Big Data Analysis in Traffic Analysis Prediction System Using Volatality Model

Pagadala Veeranjaneyulu[1], Tippani Gayathri[2], Kathi Venkata Ramana[3]

[1, 2] Faculty in Keshav Memorial Institute of Technology, Hyderabad

[3] Research Scholar in Madhav University, Rajasthan.

Abstract- *Traffic Prediction is critical as it is enhancing day by day leading to worst on road situations. Increased accidents and delays in critical applications is causing awful situations for the user The prediction model was used as an estimator to identify unusual traffic patterns. The generic model was designed using data mining techniques, multivariate regression algorithms, ARIMA and visually correlated with real-time traffic tweets. traffic observations that produce analytical dashboard providing traffic prediction and analysis.*

Keywords- Data Mining, Visualization, Traffic Analysis, Prediction Analysis, Multivariate regression.

I. INTRODUCTION

Traffic is enhancing by leaps and bounds. Giving accurate prediction regarding traffic is need of the hour. The requirement of traffic prediction is critical in time specific applications. The environment such as ambulance movement, student needs to get back to exams or any other time critical application has significant usage of this research. Proposed literature utilizes techniques of data mining and ARIMA model and Traffic observation model for time series analysis.

The data that utilized for analyzing traffic at specified time is traffic flow data. This would help in predicting future traffic. The proper action is needed to be taken for reducing the congestion. By the use of traffic flow data this reduction of congestion has to be done as future prediction can be made.

Data Mining has quickly grown with the presence of the wonder BIG Data[1]. For sure, numerous associations have begun to digitize their records, and have changed their paper-based frameworks to electronic frameworks. This change conveys a few advantages to the associations, among them time funds, a superior administration and a more tightly checking making the assignments less demanding. One of the immediate results of this change is the visit gathering of significant Data. While the Data's holders started to stress over the capacity of Data, they understood the benefits they can take from it. The Data gathered can be considered as another unformatted of structure (Raw Data) which needs to be filtered. Handling Data give a superior quality Data which contribute in request to make choice in data selection[2]. Moreover, Healthcare elements likewise choose electronic frameworks, by utilizing different strategies, among them, Electronic health Record (EHR) or Electronic Traffic Records (EMR) frameworks. It implies the executing EHR frameworks, leads to an immense measure of Data gathered by doctor's facilities, centers and other traffic suppliers.[3] At that point, the vast majority of these Datasets are most certainly not extremely very much organized and fitting for explanatory purposes. In expansion, traffic Data are generally extremely perplexing and difficult to investigate. For instance the US Healthcare framework alone as of now achieved 150 Exabyte (1 Exabyte = 8388608 Terabit) five years prior. This pattern is because of the way that multi scale Data created from people is consistently expanding, especially with the new high-throughput sequencing stages, continuous imaging, and purpose of care gadgets, also as wearable figuring and versatile traffic innovations. As needs be, Data Mining has gotten a great deal of consideration on account of its solid capacity of separating Data from Data, furthermore, winds up noticeably prevalent in Healthcare field by dint of its productive diagnostic procedure for recognizing obscure and significant Data in traffic Data[1], [4].

In Traffic Prediction, Data Mining gives a few advantages for example, discovery of the extortion in traffic coverage, accessibility of therapeutic answer for the patients at lower cost, discovery of reasons for ailments and recognizable proof of therapeutic treatment techniques. It likewise helps the Healthcare analysts for making productive Healthcare approaches, developing medication suggestion frameworks, creating traffic profiles of people and so on[1], [2], [5], [6]. Taking such a case, McKinsey gauges that enormous Data examination can empower more than 300 billion in investment funds for every year in U.S. Medicinal services, 66% of that through decreases of around 8 percent in national Healthcare consumptions. Clinical operations and R D (innovative work) are two of the biggest ranges for potential reserve funds with 165 billion furthermore, 108 billion in waste individually The result of Data Mining advancements are to give advantages to Healthcare association for gathering the patients having comparative sort of infections or traffic problems so that

Medicinal services association gives them successful medications[6]. It can likewise valuable for anticipating the length of remain of patients in healing center, for restorative conclusion and making arrangement for compelling Data framework administration. Late innovations are utilized as a part of restorative field to improve the restorative administrations in practical way. Data Mining methods are additionally used to examine the different elements that are in charge of sicknesses for instance sort of nourishment, diverse working condition, instruction level, living conditions, accessibility of unadulterated water, human services administrations, social, natural and rural variables.

In this paper, we introduce the upsides of Data Mining for traffic and the reasons make Data Mining critical to be considered in traffic Data examination. Data mining traffic Dataset with missing values is considered to be analyzed initially through Support vector machine.

II. STUDY OF EXISTING LITERATURE

Data mining approaches is the base of this literature. Analysis of existing literature provide base for proposed literature. [11] Reviewed various models and methods used within data mining. Data mining techniques development from 2005 to 2015 is reviewed and application in regards to traffic is proposed. [1] Suggests the integration of traffic data with data mining strategies used to form traffic information system. Patient traffic condition can be analyzed along with future prediction about patient's health. Hidden possibilities can be extracted using unlimited data mining techniques to make accurate health forecast. [12] Proposed multilayer perceptron in order to analyze big data corresponding to traffic. As literature deals with traffic of patients hence high degree of accuracy is desired. To accomplish the desired goal comparison of SVM and multilayer perceptron on traffic data set is made. Results of SVM in terms of classification are better as compared to multilayer perceptron. [3] Suggests data mining techniques used for analysis of diabetics. Support Vector Machine (SVM) is used for this purpose.

Genetic approach is also analyzed for diabetic's dataset in the field of data mining. Results of SVM are obtained to be better. [13] Suggests five J.48 classifiers to predict hypertension and eight other diseases. Prediction accuracy is obtained and compared against naïve bayes approach. Results in terms of J.48 are obtained to be better. [7] Suggests hybrid approach for traffic to predict diseases using Big data. Pruning based KNN is used for this purpose which used density based clustering based method integrated with KNN approach. Local outlier factor of PB-KNN is better as compared to KNN. [14] proposes SVM and neural network techniques for skin lesion detection in human body. Segmentation along with classification is performed in order to detect the diseases. [8]predict heart diseases are primary cause of death among humans in last decade. Data mining techniques are used in order to detect and predict heart diseases efficiently.

[4]proposes a mechanism through which information about patient coming for checkup at hospital is stored and algorithm is applied in order to perform predictions. Data mining algorithm considered in this approach is naïve bayes. Accuracy of prediction is obtained is significant in this case.[15] suggests intelligent heart disease prediction system. Decision tree , naïve bayes and neural network technique are used for accurate analysis and prediction of disease.

Analyzed approaches enhance performance considering datasets not including any noisy or missing values. Missing values or noisy data handling and increasing prediction accuracy is primary task of proposed approach

III. PROPOSED SYSTEM

Proposed system uses Dublin Traffic data model for time series analysis. Visualizations and Volatility, Inbound, Outbound Analysis observations.

1. DATA SOURCES

Open data sources DubLinked (2014), Wunderground (2014) and Twitter (2014) were used in this work for the visualisations. The following section is a summary of the

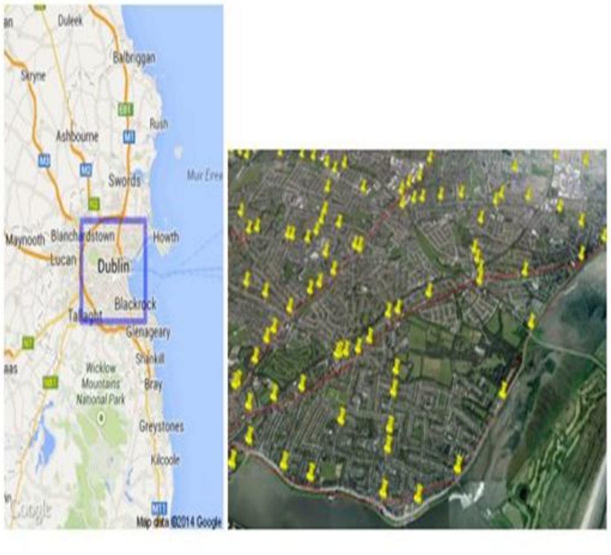


Figure 1. Dublin Traffic data

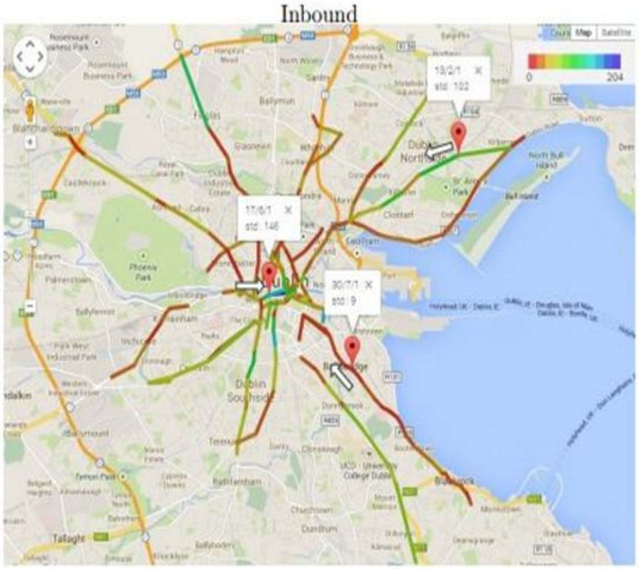
data used in the data collection and creation of the visualisations. 2.1 Traffic Data Sets The prediction model was generated from data accumulated from open source portal known as DubLinked.

2. Visualisations

The analytics dashboard enables the users to identify the patterns of traffic visually as mentioned in the introduction 1. The visualisations are generated using Google Maps, JQuery along with a Python backend.

Volatility

Volatility is a way of identifying an inconsistency in the travel time. With this users can identify areas that are prone to delays. Standard deviation can be considered a way of measuring the volatility according to a paper from Tulloch (2012). A range of colors provides the result of the standard deviation from low red equal to 0 and high purple equal 200+ see Figure



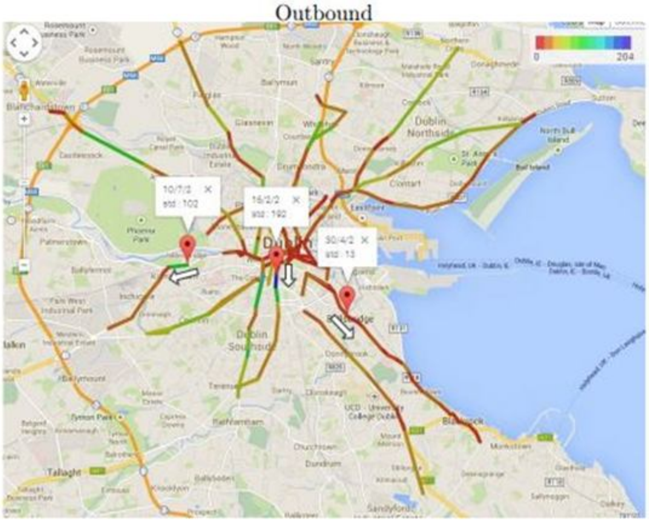


Figure 2. Inbound and Outbound observations

IV. PREDICTION MODEL

During the exploration process, the highest correlated weather stations and spatial neighbour were used in the generic estimation data model. A generic data model provides the ability to reuse the process of building the data sets that would best fit the majority of the 512 observed locations while still keeping features that improve accuracy. As a result, the prediction algorithms behaved differently depending on the influence of features. The best-performing algorithms for the least volatile road segments mentioned ?? are linear regression. Some road segments had little or no volatility. Other linear regressions, performed well that had volatility used a normalization of feature to improve accuracy.

Road segments with highly volatility with features of insignificant correlation resulted in a non-linear Support Vector Machine with Fourier transform with the highest accuracy.

Bayesian Ridge linear regression algorithm performed very well for the prediction. It demonstrates that when noise accounts for the more linear the data becomes.

In figures, ?? and ?? shows that the area of Finglas and Glasnevin is the least affected by weather and is highly volatile. Where the city centre and Clontarf are volatile and highly affected by weather conditions becomes a linear problem, see figure4.1

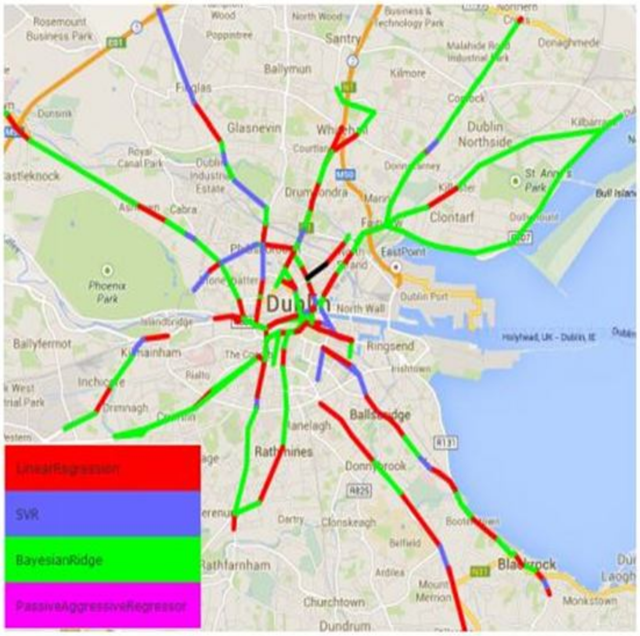


Figure 3. Off-peak and Inbound algorithm

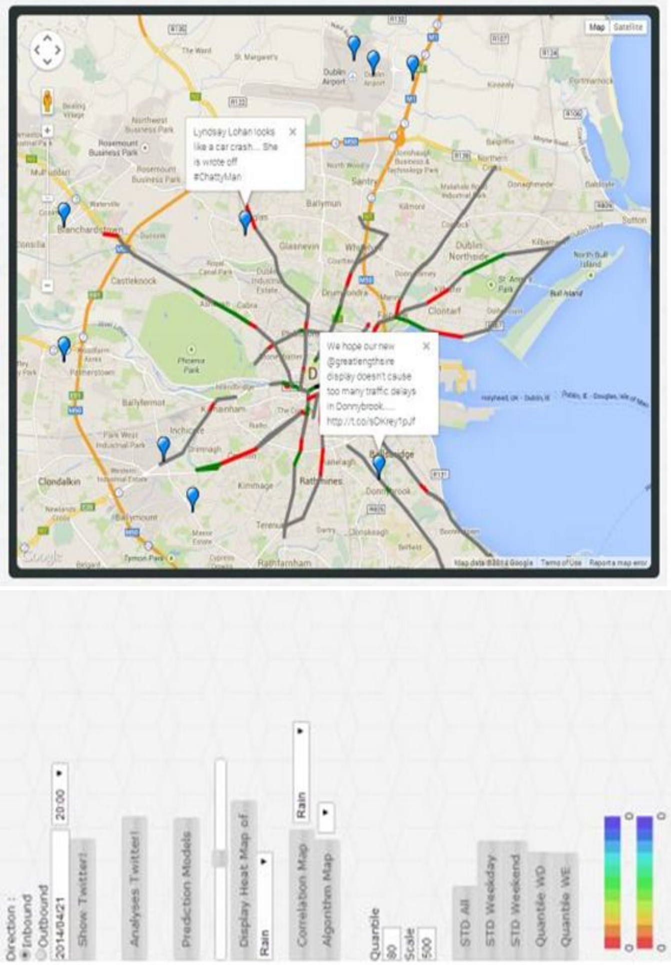
V. RESULT ANALYSIS

The classification approach worked as a proof of concept. The real-time traffic tweet could be used to provide further analysis on traffic delays. In Figure 8 the dashboard demonstrates traffic related tweets as blue markers overlayed above the road segments and its estimated result. The red lines indicate delays, the green indicate better than expected while the grey is as expected. Each tweet marker is click-able to provide more informative details on the traffic conditions. Using the buttons on the left of the dashboard will display the different elements of the visualisations show in this abstract.

Table 1. Data Volume

Data Source	*Items*	*No.of Documents*
Traffic Observations	501,402,840	8,356,714
Real-time	3,048,310	116
User	5,267	5,267

Issues such as in figure 8 the dashboard contains a some false positives, example ”Lyndey Lohan looks like a car crash.


Lyndsay Lohan looks like a car crash... She is wrote off #ChattyMan

We hope our new @greatengthsire display doesn't cause too many traffic delays in Donnybrook..... http://t.co/sOKrey1pJf

Figure 4. Dashboard Analysis

VI. CONCLUSION

The traffic prediction is critical as traffic is enhancing due to increase in on road vehicles. The proposed literature uses modified ARIMA model for prediction of traffic accurately. The results are predicted in terms of accuracy and mean square error. The accuracy is enhanced since Euclidean distance is used for determining the closest distance between the points present within the dataset. The dataset is fetched from the online source UCI. The accuracy is obtained by subtracting the actual value from the obtained value. The least error rate and enhanced accuracy proves the worth of the study. The result is compared against the existing literature involving ARIMA without KNN and Euclidean distance.

REFERENCES

[1] I. Țăranu, “Data mining in healthcare: decision making and precision,” Database Syst. J., vol. 5, no. 4, pp. 33–40, 2015.

[2] M. E. Student, C. T. Nadu, and C. T. Nadu, “Heart

disease classification and its co-morbid condition detection using WPCA genetic algorithm," pp. 287–291, 2016.

[3] C. Anusha, S. K. Vinay, H. J. Pooja Raj, and S. Ranganatha, "Medical data mining and analysis for heart disease dataset using classification techniques," Natl. Conf. Challenges Res. Technol. Coming Decad. (CRT 2013), pp. 1.09–1.09, 2013.

[4] E. Pinheiro, W. Weber, and L. Barroso, "Failure trends in large disk drive population," Proc. 5th USENIX Conf. File Storage Technol. (FAST 2007), no. February, pp. 17–29, 2007.

[5] A. Sharma and V. Mansotra, "Emerging applications of data mining for healthcare management - A critical review," 2014 Int. Conf. Comput. Sustain. Glob. Dev., pp. 377–382, 2014.

[6] K. Yan, X. You, X. Ji, G. Yin, and F. Yang, "A Hybrid Outlier Detection Method for Health Care Big Data," 2016 IEEE Int. Conf. Big Data Cloud Comput. (BDCloud), Soc. Comput. Netw. (SocialCom), Sustain. Comput. Commun., pp. 157–162, 2016.

[7] S. Sivagowry, M. Durairaj, and a. Persia, "An empirical study on applying data mining techniques for the analysis and prediction of heart disease," 2013 Int. Conf. Inf. Commun. Embed. Syst., pp. 265–270, 2013.

[8] W. E. Leland, W. E. Leland, D. V Wilson, and D. V Wilson, "On the Self-Similar Nature of Ethernet Traffic," Comput. Commun. Rev., vol. 2, no. August 1989, pp. 203– 213, 1992.

[9] S. Jain and N. Pise, "Computer aided Melanoma skin cancer detection using Image Processing," Procedia - Procedia Comput. Sci., vol. 48, no. Iccc, pp. 735–740, 2015.

[10] N. Jothi, N. A. Rashid, and W. Husain, "Data Mining in Healthcare - A Review," Procedia Comput. Sci., vol. 72, 306–313, 2015.

[11] P. Naraei, V. Street, V. Street, and V. Street, "Application of Multilayer Perceptron Neural Networks and Support Vector Machines in Classification of Healthcare Data," no. December, pp. 848–852, 2016.

[12] F. Huang, S. Wang, and C. Chan, "Predicting Disease By Using Data Mining Based on Healthcare Information System," 2012 IEEE Int. Conf. Granul. Comput. Predict., 12–15, 2012.

[13] M. A. Farooq, M. A. M. Azhar, and R. H. Raza, "Automatic Lesion Detection System (ALDS) for Skin Cancer Classification Using SVM and Neural Classifiers," 2016 IEEE 16th Int. Conf. Bioinforma. Bioeng., pp. 301–308, 2016.

[14] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," 2008 IEEE/ACS Int. Conf. Comput. Syst. Appl., pp. 108–115, 2008.

[15] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," Procedia Technol., vol. 10, pp. 85–94, 2013.

[16] I. K. A. Enriko, M. Suryanegara, and D. Gunawan, "Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient ' s Health Parameters," vol. 8, no. 12, 1843.

[17] B. Veytsman, L. Wang, T. Cui, S. Bruskin, and A. Baranova, "Distance-based classifiers as potential diagnostic and prediction tools for human diseases," BMC Genomics, vol. 15 Suppl 1, no. Suppl 12, p. S10, 2014.

[18] M. M. El-Hattab, "Applying post classification change detection technique to monitor an Egyptian coastal zone (Abu Qir Bay)," Egypt. J. Remote Sens. Sp. Sci., vol. 19, no. 1, pp. 23–36, 2016.