

Healthcare Supply Chain Optimization using Machine Learning Techniques

Ayanabha Jana, Yash Dekate, Alpanshu Kataria, Rohan Dastidar, Dr. Asnath Victry Phamila Y,
Dr. S. Geetha

School of Computer Science and Engineering
Vellore Institute of Technology, Chennai, Tamil Nadu, India

Abstract

The past two years have been challenging and have certainly not been what anyone would have imagined. During these trying pandemic times, although there was a lot of pressure for procuring the healthcare products used for treating patients suffering from COVID, this quintessential healthcare operation for other critical patients should not be neglected. Keeping in mind the above issue as well as being aware of the criticality of the pandemic, this paper attempts to ease this pressure by optimizing the Supply Chain aspect of the healthcare products other than those used for treating COVID using various machine learning, ensemble learning, deep learning, and time-series forecasting algorithms with each of them segregated under various parts of the Supply Chain Management. In this paper, the required medications are first predicted using a combination of K-Means Clustering and K-Nearest Neighbors (KNN) algorithms. Then, the Supply Chain Management is segregated into four main parts namely - the Inventory Management implemented with Multiple Linear Regression (MLR), Demand Estimation implemented with Autoregressive Integrated Moving Averages (ARIMA), Production Estimation implemented with Random Forest Regression (RFR), and Supply Management implemented with a combination of DBSCAN and Regressive Neural Networks. After implementing this solution, one could get the optimal prediction values regarding the amounts, suppliers, etc. as the case may be for each aspect of the Supply Chain.

Keywords: Healthcare, Supply Chain Management, Machine Learning, Ensemble Learning, Deep Learning, and Time-Series Forecasting.

1. Introduction

The healthcare sector is critical in terms of determining the accurate procedure of treating a patient and allocating scant resources in trying times such as the current COVID-19 pandemic. This is achieved through a supply chain system that not only manages the inventory but also forecasts future demands for medications in accordance with the patients admitted. In addition to that, the system also needs to decide on the most reliable vendor based on the cost involved and the time constraint of operations.

However, the first task that needs to be accomplished before the supply chain techniques can be developed is setting a limit on the medications for which the system needs to produce results and this is achieved by grouping the medications based on symptoms and predicting the ones which are most similar. This transformation speeds up the system in the healthcare sector.

Then, the necessary results are obtained from a singular system that is trained and optimized on patient data, especially the symptoms. This aids the medical body to tend to patients who are atrophied by a variety of diseases, not just COVID-19. The models that are built subsequently utilize both numerical and categorical attributes. Since the majority of the data is text, text analytics has been implemented in manifesting such a system because it helps to apply standard machine learning tools to accrue the required results.

1. Related Work

In [1] it is discussed that how Covid-19 has impacted the healthcare supply chain and how shortages in medicines and other healthcare equipment have made the situation worse. Due to these problems, it is very important to be able to forecast which medicines would be required in the future. The author also suggests that the forecasting techniques can be evolved as more and more data is accumulated. If one wants to rely less on technology the other method is customer surveys but it is very time consuming. Every organization has different prediction needs therefore the prediction system must be adaptable. External factors can affect the results but they are specific. Various factors need to be considered to improve the working of healthcare system like stock-outs, corruption, product diversion etc. These can be easily eliminated by the use of technology. Demand estimation with respect to healthcare supply chain involves estimating which health commodity will be purchased when, where and in what quantity. Normally forecasting is done on the basis of price, use rate and availability of funds but there are other more important factors that need to be considered while making predictions. A bad forecast will lead to shortage in supply chain and limited funds. Demand forecasting is not a new thing but scenario is completely different in the backdrop of pandemic. To understand the challenges, we have to understand the extent and nature of risk for various stakeholders. Risks further depend on changing healthcare technologies, shifting nature of diseases, international markets etc. [1] also helps us in understanding the opportunities and challenges in making predictions in healthcare supply chain. [1] also provides us information and strategies about how risks can be reduces.

[2] is a rather different type of research paper but is highly useful for people thinking of proposing their solutions related either to introducing new methods of handling supply chains or in improving the pre-existing methods to handle them. The literature survey of [2] can be divided into two parts. The first part is solely based on the extensive literature review that they have done in regards to supplying chains and hence, talks about the methodology they have used in conducting their literature review of supply chain management. They first describe the way they have selected the concerned articles and then justify it on the basis of certain rules and identification of the frequently used algorithms in various parts of the supply chain system. They then present the results, in which they first describe the research trends, common machine learning algorithms used, and finally, how these machine learning algorithms can be used in the supply chain systems. The final part of the results is the most important one. This part is the one which first actually goes on to explain to us what a supply chain actually is and then lists out its various parts or sub-domains, viz., demand or sales estimation, procurement and supply management, production, inventory and storage, transportation, distribution, and supply chain improvement. It turns out that only a combined treatment of all these parts together with machine learning would yield a successful outcome in supply chain handling or optimization using machine learning. [2] provides information about what these parts actually mean and what is their importance in the overall supply chain management and lists out various algorithms that can be used to address these. It lists that Demand or Sales Estimation can be conveniently handled by Neural Networks (NNs) and Fuzzy Logic. Procurement and Supply Management can be conveniently handled by Support Vector Machines (SVMs) and Earth System Modeling (ESM) techniques. Production can be conveniently handled by NNs and SVMs. Inventory and Storage can be conveniently handled by Naïve Bayes Classifiers (NBCs) and SVMs. Transportation and Distribution can be conveniently handled by Adaptive NNs and finally, Supply Chain Improvement can be conveniently handled with Logistic Regression (LR) and Decision Trees (DTs). Hence, all in all, optimizing the whole supply chain system is a quite complex task owing to its various parts, irregularities in handling various parts and the data associated with the sub-domains as well as the level of complexity involved in implementing solutions for some as some parts are easy, while some are not. Hence, making a complete solution for addressing supply chains would involve incorporating different algorithms to address different parts of it and ultimately, combine them all together to form the final solution of handling or optimizing the supply chain. [2] concludes with stating the future directions and uses of machine learning in supply chain management and refers to almost 130 references regarding the application of ML in SCM.

[3] discusses operational, strategic, and tactical issues involved in supply chain management from the viewpoint of analysis and modeling and puts forward the models that deal with different issues. It also discusses the challenges posed by real-world supply chain problems such as uncertainty, huge data, interactions among team members, etc. It also discusses the opportunities created by recent development in hardware, mathematical modeling, information technology, and algorithmic methods to deal with real-world problems. Different problems faced during the implementation of the supply chain model in real life are randomness of variables, combinatorial nature of real-life problems, a large volume of data, and complex interaction among variables. Different problems of supply chain management like integrated production, inventory distribution problems can be formulated as linear programming problems. Integer programming problems are a special case of linear programming problems in which the decision variable

takes only integer values. Also, in some applications variables may take 0 or 1, their binary integer programming algorithm is used. Dynamic programming is also a method for solving difficult problems by dividing them into smaller sub-problems. Then the solution of the bigger problem is derived from the solution of sub-problems. Then there is a genetic algorithm that was developed by Fulya Altiparmak for multi-product supply chain network design problems. Using this algorithm, he planned which distribution centers and plants should be opened and which distribution center should serve which customer.

[4] focuses on integrated inventory and distribution planning by taking into consideration patient safety and achieves this by formulating the problem statement as a mixed-integer optimization problem, which is an NP-hard optimization model. The main basis of [4] is to minimize the cost incurred when procuring medicine from multiple suppliers to a single hospital in a distributed manner while simultaneously managing the inventory as well. To help solve the above problem, the inventory routing problem or IRP is proposed which builds a mathematical representation that finds exactly the time to deliver to the hospital, the quantity to deliver to a hospital within a given period, and how to route vehicles such that the sum of the transportation and inventory costs is minimized without compromising the patient safety level. All these factors when combined represent the total cost as follows

$$\sum_{t=1}^T \left[\sum_{j=1}^N \sum_{v=1}^V f_{jt} x_{0jt}^v + \sum_{i=0}^N \sum_{j=0}^N \sum_{v=1}^V c_{ij} x_{ijt}^v + \sum_{i=1}^N h_i I_{it} + \max\{0, \sum_{i=1}^N (d_{it} - I_{it-1} - Q_{it})\} p_i \right] \quad (1)$$

This equation needs to be minimized based on the constraint of setting a patient safety level affecting the transport route and inventory level. It is seen that a high inventory level and frequent transportation increases patient safety but with the downfall of increasing total costs as well. To ameliorate this direct proportion to total costs, ‘service level’ which measures the capacity to have inventory present as required by patients is used as an index to assess patient safety level. However, since IRP is an NP-hard problem, only an approximate solution can be obtained by the usage of a genetic algorithm or GA where a new chromosome to represent vehicle routing and inventory optimization is developed. Representing a gene as a single unit rather than a 4D matrix, representing the sequence number of vehicles, the number of vehicles, and the transported quantity as four indexed values. The decision variables which are required to minimize the cost function are calculated as combinations of the indexed values. The fitness function which measures the accuracy of a solution is represented as an inverse of the cost function. In the proposed GA system – the number of generations, population size, the crossover rate, and the mutation rate were fixed at 1000, 100, 0.8, and 0.01 respectively. To compare this solution with the standard GA implementation, IBM’s CPLEX optimization tool was employed and it was observed that the percentage differences between the optimal solutions of CPLEX and the proposed GA stayed within 8% and the difference for execution time was 44%. The proposed GA showed better performance for larger instances of data.

2. Proposed Work

3.1. Data Source

4 datasets have been utilized in total, viz., 2021VAERSData, 2021VAERSSYMPTOMS, and 2021VAERSVAX from [5]; and SCMS_Delivery_History_Dataset from [6]. Out of these, the first 3 are preprocessed and combined together while the last one is used independently.

3.2. Data Preprocessing

3.2.1. Patient Dataset

This is the dataset that is made by pre-processing and combining the first 3 datasets listed in section 3.1. It will be used to predict medicines for the patient based on the symptoms shown by them.

3.2.2. Supply Chain Dataset

This dataset will be used to manage the supply chain aspect of procuring the medicines predicted for the patient from the dataset mentioned in section 3.2.1. It will mainly be utilized to predict and optimize Inventory Management, Demand Estimation, Production Estimation, and Supply Management for the suppliers of the concerned medicines.

3.3. Algorithms Used

3.3.1. Machine Learning Algorithms

3.3.1.1. Multiple Linear Regression

It expresses a dependent variable y in terms of multiple independent variables x . The model fits a line in n -dimensional space by finding coefficients for each of these independent variables. This process continues until the difference between predicted and actual values is minimum. The resultant estimated equation of $y=f(x)$ can subsequently be used to predict future instances of x .

3.3.1.2. K Nearest Neighbors

In general, the KNN algorithm is utilized as a classifier to allocate a new data point to a particular category based on the first 'K' nearest data points based on a distance measure where the 'K' parameter is set by the user. The category to which the maximum number of neighbors belong is the resultant category. However, in this paper, the system utilizes the data points itself rather than the category which means that the system is only interested in the nearest data points.

3.3.1.3. K-Means Clustering

It is a standard machine-learning algorithm to analyze unsupervised data, that is, when there is no starting point for making any predictions. The algorithm takes numerical data in the form of n-dimensional vectors and the number of clusters 'K' which are optimally determined by the elbow method, initializes starting medians, and then computes the distance from those medians to all other data points. The ones that are at a minimal distance are assigned to that group or cluster followed by the updation of medians by taking the arithmetic mean. This process continues iteratively until the medians don't change.

3.3.1.4. DBSCAN

To start with the explanation of this algorithm, two terms, viz., epsilon and minimum points have to be defined. Epsilon is basically the radius around a point that can contain other points of the cluster. It is a fixed value for all the points in the dataset. Minimum points are the minimum number of points that have to lie within the epsilon of any point to be considered as a cluster. This is again a fixed value for all the points in the dataset.

Now to explain this algorithm, it starts with selecting a random point of the dataset and then expanding the cluster by adding the points falling within its epsilon of the initially selected point. As soon as the minimum number of points is added, these newly added points are now a part of the cluster. These new points then themselves start adding the points falling within their epsilon and this process goes on until a single cluster is formed. At the termination of a certain iteration, if all the points have not been clustered, the algorithm is restarted by selecting a new random point from the remaining un-clustered points. This goes on until all the points are clustered and the final clusters are obtained.

3.3.2. Ensemble Learning Algorithms

3.3.2.1. Random Forest

Random forest is an ensemble learning algorithm. Ensemble learning algorithms generally combine the outputs of various smaller base learners who have trained themselves on feature samples of the main dataset and then give the final output by either taking a majority vote or an average. The initial sampling step is called the bootstrap stage and the final combining step is called the aggregation stage. Random Forest follows a similar method by using Decision Trees as the base learning model. It then predicts the final output on the basis of a majority vote in case of a classification problem and on the basis of average in case of a regression problem.

3.3.3. Deep Learning Algorithms

3.3.3.1. Regression Neural Networks

Neural Networks consist of a layered structure of artificial neurons which are the basic functional units of the neural network. The first layer is the input layer, the last layer is the output layer, and all the middle layers are hidden layers that are mainly responsible for the learning process of the neural network. Whenever an input is fed into the neural network via the first layer, the input is first pre-processed if needed and then passed on to the hidden layers via channels having certain weights. These hidden layers consist of neurons that first calculate the values by taking a weighted sum of the input channels and adding a bias to it followed by passing it through an activation function. After passing through the activation function, some of the neurons get activated and only these neurons pass their outputs to the next hidden

layer. This process is called forward propagation. Now, at the end of the forward propagation, it may so happen that the predicted output is completely wrong. It is at this moment that the wrong predictions are compared with the actual values and the calculated errors are back propagated to re-adjust the channel weights. These movements of forward propagation and backpropagation help the neural network to set the channel weights to the most optimum values. As a result, after a series of forward and back propagations, the neural network learns to predict the output value with very high accuracy and precision.

3.3.4. Time Series Forecasting Algorithms

3.3.4.1. ARIMA

Time-series data is different from other types of data because it is univariate and also evenly spaced. ARIMA model assists in forecasting the future value of such a series in successive time intervals. It performs autoregression (AR) on previous instances to predict future instances, differencing i.e., Integration (I) to make the model stationary that is eliminating backward dependencies and moving averages (MA) to advance the time series forward by taking the mean of data points in its vicinity.

3.4. Implementation of System

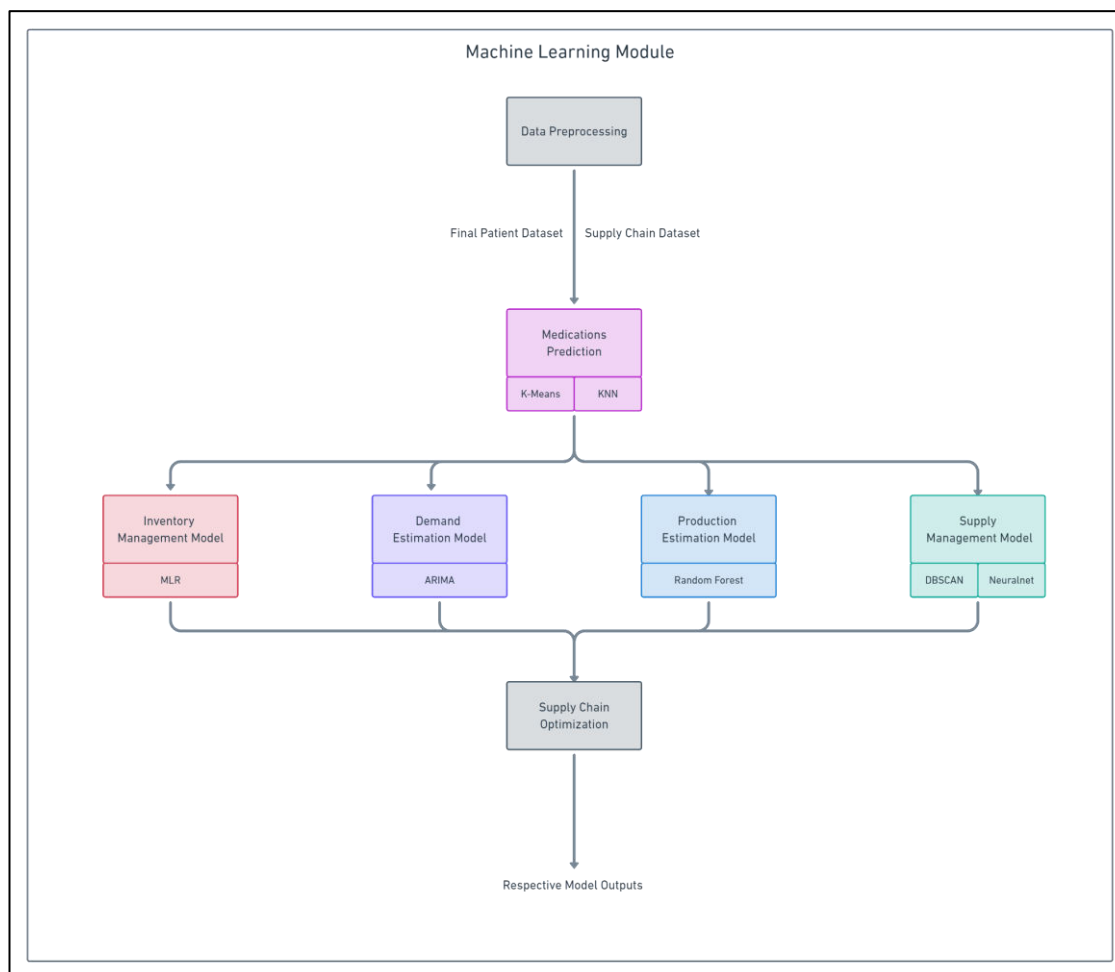


Fig. 1: Machine Learning Module Architecture

In this paper, data science and machine learning are applied on a sample dataset with the primary objective of predicting different medicines, which will be required if a person exhibits any side effects after the application of a vaccine. The objective is chosen to keep in mind the large-scale vaccination happening against the backdrop of the widespread pandemic but can be utilized as a general solution for similar situations. [6] is utilized, which in itself comprises three different datasets, namely, ‘VAERS Data’, ‘VAERS Symptoms’, and ‘VAERS Vaccine’. To convert these datasets into a single usable dataset, some data cleaning and preprocessing are done by handling all the missing values and by dropping all the unnecessary attributes. In the final dataset, the most important column i.e., ‘SYMPTOM_TEXT’, is the main dependency for predicting the required medicines. After this step, various supply chain techniques are implemented which are as follows

3.4.1. Medications Prediction

- The medications required by the patient would depend on the symptoms exhibited by the same. In the first step, a ‘Symptom2Vec’ representation of the text is generated by passing it through a TFIDF pipeline.
- Since there is a large number of meds in the dataset corresponding to the various symptoms, classifying the correct meds for a given tuple of symptoms is bound to return results with minimal accuracy.
- A better approach is thus employed where the meds are segregated into clusters based on the similarity of the TF-IDF vectors in order to generalize the medications according to intra-cluster similarity, which helps to achieve an appreciable amount of accuracy.
- Once the cluster labels have been assigned, a KNN classifier is trained on the vectors and labels, corresponding to which a test symptom tuple, when given as input, the classifier returns the k-nearest medications relevant to that instance, which serve as the predicted meds for that instance.

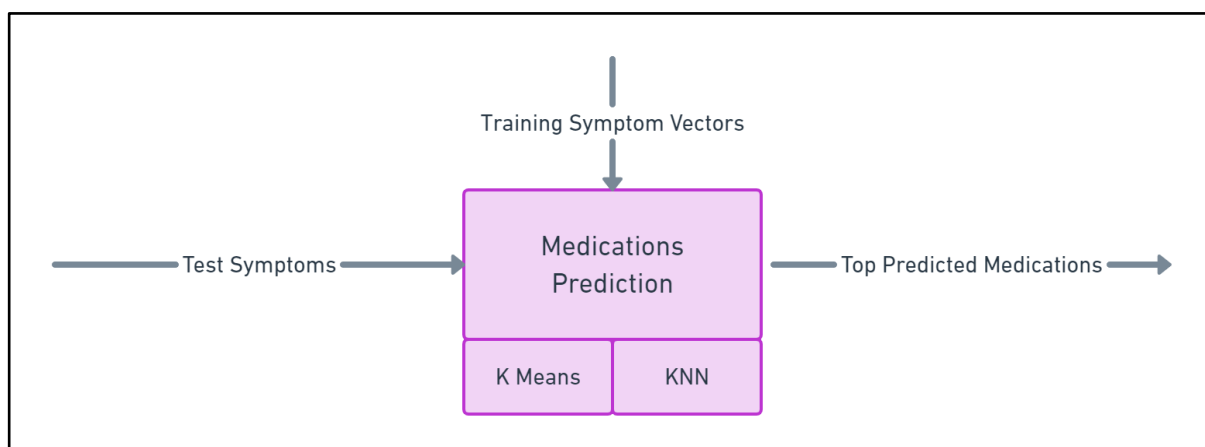


Fig. 2: High Level View of Medications Prediction

3.4.2. Inventory Management Model

- The core idea of this model is to predict how much of a particular medication actually needs to be stored in the hospital inventory.
- To predict this quantity, a multiple linear regression model is built which gives this quantity as a function of the 'weight' of the medications and a variable that stores the number of times a particular medication is predicted.
- This 'suggestion count' variable prevents the cold start problem in the regressor by assigning random seed values to each medication and updating it accordingly to the medications that get predicted.
- Once the regressor is trained, the meds which are predicted from the symptoms are used to subset test instances from [6] to be fed into the same and obtain the optimal number of medications that need to be stored in the inventory.

The model can be visualized using the following equation

$$\text{Medicine Quantity} = \theta_2 * \text{Medicine Weight} + \theta_1 * \text{Suggest Count} + \theta_0 \quad (2)$$

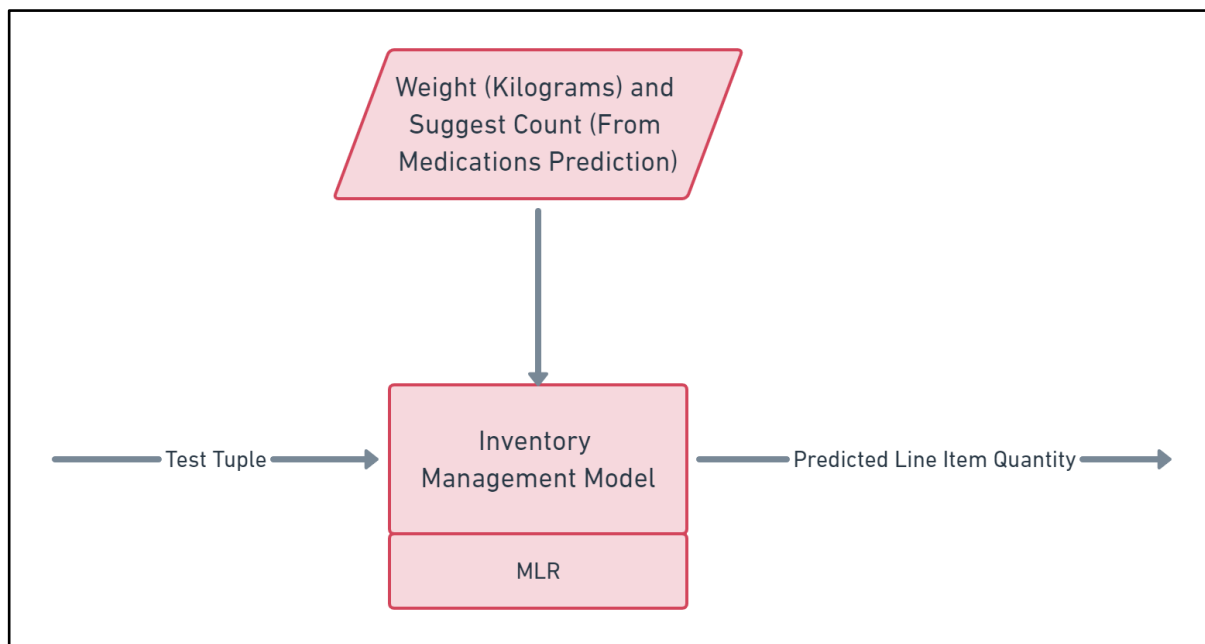


Fig. 3: High Level View of Inventory Management Model

3.4.3. Demand Estimation Model

- The aim is to obtain the item quantity and item value to be forecasted based on a time series for the medications predicted from the symptoms.
- ARIMA (Auto Regressive Integrated Moving Averages) model is used for this purpose. The item quantity and value are fed as input to the model and it trains on the previous instances of this

dataset at each iteration to give a forecast of the future which indicates the demand of the medications that will be required.

- This can be used to stock on the existing inventory in accordance with the predicted values.

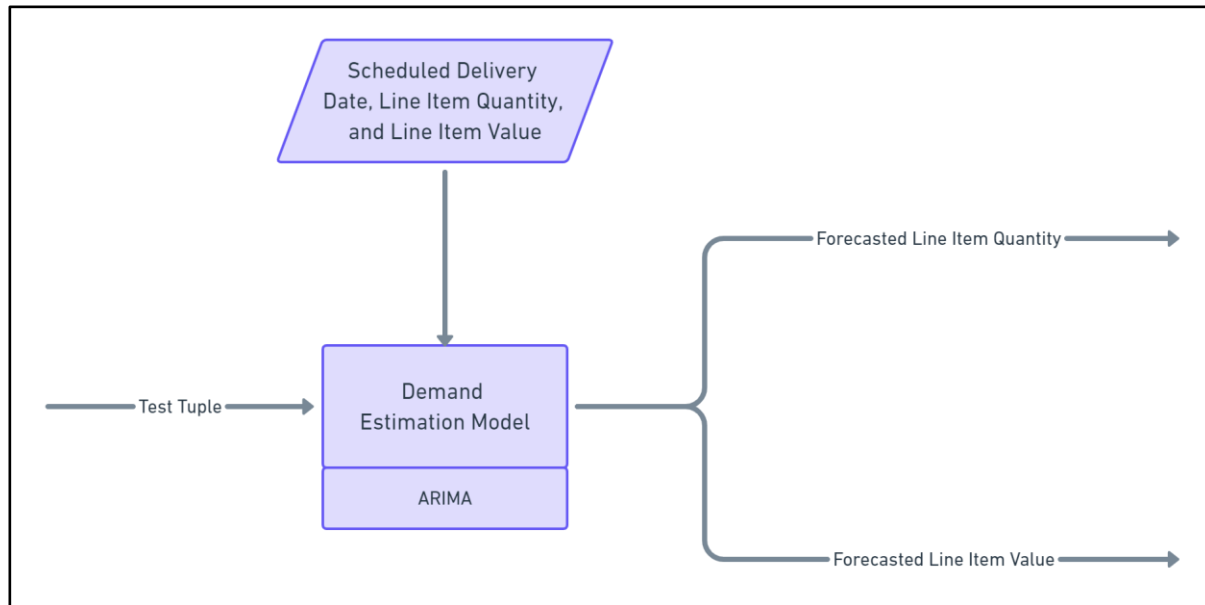


Fig. 4: High Level View of Demand Estimation Model

3.4.4. Production Estimation Model

- This model predicts the delivery time required to send the medications to the hospital, calculated as a difference between the scheduled delivery date and the date on which the vendor obtained the medications, present in [6]
- The delivery time depends on three factors – the medicine quantity, freight cost and the vendor involved
- Since the vendor is categorical data, it is label encoded for model training.
- The independent and dependent variables are fed to a random forest classifier in order to improve the accuracy of the delivery time so obtained.
- The predicted meds are then used to subset test instances for this ensemble classifier to calculate the estimated delivery time.
- Reverse label encoding on the vendor is done and then the medications are plotted against their delivery time and vendor to see which vendor needs to be chosen for a particular medication.

The model can be visualized as follows

$$\text{Delivery time} = \text{Scheduled delivery date} - \text{date of obtaining medications} \quad (3)$$

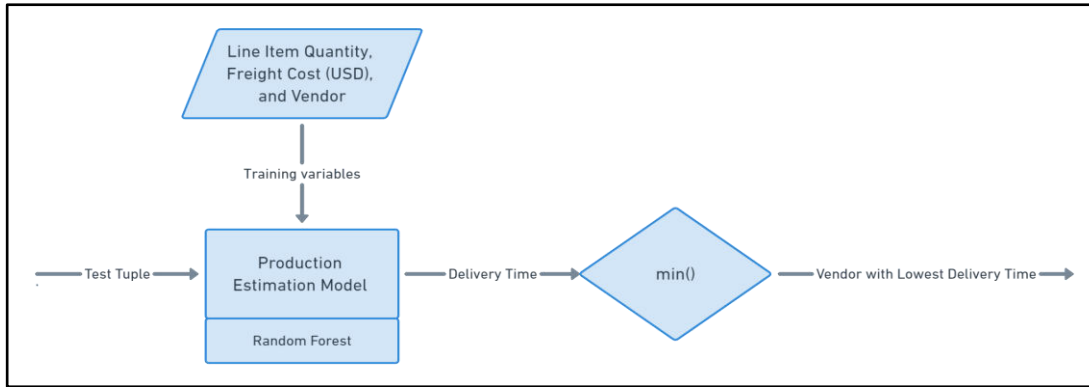


Fig. 5: High Level View of Production Estimation Model

3.4.5. Supply Management Model

- This is a culmination of all supply chain factors that make a vendor reliable like shipment mode, freight cost, shipment cost and vendor.
- An extra attribute called 'db_cluster' is added to this model which accounts for the geospatial distribution of supply chain manufacturing sites using DBSCAN as the underlying algorithm.
- The model is designed to predict the 'supply score' for each vendor, which is a linear combination of all the aforementioned attributes.
- The attributes are fed to a regressive neural network which learns the weights for the supply chain factors and returns the estimated supply chain scores.
- The test instances acquired from the predicted meds give the supply chain scores for the vendors and the model selects the one with the minimum score as the best vendor.

The model can be visualized as follows

$$\text{Shipment cost} = \frac{\text{Freight cost}}{\text{Weight}} * \text{Line Item Quantity} \quad (4)$$

$$\text{Supply score} = \text{Shipment Mode} + \text{Vendor} + \text{Freight Cost} + \text{Shipment Cost} + \text{dbcluster} \quad (5)$$

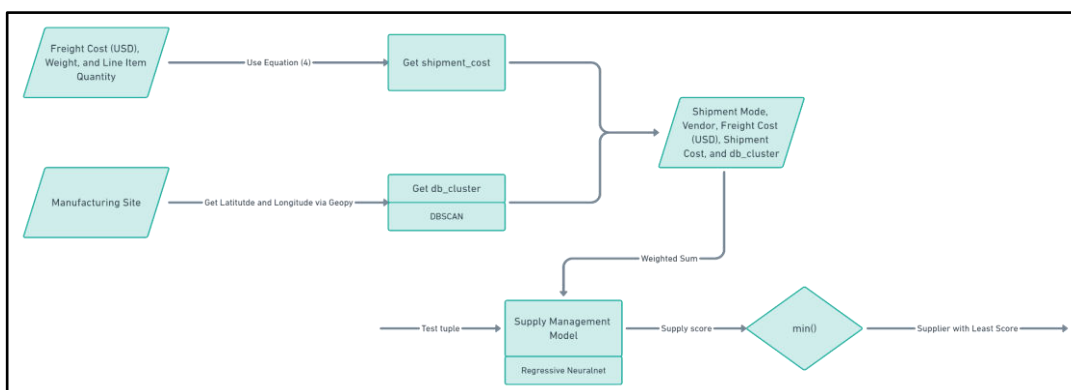


Fig. 6: High Level View of Supply Management Model

3. Results and Discussion

4.1. Medications Prediction

This part is solely concerned with predicting the appropriate medicines for the patient based on the symptoms he/she is showing. A combination of K-Means Clustering and K-Nearest Neighbors algorithm has been utilized to predict the medicines closest to the actual medicines which are used to treat the detected symptoms. The closeness that is used here proves to be quite effective because previously similar medicines have been clustered together and then KNN algorithm is used to find the appropriate medicines.

Figure 7 is the K-Means cluster assignment for the medications from [6].

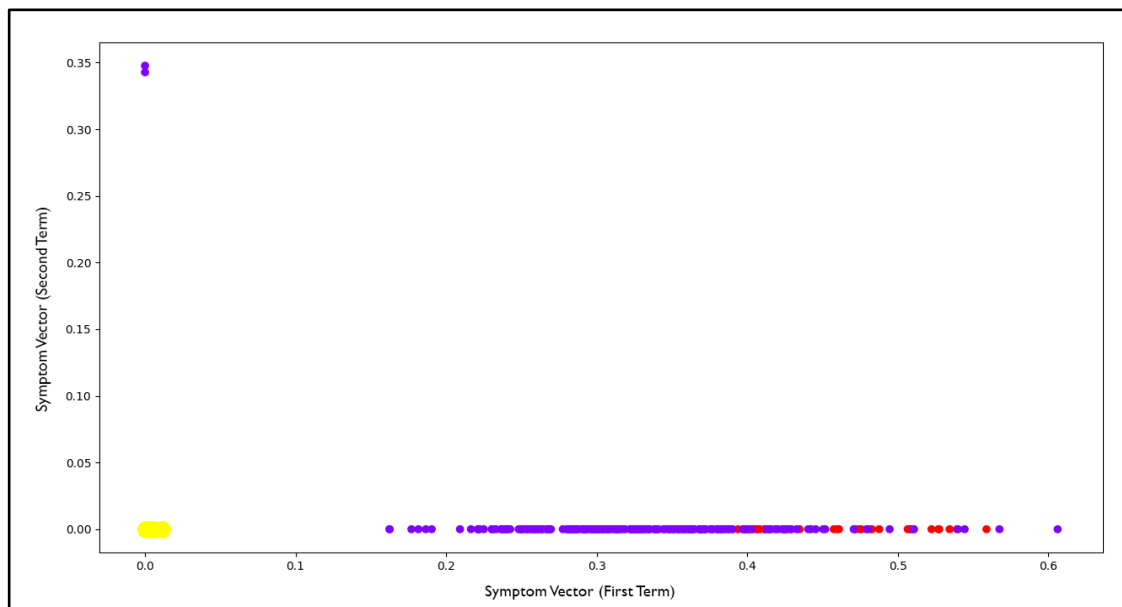


Fig. 7: K-Means Medication Clusters

For a sample input, table 1 shows the predicted medicines that are obtained after applying K-Means Clustering and KNN algorithm to the data in [6].

Index	Predicted Medicines
0	Flordipine Vitamin D 4
7875	Rampiril
7642	Sulfamethoxazole 800-160mg tab

Table 1: Predicted Medications for Input Test Symptoms

Now, the medicines in table 1 are used to address the Supply Chain aspect of the procurement by adding a column of randomly assigned medicines (which contain the predicted medicines) to the Supply Chain dataset and apply various algorithms to it for optimizing different Supply Chain techniques.

4.2. Inventory Management Model

In this model, Multiple Linear Regression algorithm is used to find the Line-Item Quantity. Below are the optimized Line-Item Quantity values predicted by the model for inventory storage.

```
[25299.98617691 18265.95742649 18288.31958614 ... 16304.16855157
25524.87183304 15841.52905083]
```

Fig. 8: Predicted Line-Item Quantities by Inventory Management Model

These are the ideal values of quantities for each medicine which should be kept in the medical inventory.

4.3. Demand Estimation Model

First of all, only those medicines which were predicted in section 4.1 are used to subset and create a new sub-dataset to train the model.

Then, ARIMA (Autoregressive Integrated Moving Average) model is used to predict the demand in terms of medicine quantity and value. Table 2 and table 3 exhibit the predicted values by the ARIMA model on the subsetted dataset. The red line denotes the predicted values and the blue line denotes the actual values.

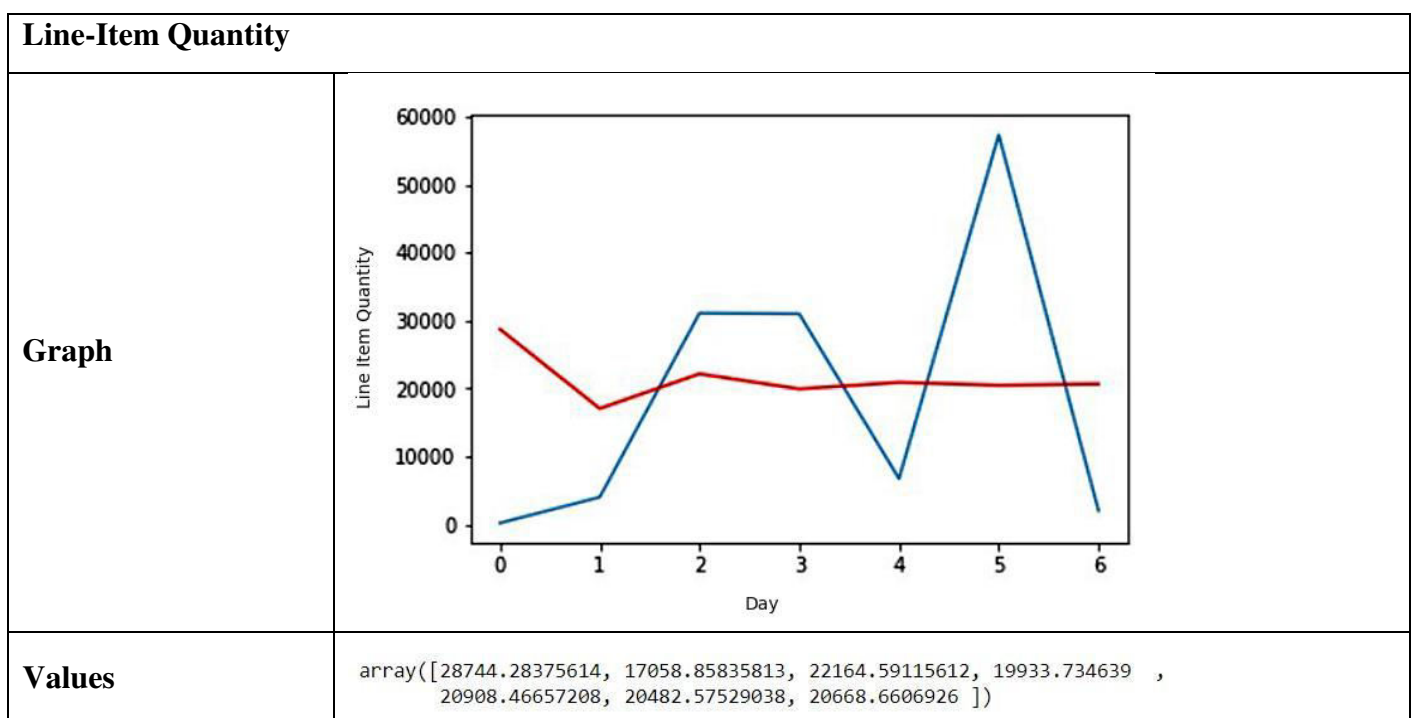


Table 2: Forecasted Line-Item Quantities by Demand Estimation Model

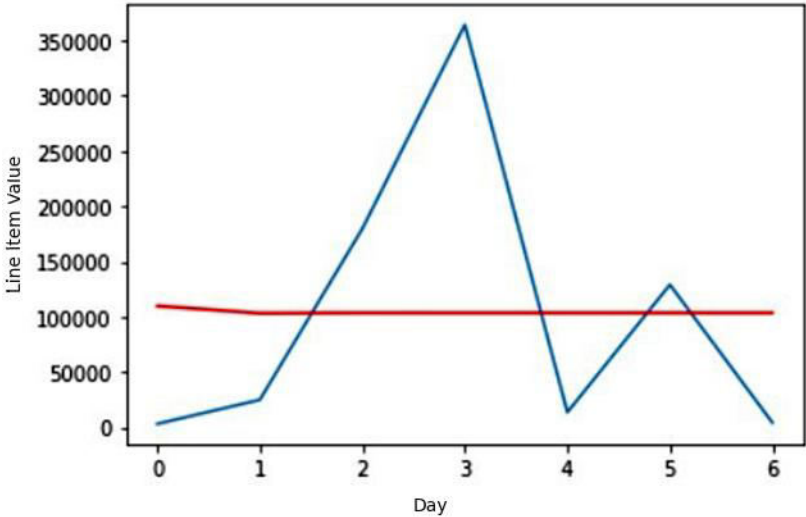
Line-Item Value	
Graph	
Values	<pre>array([109835.38660611, 103218.23368048, 103632.74168057, 103606.77630179, 103608.40281056, 103608.3009237 , 103608.30730604])</pre>

Table 3: Forecasted Line-Item Values by Demand Estimation Model

4.4. Production Estimation Model

In this model as well, the concerned data is first subsetted from [6] to get only the data related to the predicted medicines in section 4.1. After doing that, the vendor data is label encoded and Random Forest Regression is applied to predict the delivery times from where the vendors with the least delivery time can be visualized in figure 9. The lesser the height, the better the supplier.

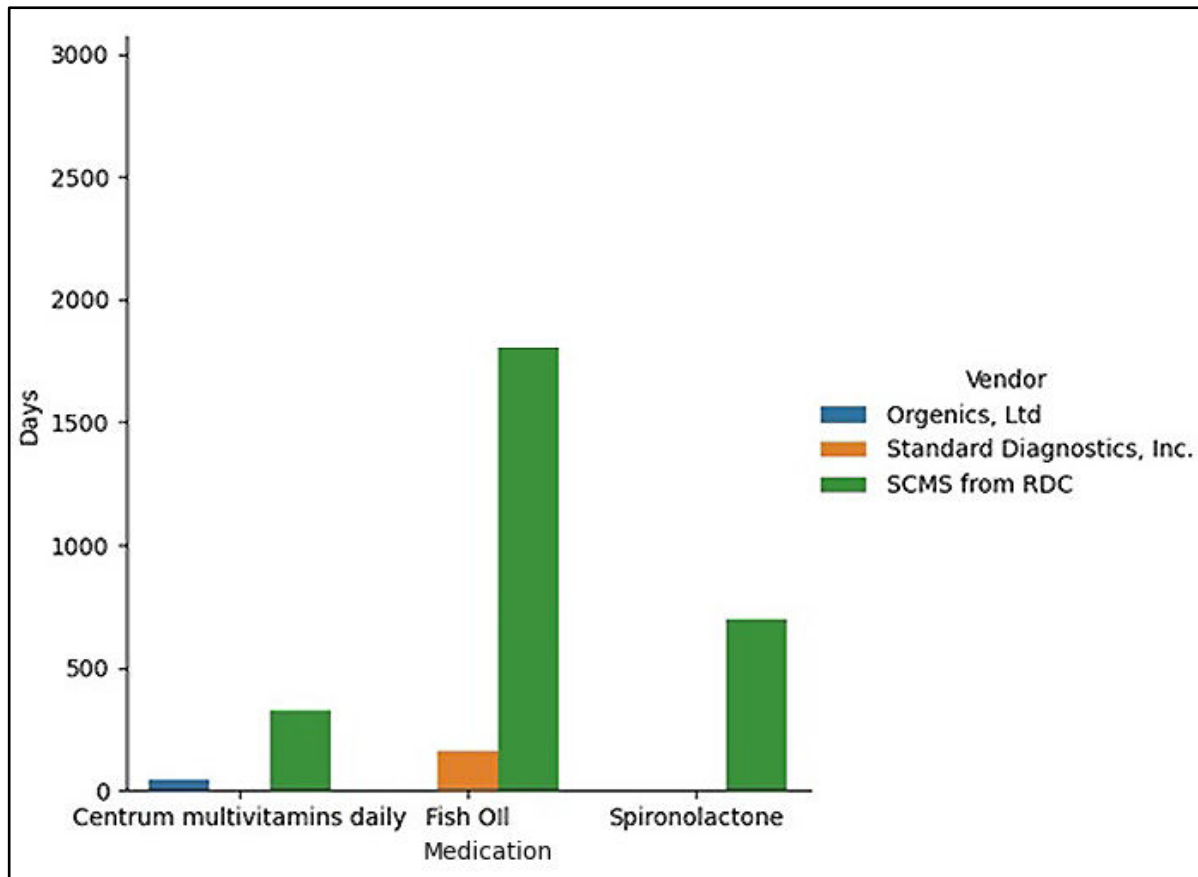


Fig. 9: Predicted Delivery Times by Production Estimation Model

4.5. Supply Management Model

In this model as well, the data concerned is first subsetting with the predicted medicines from section 4.1.

Subsequently, a 'Supply Score' needs to be predicted that tells the user about the best-suited supplier who would deliver the medicines in the least possible time based on location data. To get the location data, the Geopy library of Python is utilized to find the latitudes and longitudes of the supplier's manufacturing sites and then DBSCAN is applied to form the spatial clusters. These cluster assignments are used as an additional parameter for the input along with the other parameters mentioned in section 3.4.5. Figure 10 illustrates the spatial clustering of the manufacturing sites found by DBSCAN.

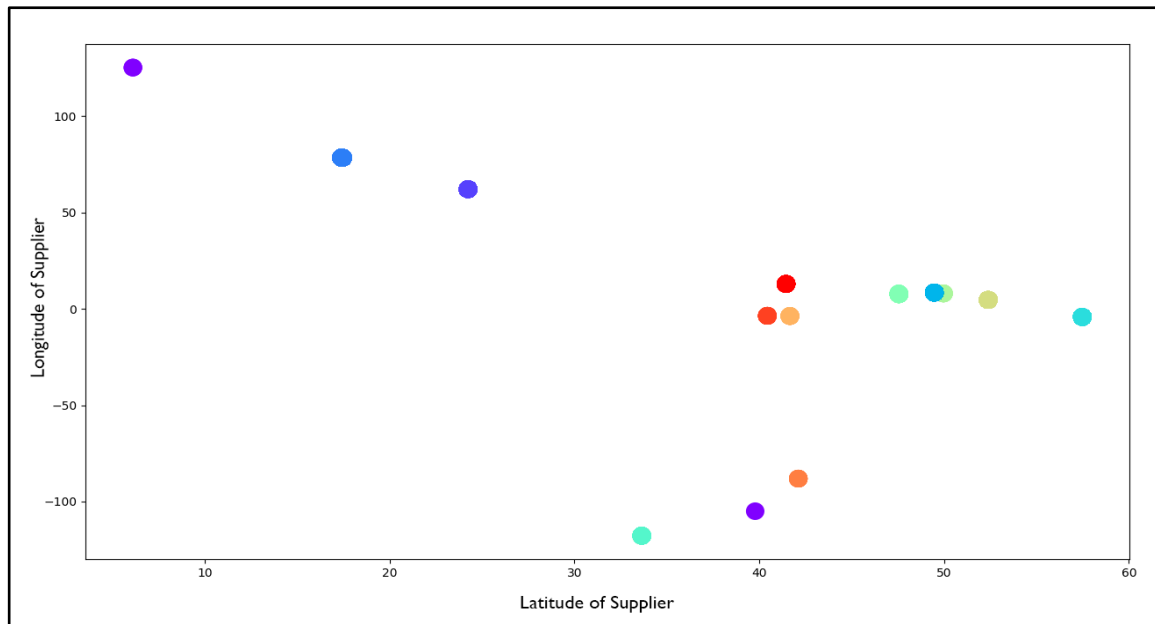


Fig. 10: DBSCAN Spatial Clusters of Medication Manufacturing Sites

Then, the parameters mentioned in section 3.4.5 are simply fed to a Neural Network which would learn the appropriate weights of the input attributes on its own and predict the label of the supplier with the least delivery time, i.e., the best supplier. Table 4 is a sample prediction of the supplier scores of all the vendors who sell the concerned medicines followed by the name of the best supplier in figure 11 for the medications predicted in section 4.1.

Supply Scores
2665.699574224041
3890.8910732334784
11159.310690467153
11162.831480272504
11166.884544390598
11162.82023109751

Table 4: Predicted Supply Scores by Supply Management Model

SCMS from RDC

Fig. 11: Predicted Best Supplier by Supply Management Model

4.6. Performance Analysis

The performance of all the Supply Chain Management models in terms of R-Squared Measure (R² Score) and Root Mean Squared Error (RMSE) have been shown in table 5 and the inference for the same has been provided in the following sections.

Supply Chain Model	Sub Model (If Any)	R ² Score	RMSE
Inventory Management	-	0.0903651	35810.7539922
Demand Estimation	Line-Item Quantity	0.9412234	227.1051092
	Line-Item Value	-0.2464457	20170.9075700
Production Estimation	-	0.1807163	898.5380967
Supply Management	-	0.9999999	2.1268791

Table 5: Performance Analysis of Every Supply Chain Model

4.6.1. Inventory Management Model

The variability in the SYMPTOM_TEXT underfits the model due to which the R² score is significantly less than 1 and the error measure is subsequently greater. However, the model does provide an approximation to similar medications by comparing the symptom vectors.

4.6.2. Demand Estimation Model

Variability in the time-series data due to different predicted medications at each iteration results in different error values after each execution. These error values can also be negative as is the case with the R² score of Line-Item Value which indicates that the proposed model doesn't follow the trend exhibited by the dataset instance, while the R² score is significantly closer to 1 and the error values are minimal in case of Line-Item Quantity which means a perfect fit.

4.6.3. Production Estimation Model

Label encoding the vendor attribute for the prediction of delivery time adds more information for training the model as both the variables have a causal relation. Moreover, since Random Forest is an Ensemble algorithm, accuracy increases slightly compared to the Inventory Model in favor of mis predicted values.

4.6.4. Supply Management Model

Since the supply score is a linear combination of multiple factors, the Neural Network is accurately able to learn weights for the prediction process and hence, this model has the highest accuracy of all the models with even the R² score being approximately equal to 1.

4. Conclusion

To conclude, first and foremost, the K-means algorithm aids to minimize the variability in the text symptom data to predict medications by assigning them to clusters by comparing their corresponding vectorized representation. Setting an optimal cluster count results in a prediction model based on distance rather than classification and this makes the system dynamic in terms of selecting the number of medications. Secondly, the model for inventory management generally provides a higher quantity value than the training data due to the presence of outliers which adds significant bias to the regression model and subsequently raises the total error. Thirdly, ARIMA model for demand estimation depends on the number of data tuples available for forecasting which is why the difference between actual and predicted values is higher. This can be rectified by constructing a subset of numerous medications for this model and evenly distributing the value among the medications actually required. It is, however, evident that the random forest model and the neural network model yields higher accuracy due to the ensemble structure in the former and the representation of the variables as a linear combination in the latter.

5. Future Work

The implementation of the system in this project can be improved upon in the following ways

- Data procurement could be done in real-time such that the system is capable of taking unknown instances and training the models accordingly to augment the accuracy rates.
- Inventory management model could be provided an extra parameter to indicate the inventory weight and allocation of medications can be further optimized using a dynamic knapsack problem approach.
- The demand model can be extended to analyze the seasonality of medication requirements i.e., the trends in the demand graph can be regressed to predict medicine requirements according to time events like peak urgencies.
- The production model can make use of path-tracking data to give a better estimate of the number of delivery days by various vendors.
- Supply score induction can be done by getting linear combinations of all necessary variables and utilizing the coefficients that are most accurate.
- Model training can be parallelized for the faster response time of results to the user.

Acknowledgment

We would like to acknowledge the scholarly articles cited as part of the literature survey for this research paper. We would also like to extend our gratitude to the authors of the references included as part of our practical analysis.

References

- [1] L. Subramanian, “Effective Demand Forecasting in Health Supply Chains: Emerging Trend, Enablers, and Blockers,” in *Logistics*, vol. 5, no. 1, p. 12, Feb. 2021. <https://doi.org/10.3390/logistics5010012>
- [2] Ni, D., Xiao, Z., & Lim, M.K, “A systematic review of the research trends of machine learning in supply chain management,” in *International Journal Machine Learning & Cybernetics*, 11, 1463–1482, Dec. 2019. <https://doi.org/10.1007/s13042-019-01050-0>
- [3] Vsrk, Prasad & Srinivas, Kolla & C.Srinivas, & Saleem, Sk.Abdul, “Supply Chain Management - Modeling and Algorithms: A Review,” in *Conference in Recent Advances in Mechanical Engineering (RAME 2017)*, Andhra University College of Engineering, Vizag, Mar. 2016.
- [4] Lee, Young Hae & Cho, Dong & Lee, Sang, “Optimization of Healthcare Supply Chain using Integrated Inventory and Distribution Planning,” in *international journal on Information*, 15, 6297-6304, Dec. 2012.
- [5] [VAERS Datasets](#)
- [6] [Supply Chain Dataset](#)