Supplier Prediction in Fashion Industry Using Data Mining Technology

Nitin Harale, Sebastien Thomassey, Xianyi Zeng
GEMTEX, ENSAIT
Ecole Centrale de Lille
Roubaix, France

***Abstract*—Selection of supplier is the integral and most challenging part of supply chain management in fashion industry as its performance is largely dependent on effective supplier selection process. Multitude of information from the customers' and market's perspective has to be considered before making the decisions in regard to supplier selection. Data mining methods have found many applications in fashion supply chain management. However, decision making related to supplier selection lacks the application of data mining methods. It warrants an automated decision support system wherein machine learning models can be trained on historical customer order data to predict best suppliers or recommend new ones depending on the degree of matching between suppliers' capabilities and the product order features. In this paper, our work revolves around the research question of how we can predict suppliers based on historical customer order data by using data mining methods. Our aim was to predict supplier by applying classification models on the historical customer order data. We applied four machine learning classification models and the research findings suggest that these models can be employed for the decision making concerning supplier prediction. This study can contribute to the development of automated decision support system which is reliable and efficient for the supplier prediction.**

***Keywords*—supplier prediction, data mining methods, machine learning, fashion industry**

I. INTRODUCTION

Fashion industry is characterized by rapidly varying product designs and styles, market trend towards small series production, short product life cycle, and growing preferences for personalized products. In recent decades, there has been a shift in traditional fashion retailing as fashion market is increasingly adopting e-commerce retailing. Given the growing data repositories in the companies' databases through e-shopping platforms, it becomes challenging to manage the load of such complex datasets, and deriving valuable insights for decision making is complex and inefficient [1]. Rapid transformation of fashion industry in the wake of advanced data management tools, adoption of e-commerce business models and Artificial Intelligence has paved the way for automated decision making in the context of various decision problems facing fashion industry today [2]. Moreover, given the growing concerns of sustainable fashion production and regulatory pressure on the part of policy organizations, data mining methods are gaining popularity in handling big data driven decision problems in fashion supply chain management [3]. Supplier selection is one of the critical decision making problems that fashion retailers always encounter as their business efficiency and profitability depends on the efficiency of their decision making related to sourcing of raw materials for the production. In other words, collaboration with reliable and potential suppliers is quite integral to the effective management of fashion supply chain. Classical mathematical methods to select the best suppliers are not appropriate to deal with complex information and criteria that form the premise for the selection of suppliers [4]. Data mining techniques have been used for solving apparel logistic management problems such as locating warehouses and manufacturing plants; inventory management; transportation management; sales management, and so on. However, data mining methods have not been used so far to solve supplier prediction problems in a big data framework of the fashion logistic management. Big data has significantly transformed today's business models. Fashion industry is following the similar trend, and decision making in a real time by studying rapidly generated data is critical to their business growth as well as for addressing sustainability related issues such as over-production, overutilization of energy and other resources, and waste management.

The rest of the paper is organized as follows: The problem statement and its context are introduced in Section II. Section III illustrates research methodology highlighting key steps involved in the model implementation. The results are illustrated in Section IV, and finally, the conclusions drawn from this study including limitations and future scope are discussed in Section V.

II. PROBLEM STATEMENT

For fashion retailers, it is indispensable to choose the best suppliers who support their business processes by providing products with high quality to their customers on the right time. The operational and business efficiency of fashion retailers largely depend upon their suppliers [5]. As fashion market is adopting e-commerce platforms and technology, fashion retailers often realize the need for automated mechanism for selecting best suppliers who will be able to fulfil their customer demands and help them to retain their loyalty. Given the recent advancements in the database management technologies and customer's online shopping preferences,

European Union's H2020 FBD_BModel research project funded this study.

978-1-7281-1566-5/19/$31.00 ©2019 European Union

fashion companies retrieve variety of information from their databases which can include customer choices, market trend, product features and demand etc. However, to utilize this information for the decision making is often complex and laden with many difficulties. Supplier selection in fashion industry is highly critical and complex process as it entails multiple qualitative and quantitative criteria, and the participation of many managers making decisions in the supply chain management of companies. Typically, data mining technologies have been used for demand forecasting, market analysis, social media analysis of brand's popularity, etc. [6]. Machine learning methods are increasingly being applied to derive customer information, their opinions about fashion products and purchasing experiences from social media platforms such as Twitter, Instangram, and Facebook [7]. Much emphasis is also being given on extracting correct information from the databases as it significantly influences the accuracy of data mining methods used for various decision making problems. However, application of data mining methods for the prediction of best suitable supplier corresponding to customers' specific demands has not been studied in academic research before. Therefore, authors in this paper aim to explore the applicability of data mining methods to predict best suitable suppliers. Data mining methods have been applied for predicting customer demands for the products, social media analytics, market research etc. There is a dearth of literature on the application of data mining methods on the supplier prediction in the fashion industry. Data mining methods are usually applied on historical data to find the pattern in it and derive insights out of it. Building on this premise, our goal is to propose a data mining based methodological framework for the supplier prediction in fashion industry.

To predict supplier for customers' specific order in future, we aim to train data mining models on historical customer order data that include information of products' customized features. Data mining models extract as much information as possible from the dataset and derive the pattern in it [8].

The overarching research question formulated in the context presented in Section I is as follows.

How to predict fashion supplier by using data mining methods on the customer order data?

The motivation behind exploring this research question emerges from the need for addressing growing complexities involved in the decision making of fashion logistic and supply chain management.

III. RESEARCH METHODOLOGY

In this section, we detail experimental set up including research framework as a first broad step; solution methodology in the context of defined research question; background of implemented data mining methods and their working mechanism; and finally, data preparation for the experimental work.

A. Research Framework

In order to predict supplier from the customer order data using data mining methods, historical customer order data provided by the European fashion company is collected and used for the analysis. The overall experimental research framework entails broad steps, starting from the formulation of the research question to historical customer order data collection; and eventually, the data preprocessing steps. The depicition of overall research framework is shown below in Fig. 1.

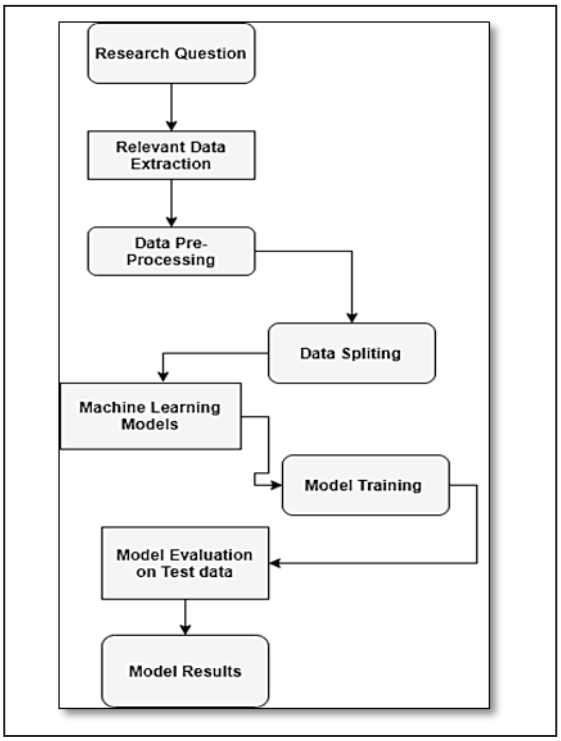

Fig. 1. Research framework

B. Data mining models and evaluation metrics

In this paper, we used historical customer order data with labels, meaning that customer order attributes were tagged with the label of suppliers who fulfilled these orders; therefore, supervised machine learning approach has been adopted in this study. Supervised learning approach mainly includes regression and classification tasks [9]. Since the suppliers are unique labels in the dataset, we used classification models to predict the suitable supplier by identifying and including appropriate predictors in the dataset. Classification models perform the task of drawing conclusions by observing data points, and subsequently predict the label value for each label in the dataset.

Classification models take inputs from various features in the dataset and gives output in terms of a label or a class for the predicted variable [10]. Classification is of two types depending on the number of classes we want to predict. If the classes to be predicted are two, it is known as binary

classification, while in case of more than two classes it becomes multi-class classification.

To derive the pattern out of qualitative features of data, rule based classifiers such as kNN, RF, NN, naïve Bayes are generally used and give better performance [11]. Therefore, we have used these models to predict suppliers.

1) K Nearest Neighbors (kNN): kNN model is a non-parametric classification method which predicts the classes by computing the Euclidean distance between the data points in the features and the new data points. kNN classifies data points into labelled classes based on the similarity between them. kNN learns the patterns in the data iteratively and identifies majority of similar 'k' nearest neighbors. The number of 'k' neighbors can be selected by users depending on the learning goal of the research problem.

2) Random Forest (RF): RF model can be considered to be the extension of Decision Trees algorithm. It builds decision tress by taking into account data points in the training dataset and gives output in terms of class label. The splitting nodes are decided based on statistical probability of assigning specific class to the data points, and furthermore it averages them in order to predict the class.

3) Neural Networks (NN): NN classifier mimics the human brain neurons which are characterized by multiple interconnected nodes of neurons, which carry information to the other nodes and ultimately produce output as a single node. This classifier is a popular technique for multi-class classification problem as it gauges interdependencies between the classes to be predicted.

4) Naïve Bayes (NB): Naïve Bayes is a probabilistic classifier, which works based on the theory of Bayes conditional probability theorem, and takes into account prior information available to calculate the probability of future event. It is used when the size of the dataset is small and when features in the dataset are not correlated.

5) Evaluation Metrics: The performance and accuracy of the classification models is evaluated based on the metrics: Confusion matrix; Precision; Recall; F1; Area under the curve (AUC); and Classification Accuracy (CA) [12].

a) Confusion Matrix: Confusion matrix presents an overall model performance in a tabular form, in which classes predicted correctly and incorrectly are summarized as follows:

True Positives (TP): correctly predicted class 1 as 1

True Negatives (TN): correctly predicted class 2 as 2

False Positives (FP): incorrectly predicted class 2 as 1

False Negatives (FN): incorrectly predicted class 1 as 2

The confusion matrix, as depicted in Fig. 2, contains rows and columns; rows represent actual classes while columns represent predicted classes.

	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

Fig. 2. Confusion matrix

b) Precision: Precision the ratio of true positive classes to the number of actual positive classes and is formulated as follow:

$$Precision = TP/(TP + FP) \quad (1)$$

c) Recall: Recall is the ratio of true positive observations to all the observations in actual class.

$$Recall = TP/(TP + FN) \quad (2)$$

d) F1 score: F1 score is defined as the harmonic mean of recall and precision. The accuracy indicates the classifier's overall performance as it takes into account both false positives and false negatives, and is calculated as follows:

$$F1\ Score = 2*(Recall * Precision) / (Recall + Precision) \quad (3)$$

e) Area under the Curve (AUC): AUC gives us an aggregated performance measure for all possible classification thresholds. AUC is the probability of a model classifying positive observation higher than the negative one. AUC value ranges between *0* and *1*.

f) Classification Accuracy (CA): Classification Accuracy of the model is the fraction of correctly predicted observations. It is calculated as follows:

$$Classification\ Accuracy = (TP+TN)/(TP + TN + FP + FN) \quad (4)$$

C. Softwares tools and libraries

We have implemented data mining classification using Python and its supported machine learning libraries. 'Windows 10' operating system with processor speed 2.90 GHz and memory speed 32 GB has been used for this experimental study.

D. Data Source

To explore the research question stated in the Section II, we collected data from the database of European fashion brand, which is specialized in various fashion products including luggage and hand bags. The European fashion company provided us with historical one-month customer order data. The dataset was relatively small and originally contained 100 elements and 35 variables. The collected data size is small given that our supplier prediction is performed considering both data and time constraints.

E. Data Description

Given that the data we used was labelled, we have adopted supervised learning methods. The attributes in the original

dataset and their description is given in the table 1. We followed ethical data guidelines to maintain the confidentiality and sensitivity of seven unique suppliers in the dataset. We anonymized supplier Id's by assigning them labels such as 'A-G' as shown in Table I.

TABLE I. CHARACTERISTICS OF DATA

Attributes	Description
Buy	Quantity of products bought
Panel	Seasonal availability of product
Proto/Dev	P= Product Prototype D= Customized product
Gender	U= Unisex; X= Male; D= Female
Business Line	Business Channels
Category	Product Category, e.g. Luggage, Bags, Accessories
Sub-Category	Sub-category of the products
Mtl	Product material, e.g. polyster, Nylon, Rubber, etc.
Supplier	Suppliers (Labelled as A, B, C, D, E, F, G)

F. Data Preprocessing

We preprocessed the dataset by removing the noise in terms of missing values and outliers. Further, we performed dataset splitting into trained set, validation set and test set by using deterministic random sampling as it ensures that model is correctly fitted to the data without biasedness. We employed '80:20' fraction to split data into train and test datasets. We then applied k-fold cross validation on trained data and the default value for "k" is set to "10". It is important to highlight that the features in the dataset are mostly categorical or nominal owing to which it is subject to encoding in order for machine learning models to be applied. We employed label encoding and subsequently one-hot encoding to transform the categorical features in the dataset into binary values.

Each classification model is then applied on trained and prediction is done on test dataset. Prediction accuracy and the model performances were evaluated according to the metrics corresponding to each model.

IV. RESULTS

Classification models to predict supplier are applied on trained dataset (80% of the original dataset); k fold crossed validation data (k = 10) and finally on test dataset (20% of the original dataset). It is observed that, on the trained dataset model performance of kNN, RF, NN gives 100% accuracy results, while Naïve Bayes model gives 86% accuracy. The accuracy values of classification models applied on trained dataset can be seen in Table II and their performances are compared in Fig. 3.

TABLE II. MODEL ACCURACY ON TRAINED DATA

Classifier	AUC	CA_Trained	F1	Precision	Recall
kNN	1	1	1	1	1
RF	1	1	1	1	1
NN	1	1	1	1	1
NB	0.99	0.867	0.88	0.922	0.867

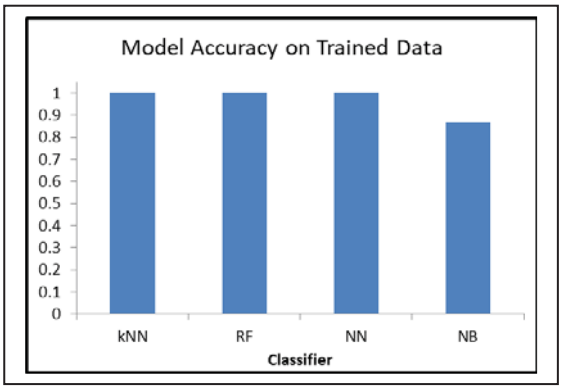

Fig.3. Model performance on trained data

It means that the models are overfitting to the trained data and can be ridden with biasedness. To address this problem, we applied k-fold cross validation on trained data with k=10. K-fold cross validation subsets data into 10 subsets, where 9 subsets are used for model training and remaining one subset is used for model testing. The accuracy of models on cross validated data is averaged over all the subsets (folds) and is shown in Table III. If we compare the models' accuracy on cross validation data, as depicted in Fig. 4, it can be seen that RF and NN models' performances are consistent with the results for trained dataset. However, their classification accuracy dropped when applied on cross validation data.

It can be observed that model accuracy is reduced after applying cross validation on trained data and it implies that models are not overfitting to the data.

TABLE III. MODEL ACCURACY AFTER CROSS VALIDATION ON TRAINED DATA

Classifier	AUC	CA_CV	F1	Precision	Recall
kNN	0.935	0.783	0.755	0.73	0.783
RF	0.951	0.883	0.865	0.86	0.883
NN	0.937	0.8	0.782	0.772	0.8
NB	0.939	0.717	0.717	0.737	0.717

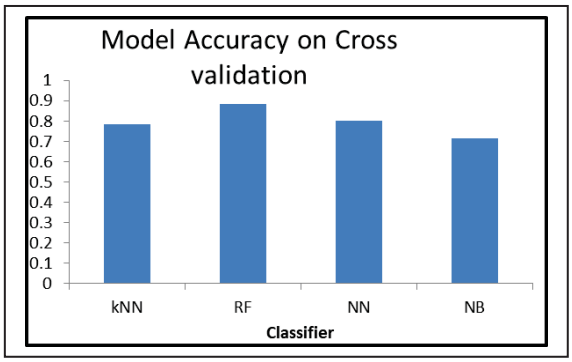

Fig. 4. Model performance on cross validation

Further, we predicted supplier by applying models on the test data, which was completely unfamiliar with the models during training as it was held out for the prediction. Test data had 14 observations in total that includes three instances for class 0 (Supplier A) and class 6 (Supplier G) respectively; two instances for class 3 (Supplier D), 4 (Supplier E) and 5 (Supplier F) respectively, one instance for class 1 (Supplier B), and 2 (Supplier C) respectively, as shown in Fig. 5.

TABLE IV. MODEL ACCURACY ON TEST DATA

Classifier	AUC	CA_Test	F1	Precision	Recall
kNN	0.905	0.714	0.684	0.768	0.714
RF	0.933	0.786	0.81	0.857	0.786
NN	0.97	0.786	0.774	0.81	0.786
NB	0.97	0.714	0.702	0.821	0.714

If we notice, confusion matrix, depicted in Fig. 5, of model results on test data, RF and NN have shown the same predictions for classes 0, 2, 3, 5, and 6, whereas classes 1 and 4 were poorly predicted by both these models. Model NB predicted classes 0, 3, 4 and 6 accurately and they weakly predicted the classes 1, 2 and 5. Lastly, model KNN has given worse prediction over all other models and it has predicted only class label 1 correctly. Classification accuracy of the models on test data is significantly lower than on the trained data. It is evident that the RF and NN models outperform KNN and NB models while predicting suppliers on the test data. Classification accuracy of the models on test data can be seen in Table IV.

The comparative performance of the classification models is depicted below in the Fig. 6.

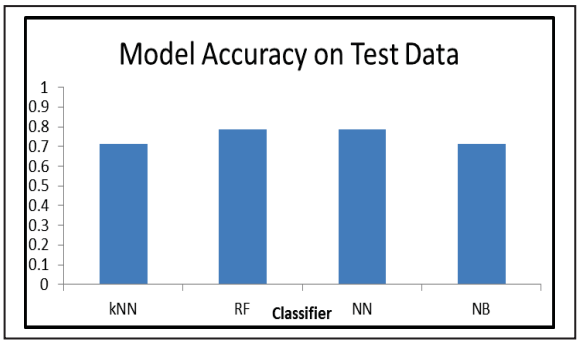

Fig. 6. Model performance on test data

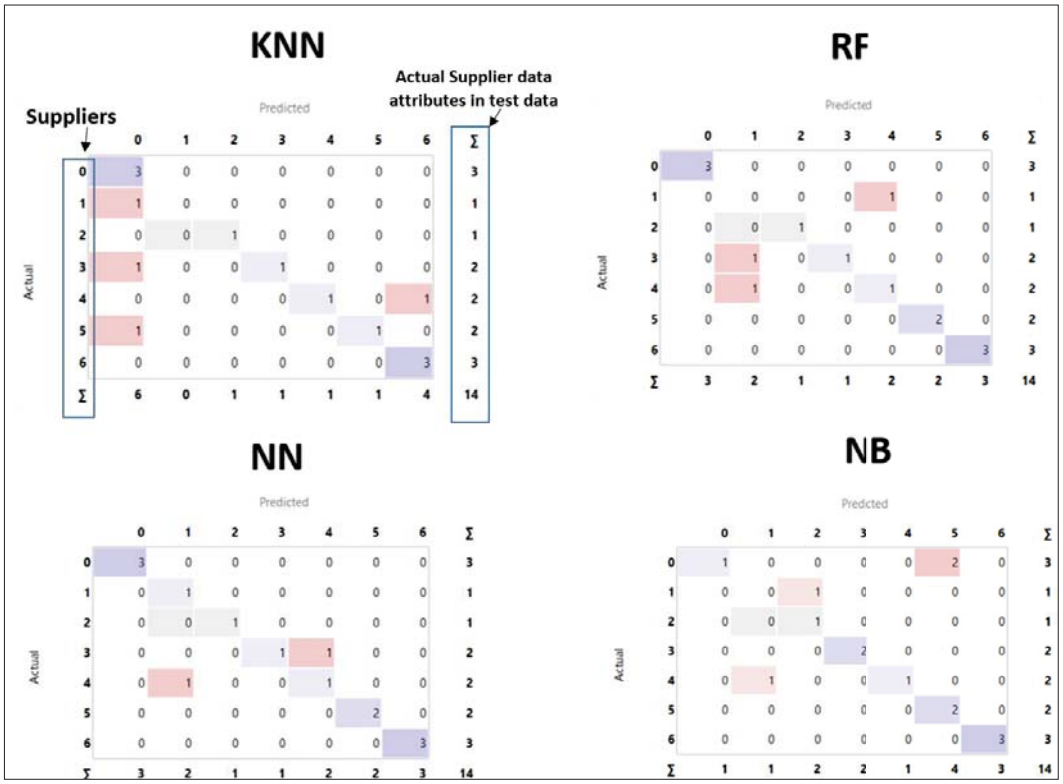

Fig. 5. Confusion matrix for test data

V. CONCLUSION

In this study, we aimed at predicting supplier for the new customer order which entailed product features. The European fashion company that deals with the difficulty of analyzing customer order data and identifying the best matching supplier can greatly benefit from our results as it can provide them with the ability and an effective tool to predict the best matching supplier for the newly received order based on the classification of suppliers vis-à-vis customer order attributes. We applied KNN, RF, NN, NB classification models on the customer order data to predict suppliers, and we found that the performance of RF and NN models is better than KNN and NB models. When we compared the model accuracy with their accuracy on trained and cross validation data, it is observed that the performance of RF and NN is consistent. It is imperative for the data mining methods that the dataset we use should be cleaned, preprocessed and transformed.

In a current fashion business framework, selection of suppliers is a manual decision making process and is based on the long term relationship with the suppliers. However, given the deep level of customization of products on the part of customers, it is difficult to make decisions with regard to suppliers as it is highly challenging and complex to study attributes in the customer order data and match them with the existing set of potential suppliers. In this study, we proposed data mining methods to solve this problem and the results from this study indicate the suitability of machine learning classification methods to address this problem. This study contributes to the automated decision making in the context of supplier selection for the customized fashion products.

The major scope for the improvement in this study lies in the selection of size of the dataset. Since our dataset is significantly small, models do not seem to perform overall very well. This can be enhanced by training classification models on historical customer order data that spans for at least 2 years so that we can ensure that training of the model is efficiently done. We believe that the limitation of this study arising from the relatively lower classification accuracy can be overcome by improving the model parameters according to the chosen size of dataset. Application of other classification models can also be explored while locating the problem of supplier prediction in a more enhanced problem context. Moreover, number of features selected for the prediction can be increased depending on the relationship between predictors and target supplier variable.

ACKNOWLEDGMENT

This study is carried out under the framework of FBD_BModel project funded by Horizon 2020 research and innovation project of European Union. The authors extend their gratitude to the European fashion company for providing the data, without which this study would not have been possible.

REFERENCES

[1] Wang L, Xianyi Z, Koehl L, Chen Y, "Intelligent fashion recommender system: fuzzy logic in personalized garment design." IEEE Trans Hum Mach Syst, 2015, 45:95–109

[2] Z. Guo, W. Wong, S. Leung, and M. Li, "Applications of artificial intelligence in the apparel industry: a review," Text. Res. J., vol. 81, no. 18, pp. 1871–1892, Nov. 2011.

[3] T.-M. Choi and B. Shen, "A system of systems framework for sustainable fashion supply chain management in the big data era," in 2016 IEEE 14th International Conference on Industrial Informatics (INDIN), 2016, pp. 902–908.

[4] M. Brandenburg, K. Govindan, J. Sarkis, S. Seuring, "Quantitative models for sustainable supply chain management: Developments and directions," European Journal of Operational Research, 2014, 233 (2) pp. 299-312

[5] M. J. Songhori, M. Tavana, A. Azadeh, and M. H. Khakbaz, "A supplier selection and order allocation model with multiple transportation alternatives," The International Journal of Advanced Manufacturing Technology, 2011, vol. 52, no. 1-4, pp. 365–376.

[6] Kim, J. K., Song, H. S., Kim, T. S., & Kim, H. K, "Detecting the change of customer behavior based on decision tree analysis," Expert Systems with Applications, 2005, 22, 193–205.

[7] C. Giri, N. Harale, S. Thomassey, and X. Zeng, "Analysis of consumer emotions about fashion brands: An exploratory study," in Data Science and Knowledge Engineering for Sensing Decision Support, 2018, pp. 1567–1574.

[8] Giudici, P., & Passerone, G., "Data mining of association structures to model consumer behaviour. Computational Statistics and Data Analysis," 2002, 38, 533–541.

[9] Bishop, C., "Pattern recognition and machine learning (information science and statistics)," 1st edn. 2006, corr. 2nd printing edn. Springer, New York.

[10] Marsland, S., "Machine learning: an algorithmic perspective," 2015,CRC press.

[11] James, G., Witten, D., Hastie, T., & Tibshirani, R.,"An introduction to statistical learning," 2013, (Vol. 6), Springer.

[12] Loh, W.-Y., "Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery," 2011, 1 (1), 14–23.