# Predicting the Number of Downloads of Open Datasets by Naïve Bayes Classifier

Barbara Šlibar

*University of Zagreb, Faculty of Organization and Informatics Varaždin, Pavlinska 2, 42 000 Varaždin, Croatia*

***Abstract*** **– Nowadays, the use of Open Data has become more common and prominent, but there are a lot of questions regarding its quality. Most of the revised researches deal with the quality of Open Data portals, rather than estimation of the open datasets quality. Therefore, the main idea of this research is lowering to the level of the dataset itself in order to assess how much such data is downloaded by end users of Open Data portals on the basis of general dataset characteristics. A model for predicting the number of downloads of open datasets based on their general characteristics was constructed using the Naïve Bayes Classifier. Based on the obtained results, it is discussed if the certain dataset character is good predictor of open dataset downloading and to what extent.**

***Keywords*** **– Dataset Characteristics; Naïve Bayes Classifier; Open Data.**

## 1. Introduction

The main aim of the paper is to determine if the identified dataset characteristics are good predictors of the open datasets downloading by the end users of Open Data portals. In order to achieve it, the main aim is decomposed into the following three aims: 1) to identify general characteristics of open datasets, 2) to determine Open Data portal from which the metadata will be downloaded, 3) to build Naïve Bayes classification model for predicting number of downloads based on identified characteristics.

Firstly, the literature review is presented. Prior to the experiment, the review of prominent researches in which the open dataset characteristics are investigated was made. Also, the literature review on the topic Naïve Bayes Classifier was made, since it had been used for predicting number of downloads of open datasets. Secondly, the experiment design is described and it includes: the identification of some general characteristics of datasets, data collection and preparation, the process of constructing the classification model itself, and the presentation of the obtained results. Finally, the findings of the paper are highlighted as well as recommendations for improvement of the conducted research.

DOI: 10.18421/TEM84-33
https://dx.doi.org/10.18421/TEM84-33

**Corresponding author:** Barbara Šlibar,
*University of Zagreb, Faculty of Organization and Informatics Varaždin, Varaždin, Croatia*
**Email:** barbara.slibar@foi.hr

*Received: 21 August 2019.*
*Revised: 02 November 2019.*
*Accepted: 07 November 2019.*
*Published: 30 November 2019.*



## 2. Nature of datasets

Identification of the open datasets characteristics is a precondition for the implementation of this research.

Therefore, in the next few paragraphs the several recent studies related to the above mentioned problem will be highlighted. The conclusion reached by the review of greater amount of papers is that scientists and experts are commonly using the term 'dataset characteristics', even if they were exploring completely different issues.

Bhatt, Thakkar, & Ganatra [1], Bhatt, Thakkar, Ganatra, & Bhatt [2], Kwon & Sim [8], Wang, Song, & Zhu [19] have examined how the dataset characteristics affect the performance of classification algorithms and some of the examined characteristics are:

- sample size - number of instances or number of rows;
- missing values - the ratio of instances having a defective value;
- continuous feature - the ratio of continuous and nominal attribute;
- functional dependency of features - the overall degree of functional independence between the attributes;
- number of features – number of attributes etc.

In contrast, Maali, Cyganiak, & Peristeras [11] and Lnenicka [10] consider metadata as dataset characteristics. Maali et al. [11] investigated metadata related to the structure, consistency, and availability, while Lnenicka [10] focuses only on structural metadata.

## 3. Naïve Bayes Classifier

Data mining is an iterative process of detecting new, valuable, nontrivial information from large amount of data by using automated or manual methods. According to Kantardzic [5] this process consists of activities that are divided into two general groups - prediction and description. In order to achieve objectives of predictive mining or descriptive mining the following data mining methods can be used: classification, regression, clustering, summarization, modelling dependencies and detect changes and deviations. Classification reveals the function of predictive learning which then classifies data or observation into one of several predefined classes [5].

The Naïve Bayes Classifier is a classification algorithm which is based on the application of the Bayesian theorem where for each attribute is assumed to be a class - conditional independent [9]. When there are a large number of attributes and classes, classification of objects is very difficult given that they require a large number of observations for probability estimation. The Naïve Bayes Classifier assumes that the influence of variable values for a particular class is independent of the values of other variables. In the studies carried out with the goal of comparison of classification algorithms, it was found that performance of Naïve Bayes Classifier are comparable to algorithms such as classification tree or neural network [9]. Murty & Devi [13] emphasized that all parameters of model can be calculated from the training set. If the class and some attribute value never appear together in the training set, the probability calculated on a frequency basis will be zero. This will cause that some other probabilities would be zero if they will be multiplied with that probability. Therefore, it is necessary to make minor corrections on all probability estimates so that no probability would be zero.

### *3.1 Background of Naïve Bayes Classifier*

Performing classification on the input vector $\vec{x} = \{x_1, \dots, x_n\}$ of the $n$ attributes includes assigning each vector $\vec{x}$ to a class $C_k$. For example, if one attempts to classify the *plant life-form* using *height of the plant* attribute, and *life cycle of the plant* attribute, the possible classes would be $C_1 = tree$, $C_2 = shrub$, $C_3 = subshrub$. The modeling of such classification function usually includes the following: 1) generating a probability distribution by modelling joint distribution $p(\vec{x}, C_k)$, 2) calculating the class probability $p(C_k|\vec{x})$ conditioned on vector $\vec{x}$ using Bayesian theorem and initial class probabilities $p(C_k)$ [14]. In the above example, the joint function is a function that models the likelihood that a class label (which specifies the plant life-form) and specific value of an attribute appear together. Probability $p(C_2 = shrub,\ height\ of\ the\ plant = 30m,\ life\ cycle\ of\ the\ plant = perennial)$ is very low because shrub can reach a height of up to 3 meters. Once the distribution is found, the Bayesian theorem can be applied to calculate probability of plant life-form based on attributes. Therefore, the highest class probability for $p(C_k|height\ of\ the\ plant = 30m,\ life\ cycle\ of\ the\ plant = perennial)$ is k=1 because tree is the only plant life-form that can grow to a height of 30 meters.

As already mentioned, the naive assumption implies that attributes are conditionally independent given the class $C_k$. In that way, the issue of dimensionality is avoided and it is allowed for the joint distribution $p(x_1, \dots, x_n, C_k)$ to be decomposed into $n+1$ factors ($n$ attributes plus class prior probability $p(C_k)$). The assumption that attributes are mutually independent is considered naive precisely because it is generally not fulfilled. However, with this typical false assumption it turned out that for a large number of attributes (e.g. $n = 150$), Naive Bayes gives good results in practice [14], [15]. For some observation $\vec{x} = \{x_1, \dots, x_n\}$ of the $n$ attribute, the Naïve Bayes predicts the class $C_k$ for $\vec{x}$ according to the following probability [14]:

$$p(C_k|\vec{x}) = p(C_k|x_1, \dots, x_n), \quad za\ k = 1, \dots, K. \tag{1}$$

The probability (1) is then introduced into the Bayesian theorem as shown in the following formula [18]:

$$p(C_k|\vec{x}) = \frac{p(\vec{x}|C_k)p(C_k)}{p(\vec{x})} = \\ = \frac{p(x_1, \dots, x_n|C_k)p(C_k)}{p(x_1, \dots, x_n)}. \tag{2}$$

By using chain rule factor, the $p(x_1, \dots, x_n|C_k)$ factor in the numerator (2) can be decomposed as follows [13], [14]:

$$p(x_1, \dots, x_n|C_k) = p(x_1|x_2 x_n, C_k) \cdot \cdot p(x_2|x_3, \dots, x_n, C_k) \cdot \dots \cdot \cdot p(x_{n-1}|x_n, C_k) \cdot p(x_n|C_k). \quad (3)$$

Furthermore, the Naïve Bayes models assume that the attribute $x_i$ is independent of the attribute $x_j$ for $i \neq j$ with respect to class $C_k$. Therefore, the previously decomposed factor in (3) can be written as [13], [14]:

$$p(x_i|x_{i+1}, \dots, x_n|C_k) = p(x_i|C_k) \Rightarrow p(x_1, \dots, x_n|C_k) = \prod_{i=1}^{n} p(x_i|C_k), \quad (4)$$

$$p(C_k|x_1, \dots, x_n) \propto p(C_k, x_1, \dots, x_n), \propto p(C_k)p(x_1, \dots, x_n|C_k), \propto p(C_k)p(x_1|C_k)p(x_2|C_k) \dots p(x_n|C_k), \propto p(C_k) \prod_{i=1}^{n} p(x_i|C_k).$$

Naïve Bayes gives the probability that a vector $\vec{x}$ belongs to the class $C_k$ as product of $n+1$ factors (class prior probability $p(C_k)$ plus $n$ conditional probabilities, or simpler probability of predictor with respect to class $p(x_i|C_k)$. The classification involves the assignment of class $C_k$ to an observation for which the value $p(C_k|\vec{x})$ is the largest [14]:

$$p(C_a) \prod_{i=1}^{n} p(x_i|C_a) > p(C_b) \prod_{i=1}^{n} p(x_i|C_b) \Rightarrow p(C_a|x_1, \dots, x_n) > p(C_b|x_1, \dots, x_n). \quad (5)$$

Therefore, the class with the highest probability for observation $\vec{x} = \{x_1, \dots, x_n\}$ can be determined by calculating $p(C_k) \prod_{i=1}^{n} p(x_i|C_k)$ for $k = 1, \dots, K$ and assigning the class with the highest value to the observation as shown in (6).

$$\hat{C} = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}} p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \quad (6)$$

where $\hat{C}$ is estimated class for $\vec{x}$ given its attributes $x_1, \dots, x_n$.

## 4. Experiment design

### *4.1 Identification of open datasets characteristics*

In this research, metadata are considered as dataset characteristics. According to Maali et al. [11], a part of the identified characteristics used in this research belongs to the general group (e.g. title, description, publisher, etc.), while some belong to the availability group (e.g. dataset URL, rating assigned to machine-readable format). A total of 11 characteristics of open datasets have been identified [7], [10], [11]:

- Name/Id,
- Title,
- Description/Note,
- Publisher,
- Update frequency,
- License,
- Dataset URL,
- Number of views,
- Number of downloads,
- Domain,
- Machine-readable format score.

### *4.2 Data collection and preparation*

The comparison of Open Data portals conducted by Lnenicka ranked Data.gov.uk as the best one based on selected indicators [10]. The Open Data Barometer is a website where on a yearly basis it is published how governments publish and use open data [16]. According to the Open Data Barometer, the Data.gov.uk is rated as the best open data portal of 2013, 2014, 2015, and 2016 [16]. In addition to that, Data.gov.uk has also been listed on Forbes' official website as one of 33 of the world's most promising free data sources that anyone can use [12].

Therefore, metadata are collected from the portal Data.gov.uk using Java Web application. The dataset usage statistics is found on the portal and it was downloaded as .csv file. The file with dataset usage statistics had been uploaded into the implemented Web application together with dataset identifier, number of downloads, and number of views. For each dataset id, the API request https://data.gov.uk/api/3/action/package_show?id={id} was sent to fetch metadata and save them to a new separate .csv file. The collected dataset contains 1049 records without header and it was last updated on 28th August 2018.

*Table 1. Identified characteristics of open datasets*

| Attribute | Conversion of retrieved metadata (yes/no) | Rule of conversion |
|---|---|---|
| Name/Id | No | - |
| Title | Yes | If the data is retrieved it is labeled as TRUE, else it is labeled as FALSE. |
| Description/Note | Yes | If the data is retrieved it is labeled as TRUE, else it is labeled as FALSE. |
| Publisher | No | - |
| Update frequency | No | - |
| License | Yes | If the retrieved data contains the keyword "open" then license is labeled as TRUE, otherwise FALSE. |
| Dataset URL | Yes | If the data is retrieved it is labeled as TRUE, else it is labeled as FALSE. |
| Number of views | No | - |
| Number of downloads | No | - |
| Domain | No | - |
| Machine-readable format score | Yes | The retrieved data for the score assigned to the machine-readable format are expressed numerically. So, they are converted into textual values as follows: 0=="BAD", 1=="NOT OK", 2=="OK", 3=="GOOD", 4=="VERY GOOD", 5==" EXCELLENT". |

Altogether, prepared dataset contains 9 categorical and 2 numerical attributes. All predictors/independent attributes are categorical, apart from the attribute Number of views which is numerical. Dependent attribute is Number of downloads and it is numerical attribute. Binominal attributes Title, Description/Note, License, and Dataset URL can take one of possible two values TRUE or FALSE, while polynomial attributes Name/Id, Publisher, Update frequency, Domain, and Machine-readable format score can take one of possible multiple different values. Value UNKNOWN is assigned to all missing values.

In the data preparation phase, the shape of distribution was checked for two numerical variables Number of views and Number of downloads. The distribution of both observed attributes is skewed right. Therefore, the log transformation is performed for both attributes to reduce the variability of the data.

### *4.3 The construction of Naïve Bayes classification model*

The survey about the most popular analytics / data science tools conducted by the KDNuggets portal in 2017 has shown that the most popular general data science platform is RapidMiner. According to results of the survey, platform RapidMiner is ranked at the fourth place. Tools were grouped into 4 sections and the results showed up that most popular section are languages. So, Python, R and SQL have taken the first three places because they are still more popular in data science community then other free/commercial solutions. The respondents were divided into 6 regions: US and Canada (41.5%), Europe (35.5%) Asia (10.1%), Latin America (6.5%), Africa and the Middle East (3, 8), Australia and New Zealand (2.7%). If the survey is compared with the one conducted in 2016 it can be noticed that in 2017, there were fewer participants from Europe and slightly more from all other regions [4]. The percentage of use of a certain tool by respondents, difference in tool usage in 2017 compared to the previous year (*positive percentage* - a tool is used more than in the previous year, *negative percentage* - a tool is less used in 2017 than in 2016), and the percentage of respondents which use only one tool among all those who use this tool are displayed in Table 2. It should be highlighted that those who are using only one toll have mostly been using RapidMiner (13.6%).

*Table 2. List of the most used analytical / data science tools in 2017*

| **Tool** | **% of usage for 2017** | **% of change in usage 2017 vs 2016** | **% of usage of a single tool** |
|---|---|---|---|
| Python | 52.6 | 15 | 0.2 |
| R | 52.1 | 6.4 | 3.3 |
| SQL | 34.9 | -1.8 | 0 |
| RapidMiner | 32.8 | 0.7 | 13.6 |
| Excel | 28.1 | -16 | 0.1 |
| Spark | 22.7 | 5.3 | 0.2 |
| Anaconda | 21.8 | 37 | 0.8 |
| Tensorflow | 20.2 | 195 | 0 |
| scikit-learn | 19.5 | 13 | 0 |
| Tableau | 19.4 | 5.0 | 0.4 |
| KNIME | 19.1 | 6.3 | 2.4 |

The assumption which was not checked during the data preparation is that Naïve Bayes Classifier can only be conducted when the predictors are categorical. So, dataset contains numeric predictors,

they should be categorized or divided regarding to numeric values [3]. According to Larose & Larose [9], there are four common binning methods of numeric predictors:

- grouping based on equal widths,
- grouping on the basis of equal frequency,
- grouping by clustering,
- grouping on the basis of the predictive value.

Equal-width grouping is not recommended for data mining because the width of the category can be largely influenced by outliers. Furthermore, the distribution of frequencies with equal class sizes assumes that each category is equally probable which is not justified in most cases. Larose & Larose [9] propose the use of the third or fourth method.

The described assumption is resolved by using k-means algorithm. During the clustering, it is important to determine the optimal number of partitions, or rather clusters. The Davies - Bouldin index (shorter DBI)

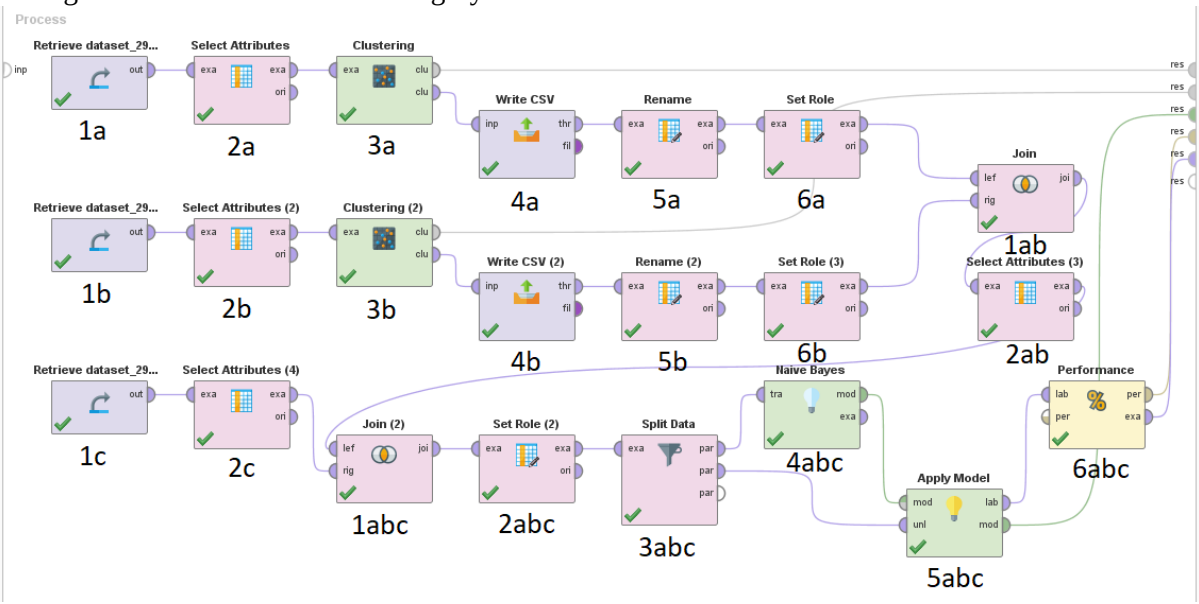


*Figure 1. The process of building the model in RapidMiner Studio*

was used to solve the problem of optimal cluster number. Partitioning dataset of N vectors into M clusters should be done in a way that the value of the DBI is as small as possible. DBI takes into account the intra cluster diversity (the mean squared error which is caused between vector of cluster centroid and its appurtenant data vectors) and the inter cluster distance (distance between the clusters) [6], [17]. Figure 1. shows the model which is created using the RapidMiner Studio and it consists of 22 steps. Table 3. shows defined steps together with their assigned labels and operators. Steps that only contain the label a (1a - 6a) and steps that contain only label b (1b - 6b) use the same operators in the same order and can take place simultaneously. The only difference in those two branches is in the selected attribute over which clustering is performed.

*Table 3. Description of the steps together with the used operators in the built process*

| **Label** | **Operator** | **Step description** |
|---|---|---|
| 1a | Retrieve | Retrieve the prepared dataset that contains two numeric attributes and load it into the process. |
| 2a | Select Attributes | Selection of numeric attribute Number of downloads which contains logarithmic values. |
| 3a | Clustering | Clustering using k-means algorithm. |
| 4a | Write CSV | An additional step for checking results obtained by clustering. |
| 5a | Rename | The result of transforming the numerical variable Number of downloads is a new attribute Cluster. Also, the result of performing clustering over attribute Number of views is a new attribute named Cluster. Therefore, due to attributes discernment, the attribute related to number of downloads is renamed to Clusters_downloads. |
| 6a | Set Role | Changing the role of attribute Clusters_downloads to the type 'regular' since join operator cannot accept attribute with role 'cluster'. |
| 1 | Retrieve | Retrieve the prepared |

| | | |
|---|---|---|
| b | | dataset that contains two numeric attributes and load it into the process. |
| 2b | Select Attributes | Selection of numeric attribute Number of views which contains logarithmic values. |
| 3b | Clustering | Clustering using k-means algorithm. |
| 4b | Write CSV | An additional step for checking results obtained by clustering. |
| 5b | Rename | The result of transforming the numerical variable Number of views is a new attribute Cluster. Also, the result of performing clustering over attribute Number of downloads is a new attribute named Cluster. Therefore, due to attributes discernment, the attribute related to number of views is renamed Clusters_views. |
| 6b | Set Role | Changing the role of attribute Clusters_views to the type 'regular' since join operator cannot accept attribute with role 'cluster'. |
| 1ab | Join | Joining attributes obtained by clustering in a table according to the attribute Id. |
| 2ab | Select Attributes | Selection of attributes Id, Clusters_views, and Clusters_downloads from the obtained set. |
| 1c | Retrieve | Retrieve the prepared dataset that contains two numeric attributes and one additional attribute Id because of merging data. Then, load them into the process. |
| 2c | Select Attributes | Selection of all attributes, except the following ones: 1) Name/Id instead of which the attribute Id is subsequently added and it is used for the identification, 2) Number of downloads, 3) Number of views. |
| 1abc | Join | Joining attributes from steps 2ab and 2c by inner join operator. |
| 2abc | Set Role | In order to define dependent variable, the role of attribute Clusters_downloads should be changed to type 'label'. |
| 3abc | Split Data | Splitting the dataset on the training subset and test subset in the ratio of 70:30. |
| 4abc | Naive Bayes | Generation of Naïve Bayes classification model and application over the training subset. |
| 5abc | Apply Model | Application of Naïve Bayes Classifier on test subset. |
| 6abc | Classification Performance | Statistical performance evaluation of classification. |

### *4.4 Research results*

Results of clustering numeric variable Number of downloads and variable Number of views will be shown in this section. Also, the results of classification by Naïve Bayes Classifier, as well as results of statistical performance evaluation of classification task will be pointed out.

One assumption of Naïve Bayes Classifier is that all input attributes should be categorical. Since attributes Number of downloads and Number of views contained numerical values they were grouped into categories by using k-means algorithm.

The RapidMiner tool through the Cluster Distance Performance operator provides the Davies - Bouldin index for determining the optimal number of clusters. So, in order to determine the optimal number of clusters of numerical attributes the following analysis was conducted (see Table 4.).

*Table 4. Values of Davies – Bouldin index for different number of clusters for attributes Number of downloads and Number of Views*

| Number of clusters (k) | DBI for Number of downloads | DBI for Number of Views |
|---|---|---|
| k = 2 | 0,158 | 0,538 |
| k = 3 | 0,432 | 0,535 |
| k = 4 | 0,472 | 0,499 |
| k = 5 | **0,442*** | 0,491 |
| k = 6 | 0,457 | 0,495 |
| k = 7 | 0,469 | 0,521 |
| k = 8 | 0,465 | 0,489 |
| k = 9 | 0,471 | 0,489 |
| k = 10 | 0,481 | 0,499 |
| k = 11 | 0,471 | **0,479*** |
| k = 12 | 0,498 | 0,490 |
| k = 13 | 0,490 | 0,488 |
| k = 14 | 0,489 | 0,493 |
| k = 15 | 0,483 | 0,501 |
| k = 16 | 0,505 | 0,503 |

** optimal number of groups of variables*

A total of 1049 observations were clustered into 5 clusters by performing k-means over numerical attribute Number of downloads.

*Table 5. Confusion matrix*

| | **true cluster_1** | **true cluster_0** | **true cluster_4** | **true cluster_2** | **true cluster_3** | **class precision (%)** |
|---|---|---|---|---|---|---|
| **pred. cluster_2** | 8 | 2 | 0 | 0 | 0 | 80.00% |
| **pred. cluster_1** | 4 | 36 | 0 | 9 | 0 | 73.47% |
| **pred. cluster_4** | 0 | 0 | 9 | 0 | 0 | 100.00% |
| **pred. cluster_3** | 1 | 21 | 1 | 97 | 23 | 67.83% |
| **pred. cluster_0** | 0 | 0 | 3 | 18 | 40 | 65.57% |
| **class recall (%)** | 61.54% | 61.02% | 69.23% | 78.23% | 63.49% | |

*Table 6. Results of clustering output attribute Number of downloads*

| Cluster label | Number of observations | Range of values (number of downloads) |
|---|---|---|
| Cluster 0 | 223 | 1362 – 6394 |
| Cluster 1 | 52 | 6675 – 139287 |
| Cluster 2 | 474 | 423 – 1340 |
| Cluster 3 | 245 | 17 - 421 |
| Cluster 4 | 55 | 0 |

By clustering values of the attribute Number of views, there were clustered 1049 observations into 11 clusters.

*Table 7. Results of clustering input attribute Number of views*

| Cluster label | Number of observations | Range of values (number of views) |
|---|---|---|
| Cluster 0 | 16 | 73733 – 204650 |
| Cluster 1 | 164 | 2389 – 3189 |
| Cluster 2 | 18 | 38666 – 69951 |
| Cluster 3 | 77 | 6179 – 9038 |
| Cluster 4 | 19 | 22016 – 32723 |
| Cluster 5 | 36 | 13923 – 21395 |
| Cluster 6 | 56 | 9187 – 13358 |
| Cluster 7 | 122 | 3217 – 4280 |
| Cluster 8 | 213 | 1809 – 2381 |
| Cluster 9 | 230 | 1400 – 1806 |
| Cluster 10 | 98 | 4353 – 6074 |

Distribution of dependent attribute Clusters_downloads using the Naïve Bayes Classifier shows that 4.9% observations from the training set is in cluster 1, 5.3% in cluster 4, 21.2% in cluster 0, 23.4% observations belong to cluster 3, and 45.2% to cluster 2. The results of applying model over test set showed that out of a total of 314 observations, 190 are accurately predicted, while 124 are incorrectly predicted. In the incorrectly predicted, there are 42 observations that had not been assigned to any class. Those are the observations in which the class and some attribute value never appeared together in the training set (the probability calculated on the basis of frequency is zero).

The 42 observations RapidMiner were not taken into account during the statistical performance evaluation of the classification task. So, accuracy and classification error were calculated on the basis of 272 observations.

Accuracy of the obtained model is 69,85% $(\frac{190}{272} \times 100)$, while the error of the model classification is 30,15% $(\frac{82}{272} \times 100)$.

## 5. Conclusion

Since the open data movement gains increasing attention in the academic community, this study was conducted in order to contribute to this research field. Based on general open dataset characteristics, the predicting model was constructed and proved to be highly accurate.

The most compelling results will be explained more in detail in the following paragraphs. The results of applying the model showed that out of a total of 314 observations from the training set, 190 of them were exactly classified. Therefore, it was expected that the accuracy of the obtained model will be high which was later confirmed by the statistical performance evaluation of classification task.

The difference between the accuracy of the classification of the two real clusters and the accuracy of the classification of their predicted clusters was noticed in the matrix of the accurate classified real values versus the accurate classified predicted values. The percentage of total accurately real values of cluster 1 (61.54%) differs from the percentage of total accurately predicted values of cluster 1 (80.00%), which can be justified by the fact that cluster 1 contains observations with a number of downloads ranging from 6675 to 139287 per dataset. Since this range is very large, the observations contained in the cluster are quite different with respect to other predictors. So, there is a wrong classification of some predicted observations. The

largest deviation between total accurately real values and total accurately predicted values is found in cluster 0. It contains only those observations or datasets that have never been downloaded (in the training set). Reason why those datasets have never been downloaded may be that dataset has recently been published on an Open Data portal. Since Release date was not observed in this study, it cannot be established with certainty whether this is the reason for the deviation.

It is concluded that the dataset characteristics are good predictors of downloading datasets themselves because of the high accuracy of the obtained model. Recommendation for future research is the inclusion of even more dataset characteristics in the existing model.

**References**


[1]. Bhatt, N., Thakkar, A., & Ganatra, A. (2012). A Survey & Current Research Challenges in Meta Learning Approaches based on Dataset Characteristics. *International Journal of Soft Computing and Engineering* (IJSCE), 2(1), 9.

[2]. Bhatt, N., Thakkar, A., Ganatra, A., & Bhatt, N. (2013). Ranking of Classifiers based on Dataset Characteristics using Active Meta Learning. *International Journal of Computer Applications*, 69(20), 31–36. https://doi.org/10.5120/12089-8269

[3]. Deshpande, B. (2012). Beware of 2 facts when using Naive Bayes classification for analytics. Retrieved from http://www.simafore.com/blog/bid/100934/Beware-of-2-facts-when-using-Naive-Bayes-classification-for-analytics

[4]. Piatetsky, G. (2017, May). New Leader, Trends, and Surprises in Analytics, Data Science, Machine Learning Software Poll. Retrieved from https://www.kdnuggets.com/2017/05/poll-analytics-data-science-machine-learning-software-leaders.html

[5]. Kantardzic, M. (2011). *Data Mining: Concepts, Models, Methods, and Algorithms, Second Edition* (Second edition). Hoboken, N.J: Wiley-IEEE Press.

[6]. Kärkkäinen, I., & Fränti, P. (2000). Minimization of the value of Davies-Bouldin index. In *Proceedings of the IASTED International Conference on Signal Processing and Communications (SPC'2000)*. IASTED/ACTA Press (pp. 426-432).

[7]. Kučera, J., Chlapek, D., & Nečaský, M. (2013). Open Government Data Catalogs: Current Approaches and Quality Perspective. In A. Kő, C. Leitner, H. Leitold, & A. Prosser (Eds.), *Technology-Enabled Innovation for Democracy, Government and Governance* (pp. 152–166). Springer Berlin Heidelberg.

[8]. Kwon, O., & Sim, J. M. (2013). Effects of data set features on the performances of classification algorithms. *Expert Systems with Applications*, 40(5), 1847–1857. https://doi.org/10.1016/j.eswa.2012.09.017

[9]. Larose, D. T., & Larose, C. D. (2015). *Data Mining and Predictive Analytics* (2 edition). Hoboken, New Jersey: Wiley.

[10]. Lnenicka, M. (2015). An in-depth analysis of open data portals as an emerging public e-service. *International Journal of Social, Education, Economics and Management Engineering*, 9(2), 589-599.

[11]. Maali, F., Cyganiak, R., & Peristeras, V. (2010). Enabling interoperability of government data catalogues. In *International Conference on Electronic Government* (pp. 339-350). Springer, Berlin, Heidelberg.

[12]. Marr, B. (2016, February 12). Big Data: 33 Brilliant And Free Data Sources Anyone Can Use. Retrieved from https://www.forbes.com/sites/bernardmarr/2016/02/12/big-data-35-brilliant-and-free-data-sources-for-2016/

[13]. Murty, M. N., & Devi, V. S. (2011). Pattern Recognition: An Algorithmic Approach. In *Undergraduate Topics in Computer Science*. Springer; 2011 edition.

[14]. McGonagle, J. (n.d.) Naive Bayes Classifier. Brilliant.org. Retrieved February 22, 2019, from https://brilliant.org/wiki/naive-bayes-classifier/

[15]. Sayad, S. (2011). Real Time Data Mining. Cambridge Ont: Self-Help Publishers.

[16]. Iglesias, C., & Robinson, K. (2016, April). Open Data Barometer – Third edition. Retrieved from https://opendatabarometer.org/doc/3rdEdition/ODB-3rdEdition-GlobalReport.pdf

[17]. Sabo, K., Scitovski, R., & Vazler, I. (2010). Grupiranje podataka: klasteri. *Osječki matematički list*, 10(2), 149–178.

[18]. Shiri Harzevili, N., & Alizadeh, S. H. (2018). Mixture of latent multinomial naive Bayes classifier. *Applied Soft Computing*, 69, 516–527. https://doi.org/10.1016/j.asoc.2018.04.020

[19]. Wang, G., Song, Q., & Zhu, X. (2015). An improved data characterization method and its application in classification algorithm recommendation. *Applied Intelligence*, 43(4), 892–912. https://doi.org/10.1007/s10489-015-0689-3