energies

MDPI

*Article*

# Enhancement of Machinery Activity Recognition in a Mining Environment with GPS Data

**Paulina Gackowiec *, Edyta Brzychczy and Marek Kęsek**

Faculty of Civil Engineering and Resource Management, AGH University of Science and Technology, 30-059 Cracow, Poland; brzych3@agh.edu.pl (E.B.); kesek@agh.edu.pl (M.K.)
* Correspondence: gackowiec@agh.edu.pl

**Abstract:** Fast-growing methods of automatic data acquisition allow for collecting various types of data from the production process. This entails developing methods that are able to process vast amounts of data, providing generalised knowledge about the analysed process. Appropriate use of this knowledge can be the basis for decision-making, leading to more effective use of the company's resources. This article presents the approach for data analysis aimed at determining the operating states of a wheel loader and the place where it operates based on the recorded data. For this purpose, we have used several methods, e.g., for clustering and classification, namely: DBSCAN, CART, C5.0. Our approach has allowed for the creation of decision rules that recognise the operating states of the machine. In this study, we have taken into account the GPS signal readings, and thanks to this, we have indicated the differences in machine operation within the designated states in the open pit and at the mine base area. In this paper, we present the characteristics of the selected clusters corresponding to the machine operation states and emphasise the differences in the context of the operation area. The knowledge obtained in this study allows for determining the states based on only a few selected most essential parameters, even without consideration of the coordinates of the machine's workplace. Our approach enables a significant acceleration of subsequent analyses, e.g., analysis of the machine states structure, which may be helpful in the optimisation of its use.

**Keywords:** sensor data; mining machinery; activity recognition; clustering; GPS data

check for updates

**Citation:** Gackowiec, P.; Brzychczy, E.; Kęsek, M. Enhancement of Machinery Activity Recognition in a Mining Environment with GPS Data. *Energies* **2021**, *14*, 3422. https://doi.org/10.3390/en14123422

Academic Editor: Nikolaos Koukouzas

Received: 6 May 2021
Accepted: 3 June 2021
Published: 10 June 2021

## 1. Introduction

Nowadays, companies are looking for innovative methods and techniques to maximise the efficiency of their operations and optimise the usage of their fixed assets. Because of progressive technological development and increasing hardware capabilities of data acquisition and storage, approaches based on the analysis of machinery data are gaining importance. The gathered data enables monitoring and ongoing insight into the operation of the utilised equipment, as well as its comprehensive and complex evaluation. Currently, very detailed and precise machine-specific data are available to companies, characterising the operation of all the main components of the machine, often recorded continuously. This constitutes a valuable opportunity to understand a machine's performance from a broader perspective, trying to discover new patterns and specific work behaviours based on the analysis of its various parameters. The discovered dependencies can be used to improve the effectiveness and increase safety by indicating the activities and states generating the highest load for the machine. As a result of the acquired knowledge, the company may undertake real changes of unfavourable work parameters and, therefore, obtain notable benefits such as reduction of fuel consumption, an extension of the machine's operating time, or minimisation of extreme and dangerous states of the machine's operation.

Increasing expectations of the business environment, and also in the mining industry, stimulate the implementation of innovative solutions in the scope of productivity, work safety, and rational use of assets. An additional factor determining the searching and development of innovative methods is the fact that heavy machinery is characterised by considerable complexity of operating parameters, often imposed by the highly demanding conditions under which it operates. These aspects result in the fact that often defining the cause of an undesirable event or explicitly indicating patterns in the specifics of the machine operation depends on many variables and conditions and can be a challenging task. The existing techniques and analytical software currently used in data analysis allow comprehensive analyses to be conducted, enabling the specificity and complexity of the analysed machine operation data to be taken into account. In view of these considerations, it has been concluded that complementing the existing methods with additional assumptions reflecting the specificity of machine operation may contribute to more accurate and promising analysis results.

One of the interesting directions of machinery data analysis is its activity recognition based on raw sensor data. Various techniques can be used in the activity recognition task, especially from a common data mining task, namely clustering. This task aims to identify groups of observations with characteristics that are as similar as possible within a particular cluster and dissimilar across different clusters. Clustering is often applied in machine learning, social network analysis, geosciences, decision making, document retrieval, and image segmentation [1,2]. In scientific literature, several clustering approaches have been developed: partitional, hierarchical or density-based methods, and other paradigms such as nearest neighbour-based clustering, fuzzy clustering, and neural network-based clustering [2,3].

The paper is devoted to enhancing activity recognition of the wheel loader operation with data from the Global Positioning System (GPS). We have assumed that broadening the analysis of the loader's operating states with the location data may improve the quality of the information on wheel loader operation and its efficiency. In this research, we aimed to fill the gap in the field of activity recognition of mining machinery based on sensor and GPS data by using clustering and classification techniques. Results of the analysis provide additional knowledge about the operating characteristics of the equipment and may contribute to a better understanding of the specific operation of the equipment and the effectiveness of the operations carried out. The main contribution of our work is a demonstration that localisation data should be taken into consideration in activities' analysis.

In our approach for activity recognition, we used density-based clustering with the DBSCAN algorithm. The variables included in the analysis are related mainly to engine performance, driving system, bucket statuses that were selected from a broader set based on an assessment of data completeness, and principal component analysis (PCA). Discovered clusters were named based on statistical analysis results and, subsequently, named activities were analysed regarding identified working areas (mine base and open pit). We also prepared a description of defined activities as a rule set to enable labelling the raw sensor data, e.g., for process monitoring needs, using tree-based classifiers.

Obtained results confirmed a statistically significant difference in distributions of variables characterising defined activities in the identified areas. Thus, during process monitoring and activities analysis, these findings can bring more in-depth knowledge about machinery operation that can be helpful in the decision-making process regarding equipment management.

The paper is organised as follows: Section 2 provides an overview of the most important scientific literature in the field of machine condition recognition and data analysis with the use of GPS data. In Section 3, we introduce the dataset used for the analysis and the methodology applied to assign data to operation areas and a brief description of data mining methods used in our research. The results of the conducted analyses are presented in Section 4, along with a discussion. Finally, in Section 5, we formulate concluding remarks.

## 2. Related Works

The application of analytical techniques based on data obtained from sensors in state, activities, and operating conditions of machinery identification is commonly discussed in the literature. Especially in recent years, numerous publications have been published that deal with the topic of evaluating and monitoring machines based on their performance data. The literature analysis indicated examples of practical implementations of analytical methods in the field of monitoring parameters of various machines [4,5], including machines working in the mining industry [6]. In addition, some examples of the application of data mining techniques for monitoring working conditions in mining areas to improve safety by identifying microseismic events (based on classification techniques) can be found in [7]. Other work proposes monitoring and evaluating mine climate conditions based on sensor data to predict potential hazards [8]. In the more detailed view, an example of applying an analytical approach to identify different types of activities during construction equipment operation is presented in [9]. The authors use a data fusion and machine learning algorithm approach applied to audio and kinematic data to monitor the operations of single pieces of equipment. In [10], the authors present the application of activity recognition to the analysis of the construction equipment operation illustrated by the example of a front-end loader based on supervised machine learning classifiers in [10]. An interesting application of an analytical approach to operational data of construction equipment is presented in the publication [11]. The authors use a recurrent neural network to recognise the activities of an excavator and a front-end loader based on synthetic data. Another example illustrating the use of operational data in the field of analysis and activity recognition for construction machinery is the reference publication [12]. Wu C. et al., carried out analyses of data extracted from a smartphone in terms of a behaviour model for operations and to identify patterns of agricultural machinery [13]. In the paper [14], the authors analysed the operation of an LHD (load, haul, dump) machine from an underground copper ore mine based on statistical analysis and temperature data in the context of maintenance activities. Langroodi et al., have proposed in their work a new Fractional Random Forest machine learning method that can be applied to machine activity recognition based on a limited dataset [15]. This method has been applied to data for excavators and rollers.

In the literature, there are also current examples of analyses based on sensor data acquired from the loader and complemented by the specifics of the device operator's work. The authors of [16] addressed the issue of analysing the characteristics of the machine operation in different working conditions for a wheel loader considering the driver's influence. The authors analysed the operation of the device based on the data of the boom head cylinder pressure and proposed a method for evaluation of the difficulty level of the operating conditions based on the radar chart and clustering analysis. The analysis of braking strategies by deep learning methods for an automatic wheel loader based on driving data and operator work specifics was undertaken in [17]. Other examples of using operational data for a wheel loader machine to optimise its performance are given in [18]. The problem of finding the optimum for the wheel loader work cycle in terms of fuel efficiency was discussed in the article [19]. The paper presented an algorithm for improving fuel efficiency and productivity of a loader, which can be applied in the operator work support or system optimisation and concept selection for loaders.

The subject of clustering sensor data that characterise the operational states of machines has been addressed in the literature in various papers. J. Amutha et al. conducted a comprehensive literature review of data clustering methods and algorithms, including classical optimisation and machine learning techniques [20]. One can find other references presenting literature research on clustering methods in the field of sensor data analysis in [2,3,21]. The area of machine operation data clustering is widely addressed in the literature, especially in the context of equipment condition diagnostics based on various techniques such as correlation-based clustering [22], clustering maintenance records of excavator buckets [23], improved K-means clustering for detecting power transformer abnormal state [24], mean shift clustering in anomaly detection for machine tools [25], and k-means clustering algorithms for mining shovel failure prognostics [26]. An approach based on methods such as time series segmentation, clustering, and classification for analysing the operating states of wheel loader machines to detect anomalies in the time series dataset was proposed in [27].

In this work, we aimed to enrich the standard analysis of the obtained clusters with location data from the GSP signal, which, combined with the map of the working area, allowed us to distinguish distinct sub-areas within the territory where the machine operates. Distinguished areas allowed for a more detailed analysis and an indication of more specific subclusters concerning the operating site. Numerous examples in the literature review use location-based data, mainly for road and pedestrian traffic or travel behaviour detection [1,28,29]. Cheng et al., highlight the benefits of using location data to understand worksite operations better and to analyse the productivity of machines working in the field effectively [30]. An example of using GPS data to analyse and infer the working of construction equipment is in [31]. This paper proposes a method to identify workstations for a group of heavy vehicles, including wheel loaders, excavators and dump trucks based on GPS data. The proposed method determines the locations of different types of workstations with a probability density function. The paper [32] presents case studies using GPS data to analyse construction equipment performance, the job site layout and to visualise the GPS data through a developed user interface.

The subject of sensor data analysis for industrial machines is an issue that has been frequently addressed in scientific publications in recent years concerning a wide range of applications, from issues related to improving efficiency, determining patterns of machine operation, or investigating anomalous states in order to extend machine lifetime. The research examples cited above are mainly concerned with the implementation of data mining techniques for detecting operating states and monitoring machine performance from sensor data, visual data, and audio data. Other examples of cited publications focus on the use of machine data to support the work of operators. In the context of machine operation analysis, the primary and common application of GPS location data is a route optimisation and ongoing fleet monitoring in management systems. Known methods of detecting machine conditions do not take into account information about the device's current location during operation in terms of more accurate detection of activities. On the other hand, data on current operating parameters are used mainly to assess the effectiveness of work or monitor abnormal operating conditions. The combination in the analysis of both the data on the defined area in which the device is located and selected parameters of the machine operation allows for developing specific rules to identify the machine's state during the working process.

The challenge in evaluating machine performance remains to define the value-added activities and separate the nonvalue-added activities. The proposed approach, using the definition of activities in the context of the current location, provides an advantage in the analysis of machine performance by allowing easy identification of workspace-specific activities in addition to the main activities. The ability to easily determine the current operating status can provide a basis for more detailed analyses of machine operation from a process analysis point of view.

## 3. Materials and Methods

In this section, we briefly present methods and materials used in our research, namely: the analysed data set, marking of location perspective in data, as well as selected data mining techniques applied in our work.

### *3.1. Data Set*

The original raw data set includes 9,810,934 observations and 432 variables covering six months of wheel loader operation; however, due to quality issues (observations only with timestamps), we had to select the best samples for further analysis (Figure 1).

For the best candidate of the sample, we selected observations from one week (156,863 observations). In the sample, we identified main groups of variables related, among others, to engine characteristics (e.g., speed, fuel pressure, fuel temperature, crankcase pressure), driving system (e.g., speed of the vehicle, acceleration pedal position, parking brake switch), bucket statuses and other variables (e.g., GPS position). From the original variable set, 208 variables were excluded from further analysis (containing 100% missing values).

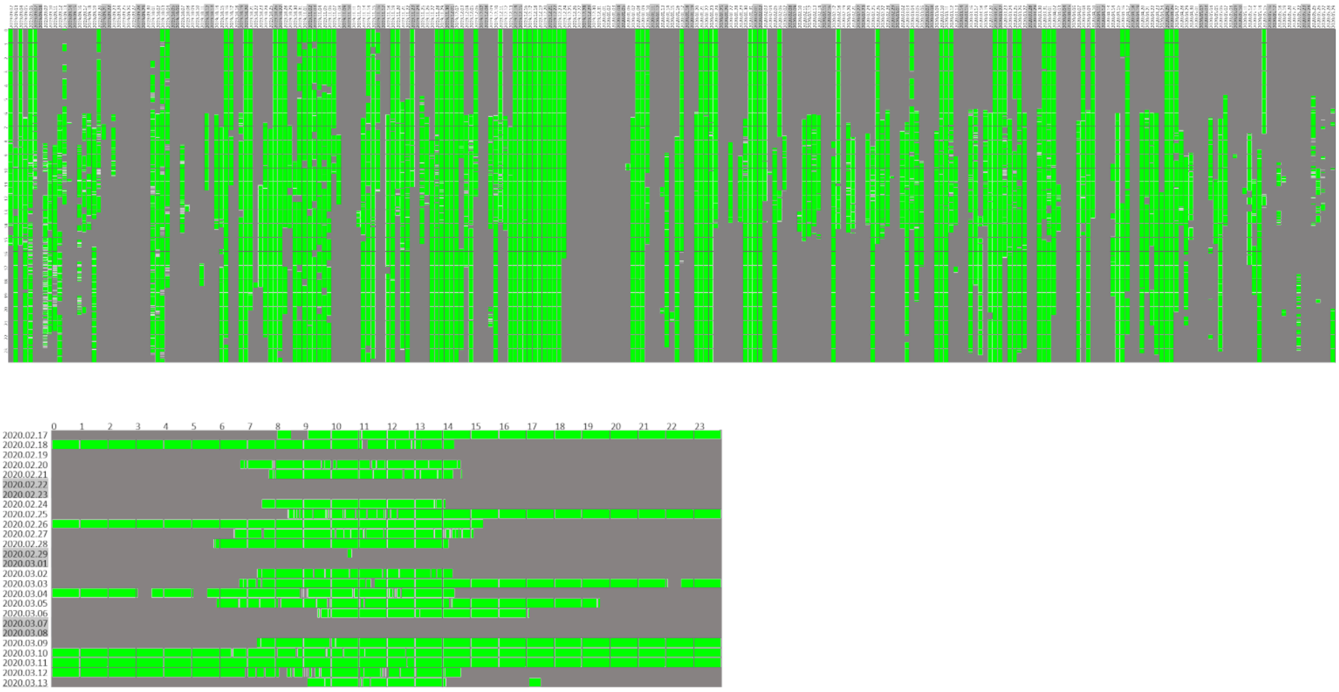


**Figure 1.** Data quality scan visualisation (grey colour denotes missing data).

The principal component analysis (PCA) method was used to select the appropriate target set of variables for analysis. PCA is a multivariate statistical technique for extracting information from a data set and representing it as new orthogonal variables, referred to as principal components. PCA is used to analyse a table of inter-correlated variables. The main premise of the PCA method is to reduce the dimension of the data used in the analysis, assuming that the reduction of multidimensionality is carried out maintaining most of the variability in the data set. This process is often carried out as a preliminary step before proceeding to further analyses [33,34].

After verification of 208 variables (some of them had a high rate of NAs), 68 variables were examined. Based on their dependency analysis and PCA analysis, we chose 19 variables with 47,093 observations as the final set used for activity recognition. Selected variables (due to IP requirements) are presented in Table 1 and in Figure 2.

**Table 1.** Fragment of variables list used in machine activity recognition.

| Item | Variable Name | Unit |
|---|---|---|
| 1 | wheel-based vehicle speed | [km/h] |
| 2 | engine speed | [rpm] |
| 3 | accelerator pedal position | [%] |
| 4 | engine fuel rate | [l/h] |
| 5 | engine fuel temperature | [ºC] |
| 6 | lifting bucket state | [-] |
| 7 | current weight | [kg] |
| 8 | parking brake switch | [-] |
| 9 | engine shutdown | [-] |

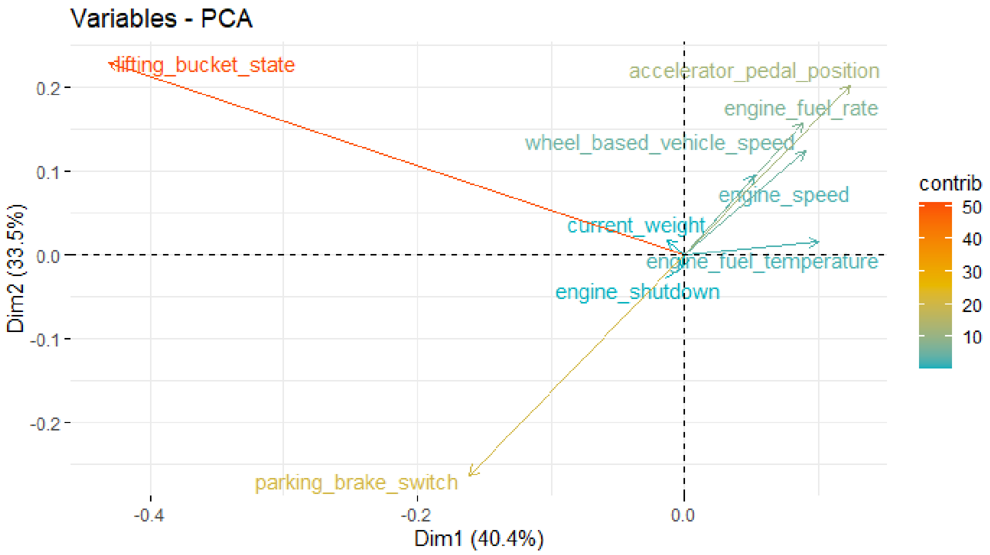


**Figure 2.** Visualisation of PCA analysis (selected variables).

Additionally, we created a variable denoting the area of wheel loader operation, based on GPS data, with the approach described in the next section.

### 3.2. *Markings of Location Perspective in Data*

The assignment of data to the working areas based on the records of GPS coordinates was performed using the PNPOLY (Point Inclusion in Polygon Test) method. The method is based on leading a ray horizontally from the tested point and then switching the in/out status at each polygon edge encountered. An odd number of intersections indicates the location of the test point inside the polygon.

The inpoly function for the R language [35], which implements the PNPOLY method, was developed based on the program code included in the work [36]. There is a point.in.polygon function in R (in the *sp* package), but it is not suitable for use in pipe mode. This was the main reason for creating the proprietary inpoly function. The function returns the true value if the point lies inside the polygon, which is set by an additional data frame with successive coordinates of forming points.

The data analysed includes the geographic coordinates of the current machine position (gps_position). The format of this record requires an appropriate transformation for the extraction of the desired coordinates.

The extraction of longitude and latitude comprised of the following operations were performed in R:

- extracting a string containing GPS coordinates,
- creating a list with three separated coordinates,
- conversion from a list to a vector,
- converting data type to numeric.

After the operations mentioned above, the GPS coordinates and the given area become input parameters to the function inpoly that returns information about belonging to the area.

Considering the geographical location of a mine and analysing the machine movement, we distinguished the mine base (1) and two mine exploitation areas (2,3). Other locations were marked as 0. The mutual position of the areas is shown in Figure 3, as well as the points describing the areas used in the inpoly function to assign the machine's current position to the working areas.

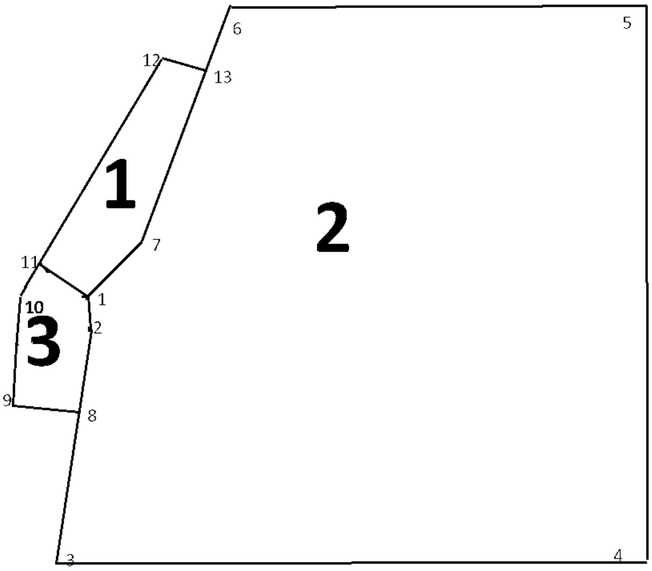


**Figure 3.** Wheel loader operation areas (1—base, 2—mine, 3—minor mine).

The distribution of identified working areas in the sample data set is presented in Table 2.

**Table 2.** Distribution of working areas in data set from one week.

| Working Area | Frequency | Percentage (%) | Cumulative Frequency | Cumulative Percentage (%) |
|---|---|---|---|---|
| 0 | 1463 | 0.9 | 1463 | 0.9 |
| 1 | 110,933 | 70.7 | 112,396 | 71.7 |
| 2 | 30,592 | 19.5 | 142,988 | 91.2 |
| 3 | 175 | 0.1 | 143,163 | 91.3 |
| NAs | 13,700 | 8.7 | 156,863 | 100.0 |

The main operation area of the analysed machine is the mine base (70% of observations). One can notice that area no 2 (the exact open pit mine) occurs almost in 20% of observations. Area no 3 is underrepresented; thus, we decided to join this data with area no 2. In the case of 1% of observations, the machine worked out of the mine area; for about 9% of observations, we could not identify the working area due to missing data.

Identified areas will be used for the enhancement of machinery activity recognition, presented in Section 4.

### 3.3. *Selected Data Mining Techniques for Clustering and Results Explanation*

Currently, for industrial enterprises, Internet of Things (IoT) systems are a primary source of vast data gathered during operations, allowing for insight into the process and its comprehensive analysis. The collected data can be used in the knowledge discovery process, which can be automated with appropriate data mining methods [37].

In general, data mining approaches can be divided into supervised, unsupervised, or reinforcement learning [38]. Among different unsupervised learning methods, clustering is one of the most popular tasks, which has the advantage of uncovering hidden, often unexpected groups in a data set without any prior knowledge or input about the partition. Cluster analysis at the stage of exploratory data analysis allows a better understanding or summary of the data [2].

Clustering is widely implemented in activity recognition tasks [21]. The primary purpose of a clustering task is to divide instances into different groups, determined by their similarity [37]. Researchers have proposed many clustering methods, such as partitioning clustering (e.g., k-mean, k-medoids, PAM, CLARANS, and CLARA algorithm), hierarchical clustering (e.g., CURE, BIRCH, and ROCK algorithm), density-based clustering (e.g., DBSCAN, OPTICS, and DBCLASD), grid-based clustering (e.g., STRING) or model-based (e.g., Self-Organized Map algorithm) [38,39]. Clustering techniques differ in several assumptions, such as the procedure for calculating similarity within and between clusters, setting the threshold for identifying cluster elements, or the methodology for grouping objects belonging to different degrees into one or more clusters [40]. Depending on the conditions and type of data, different clustering analysis algorithms result in different clusters. A comparison of the most popular clustering algorithms with the main assumptions and limitations of these methods can be found in [41,42].

Considering the characteristics of our data set, we selected the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm because of its main advantages, such as applicability to multidimensional data, robustness to noisy data, ability to discover clusters of different shapes and sizes, and the lack of requirement to determine the number of clusters in advance [40].

DBSCAN is recommended as a leading algorithm for clustering high dimensional data and is the most commonly used density-based algorithm [40].

The main assumption of density clustering approaches is determining and separating high- and low-density regions [43]. The DBSCAN algorithm was proposed by Ester, M. et al., in response to observed challenges for clustering algorithms, such as the ability to define clusters of arbitrary shape, achieving high efficiency with large data sets, and requiring little domain knowledge to determine input parameters [44]. The authors defined a cluster as a set of connected density points that is maximal with respect to density reachability. The algorithm employs two main input parameters: the neighbourhood radius—Eps (neighbourhood of a point) and the minimum number of points in the neighbourhood—MinPts to determine a density threshold; based on that, the data are aggregated into clusters. The DBSCAN algorithm with correctly specified parameters can produce clusters of any shape [45]. For a point in a cluster, a neighbourhood of a specified radius must contain at least a minimum number of points, i.e., the density must exceed the mentioned threshold, and the chosen distance function indicates the shape of the neighbourhood. There are two types of points within a cluster: core points, which are located inside the cluster, and border points, which are points on the border of the cluster [44]. The DBSCAN algorithm is also used to efficiently discover noise in the data, which is defined as a set of points in the database that do not belong to any of the clusters [44,46], so-called outliers.

In an analysis of discovered clusters, as an explanation of results, we used decision trees which enabled the formulation of rules for assigning observations to proper clusters in the analysed area. A method for solving classification problems such as decision trees is characterised primarily by its intuitiveness, simplicity of interpretation of results, and reasonable accuracy [47], enabling prediction of categorical outputs with tree or rule structures. Trees are graphical representations of the decision-making process in which the structure consists of internal nodes representing attribute tests (decision nodes) and leaf nodes corresponding to predicted class labels [48,49]. There are many algorithms for classification using decision trees, the most popular of which are: CART [50], CHAID [51], QUEST [52], and C5.0 [53,54] as a successor of the C4.5 algorithm [55], which is based on ID3 [56]. The main improvement of the C5.0 is boosting technology, allowing the addition of each sample weight to determine its importance [53].

In explanation of clustering results, we tested CART and C5.0 algorithms (with various settings). The CART algorithm is the most commonly used decision-tree technique [50] which allows the detection of structures even in complex data and the construction of accurate and reliable models [57]. Based on labelled data, CART trees enable the discovery of rules that can be used to classify new data. This method is an example of binary recursive partitioning using the GINI index. Binary partitioning can be performed repeatedly, and instances in a node can only be divided into two groups [58]. The second algorithm used to describe the obtained clusters is the C5.0 algorithm (with the rule model option). The C5.0 algorithm employs an information gain rate to build a decision tree. Information gain is determined for each feature in the data set to identify the best split points [54,59].

In order to increase the accuracy of a classifier like a classification tree, pruning techniques aim to reduce the size of the tree by eliminating the less frequent sections of the tree that are considered non-critical and have low classification ability. Pruning reduces the complexity of the tree and consequently improves the classification performance by preventing overfitting [60]. In the case of used algorithms, we tested the following parameters for pruning: cp—complexity parameter (CART) and Min Cases parameter as the smallest number of samples that must be put in at least two of the splits (C5.0). The complexity parameter (cp) denotes the minimum improvement in the model needed at each node. It is based on the cost complexity of the model. The cp parameter helps speed up the search for splits because it can identify splits that do not meet this criterion and prune them before the tree becomes too wide [61].

## 4. Results and Discussion

Our data set with 18 variables (the time variable was omitted in clustering task) was analysed with R library *dbscan* [62]. Since the DBSCAN algorithm is sensitive to Eps and MinPts settings, we ran multiple calculations with various values of parameters. The Eps parameter setting started from the analysis of the KNN plot for k = 18 (number of dimensions) (Figure 4). At first, we tested a value that corresponded to the curve's inflexion point (that is 0.5). In the beginning, we changed the value of the Eps parameter with step 0.05. When the number of clusters decreased, we adaptively decreased the Eps value. In the case of MinPts value, firstly, we assumed a number of points equal to a number of dimensions +1 (19); however, we obtained many outliers. We repeated the calculations by doubling and tripling this number. Finally, we chose MinPts as 57 points and Eps value as 0.6.

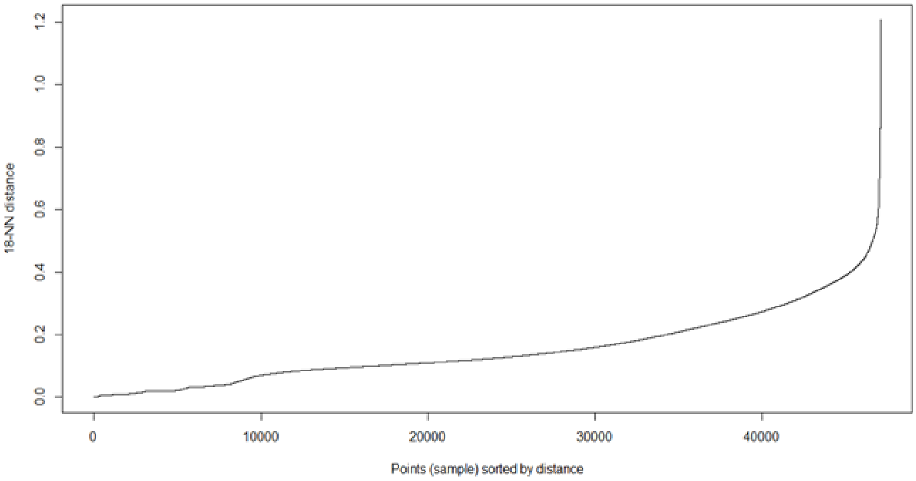


**Figure 4.** KNN plot for k = 18.

The DBSCAN algorithm, with defined parameters as above, found five clusters in our data set (Figure 5). Black dots in the figure denote outliers (not clustered observations).

The distribution of observations among discovered clusters is presented in Figure 6. Cluster 0 contains outliers—only 145 observations were collected in this cluster (0.3% of observations). The largest, cluster 1, contains almost 52% of observations (24,338), and cluster 3 contains 32% observations (15,303). Smaller clusters, no 2 and no 4, contain 8% (3763) and almost 7% of observations (3251), respectively. Finally, cluster 5 contains 0.6% of observations.

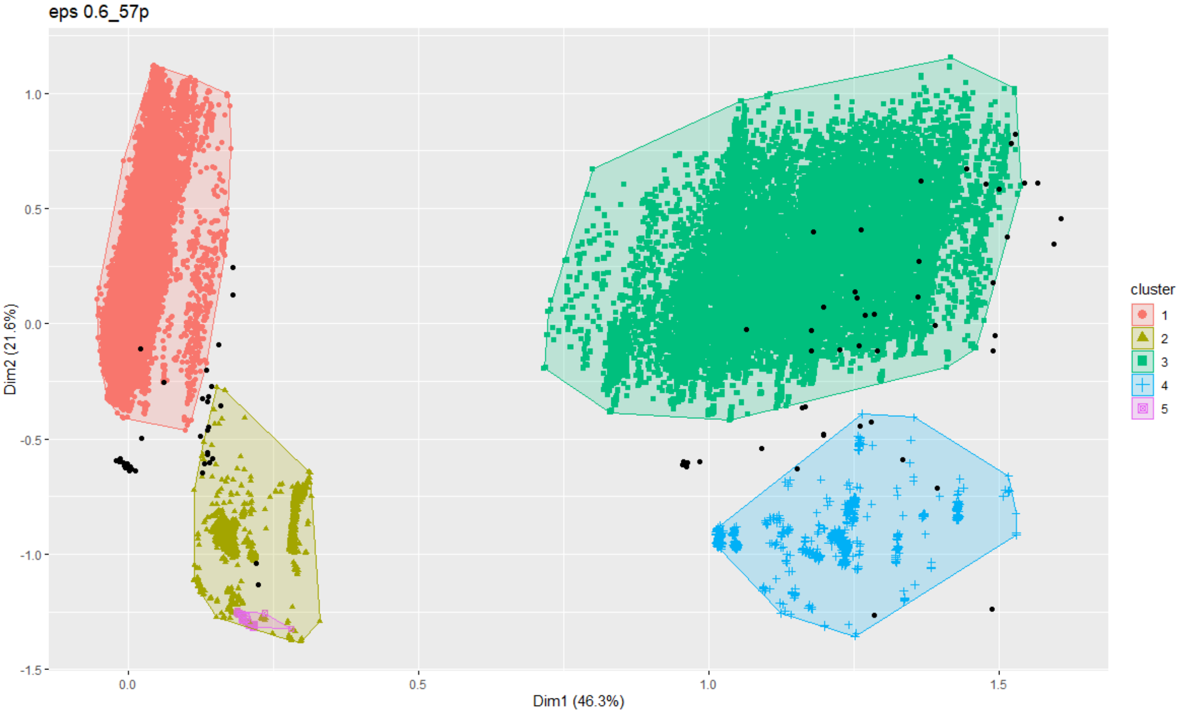


**Figure 5.** Clustering results (with outliers marked as black dots).

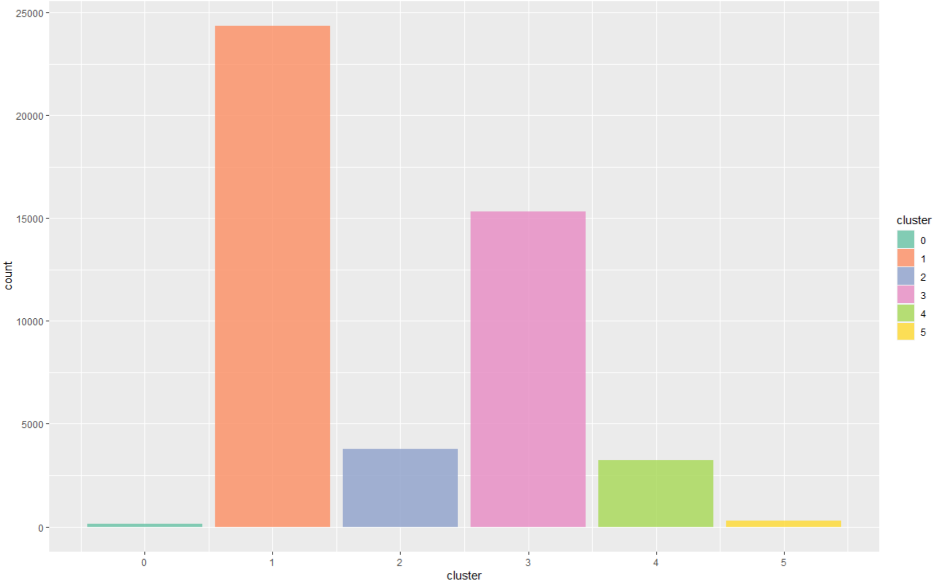


**Figure 6.** Distribution of observations in clusters.

Selected statistics of the discovered clusters are presented in Table 3.

Based on presented statistics, the following description of activities was prepared:

- Cluster 1—Moving/travelling;
- Cluster 2—Stoppage with engine ON;
- Cluster 3—Normal work with loading;
- Cluster 4—Stoppage with loading;
- Cluster 5—Engine OFF.

Distributions of discovered clusters in the identified working areas are presented in Figure 7.

**Table 3.** Basic statistics of the discovered clusters.

| Variable Name | Cluster 1 | | | Cluster 2 | | | Cluster 3 | | | Cluster 4 | | | Cluster 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Min | Median | Max | Min | Median | Max | Min | Median | Max | Min | Median | Max | Min | Median | Max |
| wheel-based vehicle speed | 0 | 5 | 21 | 0 | 0 | 0 | 0 | 4 | 20 | 0 | 0 | 0 | 0 | 0 | 0 |
| engine speed | 621 | 1128 | 2015 | 0 | 791 | 1495 | 645 | 1117 | 1699 | 0 | 793 | 1606 | 0 | 0 | 52 |
| accelerator pedal position | 0 | 29 | 72 | 0 | 0 | 57 | 0 | 29 | 72 | 0 | 0 | 53 | 0 | 0 | 0 |
| engine fuel rate | 0 | 15 | 55 | 0 | 3 | 27 | 0 | 15 | 55 | 0 | 3 | 24 | 0 | 0 | 0 |
| engine fuel temperature | 0 | 21 | 26 | 0 | 21 | 28 | 0 | 19 | 23 | 0 | 6 | 25 | 0 | 16 | 17 |
| lifting bucket state | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| current weight | 0 | 0 | 0 | 0 | 0 | 0 | −24,608 | 1072 | 29,353 | −4210 | −2233 | 6213 | 0 | 0 | 0 |
| parking brake switch | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| shutdown engine | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |

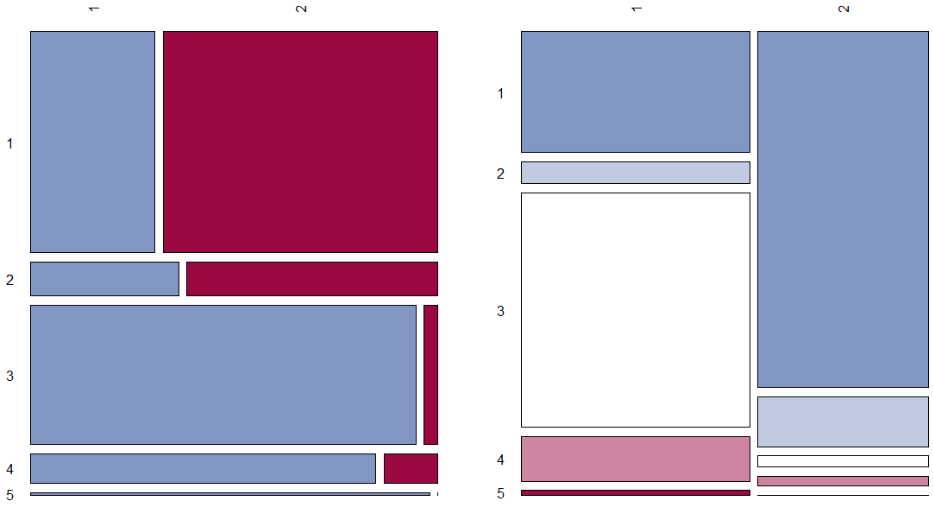


**Figure 7.** Distribution of clusters and observations in identified working areas ($x$ axis = area, $y$ axis = cluster).

As can be observed, the main activity in the mine base area (area no 1) is regarding Normal work with loading, while the dominant activity in the open pit area is Moving (Figure 7—right figure). Thus, we can conclude that the main area of efficient work is the mine base area, and usage of wheel loader in the open pit has a rather auxiliary character, e.g., raw material moving (distribution of activity no 2 and 4—Stoppage with/or without loading). Activity Engine OFF is mainly observed in the mine base area.

In further investigations, we analysed whether there was a difference between machine behaviour in areas of work. In other words, whether a place of work influences the characteristics of the discovered activity. We performed a statistical analysis of distributions of variables in each cluster versus working area.

In Table 4, we present the statistical analysis results regarding the comparison of selected numerical variable distributions in the working areas for each activity. Since none of the numerical variables in the data set is normally distributed, to test the differences within groups, we chose the non-parametric Wilcoxon test. Colours in the table indicate statistically significant test values.

Comparison of variable boxplots in each cluster and area are presented in Figure 8.

Statistical analysis of selected variables in each cluster confirmed that there is a statistically significant difference in their distributions in the defined areas for most variables. Thus, during the process monitoring and activities analysis, these differences should be taken into consideration.

**Table 4.** Results of the Wilcoxon test for group comparison.

| Variable Name | | Wheel-Based Vehicle Speed | Engine Speed | Accelerator Pedal Position | Engine Fuel Rate | Engine Fuel Temperature | Current Weight |
|---|---|---|---|---|---|---|---|
| Cluster 1 | median 1 | 5.0 | 1120 | 29.6 | 16.6 | 22.0 | 0 |
| | median 2 | 5.9 | 1131 | 28.8 | 14.6 | 21.0 | 0 |
| | *p* value | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | - |
| Cluster 2 | median 1 | 0 | 800.5 | 0.8 | 5.4 | 5.0 | 0 |
| | median 2 | 0 | 651.0 | 0 | 2.6 | 21.0 | 0 |
| | *p*-value | - | <0.01 | <0.01 | <0.01 | <0.01 | - |
| Cluster 3 | median 1 | 4.7 | 1122 | 29.6 | 15.6 | 19.0 | 1127.5 |
| | median 2 | 1.6 | 915 | 10.0 | 5.9 | 21.0 | 665 |
| | *p* value | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | <0.05 |
| Cluster 4 | median 1 | 0 | 798.2 | 0 | 4.6 | 5 | −2236 |
| | median 2 | 0 | 649.9 | 0 | 2.6 | 22 | 5996 |
| | *p*-value | - | <0.01 | >0.05 | <0.01 | <0.01 | <0.01 |
| Cluster 5 | median 1 | 0 | 0 | 0 | 0 | 16.0 | 0 |
| | median 2 | 0 | 0 | 0 | 0 | - | - |
| | *p*-value | - | - | - | - | - | - |

For a better description of discovered clusters and considering revealed differences, we attempted to discover rules enabling raw data labelling. In this task, we examined selected classifiers with CART and C5.0 algorithms.

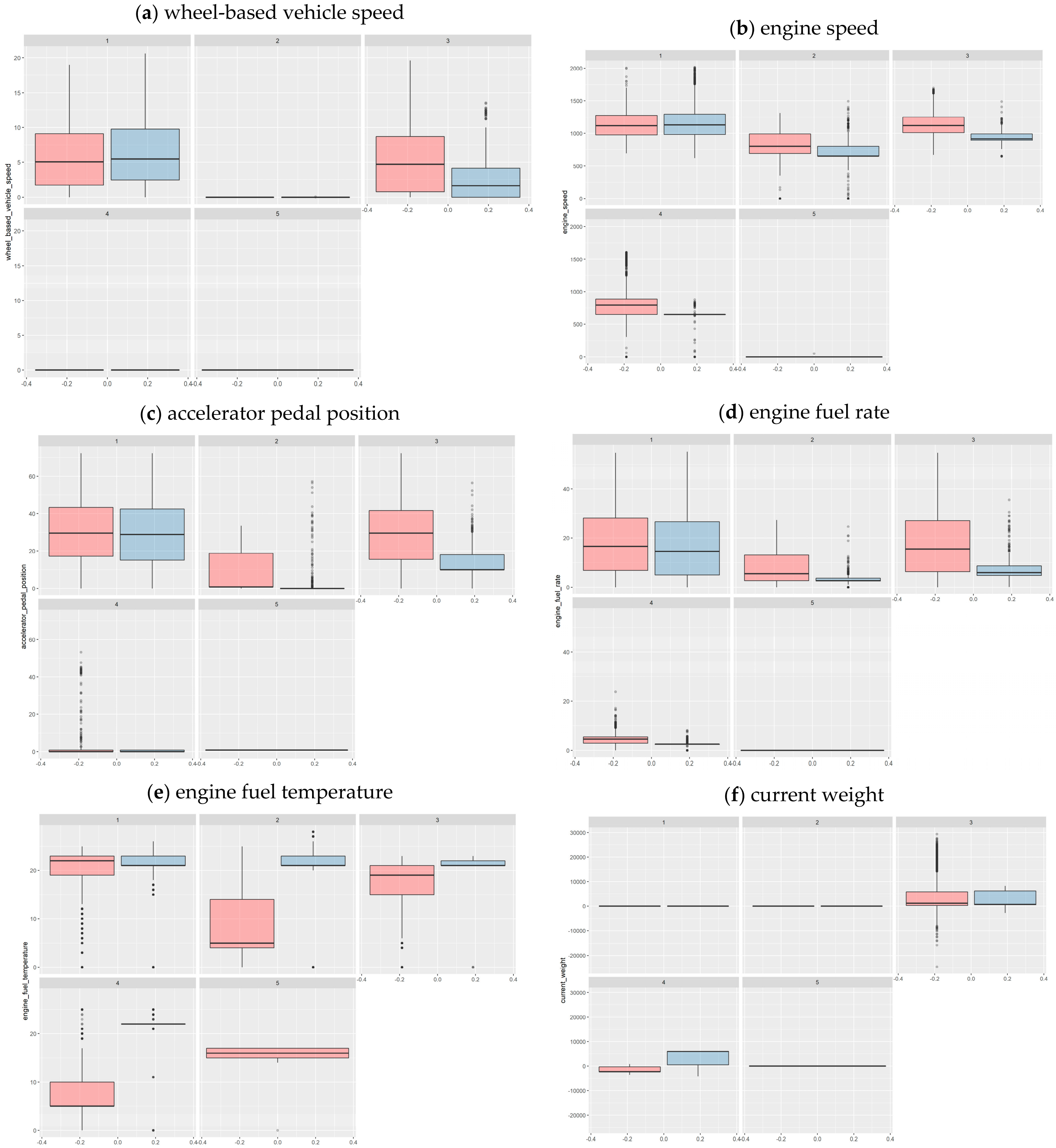
Firstly, we divided the data set randomly (without data of cluster 0, denoting outliers) into train and test subsets with the proportion of 80% (37,558 observations) and 20% (9390 observations), respectively. Subsequently, we trained classifiers and checked them on the test data set. In Tables 5 and 6, we present the main parameters and the obtained results (the most similar results obtained for the two classification algorithms are marked with bolded font).

**Table 5.** Parameters of CART classifiers.

| No of Classifier | Cp | Tree Size (Nsplits) | Accuracy on Train Dataset | Accuracy on Test Dataset |
|---|---|---|---|---|
| 1 | 0.000250 | 115 | 0.9098 | 0.9085 |
| 2 | 0.000375 | 87 | 0.9045 | 0.9043 |
| 3 | 0.000600 | 34 | 0.8890 | 0.8882 |
| 4 | 0.000750 | 30 | 0.8874 | 0.8858 |
| 5 | 0.001000 | 27 | 0.8854 | 0.8851 |
| **6** | **0.002000** | **17** | **0.8758** | **0.8765** |
| 7 | 0.002500 | 9 | 0.8602 | 0.8608 |

The simplest classification tree (no 7) obtained using the CART algorithm is presented in Figure 9.

The performed tests have shown that both algorithms have built the tree models enabling accurate prediction of the tested class: wheel loader activity in the working area. The most valuable rules, due to their confidence, obtained from comparable-in-size classifiers: CART with 18 rules (no 6) and C5.0 with 17 rules (no 5), are presented in Table 7.


(**a**) wheel-based vehicle speed

(**b**) engine speed

(**c**) accelerator pedal position

(**d**) engine fuel rate

(**e**) engine fuel temperature

(**f**) current weight

**Figure 8.** Boxplots of selected variables (**a**–**f**) in the perspective of clusters and working areas. (Red colour denotes area 1, blue colour denotes area 2).

According to Table 7, we can observe that obtained rules are characterised by a high or very high confidence level (0.73–1.0). Rules responsible for classifying machine characteristics as a particular activity are very similar for both applied classifiers; however, in some cases are different in the level of variable occurrence in the tree structure.

**Table 6.** Parameters of C5.0 classifiers.

| No of Classifier | Min Cases | Numbers of Rules | Accuracy on Train Dataset | Accuracy on Test Dataset |
|---|---|---|---|---|
| 1 | 50 | 41 | 0.9040 | 0.8954 |
| 2 | 100 | 22 | 0.8870 | 0.8825 |
| 3 | 150 | 17 | 0.8830 | 0.8797 |
| 4 | 200 | 16 | 0.8800 | 0.8778 |
| **5** | **250** | **17** | **0.8770** | **0.8729** |
| 6 | 300 | 13 | 0.8760 | 0.8732 |
| 7 | 600 | 9 | 0.8480 | 0.8436 |

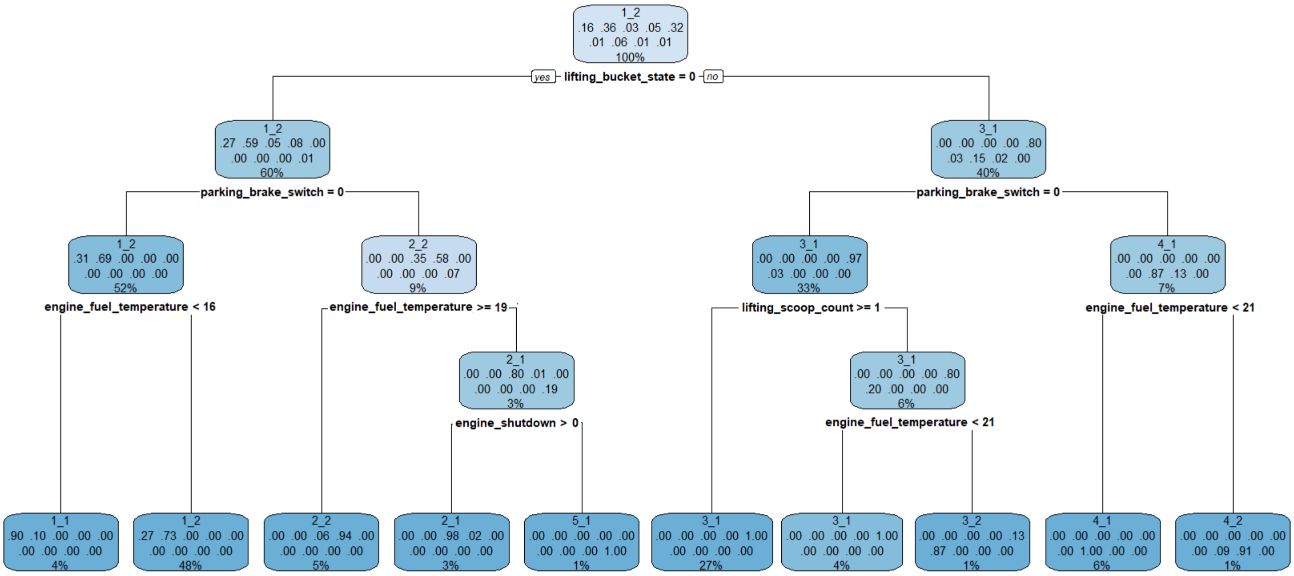


**Figure 9.** Classification tree CART.

The key machine operating parameters for assigning an observation to activity 1, identified similarly by both types of classifiers, are primarily lifting_bucket_state and parking_brake_switch (less than 0.5; in practice, only 0 or 1 are possible). Based on engine_fuel_temperature, classification to area 1 is made for temperatures less than 15/16 °C and area 2 for temperatures greater than 15/16 °C. However, in the C5.0 tree, additional rules have been formulated related to vehicle speed. The rules determining assignment to activity 2 are based on the parameters such as parking_brake_switch > 0 and lifting_bucket_state <= 0. Similar to activity 1, depending on engine_fuel_temperature, observations are classified into area 1 (temperature less than 19 °C) or area 2 otherwise. Classification to activity 3 is based, among other criteria, on the parking_brake_switch variable, whose value, in this case, equals 0. Depending on the values taken by the lifting_scoop_weight variable in the C5.0 tree, the observations are classified in area 1 (lifting_scoop_weight > 0) or area 2 when this variable takes a value less than 0. In the CART tree, another variable related to lifting is taken into consideration—lifting_scoop_weight. Assignment to activity 4 is based largely on 3 variables, parking_brake_switch, lifting_bucket_state, and engine_fuel_temperature, where the first two variables should be greater than 0, while temperature greater than 20 °C determines the classification of the observation as activity 4 in area 2, while less than 20 °C is in area 1. The rule for classifying into cluster 5 in the C5.0 tree is mainly based on variable engine_shutdown greater than 0. CART tree extends rules with additional variables. The general conclusion can be made, comparing similarities and dissimilarities of variables used in tree classifiers, that indication of the operation of wheel loader in the open pit area is related to the greater temperature of engine fuel, which can be explained by more inconvenient conditions of work as downhill rides and elevation differences that impact loading of the engine.

**Table 7.** Rules describing wheel loader activities in the working areas.

| Activity | CART | | C5.0 | |
|---|---|---|---|---|
| | Rule | Confidence | Rule | Confidence |
| 1_1 | lifting_bucket_state = 0<br>parking_brake_switch = 0<br>engine_fuel_temperature < 16 | 0.897 | engine_fuel_temperature <= 15<br>parking_brake_switch <= 0<br>lifting_bucket_state <= 0 | 0.904 |
| 1_2 | lifting_bucket_state = 0<br>parking_brake_switch = 0<br>engine_fuel_temperature >= 16 | 0.735 | wheel_based_vehicle_speed > 9.85<br>engine_fuel_temperature> 15<br>lifting_bucket_state <= 0 | 0.874 |
| 2_1 | lifting_bucket_state = 0<br>parking_brake_switch > 0<br>engine_fuel_temperature < 19<br>engine_shutdown = 0 | 0.981 | engine_fuel_temperature<= 19<br>parking_brake_switch > 0<br>engine_shutdown <= 0<br>lifting_bucket_state <= 0 | 0.983 |
| 2_2 | lifting_bucket_state = 0<br>parking_brake_switch > 0<br>engine_fuel_temperature >= 19 | 0.938 | engine_fuel_temperature > 19<br>parking_brake_switch > 0<br>lifting_bucket_state <= 0 | 0.941 |
| 3_1 | lifting_bucket_state > 0<br>parking_brake_switch = 0<br>lifting_scoop_count >= 1 | 0.998 | engine_fuel_temperature <=20<br>lifting_scoop_weight > 0<br>parking_brake_switch <= 0 | 1.0 |
| 3_2 | lifting_bucket_state > 0<br>parking_brake_switch <= 0<br>lifting_scoop_weight < 1<br>engine_fuel_temperature >= 21 | 0.868 | engine_fuel_temperature > 20<br>lifting_scoop_weight <= 0<br>parking_brake_switch <= 0<br>lifting_bucket_state > 0 | 0.873 |
| 4_1 | lifting_bucket_state > 0<br>parking_brake_switch > 0<br>engine_fuel_temperature < 21 | 0.995 | engine_fuel_temperature <= 20<br>parking_brake_switch > 0<br>lifting_bucket_state > 0 | 0.995 |
| 4_2 | lifting_bucket_state > 0<br>parking_brake_switch > 0<br>engine_fuel_temperature >= 21 | 0.909 | engine_fuel_temperature > 20<br>parking_brake_switch > 0<br>lifting_bucket_state > 0 | 0.911 |
| 5_1 | lifting_bucket_state = 0<br>parking_brake_switch >= 0<br>engine_fuel_temperature < 19<br>engine_shutdown > 0 | 0.996 | engine_shutdown > 0 | 0.966 |

Rules describing conditions (variable values) enable data labelling considering area of operation. The labelled data can be further used for visualisation and process analysis, e.g., in the machinery monitoring systems. Such labelled activities can be further analysed, among others, with process mining techniques and used for process modelling and analysis, e.g., with process maps (Figure 10).

Analysis of activities in different areas can support understanding specific conditions of operation needed to optimise equipment efficiency and its usage. For example:

- longer travelling time in the mine area can cause a loss in the effective time of working; one can consider leaving machines in the mine area and arranging for a faster means of transport for service personnel to the machine—it can result in saving working time, increasing the use of the loader, and reducing the cost of fuel used by the loader to get to the job site.

- automatic detection of the loading activity can allow for its correlation with the fuel consumption associated with this activity, and detailed data on the weight transferred in the bucket can be the basis for determining the optimal weight that should be loaded on the bucket to minimise the cost of loading.
- the more frequent activity of loading in the base area may be a reason for purchasing an additional machine.
- the analysis of changes in the structure of activities over time may show significant changes in the use of machines that will justify (demonstrate the need of) a decision on periodic leasing (renting, etc.) of machines from external entities, which will reduce operating costs.

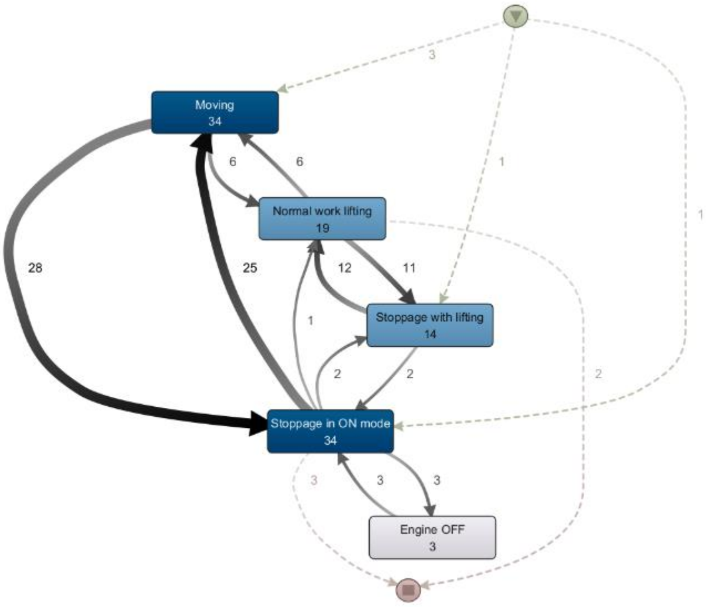


**Figure 10.** A process model of wheel loader operation.

In the further works, we plan to investigate the abovementioned issues related to a more in-depth analysis of dependencies between characterised activities and selected KPIs, as well as overall efficiency based on process-oriented analytics.

## 5. Conclusions

In this paper, an analysis of sensor data characterising the operation of a wheel loader in an open pit mine was presented. The purpose of the analysis was to effectively identify machine activities based on a real operational dataset. A subset of the entire dataset was considered in the study, including variables related to engine operation, driving system, bucket statuses and other variables, which have been clustered using the density method and the DBSCAN algorithm. We further extended the analysis with GPS data, which allowed us to divide the working area into subareas (mine base and open pit). Based on the statistical characteristics of the obtained clusters, we have named them, and together with the identified working areas, we have identified the statistical differences of variable distributions in clusters (activities) performed in various areas. The analysis results encouraged us to develop classifiers describing clusters in the form of rules, which can be helpful in raw data labelling in the future. In this part, we used selected classifiers based on the CART and C5.0 algorithms. As a result, we presented the most valuable rules for wheel loader activity recognition.

Obtained results showed that density clustering methods can provide an efficient multidimensional space search for compact and sensible clusters of observations in a real, noisy dataset. Moreover, the introduction to analysis of the location variable enabled us to identify the statistically significant differences in machinery operation in two defined working areas. These findings have also been positively validated by tree classifiers, with high accuracy rates on train and test datasets. From built classifiers, one can extract valuable rules (with high confidence factor), enabling definition/recognition of activity during process monitoring.

The identified rules require validation in a real-life environment and verification on-site during activity execution, which is planned in the next stage of our research. Then, after positive validation result, rules can be applied, e.g., in machinery or process monitoring systems.

The discovered dependencies can be used to improve the effectiveness and increase the safety of operation by indicating the activities and states generating the highest load for the machine. As a result of the acquired knowledge, the company may undertake real changes of unfavourable work parameters and, therefore, obtain notable benefits such as reduction of fuel consumption, an extension of the machine's operating time, or minimisation of extreme and dangerous states of the machine's operation.

Remarkably, distinguishing the states of a machine's work is a necessary condition for developing an algorithm that automatically determines the usage of a machine and calculates detailed indicators that consider the diversity of these states. The values of these parameters can be a premise in the decision-making process to change the utilisation of a machinery park in a mining company. In the presented analysis, consideration of the identified working areas allowed to prove the existing differences in the quantitative characteristics of activity concerning the place of machine operation. It should be noted that the analysed machine is one of many used in the mine, so automation of its activity recognition should be an integral part of the online analysis related to the machinery park.

In addition, the analysis conducted in the article is one of the steps in the development of a system that allows the calculation of the optimal operating conditions based on data collected from multiple machines. Processed data provide information about the different stages of production. They can be the basis for determining the real-time indicators of the effectiveness of used machinery and also can provide a basis for decisions about the timing of repairs or maintenance ahead of equipment failure. Awareness of such a need, strengthened by the results of data analysis, allows for planning the appropriate time and duration of activities related to the replacement of worn-out machine elements during scheduled downtime. This can help to avoid losses generated as a result of unplanned downtime.

The findings obtained from the analysis presented above can contribute to the knowledge base for MES (Manufacturing Execution System) systems in many areas specified by the international standardization organization MESA International (Manufacturing Enterprise Solutions Association International).

**Author Contributions:** Conceptualisation, E.B. and P.G.; methodology, E.B., M.K.; investigation, P.G., E.B. and M.K., data curation, M.K. and E.B.; writing—original draft preparation, P.G., E.B. and M.K.; writing—review and editing, P.G., E.B. and M.K.; visualisation, E.B. and M.K.; supervision, E.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by EIT Raw Materials, grant number 18327—project smartHUB—smart courier field data IIoT radio network and big data analytics.

**Data Availability Statement:** Data sharing not applicable.

**Acknowledgments:** In the main research, we used the following R libraries: DescTools, dbscan, ggplot2, rpart, and C5.0.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Zhou, X.; Gu, J.; Shen, S.; Ma, H.; Miao, F.; Zhang, H.; Gong, H. An automatic K-Means clustering algorithm of GPS data combining a novel niche genetic algorithm with noise and density. *ISPRS Int. J. Geo. Inf.* **2017**, *6*, 392. [CrossRef]

2. Bhattacharjee, P.; Mitra, P. A survey of density based clustering algorithms. *Front. Comput. Sci.* **2021**, 15. [CrossRef]
3. Ahmad, A.; Khan, S.S. Survey of State-of-the-Art Mixed Data Clustering Algorithms. *IEEE Access* **2019**, *7*, 31883–31902. [CrossRef]
4. Ingrao, C.; Evola, R.S.; Cantore, P.; De Bernardi, P.; Del Borghi, A.; Vesce, E.; Beltramo, R. The contribution of sensor-based equipment to life cycle assessment through improvement of data collection in the industry. *Environ. Impact Assess. Rev.* **2021**, *88*, 106569. [CrossRef]
5. Frieß, U.; Kolouch, M.; Friedrich, A.; Zander, A. Fuzzy-clustering of machine states for condition monitoring. *CIRP J. Manuf. Sci. Technol.* **2018**, *23*, 64–77. [CrossRef]
6. Polak, M.; Stefaniak, P.; Zimroz, R.; Wylomanska, A.; Sliwinski, P.; Andrzejewski, M. Identification of Loading Process Based on Hydraulic Pressure Signal. *Int. Multidiscip. Sci. GeoConf. SGEM Surv. Geol. Min. Ecol. Manag.* **2016**, *2*, 459–466.
7. Dong, L.; Wesseloo, J.; Potvin, Y.; Li, X. Discrimination of Mine Seismic Events and Blasts Using the Fisher Classifier, Naive Bayesian Classifier and Logistic Regression. *Rock Mech. Rock Eng.* **2016**, *49*, 183–211. [CrossRef]
8. Jha, A.; Tukkaraja, P. Monitoring and assessment of underground climatic conditions using sensors and GIS tools. *Int. J. Min. Sci. Technol.* **2020**, *30*, 495–499. [CrossRef]
9. Sherafat, B.; Rashidi, A.; Lee, Y.C.; Ahn, C.R. Automated activity recognition of construction equipment using a data fusion approach. In *Computing in Civil Engineering 2019: Data, Sensing, and Analytics*; American Society of Civil Engineers: Reston, VA, USA, 2019; pp. 1–8.
10. Akhavian, R.; Behzadan, A.H. Construction equipment activity recognition for simulation input modeling using mobile sensors and machine learning classifiers. *Adv. Eng. Inform.* **2015**, *29*, 867–877. [CrossRef]
11. Redlich, D.; Molka, T.; Gilani, W.; Blair, G.; Rashid, A. Data-Driven Process Discovery and Analysis. *Lect. Notes Bus. Inf. Process.* **2015**, *237*, 79–106.
12. Cheng, C.F.; Rashidi, A.; Davenport, M.A.; Anderson, D.V. Activity analysis of construction equipment using audio signals and support vector machines. *Autom. Constr.* **2017**, *81*, 240–253. [CrossRef]
13. Wu, C.; Chen, Z.; Wang, D.; Kou, Z.; Cai, Y.; Yang, W. Behavior modelling and sensing for machinery operations using smartphone's sensor data: A case study of forage maize sowing. *Int. J. Agric. Biol. Eng.* **2019**, *12*, 66–74. [CrossRef]
14. Stefaniak, P.; Śliwiński, P.; Poczynek, P.; Wyłomańska, A.; Zimroz, R. The automatic method of technical condition change detection for LHD machines—Engine coolant temperature analysis. *Appl. Cond. Monit.* **2019**, *15*, 54–63.
15. Langroodi, A.K.; Vahdatikhaki, F.; Doree, A. Activity recognition of construction equipment using fractional random forest. *Autom. Constr.* **2021**, *122*, 103465. [CrossRef]
16. Wang, S.; Hou, L.; Lee, J.; Bu, X. Evaluating wheel loader operating conditions based on radar chart. *Autom. Constr.* **2017**, *84*, 42–49.
17. Shi, J.; Sun, D.; Hu, M.; Liu, S.; Kan, Y.; Chen, R.; Ma, K. Prediction of brake pedal aperture for automatic wheel loader based on deep learning. *Autom. Constr.* **2020**, *119*, 1–12. [CrossRef]
18. Zhang, W.; Wang, S.; Hou, L.; Jiao, R.J. Operating data-driven inverse design optimization for product usage personalization with an application to wheel loaders. *J. Ind. Inf. Integr.* **2021**, *23*, 100212.
19. Frank, B.; Kleinert, J.; Filla, R. Optimal control of wheel loader actuators in gravel applications. *Autom. Constr.* **2018**, *91*, 1–14. [CrossRef]
20. Amutha, J.; Sharma, S.; Sharma, S.K. Strategies based on various aspects of clustering in wireless sensor networks using classical, optimization and machine learning techniques: Review, taxonomy, research findings, challenges and future directions. *Comput. Sci. Rev.* **2021**, *40*, 100376. [CrossRef]
21. Hennig, M.; Grafinger, M.; Gerhard, D.; Dumss, S.; Rosenberger, P. Comparison of time series clustering algorithms for machine state detection. *Procedia CIRP* **2020**, *93*, 1352–1357. [CrossRef]
22. Yoo, Y.J. Data-driven fault detection process using correlation based clustering. *Comput. Ind.* **2020**, *122*, 103279. [CrossRef]
23. Yang, Z.; Baraldi, P.; Zio, E. A novel method for maintenance record clustering and its application to a case study of maintenance optimization. *Reliab. Eng. Syst. Saf.* **2020**, *203*, 107103. [CrossRef]
24. Liang, X.; Wang, Y.; Li, H.; He, Y.; Zhao, Y. Power Transformer Abnormal State Recognition Model Based on Improved K-Means Clustering. In Proceedings of the 2018 IEEE Electrical Insulation Conference (EIC), San Antonio, TX, USA, 17–20 June 2018; pp. 327–330.
25. Netzer, M.; Michelberger, J.; Fleischer, J. Intelligent anomaly detection of machine tools based on mean shift clustering. *Procedia CIRP* **2020**, *93*, 1448–1453. [CrossRef]
26. Dindarloo, S.R.; Siami-Irdemoosa, E. Data mining in mining engineering: Results of classification and clustering of shovels failures data. *Int. J. Min. Reclam. Environ.* **2017**, *31*, 105–118. [CrossRef]
27. Krogerus, T.; Hyvönen, M.; Huhtala, K. Recognition of Operating States of a Wheel Loader for Diagnostics Purposes. *SAE Int. J. Commer. Veh.* **2013**, *6*, 412–418. [CrossRef]
28. Liao, L.; Fox, D.; Kautz, H. Extracting places and activities from GPS traces using hierarchical conditional random fields. *Int. J. Rob. Res.* **2007**, *26*, 119–134. [CrossRef]
29. Wang, Y.; Qin, K.; Chen, Y.; Zhao, P. Detecting anomalous trajectories and behavior patterns using hierarchical clustering from Taxi GPS Data. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 25. [CrossRef]
30. Cheng, T.; Venugopal, M.; Teizer, J.; Vela, P.A. Performance evaluation of ultra wideband technology for construction resource location tracking in harsh environments. *Autom. Constr.* **2011**, *20*, 1173–1184. [CrossRef]

31. Fu, J.; Jenelius, E.; Koutsopoulos, H.N. Identification of workstations in earthwork operations from vehicle GPS data. *Autom. Constr.* **2017**, *83*, 237–246. [CrossRef]
32. Pradhananga, N.; Teizer, J. Automatic spatio-temporal analysis of construction site equipment operations using GPS data. *Autom. Constr.* **2013**, *29*, 107–122. [CrossRef]
33. Abdi, H.; Williams, L.J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
34. Ringnér, M. What is principal components analysis? *Nat. Biotechnol.* **2008**, *26*, 303–304. [CrossRef] [PubMed]
35. Kęsek, M. Analysing data with the R programming language to control machine operation. *Inż. Miner.* **2019**. [CrossRef]
36. PNPOLY—Point Inclusion in Polygon Test, W. Randolph Franklin (WRF). Available online: https://wrf.ecse.rpi.edu/Research/Short_Notes/pnpoly.html (accessed on 10 April 2021).
37. Dogan, A.; Birant, D. Machine learning and data mining in manufacturing. *Expert Syst. Appl.* **2021**, *166*, 114060. [CrossRef]
38. Sunhare, P.; Chowdhary, R.R.; Chattopadhyay, M.K. Internet of things and data mining: An application oriented survey. *J. King Saud Univ. Comput. Inf. Sci.* **2020**. [CrossRef]
39. Anand, N.; Vikram, P. Comprehensive Analysis & Performance Comparison of Clustering Algorithms for Big Data. *Rev. Comput. Eng. Res.* **2017**, *4*, 54–80.
40. Valarmathy, N.; Krishnaveni, S. A novel method to enhance the performance evaluation of DBSCAN clustering algorithm using different distinguished metrics. *Mater. Today Proc.* **2020**. [CrossRef]
41. Garima; Gulati, H.; Singh, P.K. Clustering techniques in data mining: A comparison. In Proceedings of the 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 11–13 March 2015; pp. 410–415.
42. Archana Patel, K.M.; Thakral, P. The best clustering algorithms in data mining. In Proceedings of the 2016 International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, India, 6–8 April 2016; pp. 2042–2046.
43. Smiti, A.; Eloudi, Z. Soft DBSCAN: Improving DBSCAN clustering method using fuzzy set theory. In Proceedings of the 2013 6th International Conference on Human System Interactions (HSI), Sopot, Poland, 6–8 June 2013; pp. 380–385.
44. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, USA, 2–4 August 1996; pp. 226–231.
45. Hou, J.; Liu, W. Evaluating the density parameter in density peak based clustering. In Proceedings of the 2016 Seventh International Conference on Intelligent Control and Information Processing (ICICIP), Siem Reap, Cambodia, 1–4 December 2016; pp. 68–72.
46. Smiti, A. A critical overview of outlier detection methods. *Comput. Sci. Rev.* **2020**, *38*, 100306. [CrossRef]
47. Srivastava, A.; Han, E.H.S.; Kumar, V.; Singh, V. Parallel formulations of decision-tree classification algorithms. *Proc. Int. Conf. Parallel Process.* **1998**, 237–244. [CrossRef]
48. Otero, F.E.B.; Freitas, A.A.; Johnson, C.G. Inducing decision trees with an ant colony optimization algorithm. *Appl. Soft Comput. J.* **2012**, *12*, 3615–3626. [CrossRef]
49. Maleki, F.; Ovens, K.; Najafian, K.; Forghani, B.; Reinhold, C.; Forghani, R. Overview of Machine Learning Part 1: Fundamentals and Classic Approaches. *Neuroimaging Clin. N. Am.* **2020**, *30*, e17–e32. [CrossRef] [PubMed]
50. Rutkowski, L.; Jaworski, M.; Pietruczuk, L.; Duda, P. The CART decision tree for mining data streams. *Inf. Sci.* **2014**, *266*, 1–15. [CrossRef]
51. Kass, G.V. An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Appl. Stat.* **1980**, *29*, 119. [CrossRef]
52. Loh, W.Y.; Shin, Y.S. Split selection methods for classification trees. *Stat. Sin.* **1997**, *7*, 815–840.
53. Pang, S.; Gong, J. C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks. *Syst. Eng. Theory Pract.* **2009**, *29*, 94–104. [CrossRef]
54. Balamurugan, M.; Kannan, S. Performance analysis of cart and C5.0 using sampling techniques. In Proceedings of the 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, 24 October 2016; pp. 72–75.
55. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers Inc.: Burlington, MA, USA, 1993.
56. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
57. Bin, S.; Sun, G. Data Mining in census data with CART. In Proceedings of the 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), Chengdu, China, 20–22 August 2010; pp. 260–264.
58. Lewis, R.J.; Street, W.C. An Introduction to Classification and Regression Tree (CART) Analysis. In Proceedings of the Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, Francisco, CA, USA, 22–25 May 2000.
59. Guo, J.; Liu, H.; Luan, Y.; Wu, Y. Application of birth defect prediction model based on c5.0 decision tree algorithm. In Proceedings of the 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Halifax, NS, Canada, 30 July–3 August 2018; pp. 1867–1871.
60. Hssina, B.; Merbouha, A.; Ezzikouri, H.; Erritali, M. A comparative study of decision tree ID3 and C4.5. *Int. J. Adv. Comput. Sci. Appl.* **2014**, *4*, 13–19. [CrossRef]
61. Decision Trees (rpart) in, R. Available online: http://www.learnbymarketing.com/tutorials/rpart-decision-trees-in-r/ (accessed on 15 April 2021).
62. Hahsler, M.; Piekenbrock, M.; Doran, D. dbscan: Fast Density-Based Clustering with R. *J. Stat. Softw.* **2019**, *91*. [CrossRef]