



This journal provides immediate open access to its content under the [Creative Commons BY 4.0 license](#). Authors who publish with this journal retain all copyrights and agree to the terms of the above-mentioned CC BY 4.0 license

DOI: 10.2478/seeur-2021-0008

---

# **DATA MINING APPROACH IMPROVING DECISION-MAKING COMPETENCY ALONG THE BUSINESS DIGITAL TRANSFORMATION JOURNEY: A CASE STUDY – HOME APPLIANCES AFTER SALES SERVICE**

---

PhD. Candidate Hyrmet Mydyti  
Faculty of Contemporary Sciences and Technologies, South East European University  
Tetovo, North Macedonia  
hm28315@seeu.edu.mk

## **ABSTRACT**

Data mining, as an essential part of artificial intelligence, is a powerful digital technology, which makes businesses predict future trends and alleviate the process of decision-making and enhancing customer experience along their digital transformation journey. This research provides a practical implication – a case study - to provide guidance on analyzing information and predicting repairs in home appliances after sales services business.

The main benefit of this practical comparative study of various classification algorithms, by using the Weka tool, is the analysis of information and the prediction of repairs in the home appliances after sales services business. The comparison of algorithms is performed considering different parameters, such as the mean absolute error, root mean square error, relative absolute error and root relative squared error, receiver operating characteristic area, accuracy, Matthews's correlation coefficient, precision-recall curve, precision, F-measure, recall and statistical criteria. Five classification algorithms such as the Naive Bayes, J48, random forest, K-Nearest Neighbor, and logistic regression were implemented in the

dataset. J48 has proved to provide the best accuracy and the lowest error among the other examined algorithms applied to a home appliances after sales services dataset to predict repairs based on product guarantee period.

The extracted information and results of an after sales services business by using data mining techniques prove to alleviate the process of streamlining decision-making and provide reliable predictions, especially for the customers, as well as increase businesses' efficiency along their digital transformation journey.

**Key words:** Data Mining, Classification, Digital Transformation, Home Appliances After Sales Services, Decision-Making.

## INTRODUCTION

Businesses use different technologies to perform analytics, such as simple reporting and dashboarding, which are just reports to inspect and study previous performance (Bumblauskas et al., 2017). However, businesses certainly use technologies as a means of data analytics in the perspective of digital transformation (OECD-BEIS, 2018). Businesses have a common aim to use digital technologies as an opportunity to improve decision-making competency along the business digital transformation journey (Schwertner, 2017). Advanced data analytics and artificial intelligence are drivers of deep analysis and change the businesses (West and Allen, 2018). Data mining, as a confluence of statistics and machine learning (artificial intelligence), makes businesses predict future trends (Palmer et al., 2011; Sethi et al., 2016).

Digitalization enables huge developments in after sales services; with prescriptive and predictive analysis opportunities, businesses reduce life-cycle costs and optimize costs (Rudnick et al., 2020).

Businesses can identify their customers' behavioral patterns, improve the availability of production facilities, enable evidence-based corporate decision-making, and reduce innovation latency via advanced analytics (Gimpel et al., 2018). Moreover, the customer journeys do not end at the point of sale; digital technologies enable the after sales services strategies of businesses and industries revolutionize (Gimpel et al., 2018).

Business processes or business reengineering is a process that links to the enhancement of business performance. The aim of business reengineering is to attain maximum enhancement in the business (Krstic et al., 2016).

Through the case study – practical implication, this research aims to analyze how the implementation of data mining techniques, such as classification techniques, in a home appliance after sales service business can lead to an improved decision-making and increased efficiency due to business prediction analysis along the digital transformation journey of home appliances after sales service businesses. Moreover, we will provide one scenario to digitalize the processes that enhance customer experience and affect business decision-making. The findings of this study provide guidance for businesses aiming to provide improved decision-making and optimized strategic goals along the digital transformation journey.

Our research question is as follows: (i) RQ1. How can a proposed data mining model contribute in the after sales service field to a business that considers about starting to use data mining aiming to improve decision-making and enhance customer experience? The question has been answered by implementing the data mining approach and making it a model for other businesses with the same concern. After the introduction, the definition of home appliance after sales services is discussed, followed by the implementation of the case study in the home appliance after sales service field.

The rest of this paper is organized as follows: The ‘Related work’ section gives an overview of home appliances after sales services, data mining classification algorithms that will be implemented in our dataset and evaluation metrics. In ‘Methodology’, we proceed with the implementation of the case study in the field of home appliance after sales services. In ‘Results and Discussion’, the results and outcomes have been presented and discussed. In ‘Conclusion’, the research conclusion of the study and discussions on future work have been provided.

## **RELATED WORK**

### **Home Appliance After Sales Services**

Durugbo (2020) carries out a systematic literature review on after sales services and aftermarket support. Principally and generally, after sales is the “period of time during which the seller or manufacturer guarantees to the buyer assistance, maintenance, or repairs of what has been purchased”. The aim of after sales services is to uphold alternatives for extended warranty that ensures product reliability and lessened servicing costs – besides the base guarantee that repairs product faults or failures. The goal of the after sales services business is to obey the warranty (guarantee) i.e. a temporal contractual assignment of the original equipment manufacturers (OEM) to offer uphold for an equipment.

Figure 1 presents the main after sales activities such as the “provision of field technical assistance, spare parts distribution, and customer care and accessories sale” (Durugbo, 2020). In addition, secondary markets are markets for complementary products that are frequently purchased consequent to the purchase of other, related products. The role of the figure is meant to present the standard activities of a home appliances after sales services business.

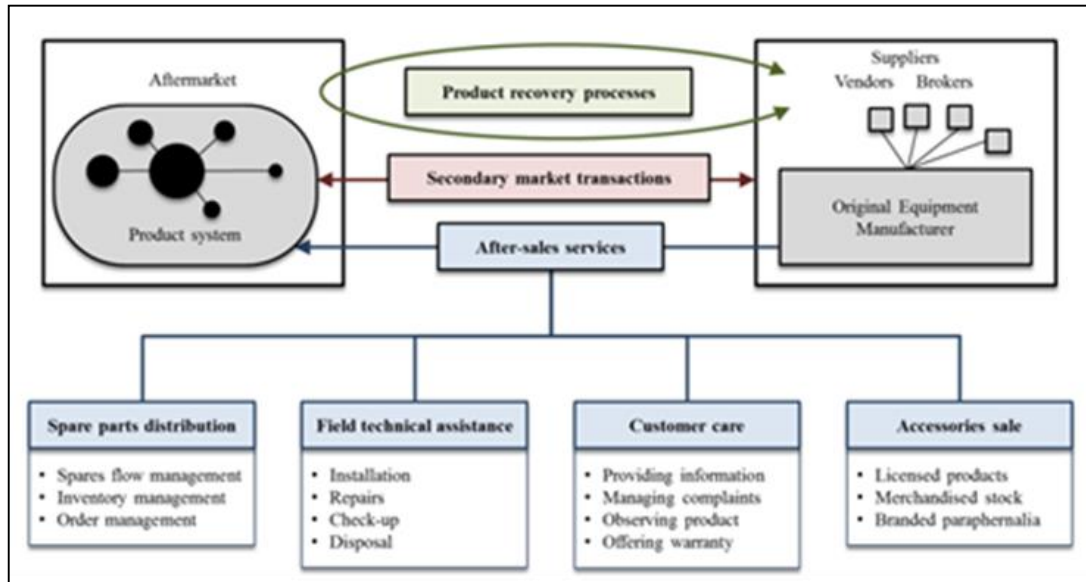


Figure 1 - After sales Service Business Processes (Durugbo, 2020)

Murali et al. (2016) define after sales services as activities that occur after the purchase of the product by customers and are dedicated to supporting customers in the usage and disposal of goods. The home appliance business is one of the most significant businesses globally and home appliances are one of the steadiest categories for customer satisfaction. Home appliances are comparatively expensive supplies and may require short and long-term financial consequences that consumers may find challenging to handle in terms of household budgets.

Esmailpour (2016) highlight “after sales service as one of the success factors of companies”. Home appliances after sales services comprise all issues of businesses selling goods to the customers such as “services warranty, commodity services shipping, installation services, supplying part goods, repairs services”. One of the problems faced by the buyers is “the failure to deliver adequate after sales service guarantee for durable home appliances”.

### Classification Algorithms

Ragab et al. (2014) present classification as amongst the most used data mining algorithms that is principally utilized to investigate a distinct dataset and gets each instance of it and designates this instance to a distinct class. Classification is applied to gain models that properly outline significant data classes. In the first stage, the model is built by executing a

classification technique on the training dataset. In the final stage, the gained model is tested opposed to a preset test dataset to evaluate the performance. Consequently, classification technique designates class label for the dataset whose class label is unrecognized.

Classification algorithms, as illustrated in Figure 2, include decision trees, statistical classifiers, k-nearest neighbours (KNN), Naïve Bayes, support vector machines, artificial neural networks (ANN), and others.



Figure 2 - Classification Algorithms used with Data Mining Technology

Classification algorithms have a wide range of applications in different aspects such as social network analysis, credit card rating, artificial intelligence, etc. (Jadhav and Channe, 2016).

### **Decision Tree Algorithm (J48)**

Kaur and Chhabra (2014) outline decision tree algorithm as a way to discover how the vector of attributes functions for a number of instances. The newly generated instances are found based on the training instances of the classes. The decision tree algorithm produces the rules for the prediction of the class of value of target variable. Through the decision tree classification algorithm, the critical data distribution is readily comprehensible. J48 algorithm, in the Weka tool, is an application of the C4.5 algorithm. The J48 algorithm produces the rules from which distinct data identity is produced. Both the discrete and continuous attributes are handled by this algorithm.

Jadhav and Channe (2016) study the advantages and disadvantages of the decision tree algorithm. The authors mention the advantages of the algorithm, stating that it simple and fast, it offers high accuracy, representation is easy to understand, it supports incremental learning, it takes the less memory and it can also deal with noisy data. The identified disadvantages of the algorithm include long training time, significantly more complex representation for some concepts due to replication problem and a problem of over fitting.

### **Random Forest Algorithm**

Avand et al. (2019) assert that the random forest technique is a trendy, tree-based method that comprises a multitude of regression trees and classification. In addition, random

forest is a nonparametric method for constructing the data of decision tree methods. The variations in the results of each tree are the key issue experienced by this method. A random forest approach is suggested, in order to decrease these variations and to decrease the estimation of variance. In addition, this is a mixture of a number of decision trees that integrate multiple bootstrap samples from the data, and several entry parameters randomly take part in the model of the tree.

Moreover, the authors, Ali et al. (2012), in their paper, include the advantages of the random forest algorithm, stating that it overcomes the problem of over fitting; it is less sensitive to outlier data; it sets the parameters easily and eliminates the need for pruning the trees; and it represents variable importance, and it generates automatically the accuracy.

### **Naïve Bayes algorithm**

Saputra et al. (2018) consider the Naïve Bayes algorithm as a probability-based classification method that is frequently used, although this algorithm pretends all attributes are independent. Thus, the Naïve Bayes algorithm functions by predicting probability that a given data instance pertains to a particular class. The Naïve Bayes algorithm can function well in real life. In addition, the Naïve Bayes algorithm does not demand a large quantity of training data to assign parameter estimation of the classification process.

Jadhav and Channe (2016) study the advantages and disadvantages of Naïve Bayes algorithm. The authors identify the advantages of the algorithm, such as it needs short calculation time for training; it enhances the algorithm performance by eliminating the insignificant features; and it offers useful results, and high accuracy. The identified disadvantages of the algorithm are that it involves a huge amount of data (tuples) to gain good performance and is less accurate in comparison to other classifiers on some datasets.

### **Logistic regression algorithm**

Zhu et al. (2019) assert that the “Logistic regression algorithm is applied when the aim is to classify data items into categories. In logistic regression, the target variable is binary, which signifies that it only comprises data classified as 1 or 0. The application of the Logistic regression model has featured dominant in many domains”.

Ray (2019) introduces logistic regression advantages and disadvantages. The advantages of the algorithm are the following: ease of implementation, computational efficiency, and simplicity of regularization. On the other hand, the disadvantages of the algorithm are as follows: incapability to figure out non-linear problem as its decision surface

is linear, disposed to over fitting, will not function fine unless all independent parameters are discovered. Logistic Regression can be implemented in different fields, such as: predicting the risk of increasing a given disease, predicting probability of faults of a process or system (in engineering), etc.

### **K-Nearest Neighbour (KNN) Algorithm**

Moreover, Avand et al. (2019) affirm that the K-Nearest Neighbour (KNN) algorithm is a method for classifying an unrecognized ensemble if we possess data with particular properties (X) and the value of the relationship (Y). A non-parametric learning algorithm is the KNN Classifier. In the classification frame, the algorithm estimates the distance of the target point to the nearest points in accordance with the value given for K, and earn conformity with the maximal amount of votes of these neighboring points.

Jadhav and Channe (2016) elaborate the advantages and disadvantages of the KNN algorithm. The authors mark the advantages of the algorithm, stating that it is mere to comprehend and apply, training is very fast, it is persistent to noisy training data, and offers good performance and accuracy. The identified disadvantages of algorithm are lazy learners, which incur expensive computational costs, sensitivity to the local structure of the data, memory limitation and delay in running.

### **Evaluation Metrics**

Hossin and Sulaiman (2015) assert that evaluation metric plays an important role in attaining the optimal classifier during the classification training. A selection of appropriate evaluation metric is a significant part for discriminating and acquiring the optimal classifier.

Çığışar and Ünal (2019) elaborate in detail the criteria of the classifier's performance; they rank six main criteria as follows: classification accuracy, speed, robustness, scalability, interpretability and rule structure. Classification accuracy is referred as the "capability of the model to correctly predict the class label, which is stated as a percentage". Speed is defined by "the time required to construct the model". Robustness is defined as the "capability to predict the model correctly despite the fact that the data has noise and missing values". Scalability is denoted as the "capability of a model to be accurate and productive while dealing with an increasing amount of data". Interpretability is indicated as the "degree of understanding and insight provided by the model". Finally, rule structure is explained as "the comprehensibility of the rule structure of the algorithm".

The author, Verma (2019), presents an overview on evaluation metrics. Evaluation metrics, such as “true positive rate (TPR), false positive rate (FPR), precision, recall, F-measure, Matthews’s correlation coefficient (MCC), receiver operating characteristic (ROC) area, precision recall curve (PRC) area”, are considered important when evaluating classification algorithm applied on a dataset.

“TPR is the measure of exactness; more specifically, it is the percentage of tuples in a test dataset that are assigned as positive. Recall is the measure of completeness; more specifically, it is the percentage of positive tuples in a test dataset that the classifier is assigned as positive. F measure is the harmonic mean of precision and recall. ROC curves are a visual tool that helps us in comparing two classification models. An ROC curve for a certain model brings out the trade-off between the TPR and the FPR. For a certain test dataset and a model, TPR is the ratio of positive tuples that the models label correctly; FPR is the ratio of negative tuples that are mislabeled as positive. PRC area is the area under the precision-recall curve that measures the accuracy of the classification model on a class dataset” (Verma, 2019).

The metrics for classification evaluations presented in the following table:

Metrics	Formula
Accuracy	$\frac{(tp + tn)}{(tp + fp + tn + fn)}$
Error Rate	$\frac{(fp + fn)}{(tp + fp + tn + fn)}$
Precision (p)	$\frac{tp}{(tp + fp)}$
Specificity	$\frac{tn}{(fp + tn)}$
Recall (r)	$\frac{tp}{(tp + fn)}$
Sensitivity	$\frac{tp}{(tp + fn)}$
F Measure	$\frac{(2 * p * r)}{(p + r)}$
MCC	$\frac{(tp * tn) - (fp * fn)}{[(tp + fn) * (tp + fp) * (tn + fp) * (tn + fn)]^{1/2}}$
<b>Note:</b> <b>tp</b> - true positive, <b>fp</b> - false positive, <b>fn</b> - false negative, <b>tn</b> - true negative	

Specificity is utilized as a measure of the proportion of negative patterns that are precisely classified. Sensitivity is utilized as a measure of the proportion of positive patterns that are precisely classified. Precision is utilized as a measure of the positive results that are precisely predicted from the total predicted results in a positive class. In addition, the accuracy



measures the proportion of correct predictions to the total number of instances. The MCC coefficient measures the quality (rate) of binary classifications (Hossin and Sulaiman, 2015; Halimu et al., 2019).

## **METHODOLOGY**

The methodology followed in this study started with data preparation from the home appliances after sales service company's dataset. A company which operates in the home appliances after sales service in Kosovo, repairs electronic products including a large variety of home appliances. The company also offers services such as installation of home appliances including air conditioners, washing machines, dishwashers, etc. The dataset used in this project has been taken from a company offering home appliances after sales service. The dataset obtained is maintained and updated by the company's customer care department. For this study, approximately data from 62K cases were extracted from the database.

Moreover, the methodology is followed with the comparison of several types of classification methods. The aim of this comparison of classification methods is to find the best algorithm amongst different method, which would be useful for this case study. The classification algorithms such as Naive Bayes, J48, random forest, K-Nearest Neighbor, and logistic regression are implemented in the dataset to predict repairs based on product guarantee period in this case study. The feature selection techniques, such as correlation-based attribute evaluation, information gain attribute evaluation and wrapper attribute evaluation, are applied to lessen the attributes of the class variables by removing the redundant and non-predictive information features from the dataset. It resulted that the high impact attributes to predict repairs are guarantee, authorized service, brand, action, category, repaired by age, days to repair and sum.

### **Data Mining Tool - Weka**

Attwal and Dhiman (2020) explore the data mining tool – Weka. The Weka tool is a Java-based open source platform and is an incorporation “of machine learning algorithms and data preprocessing tools”. The Weka workbench includes application of algorithms for implementing the key data mining, supervised and unsupervised tasks. The data mining tool – Weka enables both – graphical user interface and a command line interface. The Weka 3.8.4 implementation were applied for the classification algorithms of this case study.

### Challenges Faced By the Company as a Case Study

The business does not implicate any data mining tools to perform any advanced analytics. Thus, there is a lack of mining of data.

Given the importance of result delivery in improving decision-making and the productivity, the implementation of the analytics systems and intelligent applications endure a challenge for the home appliances after sales service businesses. The next challenge is to extract and capture reliable and actionable useful information and knowledge from large data sources with minimal human interaction. More importantly, the other challenge is integration and connection of delivery to the business logic with impact on decision-making.

### Data Selection and Pre-Processing

All the data are pre-processed through Weka. Through the data mining platform, Weka, and its techniques, is the aim to successfully analyze and bring the understanding of the hidden patterns in business data. In the existing workflow of the business, customers' requests are collected from the customer service department. The home appliances after sales service business' dataset is used in CSV format and with 62K observations and eight variables. Since, data pre-processing is an important stage of the data mining process, the selected data are cleaned from inconsistent and erroneous data, missing data and outliers. The used attributes include: authorized service (nominal), brand (nominal), category (nominal), repaired by age (numeric), guarantee (nominal), action (nominal), sum (numeric), guarantee (nominal) and days to repair (numeric).

The structure of dataset: the attributes have been presented in the following table:

Attribute	Description	Values
Guarantee	Whether the product has a guarantee or not.	Yes or No
Authorized Service	Whether the product is serviced by own service center or by subcontractor	Service center, Subcontractor
Brand	Brand means the manufacturer of the product and has 10 different values.	Brand 1 - 9, Other Brands
Action	The type of service performed to the product	Servicing, Installation, Change
Category	Category of the product is as major domestic appliances, small domestic appliances, personal care, TV, mobile, personal care, air conditioning, etc., and has nine different values	Category 1 – 9
Repaired By Age	The age of the product and is a numeric value	Numeric Value
Days to Repair	The number of days the service is completed since the time requested for service	Numeric Value
Sum	The value to perform the service and is a numeric value – if the service of the product is in guarantee, the sum is zero	Numeric Value

The objective of this project is to analyze how the implementation of classification techniques can lead to an improved decision-making and increased efficiency due to business prediction analysis.

## RESULTS AND DISCUSSION

The increasing need of data mining applications based on classification in the home appliances after sales service field has revealed the need for practicing data-mining algorithms for better decision-making. The results and discussions that were obtained from the comparisons of classification algorithms conducted in the case study have been presented in detail below.

The main aim of the algorithm comparisons is to find the best algorithm to tackle issues or solve problems in home appliances after sales services business. Five classification algorithms have been comparatively tested to find the finest algorithm for classification. The dataset has been used for experimental purposes containing 62k instances with seven attributes and 1 class attribute (Guarantee attribute) to test and justify the differences among classification algorithms. The test option chosen is 10 folds cross validation, which implies that the dataset is divided into 10 sections. The first nine sections are utilized to train the algorithm, and the 10th section is utilized to assess the algorithm. Cross validation is an important and useful feature to fine-tune data mining model. The comparative analysis linked to classification measures involved recall, precision, F Measure, Matthews's correlation coefficient (MCC), precision-recall curve (PRC), ROC curve, FPR, and TPR have been extracted. The Weka is used for simulation.

Performance measurement results (comparison of classification algorithms) have been presented in the following table:

Algorithm	Performance Measure							
	TPR	FPR	Precision	Recall	F-measure	MCC	Area	
							ROC	PRC
<b>J48</b>	0.953	0.288	0.950	0.953	0.951	0.709	0.859	0.940
<b>RandomForest</b>	0.949	0.328	0.946	0.949	0.946	0.680	0.911	0.957
<b>IBk</b>	0.946	0.359	0.943	0.946	0.943	0.660	0.882	0.948
<b>Logistic</b>	0.914	0.773	0.901	0.914	0.887	0.305	0.833	0.924
<b>NaïveBayes</b>	0.905	0.785	0.876	0.905	0.879	0.223	0.797	0.905

Comparison of classification algorithms table recapitulates the results of the simulation of five classifier algorithms used and their corresponding performance measures for the home appliances after sales service of the customer care dataset. It displays that J48, RandomForest and K-Nearest Neighbor (IBk known in Weka) provide the best performance among the others algorithms studied. The other useful statistical characteristic, receiver operating characteristic (ROC) area, for which the value has to be over 0.5 of all the algorithms, indicates the existence of statistical dependence. The two most popular metrics are overall accuracy and Kappa Coefficient. Consequently, the application of J48 algorithm is useful for home appliances after sales services business, respectively in bringing value to business in enhancing customer care services and in alleviating decision-making.

Classifier algorithm results have been presented in the following table:

<b>Algorithm</b>	<b>Accuracy (%)</b>	<b>Kappa Statistic</b>	<b>MAE</b>
<b>J48</b>	95.2604	0.7071	0.0844
<b>RandomForest</b>	94.8711	0.6755	0.0787
<b>IBk</b>	94.6373	0.6536	0.0789
<b>Logistic</b>	91.3985	0.2207	0.1334
<b>Naïve Bayes</b>	90.4935	0.178	0.1253

Classifier algorithm results recapitulate the results of the five-classifier algorithms in terms of accuracy, Kappa statistic and mean absolute error (MAE). It shows that J48, RandomForest and IBk give the best accuracy among the other algorithms examined. Logistic and Naïve Bayes algorithms give the lowest accuracy. Moreover, it shows that J48, RandomForest and IBk give the lowest error among the others algorithms examined. The Naïve Bayes algorithm gives the highest error. The value of the Kappa statistic of J48 is 0.7071, which indicates the existence of statistical dependence and is high. The value of the accuracy of J48 is 95.2604%, which indicates high value of the percentage of correctly classified instances for all instances.

Mean absolute error (MAE), root mean square error (RMSE), relative absolute error (RAE) and root relative squared error (RRSE) Metrics have been presented in the following table:

Algorithm	MAE	RMSE	RAE	RRSE
<b>J48</b>	0.0844	0.2058	49.0169 %	70.1451 %
<b>RF</b>	0.0787	0.208	45.7368 %	70.8913 %
<b>IBk</b>	0.0789	0.2187	45.8332 %	74.5325 %
<b>Logistic</b>	0.1333	0.2536	77.4472 %	86.4468 %
<b>Naïve Bayes</b>	0.1253	0.2819	72.7974 %	96.0723 %

The tabular representation in the table above shows that the decision tree algorithm (J48) has the least errors compared to the other algorithms. The errors presented are MAE, RMSE, RAE and RRSE.

The model is evaluated based on its predictive performance. J48 has proved to provide the highest accuracy rate and the lowest error among the other examined algorithms applied to a home appliances after sales services dataset to predict repairs based on product guarantee period. J48 algorithm is the best algorithm based on the MAE, RMSE, RAE and RRSE, receiver operating characteristic area, accuracy, Matthews's correlation coefficient, precision-recall curve, precision, F-measure, recall and statistical criteria. Consequently, in this phase, it is determined that the decision tree (J48) data mining algorithm performs best and can be used to generate the predictive model and can predict repairs based on product guarantee period. Furthermore, all the details from the algorithms are visually elaborated through the outputs generated by applying different binary classifiers.

```

Correctly Classified Instances      59234              94.8791 %
Incorrectly Classified Instances    3197                5.1209 %
Kappa statistic                    0.676
Mean absolute error                 0.0787
Root mean squared error             0.208
Relative absolute error             45.7368 %
Root relative squared error         70.8913 %
Total Number of Instances          62431

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0.981    0.361    0.963     0.981    0.972     0.680    0.911    0.984    Yes-guarantee
      0.639    0.019    0.783     0.639    0.704     0.680    0.911    0.696    No-guarantee
Weighted Avg.   0.949    0.328    0.946     0.949    0.946     0.680    0.911    0.957

=== Confusion Matrix ===

      a      b  <-- classified as
55438  1054 |    a = Yes-guarantee
 2143   3796 |    b = No-guarantee

```

Figure 3 – Output generated by applying Random Forest binary classifier to dataset

In figure 3, experiment results show the efficiency and effectiveness of Random Forest algorithm in predicting the repairs based on product's guarantee period. The model achieved 94.8791% accuracy. The confusion matrix for training dataset shows the accuracy of the predicted classes. The training data consist of 62,431 instances - with guarantee 56,492 and

with no guarantee 5,939. As a result, the generated tree classified 59,234 instances correctly out of 62,431 instances.

```

Correctly Classified Instances      59083          94.6373 %
Incorrectly Classified Instances    3348          5.3627 %
Kappa statistic                    0.6536
Mean absolute error                0.0789
Root mean squared error            0.2187
Relative absolute error            45.8332 %
Root relative squared error        74.5325 %
Total Number of Instances         62431

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0.982    0.395    0.959    0.982    0.971     0.660    0.882    0.981    Yes-guarantee
      0.605    0.018    0.782    0.605    0.682     0.660    0.882    0.640    No-guarantee
Weighted Avg.   0.946    0.359    0.943    0.946    0.943     0.660    0.882    0.948

=== Confusion Matrix ===

      a      b  <-- classified as
55487  1005 |      a = Yes-guarantee
 2343   3596 |      b = No-guarantee

```

Figure 4 – Output generated by applying IBk binary classifier to dataset

In figure 4, experiment results show the efficiency and effectiveness of IBk algorithm in predicting the repairs based on product guarantee period. The model achieved 94.6373% accuracy. The confusion matrix for training dataset shows the accuracy of the predicted classes. The training data consist of 62,431 instances - with guarantee 56,492 and with no guarantee 5,939. As a result, the generated tree classified 59,083 instances correctly out of 62,431 instances.

```

Correctly Classified Instances      57062          91.4001 %
Incorrectly Classified Instances    5369          8.5999 %
Kappa statistic                    0.2209
Mean absolute error                0.1333
Root mean squared error            0.2536
Relative absolute error            77.4472 %
Root relative squared error        86.4468 %
Total Number of Instances         62431

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0.995    0.853    0.917    0.995    0.954     0.306    0.833    0.964    Yes-guarantee
      0.147    0.005    0.742    0.147    0.245     0.306    0.833    0.547    No-guarantee
Weighted Avg.   0.914    0.772    0.901    0.914    0.887     0.306    0.833    0.925

=== Confusion Matrix ===

      a      b  <-- classified as
56189   303 |      a = Yes-guarantee
 5066   873 |      b = No-guarantee

```

Figure 5 – Output generated by applying Logistic binary classifier to dataset

In figure 5, experiment results show the efficiency and effectiveness of Logistic algorithm in predicting the repairs based on product guarantee period. The model achieved

91.4001% accuracy. The confusion matrix for training dataset shows the accuracy of the predicted classes. The training data consist of 62,431 instances - with guarantee 56,492 and with no guarantee 5,939. As a result, the generated tree classified 57,062 instances correctly out of 62,431 instances.

```

Correctly Classified Instances      56496          90.4935 %
Incorrectly Classified Instances    5935           9.5065 %
Kappa statistic                    0.178
Mean absolute error                 0.1253
Root mean squared error             0.2819
Relative absolute error             72.7974 %
Root relative squared error         96.0723 %
Total Number of Instances          62431

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.986   0.866   0.915     0.986   0.949     0.223   0.797   0.966   Yes-guarantee
                0.134   0.014   0.501     0.134   0.211     0.223   0.797   0.328   No-guarantee
Weighted Avg.   0.905   0.785   0.876     0.905   0.879     0.223   0.797   0.905

=== Confusion Matrix ===

  a    b  <-- classified as
55703  789 |  a = Yes-guarantee
 5146   793 |  b = No-guarantee

```

Figure 6 – Output generated by applying Naïve Bayes binary classifier to dataset

In figure 6, experiment results show the efficiency and effectiveness of Naïve Bayes algorithm in predicting the repairs based on product guarantee period. The model achieved 90.4935% accuracy. The confusion matrix for training dataset shows the accuracy of the predicted classes. The training data consist of 62,431 instances - with guarantee 56,492 and with no guarantee 5,939. As a result, the generated tree classified 56,496 instances correctly out of 62,431 instances.

```

Correctly Classified Instances      59472          95.2604 %
Incorrectly Classified Instances    2959           4.7396 %
Kappa statistic                    0.7071
Mean absolute error                 0.0844
Root mean squared error             0.2058
Relative absolute error             49.0169 %
Root relative squared error         70.1451 %
Total Number of Instances          62431

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.981   0.316   0.967     0.981   0.974     0.709   0.859   0.973   Yes-guarantee
                0.684   0.019   0.790     0.684   0.733     0.709   0.859   0.620   No-guarantee
Weighted Avg.   0.953   0.288   0.950     0.953   0.951     0.709   0.859   0.940

=== Confusion Matrix ===

  a    b  <-- classified as
55412 1080 |  a = Yes-guarantee
 1879  4060 |  b = No-guarantee

```

Figure 7 – Output generated by applying J48 binary classifier to dataset

In figure 7, experiment results show the efficiency and effectiveness of J48 algorithm in predicting the repairs based on product guarantee period. The model achieved 95.2604% accuracy. The confusion matrix for training dataset shows the accuracy of the predicted classes. The training data consist of 62,431 instances - with guarantee 56,492 and with no guarantee 5,939. As a result, the generated tree classified 59,472 instances correctly out of 62,431 instances. These results recommend that among the five algorithms tested, the J48 classifier has the potential to remarkably enhance the classification methods for use in home appliances after sales services field.

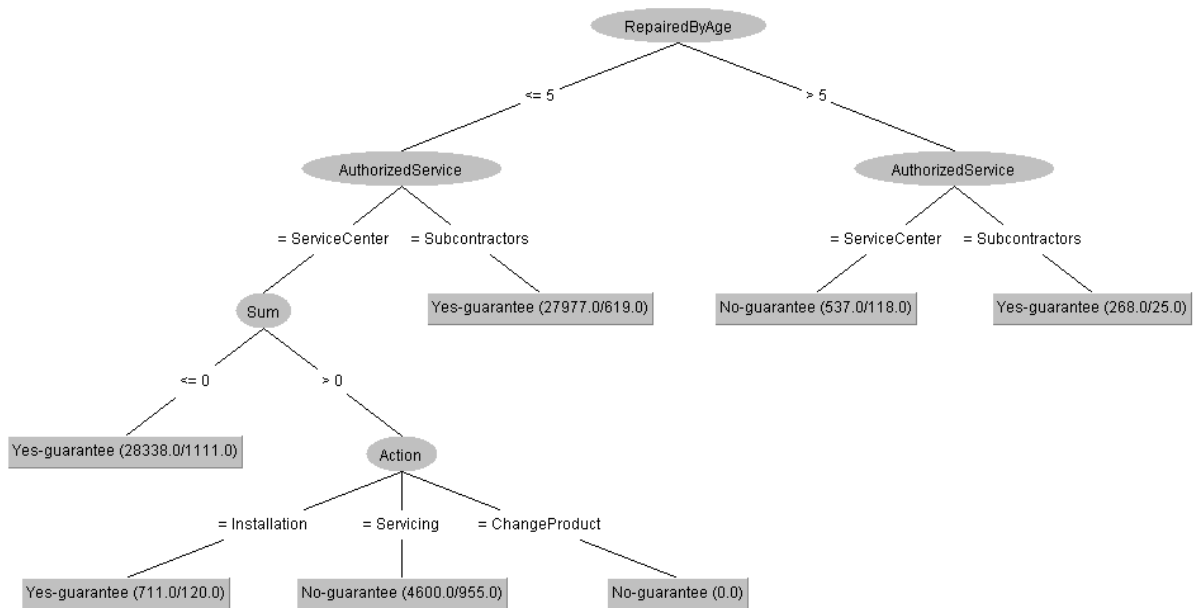


Figure 8 - Decision Tree (J48) visualization

Figure 8 shows the decision tree (J48) visualization. The decision tree displayed includes repaired by age, authorized service and sum, as root, and class for leaf node, which is the product guarantee. The leaf node, the product guarantee, has been presented by rectangle and root nodes, repaired by age, authorized service and sum, have been presented by an oval. The prediction is classified as binary (two-class) classification since the class attribute has two output values, in which the guarantee of products can be identified with or with no guarantee. This case study addressed a gap and constructed a binary classification model - decision tree algorithm in predicting the possible repairs of home appliances after sales services based on the product guarantee period.

The decision tree predicts that the subcontractors, under control of after sales services authorized company, provide services and repairs the products under guarantee. The authorized company happens to ask for services from subcontractors for the products with guarantee when



the load of repairs is huge or unexpected due to different situations and conditions. The authorized service is obliged to offer services for the products with guarantee, and consequently requests assistance from subcontractors.

Moreover, if the value of the age of the products is equal or less than five years and with sum (payment) zero, the decision tree predicts the repairs of products with guarantee. If the value of the age of the products is greater than five years, the decision tree predicts the repairs of products without guarantee.

Furthermore, if the value of the age of the products is equal or less than five years and the value of the sum (payment) is greater than zero, the decision tree predicts the repairs of products without guarantee and predicts the installation of products under guarantee. Installation of product is an extra service, which is paid, and usually the product is under guarantee.

The binary prediction model anticipates the repairs based on the guarantee period. The longer the guarantee period of products, the higher the possibility of the need of servicing. The high-predicted value of the age of the repairs under guarantee indicates not to be a beneficial indicator. In general, home appliances after sale services activities are independent activities from the sales activities businesses. Consequently, the interest of home appliances after sales services business is not to have a longer guarantee period since it affects profit of the business and customer experience. The first recommendation based on the outcome of the predictive data mining model is custodian, to be utilized in time of setting the deals with manufacturers and retailers.

Moreover, the after sale service can offer a service to the customers that own the products with expired guarantee. The second outcome of the predictive data mining model indicates the low servicing of expired guarantee products. Herewith, the second outcome of this model indicates the need for customer experience enhancement in order to improve services by an authorized after sales services business. The second recommendation based on the model outcome, the business can inform own customers with expired guarantee products that can get the service from an authorized after sales services business.

Consequently, based on all the parameters evaluated for the best algorithm, the binary prediction model can be used for the prediction of product's guarantee period in home appliances after sales service business. The model describes the relationships between the variables and the outcome. This binary prediction model helps decision-makers to identify gaps

and take appropriate decisions. Furthermore, the recommendations intend to improve business services, result in profit, and influence in decision-making processes and enhance customer experience of the home appliances after sales service business along the business digital transformation journey.

## CONCLUSION

The main goal and the contribution is the demonstration of the implementation of the data mining approach in the field of business home appliances after sales service, as a model for other businesses with the same concern. Moreover, the paper provides an overview of home appliances after sales service business processes, data mining techniques, in particular classification techniques, together with several important aspects of home appliances after sales services, and digital transformation as the essential trend by affecting the whole business.

The implementation of the data mining approach deals with discovering the best classification technique through the practical comparison of several classification algorithms. The data mining tool used is Weka and eight attributes applied include: authorized service, brand, category, repaired by age, guarantee, action, sum, guarantee and days to repair. Naive Bayes, J48, random forest, K-Nearest Neighbor, and logistic regression classification algorithms have been implemented in the dataset to predict repairs based on guarantee period in this case study. J48 has proved to provide the best accuracy among the other examined algorithms. J48 has also proved to give the lowest error among the other examined algorithms.

According to the outcomes, the constructed decision tree will be as a sample predictive binary model for other businesses with the same concern. We obtained a practical sample of the implementation of the data mining approach on business prediction in the field of the home appliances after sales service. The benefit of the constructed binary prediction model is the impact on enhancing customer experience and increasing efficiency due to business prediction analysis.

Through the presented recommendations, managers of home appliances after sales services will be aware of data mining approach implementation benefits. The main aim of the recommendations is to improve the decision-making competency along the business digital transformation journey.

Finally, the data mining approach enables to make strategic decision and data mining technologies play a significant role to improve business performance. This paper helps particularly home appliances after sales service businesses, as key beneficiaries, understand the positive aspects of data mining along their business digital transformation.

Our future goal is to conduct further studies to examine additional factors and dimensions. Therefore, future research can collect and analyze data using a larger sample size, by using the advantages of the data mining approach, and obtain feedback from different type of businesses to increase the chance of findings. In addition, the future work will advance the practical perspective.

## REFERENCES

1. Ali, J., Khan, R., Ahmad, N. and Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, 9(5), 272.
2. Attwal, K. P. S. and Dhiman, A. S. (2020). Exploring Data Mining Tool-Weka And Using Weka To Build And Evaluate Predictive Models. *Advances and Applications in Mathematical Sciences*, 6(19), 451-469.
3. Avand, M., Janizadeh, S., Naghibi, S. A., Pourghasemi, H. R., Khosrobeigi Bozchaloei, S. and Blaschke, T. (2019). A comparative assessment of random forest and k-nearest neighbor classifiers for gully erosion susceptibility mapping. *Water*, 11(10), p. 2076.
4. Bumblauskas, D., Nold, H., Bumblauskas, P., and Igou, A. (2017). Big data analytics: transforming data to action. *Business Process Management Journal*, 703-720.
5. Çığsar, B. and Ünal, D. (2019). Comparison of data mining classification algorithms determining the default risk. *Scientific Programming*, 2019.
6. Durugbo, C. M. (2020). After-sales services and aftermarket support: a systematic review, theory and future research directions. *International Journal of Production Research*, 58(6), 1857-1892.
7. Esmaeilpour, M. (2016). Analyzing after Sales Services in House Appliances Products and Measuring Customers Satisfaction: A Survey in Bushehr, Iran. *Journal of Harmonized Research in Management*, 2(2), 204-215.
8. Gimpel, H., Hosseini, S., Huber, R. X. R., Probst, L., Röglinger, M. and Faisst, U. (2018). Structuring Digital Transformation: A Framework of Action Fields and its Application at ZEISS. *Journal of Information Technology, Theory and Application*, 1(19), p. 3.
9. Halimu, C., Kasem, A. and Newaz, S. S. (2019, January). Empirical comparison of area under ROC curve (AUC) and Mathew correlation coefficient (MCC) for evaluating machine learning algorithms on imbalanced datasets for binary classification. In *Proceedings of the 3rd international conference on machine learning and soft computing* (pp. 1-6).
10. Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2), 1.
11. Jadhav, S. D. and Channe, H. P. (2016). Comparative study of K-NN, naive Bayes and decision tree classification techniques. *International Journal of Science and Research (IJSR)*, 5(1), 1842-1845.
12. Kaur, G., & Chhabra, A. (2014). Improved J48 classification algorithm for the prediction of diabetes. *International journal of computer applications*, 98(22).
13. Krstic, J., Jovanov, G., Radovanovic, R., Ljusic, M. and Nikolic, M. (2016). Process of Business Reengineering from the Aspect of E-Business. *Journal of Textile Science & Engineering*, 272(6), p. 2.
14. Murali, S., Pugazhendhi, S. and Muralidharan, C. (2016). Modelling and investigating the relationship of after sales service quality with customer satisfaction, retention and loyalty—a case study of home appliances business. *Journal of retailing and consumer services*, 30, 67-83.
15. OECD and the United Kingdom Department for Business, Energy and Industrial Strategy (BEIS). (2018). Implications of the Digital Transformation for the Business. <http://www.oecd.org/sti/ind/digital-transformation-business-sector-summary.pdf>.

16. Palmer, A., Jiménez, R. and Gervilla, E. (2011). Data mining: Machine learning and statistical techniques. *Knowledge-Oriented Applications in Data Mining, Prof. Kimito Funatsu (Ed.)*, 373-396.
17. Ragab, A. H. M., Noaman, A. Y., Al-Ghamdi, A. S. and Madbouly, A. I. (2014, June). A comparative analysis of classification algorithms for students college enrollment approval using data mining. In *Proceedings of the 2014 Workshop on Interaction Design in Educational Environments* (pp. 106-113).
18. Ray, S. (2019, February). A quick review of machine learning algorithms. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)* (pp. 35-39). IEEE.
19. Rudnick, M., Riezebos, J., Powell, D. J. and Hauptvogel, A. (2020). Effective after-sales services through the lean servitization canvas. *International Journal of Lean Six Sigma*, 5(11), 943-956.
20. Saputra, M. F. A., Widiyaningtyas, T. and Wibawa, A. P. (2018). Illiteracy classification using K means-Naïve Bayes algorithm. *JOIV: International Journal on Informatics Visualization*, 2(3), 153-158.
21. Schwertner, K. (2017). Digital transformation of business. *Trakia Journal of Sciences*, 15(1), 388-393.
22. Sethi, S., Malhotra, D., and Verma, N. (2016). Data mining: current applications & trends. *International Journal of Innovations in Engineering and Technology*, 6(4), 586-589.
23. Verma, A. (2019). Evaluation of classification algorithms with solutions to class imbalance problem on bank marketing dataset using WEKA. *International Research Journal of Engineering and Technology*, 5(13), 54-60.
24. West, D. M. and Allen, J. R. (2018). How artificial intelligence is transforming the world. Report, Brookings Institution.
25. Zhu, C., Idemudia, C. U. and Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, 17, 100179.