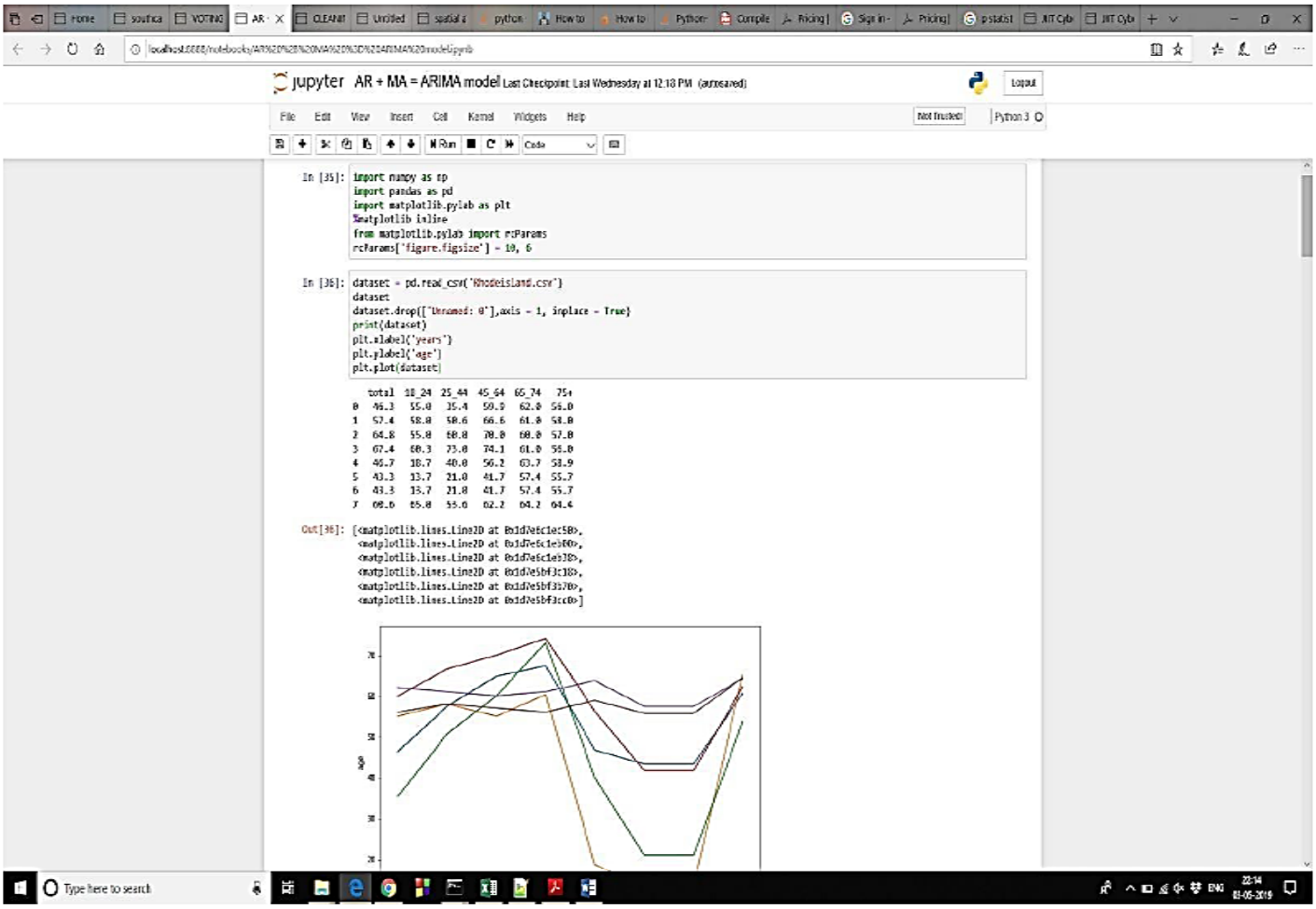

Figure 3: The ARIMA Model

Our project's major goals are to encourage voting among various age groups depending on projected characteristics. Understanding the influence [39] of the state's neighbours throughout the years depicted in figure 7. Analysing the entire voting scenario from previous years can aid in the forecast of voters' voting behaviour in the future [40]. Understanding the fundamental reason of the weaker areas and making improvements for the future.

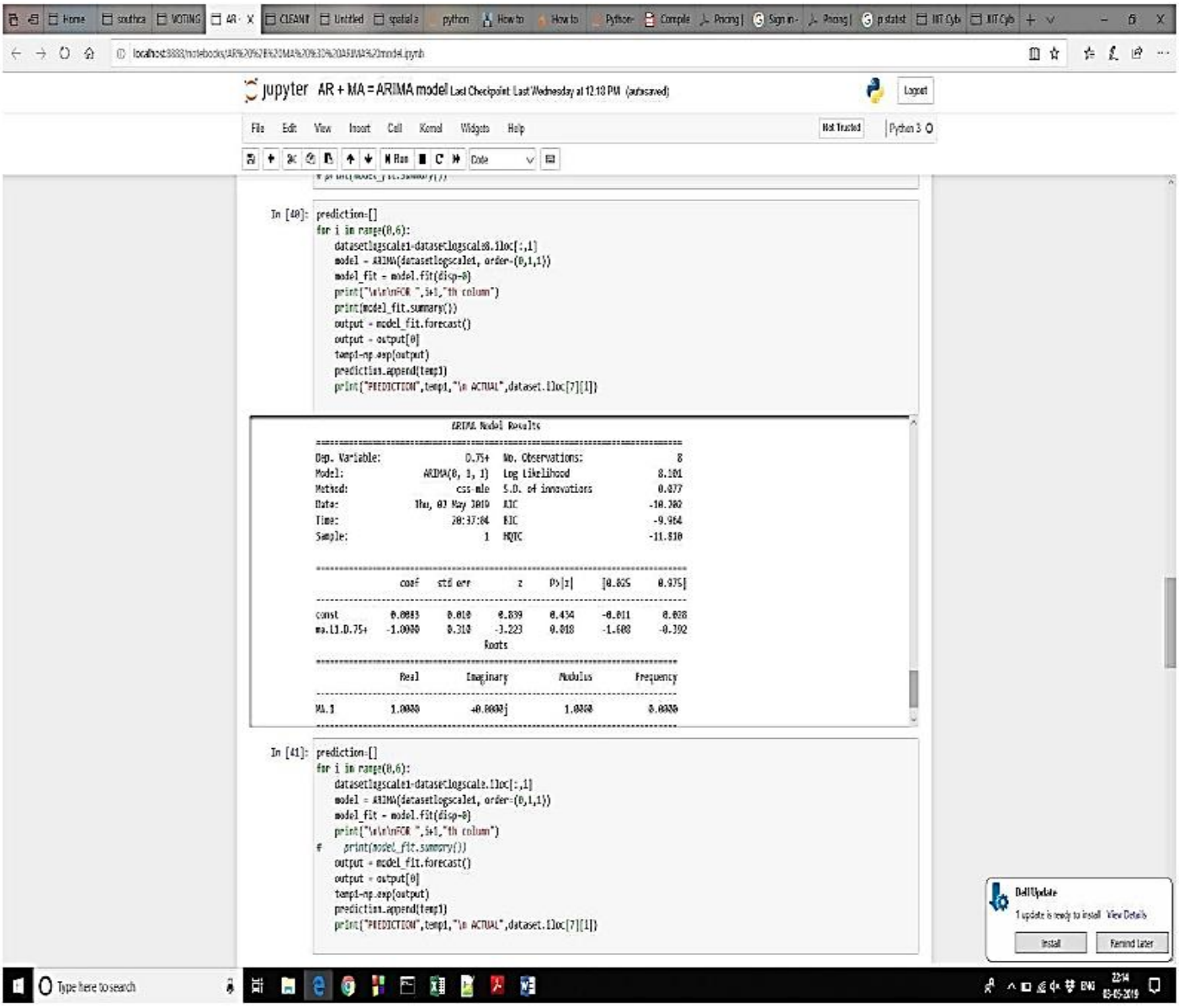

Figure 4: The ARIMA Model Prediction

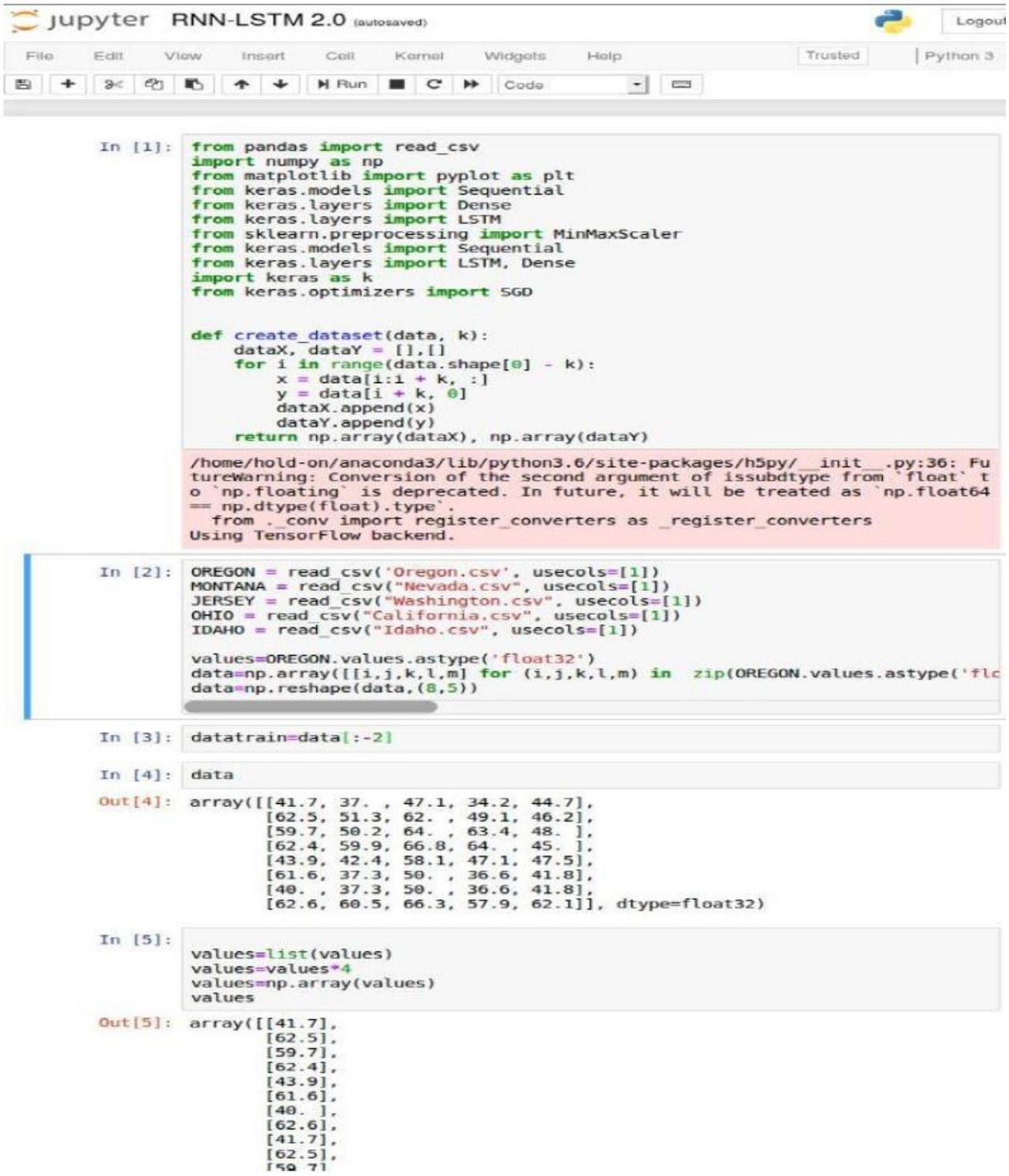

Figure 5: The RNN-LSTM

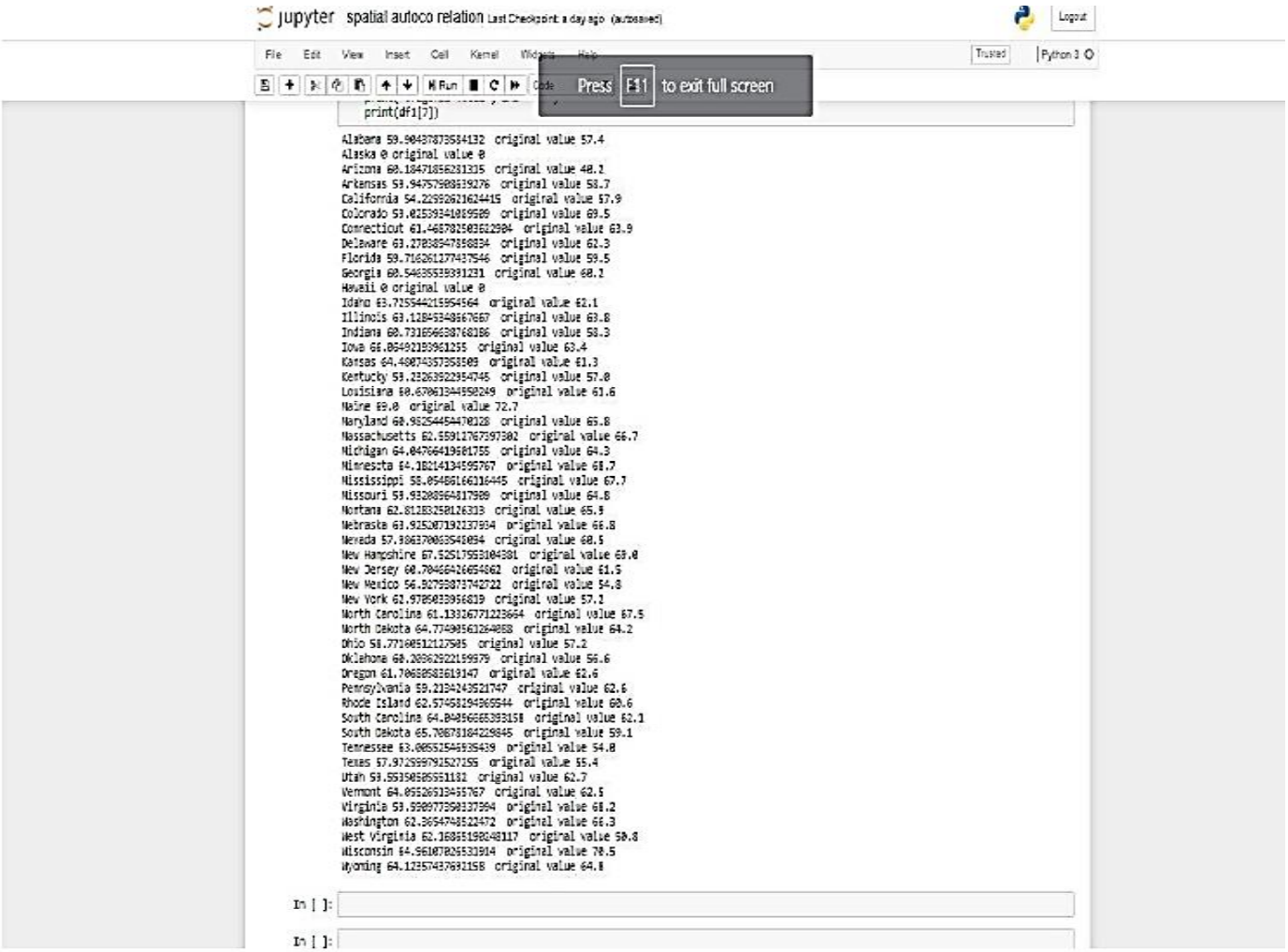

Figure 6: The Spatial Auto Correlation

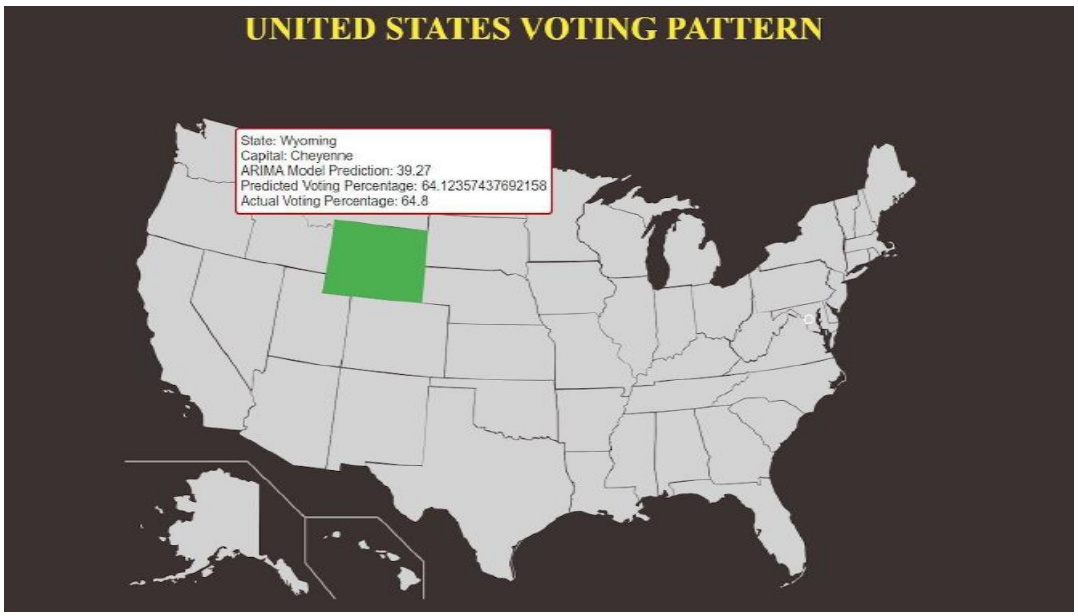

Figure 7: The USA Voting Pattern

VII. CONCLUSION

Political geography requires an understanding of where and why a country's political transition is taking place. While voting is a personal choice, it is influenced by a number of sociological, cultural, and geographic factors that have been proposed to constitute 'cultural fields' that influence human decision-making. India is the world's most populous democracy. India's true power is in the hands of its people. Citizens have the right to vote, which they use to elect their government. However, voting is a basic but complex process that can be improved in numerous ways. Our main goal is to use some current voting data from a region to train and predict the major voting population in various states around the country based on their geographical influence. This will aid in the analysis of the current situation as well as assisting the government in creating awareness in places where it is missing.

REFERENCES

[1] Michaud, J., Mäkinen, I.H., Szilva, A. et al. "A spatial analysis of parliamentary elections in Sweden 1985–2018", Appl Netw Sci 6, 67, 2021

[2] Verba S, Nie NH. Participation in America: political democracy and social equality. University of Chicago Press ed. Chicago: University of Chicago Press; 1987

[3] Dalton RJ. The potential of big data for the cross-national study of political behavior. Int J Sociol. , 46: 8–20, 2016

[4] Franch F. (Wisdom of the Crowds)2: 2010 UK Election Prediction with Social Media. J Inf Technol Polit., 10: 57–71, 2013

[5] Mavragani A, Tsagarakis KP. YES or NO: Predicting the 2015 GReferendum results using Google Trends. Technol Forecast Soc Change., 109: 1–5, 2016

[6]Ceron A, Curini L, Iacus SM. iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content. Inf Sci.; 367–368:, pp 105–124, 2016

[7] Barclay FP, Pichandy C, Venkat A, Sudhakaran S. India 2014: Facebook "Like" as a Predictor of Elec- tion Outcomes. Asian J Polit Sci., 23: 134–160, 2015

[8] Andersen R, Tilley T and Heath AF, "Political Knowledge and Enlightened Preferences: Party Choice Through the Electoral Cycle", British Journal of Political Science , 35: 285-302, 2005

[9] Tsakalidis Adam, et al., predicting elections for multiple countries using twitter and polls, IEEE 2015

[10] Hans Ulrich Buhl, 2011, From Revolution to Participation: SocialMedia and the Democratic Decision-Making Process, BISE Editorial, 2011

[11] Unankard, X. Li, M. Sharaf, J. Zhong, and X. Li, "Predicting elections from social networks based on sub-event detection and sentiment analysis," Web Information Systems Engineering WISE 2014, vol. 8787, pp. 1–16, 2014

[12] Yusuf Perwej, Kashiful Haq, Firoj Parwej, M. M. Mohamed Hassan ," The Internet of Things (IoT) and its Application Domains", International Journal of Computer Applications (IJCA) ,USA , ISSN 0975 – 8887, Volume 182, No.49, Pages 36- 49, 2019, DOI: 10.5120/ijca2019918763

[13] Fotheringham, S. A., Charlton, M. & Demšar, U., Looking for a Relationship? Try GWR. In: H. J. Miller & J. Han, eds. Geographic Data Mining and Knowledge Discovery. s.l.:CRC Press, pp. 227-252, 2009

[14] Anselin, L., Thirty years of spatial econometrics. Papers in Regional Science, 89(1), pp. 3-25, 2010

[15] Fazekas Z and Méder ZZ,"Proximity and directional theory compared: Taking discriminant positions seriously in multi-party systems", Working Paper, 2012

[16] Tapp, A. F.,. Areal Interpolation and Dasymetric Mapping Methods Using Local Ancillary Data Sources. Cartography and Geographic Information Science, pp. 215-228, 2010

[17] Makazhanov A, Rafiei D, Waqar M. Predicting political preference of Twitter users. Soc Netw Anal Min. 4: 1–15, 2014

[18] Ceron A, Curini L, Iacus S. Using social media to fore-cast electoral results: A review of state-of-the-art. Ital J Appl Stat., 25, pp 237–259, 2015

[19] David E, Zhitomirsky-Geffet M, Koppel M, Uzan H. Utilizing Facebook pages of the political parties to automatically predict the political orientation of Facebook users. Online Inf Rev. 40, pp 610–623, 2016

[20] Shira Fano and Debora Slanzi , Using Twitter Data to Monitor Political Campaigns and Predict Election Results. Springer international publishing AG, 2017

[21] Volkova S, Bachrach Y, Armstrong M, Sharma V. Inferring Latent User Properties from Texts Published in Social Media. AAAI., pp. 4296–4297, 2015

[22] Suarez Hernandez A, et al., predicting political mood tendencies based on twitter data, 2017

[23] X. Zhang, X. Zhu, B. She and S. Bao, "The spatial data integration and analysis with China Geo-Explorer", 17th International Conference on Geoinformatics Geoinformatics 2009 Conference, 2009

[24] L. Anselin, Y. W. Kim and I. Syabri, "Web-Based Analytical Tools for the Exploration of Spatial Data", Journal of Geographical Systems, vol. 6, pp. 197-218, 2004

[25] Yusuf Perwej, "An Experiential Study of the Big Data", International Transaction of Electrical and Computer Engineers System (ITECES), USA, Science and Education Publishing, Volume 4, No. 1, Pages 14-25, 2017, DOI: 10.12691/iteces-4-1-3

[26] Yusuf Perwej, Asif Perwej, Firoj Parwej, "An Adaptive Watermarking Technique for the copyright of digital images and Digital Image Protection", International journal of Multimedia & Its Applications (IJMA), USA , Volume 4, No.2, Pages 21- 38, 2012, DOI: 10.5121/ijma.2012.4202

[27] S. Deepajothi and S. Selvarajan:, "A Comparative Study of Classification Techniques On Adult Data Set", International Journal of Engineering Research Technology (IJERT), vol. 1, no. 8, 2012

[28] Yusuf Perwej, "An Evaluation of Deep Learning Miniature Concerning in Soft Computing", International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Volume 4, Issue 2, Pages 10 - 16, 2015, DOI: 10.17148/IJARCCE.2015.4203

[29] Z. Wang, W. Yan and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline", Proc. Int. Jt. Conf. Neural Networks, vol. 2017-May, pp. 1578-1585, 2017

[30] S. Yao, S. Hu, Y. Zhao, A. Zhang and T. Abdelzaher, "DeepSense: A unified deep learning framework for time-series mobile sensing data processing", 26th Int. World Wide Web Conf. WWW 2017, pp. 351-360, 2017

[31] G. Zhang, Time series forecasting using a hybrid arima and neural network model, Neuro-computing, vol. 50, pp. 159-175, 2003
[32] A.M. Alonso and C. Garcia-Martos, "Time Series Analysis - Forecasting with ARIMA models", Universidad Carlos III de Madrid Universidad Politecnica de Madrid, 2012
[33] S. Atique, S. Noureen, V. Roy, V. H. Subburaj, S. Bayne and J. Macfie, "Forecasting of total daily solar energy generation using ARIMA: a case study", 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC) (IEEE CCWC 2019), Jan. 2019
[34] Yusuf Perwej, "Recurrent Neural Network Method in Arabic Words Recognition System", International Journal of Computer Science and Telecommunications (IJCST), UK, London Volume 3, Issue 11, Pages 43-48, 2012
[35] Yusuf Perwej, "The Bidirectional Long-Short-Term Memory Neural Network based Word Retrieval for Arabic Documents", Transactions on Machine Learning and Artificial Intelligence, (UK), ISSN 2054-7390, Volume 3, Issue 1, Pages 16 - 27, 2015, DOI: 10.14738/tmlai.31.863
[36] I. Moghram and S. Rahman, "Analysis and evaluation of five short-term load forecasting techniques", IEEE Transactions on Power Systems, vol. 4, no. 4, pp. 1484-1491, Nov 1989
[37] P. J. Brockwell and R. A. Davis, Time Series: Theory and Methods, Springer, 2009
[38] Akhtar N," Perceptual evaluation for software project cost estimation using ant colony system", Int J Comp Appl 81(14):23–30, 2013
[39] Tadayoshi Kohno, Adam Stubblefield, Aviel D. Rubin, Dan S. Wallach,"Analysis of an Electronic Voting System", Johns Hopkins University Information Security Institute Technical Report, TR-2003-19, July 23, 2003
[40] A. Pajala, A. Jakulin, and W. Buntine, "Parliamentary group and individual voting behavior in Finnish Parliament in year 2003: A group cohesion and voting similarity analysis, " 2004