International Journal of Mathematical, Engineering and Management Sciences
Vol. 8, No. 3, 353-373, 2023
https://doi.org/10.33889/IJMEMS.2023.8.3.021

Integrated Model for Predicting Supply Chain Risk Through Machine Learning Algorithms

Saureng Kumar
Electronics & Computer Discipline,
Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India.
Corresponding author: skumar@pp.iitr.ac.in

S. C. Sharma
Electronics & Computer Discipline,
Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India.
E-mail: scs60fpt@iitr.ac.in

(Received on September 30, 2022; Accepted on February 05, 2023)

Abstract
The machine learning model has become a critical consideration in the supply chain. Most of the companies have experienced various supply chain risks over the past three years. Earlier risk prediction has been performed by supply chain risk management. In this study, an integrated supply chain operations reference (ISCOR) model has been used to evaluate the organization's supply chain risk. Machine learning (ML) has become a hot topic in research and industry in the last few years. With this motivation, we have moved in the direction of a machine learning-based pathway to predict the supply chain risk. The great attraction of this research is that suppliers will understand the associated risk in the activity. This research includes data pre-processing, feature extraction, data transformation, and missing value replacement. The proposed integrated model involves the support vector machine (SVM), k nearest neighbor (k-NN), random forest (RF), decision tree (DT), multiple linear regression (MLR) algorithms, measured performance, and prediction of supply chain risk. Also, these algorithms have performed a comparative analysis under different aspects. Among the other algorithms, the random forest algorithm achieves an accuracy of 99% and has accomplished superior results with a maximum precision of 0.99, recall of 0.99, and F-score of 0.99 with 1% error rate. The model's prediction indicates that it can be used to find the supply chain risk. Finally, the limitation and the challenges discussed also provide an outlook for future research direction to perform effective management to mitigate the risk.

Keywords- Risk prediction, Supply chain risk management, Supply chain operations reference, Machine learning, Customer demand.

1. Introduction

Organizational supply chain risk has become a predominant risk in today's supply chain management. Orenstein and Raviv (2022), resulting in a critical market for a product with a short life cycle, lead time, and increased supply chain risk factors (Kumar and Barua, 2022). These factors are supply risk, process risk, and demand risk. Many companies do not comply with the risk due to a lack of time, resources, data consistency, system compatibility, and technical difficulty in integrating risk management software. These risks are directly associated with customers and suppliers. it directly impacts the organization's productivity. In order to increase the productivity, the several data mining techniques, supervised machine learning (SML) and artificial intelligence (AI) techniques is used for resilient performance. Because, data mining removes hidden predictive planning information and returns the information datasets to resilient supply chain performance. The supervised machine learning (SML) approach that assumes a link between class level and features (Cavalcante et al., 2019) for the classification. Building a training dataset, for creating a module, and testing process is used to examine the predictive efficiency and resilient supply chain. Presently, the AI method has been able to delegate resilient in the field of supply chains. The authors Benjaoran and Dawood (2005) utilized SVM for categorizing the various risk within the supply chain (Baryannis et al., 2019). The

author claims the SVM is an effective approach to monitoring the effect in the supply chain (García et al., 2012).

Over the past three years, supply chain risk event has increased, a digital survey was conducted through a google form for the supply chain complexity are shown in Figure 1. The Supply chain operations reference (SCOR) is a diagnostic tool for the process reference model. This will enables next-generation supply chain research for the organization's performance, effectiveness, and implementation (Huang et al., 2005). SCOR is a step-by-step procedure consisting of three main processes (source, make, deliver) that help the organization to optimize the supply risk, process risk, and demand risk to enhance the productivity. In the last few years, there has been a growing interest in AI-based supply chain research (Cavalcante et al., 2019) to optimize risk and productivity. Managing these risks in the supply chain plays a crucial role in the organization. Several publications have appeared in recent years to overcome the risks, a computational models such as artificial neural networks (Teuteberg, 2008), deep neural networks (Wichmann et al., 2020), machine learning (Baryannis et al., 2019), and deep learning (Amani and Sarkodie, 2022) are used in the supply chain to mitigate the risks.

But most of the researchers only focused on risk assessment through Artificial intelligence, our study is based on ISCOR based risk prediction model for the organizational supply chain through support vector machine, K-nearest neighbor, random forest, decision tree and multiple linear algorithms.
The contribution and novelty of our studies are as follows:

- A SCOR based Integrated model is proposed for risk prediction.
- The various supply chain risk has analyzed and optimized the organization’s performance.
- ML algorithms has applied to manage the supply chain risk and measuring the performance parameter.
- Perform comparative analysis under different techniques of the ML algorithms.

This analysis was performed in two steps. Firstly, read the data and check the data with the process area risks, and then assess the risks. Secondly, the applicability of a various machine learning algorithms that predicts the F score, precision, error rate, accuracy, and recall. For this research, we taken a risk dataset from various sources studies on various supply chain risk of the organization as illustrated in Table 2, online and offline discussion and survey form for risk identification for the prediction. We found our algorithm outperformed with 99% accuracy.

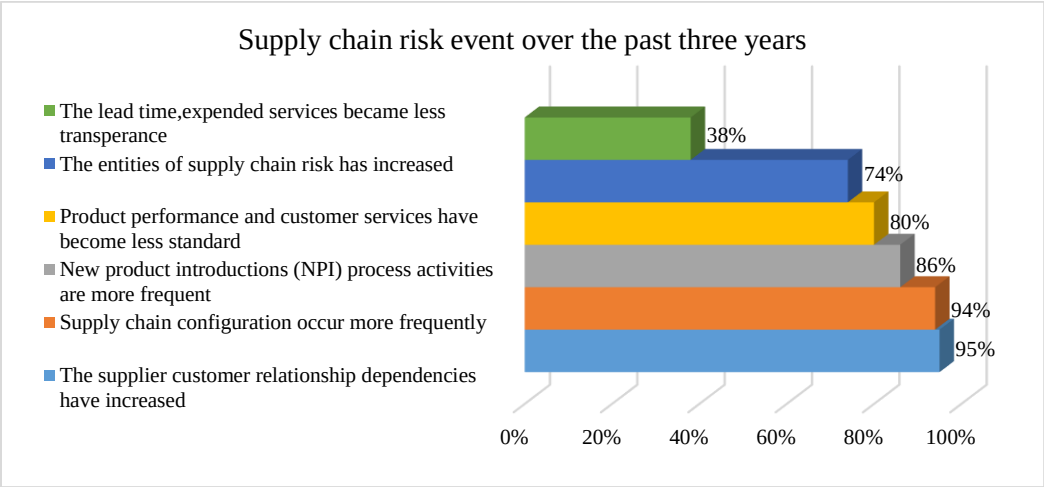

Figure 1. Supply chain complexity over the past three years.

The remainder of the paper is organized as follows. The next section presents an introduction to the SCOR model. We have given a sources study on various supply chain risk of the organization in the literature review section and discussed the area of applicability in the subsequent section. Furthermore, we proposed an integrated model for the risk prediction in a broader area that defines our research work and the scope of the study. Then we briefly discuss the different machine-learning algorithms used in this work. In the penultimate section, we summarized the performance measure of the algorithms and the final result and discussion, which concludes in this paper.

2. The SCOR Model

The SCOR model was early adopted by the supply chain council (SCC) in 1996 to evaluate the organization's strategic decision-making. The SCOR model is a process with a reference management tool for effectively assessing an organization's supply chain. Specifically, the model provides the standard description of the supply chain process in a unique format. The model incorporates the organization's process re-engineering, benchmarking, and measurement into its framework. This framework repeatedly focused on three areas of the supply chain, i.e., source, make, and deliver, spanning from supplier, manufacturer, distributor, retailer, and customer.

The model itself is organized into three primary processes.
Source: The source processes describe the activities related to the supply and acquisition of goods and services to meet the actual demand as per the customer's need. These include issuing purchasing orders, receiving, inspecting, holding, scheduling delivery, validating orders, and accepting suppliers' invoices.

Make: The making process includes a finished product to meet as per the plan and actual demand by the customer. These processes make to order (MTO), make to stop (MTS), assemble to order (ATO), configure to order (CTO) and engineer to order (ETO). These include manufacturing, assembly, testing, maintenance, repair, overhaul, packing, recycling, and eventually releasing the finished goods.

Delivery: An operation involved in the delivery of goods and services. This includes customer order management, storage, warehouse, transportation, and distribution management.

In every stage, planning is required to operate in the forward and reverse supply chain. In each step a risk is associated. The category of risks in the SCOR management process in illustrated the Figure 2. Due to a lack of sufficient market intelligence, process, and information system, various risks that may occur during the supply chain from supplier to the customer that wouldn’t be effectively predicted and mitigated these risks.

To better identification of different risks in the supply chain. Other perspectives need to be understood, like supply risk, process risk, and demand risk. A supply risk means that an organization could face loss due to material flow interruption within the supply chain that impact the production held up. The process risk is the potential risk leading to manufacturer process difficulties in a capacity bottleneck, lead time, quality, human error, output, etc. Demand risks are potential losses due to the gap between unpredictable forecasts to the actual demand. Failure inaccurately predicted demand could disrupt a sub-optimal performance like sales loss due to insufficient inventory. To identify this risk, we have applied a machine learning algorithm in our study, in order to measure the risk performance in the supply chain.

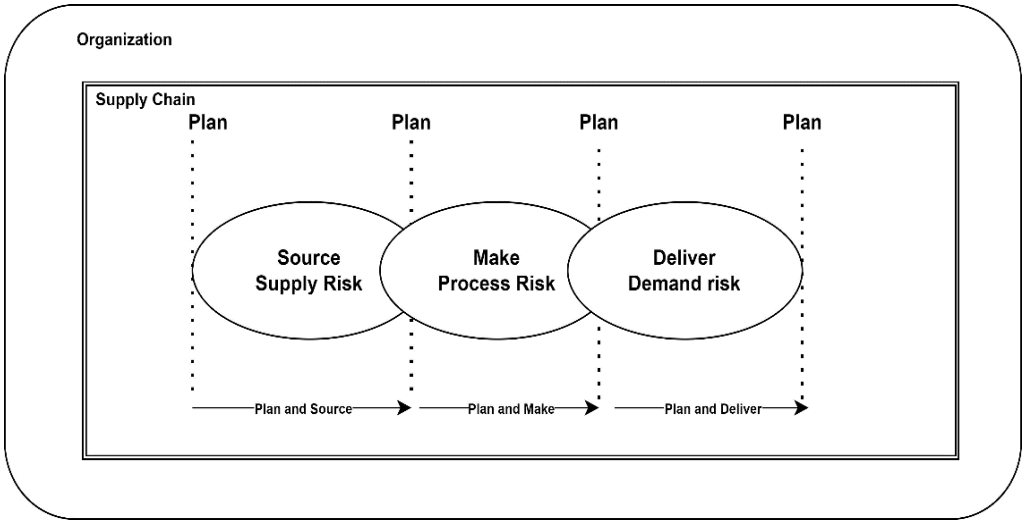

Figure 2. Categories of risk in the SCOR management process.

3. Literature Review

We want to perform the research work to achieve our goal, so that the researcher gets advantages from our research despite of any difficulty. Therefore, we explore the research problem by reading some research work. These research works were published in a high-impact journal. The research work was classified according to their respective scope of application, a method used, considered risk, and study are summarized in Table1.

Table 1. Summary of studies on various risk in supply chain.

Authors	Approach/Method	Scope of Application	Considered Risk	Limitations/ Future Research
Wichmann et al. (2020)	Deep Learning	Supply Chain	Supply risk	The study was performed on a small dataset. Transform attention mechanism model was not considered in their research.
Baryannis et al. (2019)	Data-driven AI technique	Production	Supply risk	A more feature-rich dataset for AI techniques (Neural network, Deep learning) should be taken to ensure the interpretability for the risk assessment model and response.
Cavalcante et al. (2019)	Machine learning & Simulation	Transport	Supply risk	The study was performed on a small dataset, and Simulation can be stretched out, including product variability, distribution costs, and other expenses. Design of risk mitigation and supplier base haven't been focused on in their research.
Benjaoran and Dawood (2005)	Artificial Intelligence Planner (AIP)	Production	Process risk	The research was based on concrete manufacturer through supply of bespoke concrete products. Offsite components are required for effective progress and reduce the waste bespoke product.
Hassan (2019)	Machine Learning	Supply Chain	Supply risk	The research was performed on the unstructured textual data, but parameterization of the kernel functions was not considered that could help the researcher find out risks or opportunities.
Layouni et al. (2014)	State of the art method	Production	Demand risk	The research was limited to simple defect geometry. Other defects like Stoichiometric defect, Frenkel defect, and Schottky defect were not considered in their study.

Table 1 continued…

Rodriguez-Aguilar and Marmolejo-Saucedo (2019)	Statistical Modelling	Supply Chain	Supply risk	Loss of information may occur from grouping the observations.
Yong et al. (2020)	Machine Learning	Supply Chain	Supply risk	Study limited to enterprise or agency only. However, other parameters like in transit supplier information, system log record estimate and analysis, waste management etc., must be collected and analyzed through ML.
Alfian et al. (2020a)	RFID Technology and Machine learning	Transport	Demand risk	To track perishable food supply, a RFID technology is incorporated for data recorder, However, Miss-reads value creates data corruption or failure to record data not addressed in this study.
Brintrup et al. (2020)	Supply chain data analytics	Transport	Supplier risk	In Bayesian models, Decision Trees were not performed in their study.
Alfian et al. (2020b)	Raspberry pi-based sensors	Transport	Demand risk	Improvement on the level of digitalization, food advocacy, source, and safety. Research is limited to humidity and temperature parameter.
Blackburn et al. (2015)	Predictive analytics	Production	Demand risk	Non-negative demand data to be taken into account for the specific application.
Bouzembrak and Marvin (2019)	Bayesian Network (BN) approach	Transport	Transport risk	Expert knowledge needs to be included in the model for better accuracy.
Constante-Nicolalde et al. (2020)	Machine Learning	Supply Chain	Quality Risk	False-positive control probability detection, clickstream analysis.
Fu and Chien (2019)	UNISON data-driven framework	Production	Demand risk	Research can be extended in the direction of Intermittent or lumpy demand management.
Lau et al. (2005)	Neural network	Procurement	Information Risk	The research is based on support procurement decisions and studies unable to perform the sensibility characteristic like the environmental change.
Pereira and Frazzon (2019)	Predictive approach	Supply Chain	Sales risk	Product to be improved by the insertion of more product parameters.

The potential of applying machine learning algorithms to predict supply chain risk has been recently considered. As we can see, most of the research papers has identified the risk, with only a few studies considering the supply chain operations reference (SCOR) model to identify the various risk events (Huo and Zhang, 2011), risk factors (Ríos et al., 2019), and the risk causes(Tama et al., 2019). To assess these risks, we have applied a machine learning algorithm in our research, and also, we prioritized these risks such as low, medium, high and severe to improve the supply chain performance.

4. Integrated Model for Risk Prediction

In this section, we have considered five different ML algorithms. Figure 3 shows the overall working process for the proposed model. The risk prediction involves two major stages risk identification and preparation of a risk register. With this model, the supplier (source), the manufacturer (make), and the distributor (deliver) will identify the associated risk in the process area, and the organization has to offer some helpful information to mitigate the risk. Furthermore, five ML models have been applied to train the model such as Multiple Linear Regression, Decision Tree, K Nearest Neighbor, Random Forest, and Support Vector Machine. These models are most commonly used for predicting the class of a given data point. Therefore, we used them in this study. The integrated SCOR model helps to evaluate the performance of the organization. The integrated SCOR model is discussed in Section 4.1, and the rest of the study is offered in the next section.

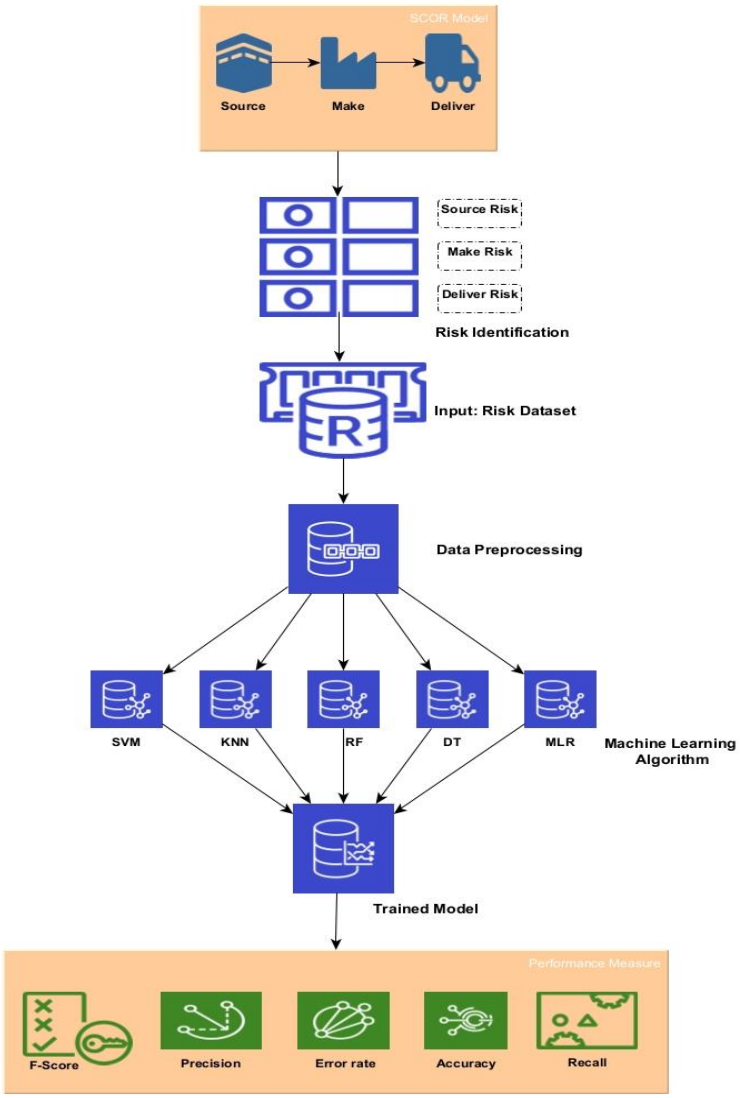

Figure 3. Integrated SCOR model for supply chain risk prediction.

4.1 Integrated SCOR Model

Integrated model for supply chain operation reference (ISCOR) is the process reference model that is used to improve and communicate the supply chain risk between the customer and supplier inside an organization evaluate. The ISCOR is based on supply chain risk experience in the various organization's practitioners. ISCOR divides the supply chain processes into three entities i.e., source, make and deliver. Source or supplier which supply the raw material, product or services to the organization as per the customer need. Make or manufacturer involves in the manufacturing or production and the delivery manages the transportation of the product. During the whole supply chain, the risk was identified. A brief detail of the attributes is shown in Table 4 were utilized in the proposed model. In order to evaluate the risk, it is necessary to identify the severity of impact and risk score. So, adopting an integrated model for supply chain operation reference is one of the paper's main objectives to provide a way for determining the risk scores of the

attributes and measures of supply chain risk. Further, the data preprocessing method converts the raw dataset into a suitable machine learning model. After that a different machine learning algorithm (SVM, KNN, RF, DT, MLR) has applied to train the model and measure the prediction performance of supply chain risk. In order to measure the performance, we have taken decision outcomes into five categories by considering the possible decision (true positive, false positive, true negative, false negative).

- True positive: Activity sample of actual value and predicted value both are positive.
- False positive: Activity sample of actual value is negative and the predicted value is positive.
- True negative: Activity sample of actual and predicted value both are negative.
- False negative: Activity sample of actual value is positive and predicted value is negative.

$Fscore$ is generally used to test the performance at a threshold. It is the harmonic average of system's precision and recall value. The $Fscore$ value ranges between 0 and 1. The value closer to 1 indicate the system model outer perform.

$$Fscore = 2 \times (Precision \times Recall)/(Precision + Recall).$$

Precision is the exactness of the model. It quantifies number of positive classes in the activity sample actually belongs to the positive class. However, the Recall quantifies number of positive classes prediction made from all positive class sample.

$$Precision = \frac{(True\ positive)}{(True\ positive + False\ positive)}.$$

$$Recall = \frac{(True\ positive)}{(True\ positive + False\ positive)}.$$

Error rate is the instance of activity sample for which model made incorrect prediction.

$$Error\ rate = \frac{(False\ positive + False\ negative)}{(Positive + Negative)}.$$

Overall accuracy of the model is the average of correct predictions over the total prediction. The accuracy of an activity sample can be calculated as given below:

$$Accuracy = \frac{(True\ positive + True\ negative)}{(False\ positive + False\ negative + True\ neagative + False\ positive)}.$$

The performance evaluation of these ML algorithms is illustrated in Table 5.

4.2 Risk Identification

Risk identification is the primary step in the supply chain risk management (SCRM) process. The author (Huo and Zhang, 2011) has addressed the process risk in the supply chain described by the SCOR model for retail enterprises and investigated the various phases like planning, purchasing, sales, delivery, and return. They suggested different associated risk factors in the planning, procurement, sales, and distribution processes to analyze customer satisfaction. Another author (Zhu et al., 2019) analyze the source and driving factor of the supply chain risk. Especially in the dairy supply chain. The study adopted value-focused process engineering (VFPE) that identifies the functional risk objective in the supply chain process. The case of increasing supply chain sustainability through driven factors such as social supply chain risk and how to mitigate this risk (Mani et al., 2017).

- Plan risk is caused by inadequate planning in the supply chain process, which results in ineffective management.
- Supply risk is caused by material flow interruption within the supply chain and the inability to meet customer demand.
- Process risk is caused by a flawed or lack of an exemplary manufacturing process, which reduces failure rate and defectiveness.
- Deliver risk is caused by disruptions or delays in the delivery of goods, which amounts to producing products and on-time delivery.

We further categorized the risk event into three process areas source, make, and deliver the result are summarized in Table 2.

Table 2. Process area-wise risk identification.

Process Area	Considered Risk	Risk Identification	Likelihood	Severity	Risk Score	Risk Prediction (L, M, H, S)
Plan	Planning Risk	Lack of production planning schedule	1	2	2	L
		Incompatibility allocation of human resources	2	1	2	L
		Lack of raw material planning	2	2	4	L
		Discrepancies between actual inventory and inventory capture in record	3	1	3	L
Source	Supply Risk	The supplier unable to fulfil the customer requests	3	2	6	M
		Increase in raw materials price	3	2	6	M
		The supplier's packaging cannot meet the demand	3	3	9	H
		Increase in the packaging price	3	1	3	L
		The supplier shut down the manufacturing	2	2	4	L
		The supplier reported physical damage to the product	2	1	2	L
		The supplier cannot meet the demand due to a monopoly situation	2	2	4	L
		Price increases for quality of material	3	2	6	M
		Difficulty in the selection of raw material	3	3	9	H
		Increase in the supply market price	3	1	3	L
		Delay in delivery of raw materials due to demand variation	4	1	4	L
		Delay delivery in packaging	2	2	4	L
		Delay delivery due to delivery date change	4	1	4	L
		Delay delivery due to plant shutdown	1	2	2	L
		The manufacturer unable to fulfilment of the customer's order	4	3	12	H
		Substandard raw material	4	2	8	M
		Change in packaging technology	4	2	8	M
		The packaging depends on supplier concentration	4	4	16	S
		Lack of packaging equipment	2	3	6	M
		Lack of skilled manpower for packaging	4	3	12	H
		Packaging damaged at cargo unloading	2	2	4	L
		Strategic alignment does not match with the supplier	2	2	4	L
Make	Process Risk	Shortened lead time production from the predetermined schedule	4	3	12	H
		A technical problem occurs during the production process	3	3	6	M
		Personal training is not appropriate	1	3	3	L
		Machine damage produces rejected products	3	2	6	M
		The production area is not clean	1	1	1	L
		The production process flow and the quality control flow not match the standard	2	2	4	L
		Reject products passed inspection	1	2	2	L
		Additional labour costs to segregate the foreign objects from the product	2	2	4	L
		Improper process control in the packaging process	2	2	4	L
		Production results are not matched with targeted	3	3	9	H
		Failure to maintain the storage conditions in the warehouse	3	2	6	M

Table 2 continued…

Deliver	Demand Risk	Inadequate logistic information to the customers	3	3	9	H
		Product damage due to distribution system problem	3	4	12	H
		Sudden external environment and policy changes in the shipping process	4	4	16	S
		Lack of visibility and inaccurate understanding of the delivery to the customer	4	2	8	M
		Late payment for shipping from the customer	2	2	4	L

The risk score has quantified as.

$$Risk\ Score(R_s) = Severity\ of\ Impact \times likelihood\ of\ occurrence.$$

The risk score has used to categorized the above risk (For Low $R_s \leq 4$, For Medium $R_s \leq 8$, For High $R_s \leq 12$, and For Severe $R_s \leq 16$).

In addition to that, the researcher(Baryannis et al., 2019b) has applied a machine learning techniques and their applicability to identify the supply chain risk and optimize the organization's performance.

4.3 Risk Register

A risk register is a risk management strategy to identify the potential risk, cause, and response to implement the supply chain performance through the risk prediction. It gives a visual reference to determine the risk that can be immediate corrected. Table 3 shows the (4×4) risk identification matrix for the different risk categories such as low, medium, high, and severe risk. This risk is based on the frequency and severity of impact. The almost certain or dangerous risk that occurs with a significant frequency can be severe and need immediate attention. However, unlikely events occur with little or even no impact organization's profitability.

Table 3. Likelihood and severity of impact model.

	Severity of Impact				
	Frequency	Minor	Moderate	Significant	Severe
Likelihood of occurrence	Almost Certain	M	M	H	S
	Likely	L	M	H	H
	Moderate	L	M	M	M
	Unlikely	L	L	L	M

Risk Prediction: L-Low, M-Medium, H-High, S-Severe.
Likelihood of occurrence: Unlikely-1, Moderate-2, Likely-3, Almost Certain-4.
The severity of Impact: Minor-1, Moderate-2, Significant-3, Severe-4.

Hence, the process leads to constructing a functional-based risk matrix based on driven factors for the risk prediction index system to determine how likely they are to occur and the severity of impact.

4.4 Data Pre-processing

Data pre-processing is a data mining technique that can transform the data and make it suitable for machine learning models. The collected data has got some missing values. This technique finds the missing value data from the collected data and replaces this missing value. The missing or null value is replaced with the median value, also known as data cleaning. Through the data cleaning process, we solved the missing value problem. Then after the pre-processed data was fed into the ML model and classified the risk through this, we completed the data transformation.

4.5 Machine Learning Algorithm

A Machine learning algorithm is the branch of artificial intelligence that can be applied to enhance the intelligence capability in various real-world applications. Different machine learning models have available to perform the specific task based on experience or data on a large scale. In our research, we have used SVM, KNN, RF, DT, and MLR to predict the supply chain risk.

4.5.1 Support Vector Machine

In recent years, SVM has gained recognition in the scientific and engineering area. It is flexible enough for both data learning regression as well as classification. It is used for both one class and binary classification problems. A support vector machine predicts and classifies data in charge of building the best hyperplane in an infinite-dimensional space. The best hyperplane builds the most significant margin used for linear and nonlinear classification and outlier detection (Debruyne, 2009). To solve the binary classification problem, the trained dataset T_d are as follows:

$$\text{Trained Dataset } (T_d) = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_n, y_n)\}, y_i \in \{-1, 1\}.$$

$x_i \in R^d$, where x_i denotes datapoint & equivalent y_i Represents design label n indicates the number of components in the trained dataset. Figure 4. illustrates the optimal hyperplane using the SVM algorithm. Our goal is to maximize the soft margin.

These are two standard formulations for soft margin by resolving the optimization process.

$$\text{Minimize}\left\{\frac{1}{2}\beta'\beta + C\sum_{i=1}^{n}\xi_i\right\}, \xi_i \geq 0 \quad (1)$$

Such that

$$y_i(\beta^T x_i + b) \geq 1 - \xi_i \quad \text{where, i=1, 2, n} \quad (2)$$

C denotes the penalty parameter, ξ indicates the positive slack variable β represents the normal vector, and b indicates scalar quantity. C keeps the allowable value decreased by the Lagrangian multiplier α_i that could attend Karush-Kuhn-trucker state (Chao and Horng, 2015) where optimally $\alpha_i >, 0$ then the x_i equivalent data called support vector (SV) and SVM tries to make decision boundary with optimal hyperplane variable w and b in the succeeding formula.

$$f(x) = sgn\left(\sum_{i=1}^{n}\alpha_i y_i\, x_i^T + b\right) \quad (3)$$

This equation leads directly by its unrestricted dual formation

$$Max_\alpha = \sum_{i=0}^{n}\alpha_i \; - \; \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\,\alpha_j x_i x_j y_i y_j \quad (4)$$

Subject to the constraints

$$\sum_{i=1}^{n} y_i \alpha \alpha_i = 0 \quad (5)$$

$$C \geq \alpha_i \geq 0.$$

The final set inequality $C \geq \alpha_i \geq 0$ shows box constraints, and the gradient equation for b gives the solution b in terms of a set of nonzero α_i which corresponds to the support vector, and the w denotes linear integration of the trained vector.

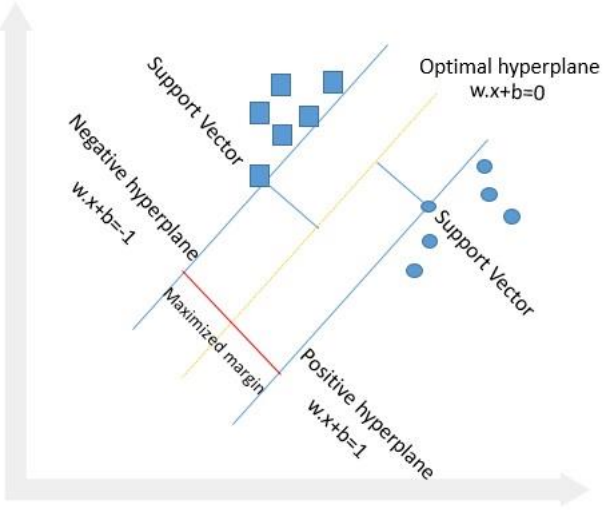

Figure 4. Optimal hyperplane using SVM.

$$w = \sum_{i=1}^{n} \alpha_i \, y_i x x_i \tag{6}$$
$$b = \frac{1}{A_{sv}}(w x_i - y_i).$$

Here, A_{sv} is the amount of SV. Proper feature mapping can make nonlinear to linear by using SVM on the feature space$\{\emptyset(x_i)\}$ only need$\emptyset(x_i)^T \emptyset(x_j)$.

$$f(x) = \text{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \, k(x_{i,} x_j) + b\right) \tag{7}$$

where, kernel function $k(x_{i,}x_j) = \emptyset(x_i)^T \emptyset(x_j)$. Although the linear classifier gives us a nonlinear classifier with a polynomial of infinite power, the radial basis function (RBF) is a powerful kernel for any complex dataset, for its reliable and accurate efficiency, is given by

$$k(x_{i,}x) = \exp\left(-\frac{\|x_i - x\|^2}{2\sigma^2}\right) \tag{8}$$

Here $x_{i,}x$ are vector points in the fixed dimensional space.

$$Max_\alpha = \sum_{i=0}^{n} \alpha_i \; - \; \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \, \alpha_j y_i y_j \exp\left(-\frac{\|x_i - x\|^2}{2\sigma^2}\right) \quad \text{by using equation (4)}$$
$$C \geq \alpha_i \geq 0.$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{9}$$

The final set of inequalities, $C \geq \alpha_{,i} \geq 0$ shows C keeps the allowable values of the Lagrange multipliers αj in a box constraint. And the gradient equation solution b gives a set of nonzero α_i.Which correspond to the support vectors.

4.5.2 K-nearest Neighbor

K-NN is the non-parametric supervised learning technique, and it has not made any presumption on the elementary dataset. It is used for both classification and regression problems. KNN classification needs a distance metric d and positive integer K for training data to compute substitution predictions. Alternatively,

this model classifies the new observations by using a predict method (Taunk et al., 2019). For real value vector space, it uses Minkowski distance, but for the continuous value, it uses the Euclidean distance to calculate its nearest neighbors. The Euclidean distance between points x_i and y_i is represented as a vector's length.

The few conditions that must satisfy the following criteria
(i) C1:$d(x, y) \geq 0$ (non-negativity).
(ii) C2: $d(x, y) = 0 \quad if\ and\ only\ if\ x = y$ (identity).
(iii) C3: $d(x, y) = d(y, x)$ (symmetry).
(iv) C4: $d(x, y) + d(y, z) \geq d(z, x)$ (triangle inequality).

$$d_E = (\sum_{i=1}^{n}(|x_i - y_i|^q))^{\frac{1}{p}} \tag{10}$$

p = 1, when p is set to 1, we get Manhattan distance.
p = 2, when p is set to 2, we get Euclidean distance.

The above formula for Minkowski distance.

4.5.3 Random Forest

Random forest is a popular supervised learning technique widely used in decision tree predictor, classification, and regression problems. This algorithm is mainly designed to develop the model quickly. Despite growing interest in the context of practical application, the RF provides an excellent indicator, more precise, fast, and straightforward algorithm compared to other ensembles (Biau, 2012). The general component of random forest is defined as $\{h(x, \theta_k),\ \ k = 1 \ldots \ldots, L\}$ where θ_k is an arbitrary random vector parameter and x is the input vector parameter and k suggest a number of the decision tree. The classification contains many trees, and each tree uses a random vector parameter, randomly selecting a dataset as the training set. If the sample feature is greater than the segmented features, then the predictive class is estimated based on the discrimination function in an indecision tree node tree.

$$h(x)=argmax_y \sum_{i=1}^{k} I(h_i(x, \theta_k) = \gamma) \tag{11}$$

Here $I(\cdot)$ is the indicator function, $h(\cdot)$ is the decision tree,γ is the output parameter, and the value of $argmax_y$ depends on the value of $h(x)$. The instance in which the sample is not being selected from k group data out of the bag. Arbitrary known as out of the bag (OOB). This OOB instance is utilized to identify model efficiency and classify the unknown data.

4.5.4 Decision Tree

A decision tree is the supervised machine learning algorithm most widely used for real-time decision-making challenges based on the specific data parameter. This data parameter has continuously split into different branches and leaves. These branches and leaves show various features concerning a particular situation that needs a small pre-processing. But it can support categorical features without pre-processing (Kim and Hong, 2017). The regression tree gives a numeric response. Assume the trained data set D_s and attribute A, then the equation D_s=$\{x, y\}$ they will split into two or more disjoint subset branches $D(L, i)$. Where L denotes the layer number and the i represent the individual subset. If all the label in the subset belongs to the same class, it is called a pure, and the node declared as the leaves node leads to tree termination; otherwise, the data parameters must continue the splitting criteria. In that case, the leaves node may be impure, which can be calculated by the formula.

$$I(n) = -\sum_{i=1}^{q}(P_i . \log_2 P_i) \tag{12}$$

where, P_i reflects population node θ_i as a class $(ie, i \in \{1,2,3 \ldots \ldots q\})$ the equation. (12) shows the level of predictivity for each node.

4.5.5 Multiple Linear Regression

Multiple linear regression is one of the supervised learning techniques used to predict a single variable by using two or more predictor variables to predict the outcome. The model used the least square concept to estimate the regression coefficient in equation 13.

The hypothesis model for multiple linear regression with the n observation is given as

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \ldots\ldots\ldots + \theta_n x_n \quad (13)$$

It is important to note that the feature of the hypothesis will be on the same scale. We cannot make a model with features that vary between 1 to 10000 in the range of 0.1 to 0.01. Hence, we need feature scaling before making the hypothesis. This can be achieved through the mean scaling concept and calculated through the mean normalization formula.

$$x = \frac{x_1 - \mu_1}{s_1} \quad (14)$$

Here x is the mean normalization, μ_1 is the average value and x_1 is the training set and s_1 is the range of values, then the cost function will be modified to

$$c(\theta_0, \theta_1, \theta_2 \ldots . \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} \left(\left(h_o(x^i) - y^i \right)^2 \right) \quad (15)$$

Therefore, gradient descent updates the values of parameters after each iteration, and this will prefer a large dataset over the normal equation.

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} c(\theta_o, \theta_1 \ldots \ldots . \theta_n) \; where \; j = 0,1,2 \ldots . n \quad (16)$$

$$x = [x_o, x_1, x_2, x_3 \ldots \ldots \ldots . . x_n] \text{ and } \theta = [\theta_0, \theta_1, \theta_2 \ldots . \theta_n].$$

Here θ and x vary in the range 0 to *n*, so the individual vector of our hypothesis will be formed.

$$h(x) = \theta^T x \quad (17)$$

To find the minimum value of θ that will reduce the cost function, then the optimal value will become

$$\theta(minima) = (X^T X)^{-1} X^T Y \quad (18)$$

This gradient descent model performs the most accurate result in the larger dataset.

5. Performance Validation

This section describes the performance of the Multiple Linear Regression (MLR), Decision Tree (DT), K Nearest Neighbor (KNN), Random Forest (RF), and Support Vector Machine (SVM) algorithm applied to the integrated risk assessment model data set. We collect the risk data from some articles and journals as we communicate with the organization to gather the associated risk data. For this research, dataset were generated through simulation of a probabilistic risk assessment adopted model to assess the risk and various risk were considered from organization's risk, journals and risk management manuals for the past ten years from multiple suppliers, vendors, manufacturers, and retailer shops, for example, inventory risk data, production risk data, distribution risk data, visibility risk at each point of sale (POS). We have generated

119734 based on the three risks (source, make, deliver) factors. Note that the data has been taken separately for each process area like source, make and deliver for the supply chain. Among the 119734 datasets, the size distribution for the risk category has been categorized as per the risk score for low, medium, high and severe are 60047, 29817, 7537 and 22333, subsequently, as given in the Figure 5. In addition, 50.15% of instances fall under the low category, 24.90% fall in the medium category, 6.29% fall in the high category, and 18.65% fall in the severe category, as shown in Figure 6.

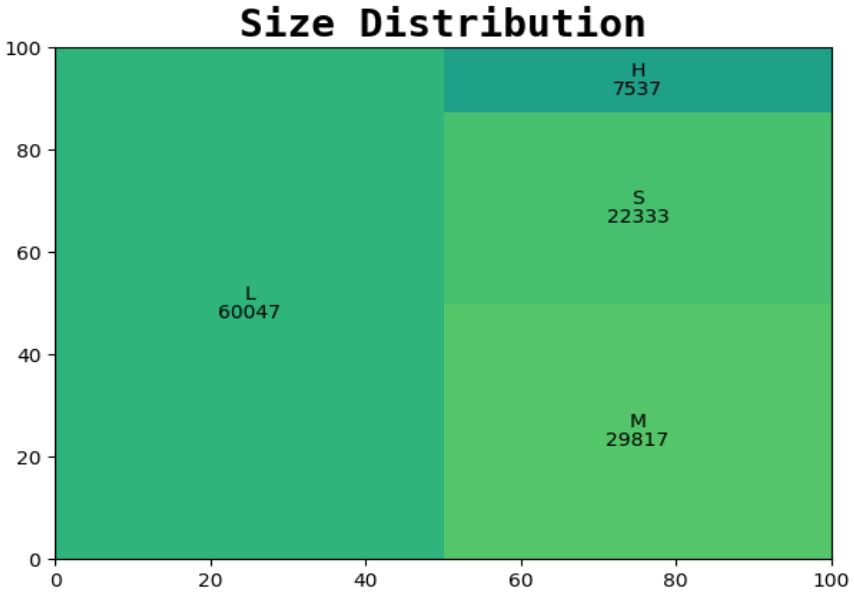

Figure 5. Size distribution of risk category.

For evaluation, we divide the dataset into two parts 75% of the data has been taken for training, and 25% of data has been taken for testing the model. The attribute involved in the dataset is shown in Table 4. In addition to that, the heat map regarding the relevant features that exist in the dataset is shown in Figure 7. In order to reduce the number of features and preserve the discriminatory information before the classification, linear discriminant analysis is used as shown in Figure 8, and the histogram plot for the source, make, and deliver is shown in Figure 9. More specifically, Figure 9(a) shows the histogram plot for source, Figure 9(b) shows the histogram plot for make and Figure 9(c) shows the histogram plot for deliver. With data visualization through the normalization, we completed the data transformation. The detailed comparative performance measure with all the models is shown in Table 6.

Table 4. Attribute description.

S. No.	Feature	Data type	Feature description	Prediction
1	Plan	Numerical	Plan Risk	Low (L)
2	Source	Numerical	Supply Risk	Medium (M)
3	Make	Numerical	Make Risk	High (H)
4	Deliver	Numerical	Demand Risk	Severe (S)
5	Source	Numerical	Information Risk	Medium (M)
6	Make	Numerical	Sales Risk	High (H)
7	Source	Numerical	Quality Risk	Severe (S)
8	Deliver	Numerical	Transport Risk	Severe (S)

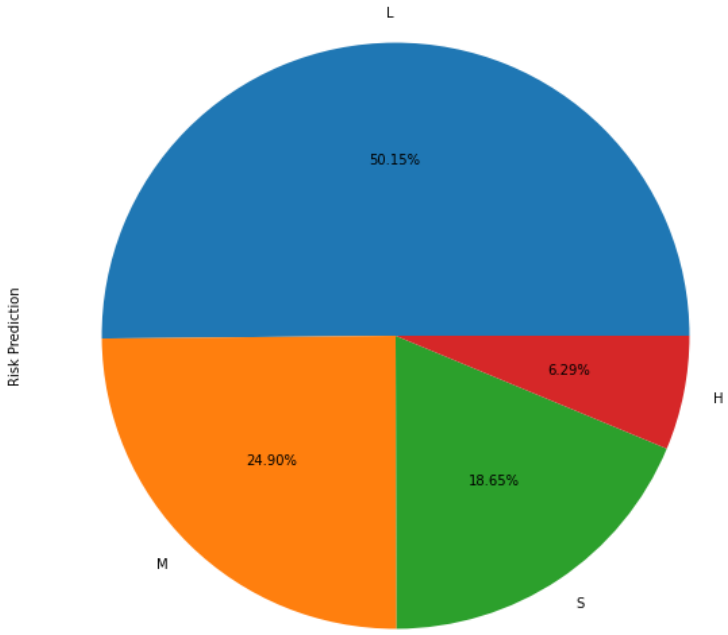

Figure 6. Pie chart demonstrating risk prediction.

6. Results and Discussion

6.1 Classification Analysis

The goal of this work is to predict and classify the source (supply risk), make (process risk), and deliver (demand risk). We found the distribution of this risk dataset had a severe class imbalance i.e.; low class risk is very high number of sample and the high class of risk was very low number of samples. In order to resolve the class imbalance data, we adopted an oversampling of high-class risk i.e.; duplicating the high-class risk through this we increase our model predictability. The classification result is shown in Table 5.

Table 5. Classification result (Mean, Standard deviation).

Measure Summary	Supply chain processes		
	Source	Make	Deliver
Count	119734	119477	119404
Mean	61.75681	34.02731	165.8058
Standard deviation	9.944863	8.149447	6.737651

6.2 Correlation Analysis

Heat Map is a covariance matrix with two dimensional, warm-to-cool visualization of complex statistical data that strengthen the relationship between the identified source, make and deliver related risk. The heat map shown in Figure 7. provides a correlation score between -1 to 1. The X-axis denotes the positions of predicted risk and Y-axis denotes the real risk. For instance, we can see that in position (source, deliver) the value is 0.39, which is close to zero hence, there is no linear trend between the source and deliver and the position (deliver, deliver) the value is 1 the correlation is more positive. However, the position (make, deliver) the value is -0.003, which is closer to -1 which shows the correlation coefficient of make increases with the decrease of deliver and vice versa.

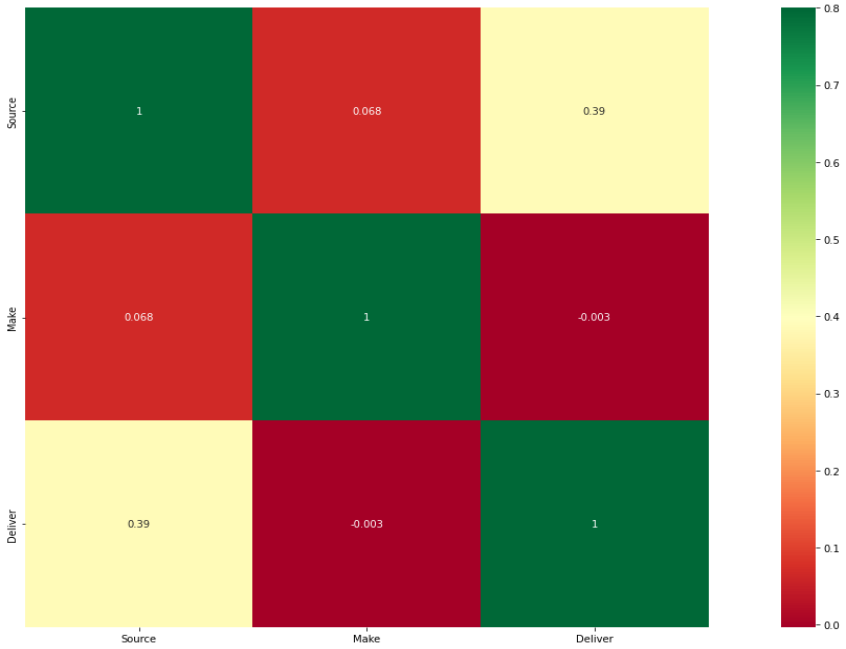

Figure 7. Heat map for the risk dataset.

6.3 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is the most common feature extraction technique representing data in low-dimension space. When the input sample is extensive, it transforms into a reduced set of features. It Computes the mean vectors for various classes of the dataset. The simulation result of the proposed model has illustrated in Figure 8. However, the effectiveness of the algorithm was determined using the performance evaluation.

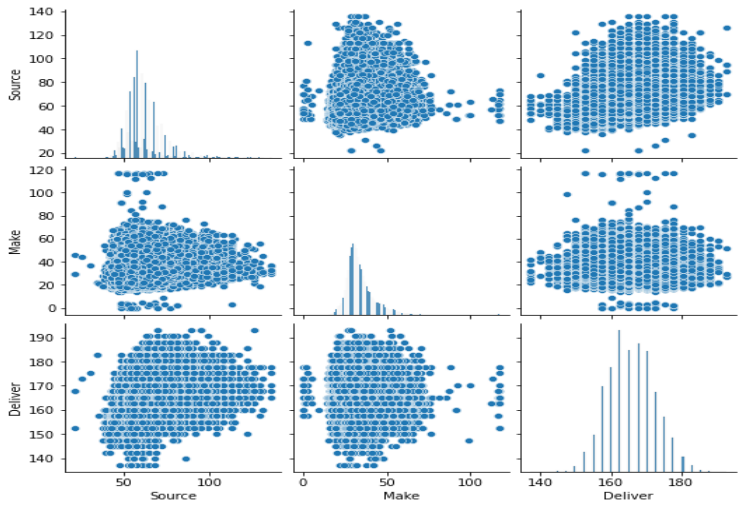

Figure 8. Illustration of linear discriminant analysis for the model.

6.4 Performance Evaluation and Comparison of ML Algorithms

To measure the performance of the organizational supply chain risk (supply risk, process risk, demand risk), and the risk assessment was based on the severity and impact model. The performance is evaluated using several machine-learning algorithms (support vector machine nearest neighbor, random forest, decision tree and multiple linear regression) were applied to measure the performance evaluation in terms of accuracy, error rate, F-score, precision, recall as shown in Figure 10. However, All the algorithms performed well.

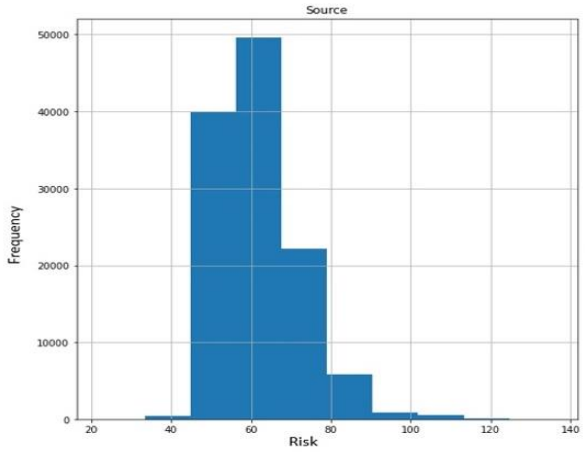

Figure 9 (a). Histogram Plot for Source

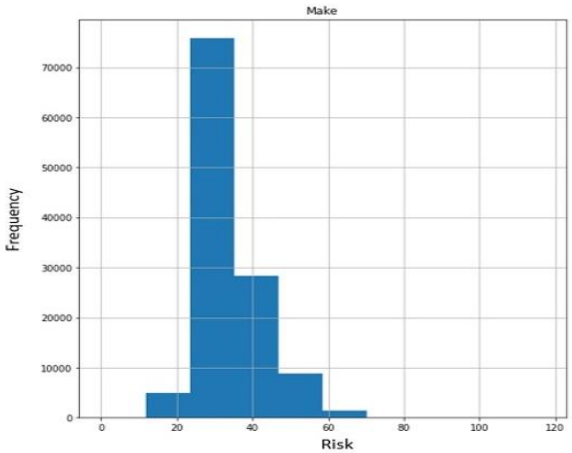

Figure 9 (b). Histogram Plot for Make

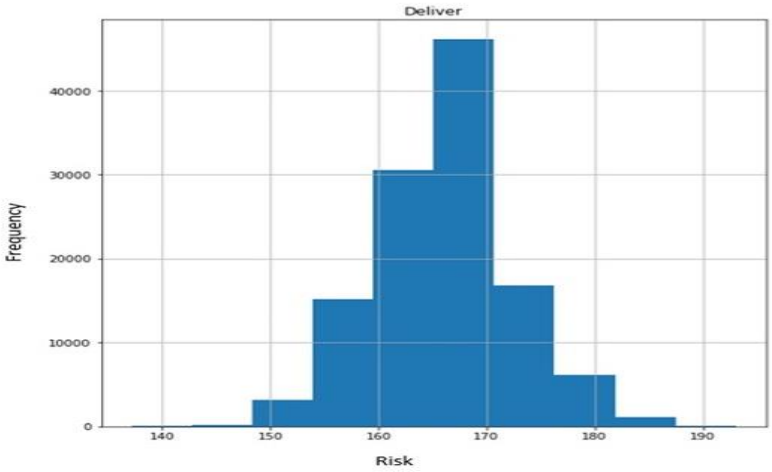

Figure 9 (c). Histogram Plot for Deliver

Figure 9. (a), (b), (c)- Histogram plot for the model (Source, make and deliver).

To facilitate discussion, Table 6 contains performance evaluation on various algorithms. Random forest (RF) outperforms among all the algorithms. The RF shows the highest accuracy of 99%, with the lowest error rate of 0.01. It accounts for some drawbacks, such as the need for more computational power, time, and resources that build many decision trees to determine the classification. However, the model faces interpretability with feature importance and proximity plot. The SVM performs 97% accuracy with 0.024 error rate because of the dataset is not easily separable as SVM assumes a hyperplane separating the data

points. Although SVM shows lower performance than RF, it can also produce a better result by using feature engineering techniques, fitting pre-processing, and SVM with the nonlinear kernel (e.g., RBF). We found 98% accuracy in K-NN, while the model is non-parametric and has no assumption for the underlying distribution of data. It was also observed that the decision tree algorithm obtains an accuracy of 86% with a 0.136 error rate and multiple linear regression unable to handle the collinearity issue between the independent variables. So, it acquired 66% accuracy and 0.338 error rate.

Table 6. Performance evaluation on various algorithm.

Algorithms	Performance Evaluation				
	F-score	Precision	Error rate	Accuracy	Recall
Support Vector Machine	0.98	0.97	0.024	0.97	0.97
K-Nearest Neighbor	0.96	0.96	0.04	0.96	0.96
Random forest	0.99	0.99	0.01	0.99	0.99
Decision Tree	0.86	0.94	0.136	0.86	0.8
Multiple Linear Regression	0.56	0.49	0.338	0.66	0.66

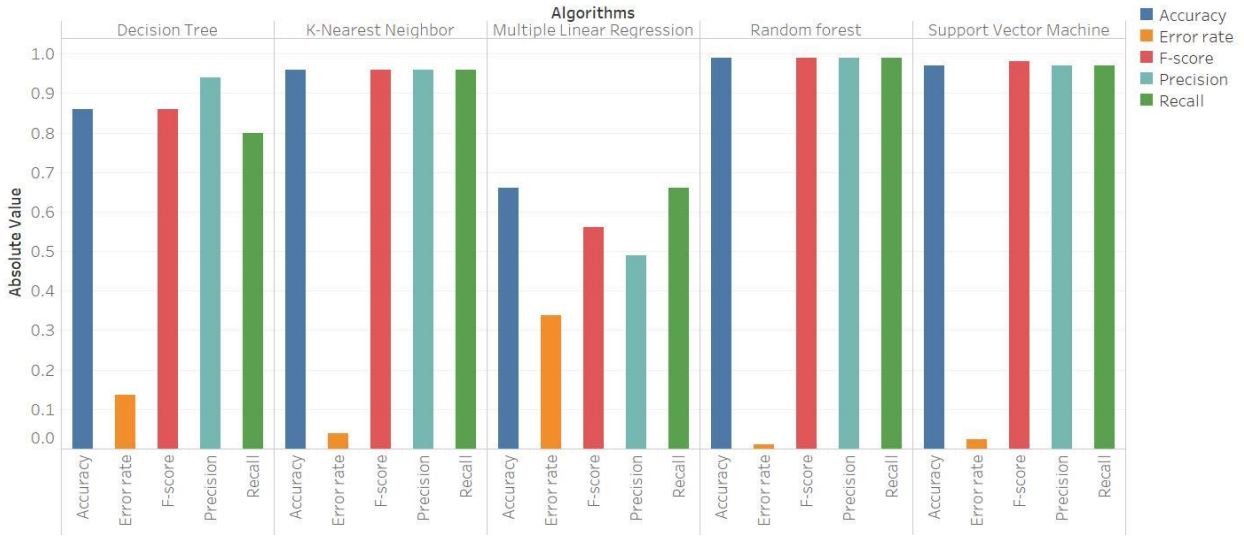

Figure 10. Comparative analysis under different techniques of the ML algorithms.

7. Conclusion

Through this work, we have proposed an Integrated SCOR model-based risk prediction through the different ML algorithms so that the organization proactively identified the risk events and their prediction from the supplier to the customer during the supply chain. This can help the organization mitigate this risk and optimize the performance of the supply chain. The proposed model explicitly considers five supervised machine learning algorithms for supply chain risk management. The results of this study shows that the model can achieve good performance with machine learning techniques. It was also observed that risk varied from organization to organization. Hence, we faced difficulty in working with an imbalanced dataset. The risk prediction over the performance evaluation requires a trade off in terms of recall (67% decrease in prediction performance) and in terms of precision (49% decrease). However, the proposed work is limited to the integrated SCOR-based risk prediction but can also be applied in the domain of credit risk and stock

market risk to identify the market price and volatility. Our future work will be based on a clustering algorithm with a different classification algorithm in the reverse logistics of the SCOR supply chain model.

Conflict of Interest

The authors confirm that there is no conflict of interest to declare for this publication.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors would like to thank the editor-in-chief, guest editors and anonymous reviewers for their comments that help improve the quality of this work.

References

Alfian, G., Syafrudin, M., Farooq, U., Ma'arif, M.R., Syaekhoni, M.A., Fitriyani, N.L., Lee, J., & Rhee, J. (2020a). Improving efficiency of RFID-based traceability system for perishable food by utilizing IoT sensors and machine learning model. *Food Control*, *110*, 107016. https://doi.org/10.1016/j.foodcont.2019.

Alfian, G., Syafrudin, M., Fitriyani, N.L., Rhee, J., Ma'arif, M.R., & Riadi, I. (2020b). Traceability system using IoT and forecasting model for food supply chain. In *2020 International Conference on Decision aid Sciences and Application (DASA)* (pp. 903-907). IEEE. Sakheer, Bahrain.

Amani, M.A., & Sarkodie, S.A. (2022). Mitigating spread of contamination in meat supply chain management using deep learning. *Scientific Reports*, *12*(1), 5037. https://doi.org/10.1038/s41598-022-08993-5.

Baryannis, G., Dani, S., & Antoniou, G. (2019). Predicting supply chain risks using machine learning: The trade-off between performance and interpretability. *Future Generation Computer Systems*, *101*, 993-1004. https://doi.org/10.1016/j.future.2019.07.059.

Benjaoran, V., & Dawood, N. (2005). An application of artificial intelligence planner for bespoke precast concrete production planning: A case study. In *Proceedings of the 13th Annual Conference of the International Group for Lean Construction* (pp. 19-21). Sydney, Australia.

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, *13*(1), 1063-1095.

Blackburn, R., Lurz, K., Priese, B., Göb, R., & Darkow, I.L. (2015). A predictive analytics approach for demand forecasting in the process industry. *International Transactions in Operational Research*, *22*(3), 407-428. https://doi.org/10.1111/itor.12122.

Bouzembrak, Y., & Marvin, H.J. (2019). Impact of drivers of change, including climatic factors, on the occurrence of chemical food safety hazards in fruits and vegetables: A Bayesian network approach. *Food Control*, *97*, 67-76. https://doi.org/10.1016/j.foodcont.2018.10.021.

Brintrup, A., Pak, J., Ratiney, D., Pearce, T., Wichmann, P., Woodall, P., & McFarlane, D. (2020). Supply chain data analytics for predicting supplier disruptions: A case study in complex asset manufacturing. *International Journal of Production Research*, *58*(11), 3330-3341. https://doi.org/10.1080/00207543.2019.1685705.

Cavalcante, I.M., Frazzon, E.M., Forcellini, F.A., & Ivanov, D. (2019). A supervised machine learning approach to data-driven simulation of resilient supplier selection in digital manufacturing. *International Journal of Information Management*, *49*, 86-97. https://doi.org/10.1016/j.ijinfomgt.2019.03.004.

Chao, C.F., & Horng, M.H. (2015). The construction of support vector machine classifier using the firefly algorithm. *Computational Intelligence and Neuroscience*, *2015*, 1-8. https://doi.org/10.1155/2015/212719.

Constante-Nicolalde, F.V., Guerra-Terán, P., Pérez-Medina, J.L. (2020). Fraud prediction in smart supply chains using machine learning techniques. In: Botto-Tobar, M., Vizuete, M.Z., Torres-Carrión, P., León, S.M., Pizarro Vásquez, G., Durakovic, B. (eds) *Applied Technologies. ICAT 2019*. Communications in Computer and Information Science (vol 1194). Springer, Cham. https://doi.org/10.1007/978-3-030-42520-3_12.

Debruyne, M. (2009). An outlier map for support vector machine classification. *The Annals of Applied Statistics*, *3*(4),1566-1580. https://doi.org/10.1214/09-AOAS256.

Fu, W., & Chien, C.F. (2019). UNISON data-driven intermittent demand forecast framework to empower supply chain resilience and an empirical study in electronics distribution. *Computers & Industrial Engineering*, *135*, 940-949. https://doi.org/10.1016/j.cie.2019.07.002.

García, F.T., Villalba, L.J.G., & Portela, J. (2012). Intelligent system for time series classification using support vector machines applied to supply-chain. *Expert Systems with Applications*, *39*(12), 10590-10599. https://doi.org/10.1016/j.eswa.2012.02.137.

Hassan, A.P. (2019). Enhancing supply chain risk management by applying machine learning to identify risks. In: Abramowicz, W., Corchuelo, R. (eds) *Business Information Systems*. Lecture Notes in Business Information Processing (vol 354). Springer, Cham. https://doi.org/10.1007/978-3-030-20482-2_16.

Huang, S.H., Sheoran, S.K., & Keskar, H. (2005). Computer-assisted supply chain configuration based on supply chain operations reference (SCOR) model. *Computers & Industrial Engineering*, *48*(2), 377-394. https://doi.org/10.1016/j.cie.2005.01.001.

Huo, H., & Zhang, J. (2011). Research on retail enterprise supply chain risk identification based on SCOR. In *International Conference on Management Science and Industrial Engineering (MSIE)2011* (pp. 1302-1305). IEEE. Harbin.

Kim, K., & Hong, J.S. (2017). A hybrid decision tree algorithm for mixed numeric and categorical data in regression analysis. *Pattern Recognition Letters*, *98*, 39-45. https://doi.org/10.1016/j.patrec.2017.08.011.

Kumar, S., & Barua, M.K. (2022). Modeling and investigating the interaction among risk factors of the sustainable petroleum supply chain. *Resources Policy*, *79*, 102922. https://doi.org/10.1016/j.resourpol.2022.102922.

Lau, H.C., Ning, A., Pun, K.F., Chin, K.S., & Ip, W.H. (2005). A knowledge-based system to support procurement decision. *Journal of Knowledge Management*, *9*(1), 87-100. https://doi.org/10.1108/13673270510582983.

Layouni, M., Tahar, S., & Hamdi, M.S. (2014). A survey on the application of neural networks in the safety assessment of oil and gas pipelines. In *2014 IEEE Symposium on Computational Intelligence for Engineering Solutions* (*CIES*) (pp. 95-102). IEEE. Orlando, FL, USA.

Mani, V., Delgado, C., Hazen, B.T., & Patel, P. (2017). Mitigating supply chain risk via sustainability using big data analytics: Evidence from the manufacturing supply chain. *Sustainability*, *9*(4), 608. https://doi.org/10.3390/su9040608.

Orenstein, I., & Raviv, T. (2022). Parcel delivery using the hyperconnected service network. *Transportation Research Part E: Logistics and Transportation Review*, *161*, 102716. https://doi.org/10.1016/j.tre.2022.102716.

Pereira, M.M., & Frazzon, E.M. (2019). Towards a predictive approach for omni-channel retailing supply chains. *IFAC-Papers on Line*, *52*(13), 844-850. https://doi.org/10.1016/j.ifacol.2019.11.235.

Ríos, J.R., Duque, D.F.M., & Gómez, J.C.O. (2019). Operational supply chain risk identification and prioritization using the SCOR model. *Ingenieria y Universidad*, *23*(1), 1-12.

Rodriguez-Aguilar, R., & Marmolejo-Saucedo, J.A. (2019). Structural dynamics and disruption events in supply chains using fat tail distributions. *IFAC-Papers on Line*, *52*(13), 2686-2691. https://doi.org/10.1016/j.ifacol.2019.11.613.

Tama, I.P., Yuniarti, R., Eunike, A., Hamdala, I., & Azlia, W. (2019). Risk identification in cassava chip supply chain using SCOR (Supply Chain Operation Reference). In *IOP Conference Series: Materials Science and Engineering* (Vol. 494, No. 1, p. 012050). IOP Publishing. Malang, Indonesia. https://doi.org/10.1088/1757-899X/494/1/012050.

Taunk, K., De, S., Verma, S., & Swetapadma, A. (2019). A brief review of nearest neighbor algorithm for learning and classification. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)* (pp. 1255-1260). IEEE. Madurai, India.

Teuteberg, F. (2008). Supply chain risk management: A neural network approach. *Strategies and Tactics in Supply Chain Event Management* (pp. 99-118). Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-73766-7_7.

Wichmann, P., Brintrup, A., Baker, S., Woodall, P., & McFarlane, D. (2020). Extracting supply chain maps from news articles using deep neural networks. *International Journal of Production Research*, *58*(17), 5320-5336. https://doi.org/10.1080/00207543.2020.1720925.

Yong, B., Shen, J., Liu, X., Li, F., Chen, H., & Zhou, Q. (2020). An intelligent blockchain-based system for safe vaccine supply and supervision. *International Journal of Information Management*, *52*, 102024. https://doi.org/10.1016/j.ijinfomgt.2019.10.009.

Zhu, Q., Liu, L., & He, Y. (2019). Application of process analysis based on value objective improvement in risk identification of supply chain. In *2019 Chinese Automation Congress (CAC)* (pp. 4213-4218). IEEE. Hangzhou, China.