Upgrading the Business Intelligence System by Implementing the Decision Tree Model in the R Software Package

Jordan ATANASIJEVIC*, Danijela MILOSEVIC
University of Kragujevac, Faculty of Tehnical Sciences Cacak,
Svetog Save 65, Cacak, Serbia
jodzavogi@hotmail.com (**Corresponding author*), danijela.milosevic@ftn.kg.ac.rs

Abstract: Business Intelligence is the key and basis of a modern understanding of management. The organizations that are able to manage their data resources, information and knowledge are more successful and competitive than the others. These organizations, as a rule, rely on modern strategic management concepts and develop business intelligence systems. Certain changes have taken place in the world of research in recent years, and open-source software packages are now most commonly used for statistical surveys. The R package in particular is gradually becoming the dominant platform for companies that are unable to spend too much on software. The R package has gathered a huge community of enthusiasts who are constantly building upon the latest developments in statistics and data mining at no cost to the end user. The main problem related to this approach lies in the data sources, because the R package is not able to store large amounts of data, as it was designed for data analysis and the respective data is mostly stored on other platforms. The aim of this paper is to find a common solution and correlation between BI, which is based on data warehouses and programs for statistical data processing, and the open-source R package on the other hand in order to obtain timely information in the shortest possible time, according to different criteria, by applying the decision tree model. Decision makers will be able to use the proposed solution to make decisions with confidence even if they don't possess the pertinent IT knowledge.

Keywords: R, Business Intelligence, Decision tree, Machine learning.

1.Introduction

It is well known that the amount of data required by organizations grows by the dayjust like the amount of data being generated, and the processing times are getting shorter, which means the relation between the requiredamount of data and decisiontimesis developing in a disproportionate manner. The pressure exerted on the systemshifts to people who work with the data which are becoming increasingly fatigued. With this in mind on the one hand and the needs of the management team that require fast and accurate information on the other one, it is imperative to find a way to integrate and organize all parties involved in a workable way.

The database market is by far the most represented Structured Query Language database today (SolidIT, 2018). The term SQL databases refers, in practice to, relational and object-relational databases. Given that Database Management Studio and the data model represent the two most important aspects of a database, the term relational database is used for those databases that are based on a relational model (and therefore employed by a relational Database Management Studio). If one takes into acount the advantages of the relational model, while respecting the trend of modern business one realizes there is a need for an easier understanding of stored data and for the development of tools for processingit. These circumstances have led to the use of new database tools characterized by flexible and easilyvariable schemas, tools that prioritize accessibility and visualization (Inmon et al., 2010).

One of the first questions is whether it is necessary to change existing modus operandi and implement a new solution or whether they just need to be refined. In doing so, one should not forget what leadership really is. They use "fresh" information, at any time, from any location, with minimal costs, maximum quality and speed, for self-reporting.

The aim of the paper is to improve the decision making process, whose product will be a new data resource with high utilization in symbiosis with the business intelligence, mainly concerned with the analysis of the past events and predictive analytics, which is a prediction in terms of the likelihood of an event.

At the end of the last century, other types of databases, besides SQL were also in use, although SQL databases were still dominantly represented (Atzeni et al., 2013), even if to a much lesser extent. At that time, choosing the type of database that organizations would use was not too complex. Only in the small number of cases when there was a need for specific types of databases, such as object databases or XML databases, organizations generally opted for SQL databases. Over time, new requirements were put on SQL databases, and expectations related to the deployment of

https://doi.org/10.24846/v29i2y202009

databases have become more complex. Some of the emerging needs have been addressed through the addition of the Star Scheme, which was meant for designing the data warehouse more adequately. The expansion of large amounts of data of varying degrees of structure, which were available from different sources (not only traditional sources, but also most recent ones like social networks), further emphasized the need for a faster processing of variable structure data available in different formats.

The employment of modern tools allows one to use, adapt and build knowledge in real time, to solve problems and make new discoveries. The greatest challenge in realizing one's full potential lies in one's ability to harness the potential of data involved, with regard to which it can be assurned that hidden knowledge exists so that so that anybody could benefit from it.

The data that needs to be collected and processed for further analysis is probably already stored somewhere. Most data is unusable due to its low quality and timeliness, but its availability imperatively prompts companies to find the right data, save it, process it and ultimately analyze it in real time. The time variable plays a key part because in the modern business world, timely and accurate information is expected regardless of the exponential rise in the amount of data that needs to be processed.

The basic idea is to make the decision-making easier by enabling decision-makers to use existing data to receive real-time answers to their questions which would otherwise require much more time and the use of specialized IT services. They can obtain these answers by using established machine-learning algorithms implemented in the user-friendly Shiny application.

In this sense, the paper presents an an an attempt to overcome the previously identified shortcomings by applying the software package R, in order to improve the business intelligence system, using already collected, processed and prepared data, which the organization already uses. The main motivation for the development and implementation of such a solution is the question of whether there is an open-source program that must be extremely fast and accurate to be able to improve the existing way of doing business, over the existing data set. Such an answer is given in this paper.

This paper is structured as follows. In Section 2, the current state of the software is presented based on computations and research findings. Section 3 deals with general concepts related to business intelligence. Section 4 describes the decision trees, one of the most popular approaches to classifying data, as well as the benefits of using them. Section 5 presents the process of data collection which preceedsthe method of creating a decision tree model in software package R, which is described in Section 6. The conclusion and analysis of this research paper are illustrated in Section 7.

2. Related Works

(Safavian & Landgrebe, 1991) suggest that the greatest advantage of the aforementioned model is its ability to break down complex decision-making processes into a collection of Simple Decisions, thereby providing a solution that will be easier to interpret. According to them the main disadvantage of the employedmethods is that each node considers only one characteristic. They also discuss whether it is necessary to use extreme values in the analyses that are carried out and how important they are for the design of decision trees. They consider flexibility as one of the main features of the decision tree i.e., its ability to use different subsets of characteristics and decision rules.

(Jatain & Ranja, 2017) define R software as a tool for statistical model creation, data analysis and prediction of future states. However, given the huge amount of data that is generated daily and that must be processed and analyzed, many organizations refrain themselves from introducing R softwarein order to analyze business processes. Decision trees when grouped together are one of the key methods for predicting future states, outcomes and trends. This work presents decision making trees as a method not for describing data, but for making decisions. Decision trees rather represent inputs for later decision making.

(Keer & Misra, 2020) used R Studio as a tool to examine workplace and performance data. The analyzed data helps companies predict employee behaviour, talent gaps and potential turnover, as well as predict areas of inefficiency.

(Wang et al., 2016) researched this idea and it prioritizes and recommends R package over Pyton

and Excel. The study confirmed the assumption that R processes data much faster than Excel when data files are larger than 1 MB. The authors state that R can also handle very large data files in an efficient manner. R created for statistical processing is very suitable for working with vectors and matrices. The rich functions peculiar to R software are suitable for processing missing values, transforming data, and for creating subsets of data required for the specific data analysis. For decision makers, R packages work more like a processing black box, which may suggest the further advancement of business intelligence towards the creation of an application that would enable users to process data and make decisions independently, without any help from developers.

(Jishag et al., 2020) used a different approach for solving problems which involved predicting the stock market by combining two different components: sentiment analysis on stock-related news reports and historic data analysis. The primary aim of their study was to construct an efficient model for predicting trends in the stock market while minimizing errors and maximizing prediction accuracy. This model achieved notably better accuracy compared to models presented in the past studies.

(Höppner et al., 2020) used decision tree for customer retention campaigns which increasingly rely on predictive models for detecting potential churners in a vast customer base. From the perspective of machine learning, the task of predicting customer churn can be presented as a binary classification problem. Classification algorithms are built to use historic behavior data to accurately predict the probability of a customer defecting. Their technique, called ProfTree, uses an evolutionary algorithm for learning profit-driven decision trees. In a benchmark study with real-life data sets from various telecommunication service providers, it is shown that ProfTree achieves significant profit improvements compared to classic accuracy driven tree-based methods.

(Blanco-Justicia et al., 2020) focused on explaining black-box models by using decision trees of limited depth as a surrogate model. Specifically, they proposed an approach based on microaggregation to achieve a trade-off between the comprehensibility and the representativeness of the surrogate model on the one hand and the privacy of the subjects used for training the black-box model on the other hand. They have presented an approach based on decision tree that allows deriving explanations of machine learning decisions while controlling their accuracy, fidelity, representativeness, comprehensibility and preservation of privacy.

(Nguyen et al., 2020) considered that deep neural networks (DNNs) which are commonly labelled as black boxes are hard to interpret; thus, hindering human understanding of DNNs' behavior. A need exists to generate a meaningful sequential logic for the production of a specific output. Decision trees exhibit better interpretability and expressive power due to their representation language and the existence of efficient algorithms to generate rules. Growing a decision tree based on the available data could produce larger than necessary trees or trees that do not translate to general problems well. In their paper, they introduced two novel multivariate decision tree algorithms for deriving rules from a DNN.

(Grossman et al., 2020) studied both deterministic and randomized decision trees and provided various characterizations and barriers for more general results. They introduced a new measure of complexity called unlabeled-certificate complexity, appropriate for graph properties and other symmetrical functions, where only information about the structure of the graph is known to the competing algorithm.

(Fotache & Strimbei, 2015) manage big data by dealing with three types of operations: data collection, data storage and data processing. In accordance with that, the key factors for analyzing large volumes of data are databases and statistical packages. They state that there is a large supply of statistical packages dedicated to data analysis. Some of them are commercial and they generally provide a wide range of statistical functions and options with highly customizable interfaces for the average non-developer user. However, these software packages are also known for their cost. Small and medium-sized businesses cannot afford to spend up to several thousands of dollars for a small number of licenses. Of course, pricing and licensing systems vary, but to the best of one's knowledge, pricing is still the most common barrier to their use.

According to (Ferrari & Russo, 2016), a report is really just a set of data that is visually organized

so as to make it easier for users to get informed. The authors stress that a basic data report is not sufficient for drawing useful conclusions from the report, which ultimately requires a more detailed description and a more appropriate type of visualization.

There has been a change in the world of research in recent years, and now open-source software packages are mostly used for statistical surveys, while R is gradually becoming the dominant platform for companies that are unable to spend too much on software. R has gathered a huge community of enthusiasts who are constantly building upon the latest improvements in the areas of statistics and data mining, at no cost for the end user. The problem with this approach lies in the data sources, because R is unable to storelarge amounts of data, it is only suitable for data analysis, as it is mostly locatedon other platforms.

Using a decision tree and logistic regression (Grandhe, Damarla & Mohammad, 2019) predict the survival of passengers based on the gender factor, the respective travel class and age of passengers for the Titac crash data, taken as an example. Predictive analysis is employed for many applications for predicting passenger behavior based on past events and the achieved results. A manual analysis oflarge datasets is a very difficult and time-consuming process, and the success of the prediction tends to be low when an analysis is performed manually. A process which involved data analysis and descriptive analysis showed that prediction using the decision tree over test data, featured a 75% success rate and one using linear regression a 78% success rate. The research was carried out for the software package R, but it did not display graphical interface for the user, the authors rather indicated the lines of code based on which graphs were generated to visualize the data.

(Wickham, 2019) states that a great deal of effort has been made recently to advance the role of statistics in data science. The author believes that statistics is a key part of data science, but also that at the same time, moststatistical areas are at high risk of becoming irrelevant. The author also believes that there are three main steps in working with data: collecting, storing and analyzing data. The end product of analysis is not a model, but rathera rhetorical product. An analysis is pointless unless it convinces someone to take action. In business, this usually means convincing several executives, who have little statistical knowledge, to make a decision.

Since each organization processes a large amount of data, which can be of different types and often contain patterns that can be useful if discovered,it is clear that creating a tool that can be used to uncover these patterns could bring great benefits. If such a tool can be created by using open-source software, such as R, instead of business intelligence software effective predictionsof any phenomena can be much more cost-effective (Manohar at al, 2015). When a classifier reads data from a processing database, it applies the ID3 Algorithm to the same data to generate a set of rules. Rules are presented as if-then conditions, but they can also be made more sophisticate by using the decision tree model. These rules can then be applied to test data, in order to verify its performance. The classifier can be employed as a web application in order to increase accessibility. Also, the web application will provide support for a diverse array of databases such as MySQL, Oracle, Microsoft SQL ServerManagement Studio, etc. and there may be additional support for various file formats such as .txt, .csv, .arff, etc. Such a classifier can be of great help in various fields such as healthcare, weather prediction, stock market forecasts, sales and marketing, etc. Based on the aforementioned work, one considered the possibility of implementing the decision tree model for the software package R by using the Shiny package.

3. Business Intelligence

The term business intelligence was first used in 1996 for disclosing categorization of data,both required and reported, that could help the management teamsynthesize valuable information based on which they could record their business decisions.

(Gupta & Sagar, 2020) define Business intelligence as an arrangement of strategies, designs, and innovations that change crude information into significant and helpful data used to empower more compelling vital and operational experiences and basic leadership. Decisions support systems (DSSs) assist in translating raw information into further understandable forms to be used by the high management executives. Business intelligence apparatuses are utilized to make DSS which separate required information

from an extensive database to produce easy-to-use outlines for basic leadership.

Business intelligence means business reporting, ie. that decision-makers constantly have access to real time data that would otherwise need more time to be obtained if it was specifically managed by an IT service.

Machine learning, on the other hand, refers to algorithms that are related to mathematics, statistics, the physical sciences, and serve to extractingcausality from data, and causality is anything but a coincidence.

The purpose of implementing business intelligence is to translate data generated at high speed into new values, both for business world and for society as a whole. This is a concept that involves the employment of tools for collecting, processing, storing and analyzing data in real time. These days, people are still not able to harness the full potential of all data, but they may intuitively feel that it has a value that everyone could benefit from.The business intelligence phenomenon can be viewed from two perspectives -the macro and micro perspective. From a macro point of view, business intelligence denotes a complex aggregated category, which is created by a systematic but not targeted collection of data on macroeconomic trends in a particular geopolitical environment. It also involves organizing and structured logging as well as logical-computational processing meant for detecting different trends. Today, the attention of knowledge engineers is increasingly triggered by the phenomenon of business intelligence viewed from a micro perspective. In this case, it is about discovering hidden knowledge based on business data, which an organization collects routinely, while performing its day-to-day business transactions.

Business intelligence is a relatively young term for which there are several definitions, but each comes down to business intelligence being a process of gathering data, and transforming that data into information that is useful for making important decisions. Business intelligence as a term is most commonly used in order to refer to a computerized decision support. The business intelligence system is a part of an organization's information system purposefully developed to support its management. The management requires comprehensive and timely insight into the performance indicators of the organization in order to make reliable and accurate decisions. According to modern theories, this insight should be available to as many employees as possible, who, when they receive it, can be expected to become more efficient and accountable for the achieved results. Business intelligence techniques (data warehousing, reporting, Online analytical processing, data mining, dashboards, etc.) extract data from an existing information system and transform it into a decision-making format. The implementation of BI techniques increases the value of the organization's existing information system, which is why interest in business intelligence systems is high and still growing.

The analysisof past data from data warehouses can be performed in order to greatly accelerate future business developments. The information that this analysisprovides to us is a key driver of a successful or failed business. As (Covington, 2016) stated in today's technologically-shaped world, something like information can do a good job or completely destroy a business, it just depends on how the information is used.

Since there is no universal definition of the term business intelligence, different authors define it in different ways. One of the most commonly used and general definitions is the following: "Business intelligence is the use of data that leads to better business decisions. It is about accessing, analyzing and discovering new opportunities."

4. Decision Tree

Decision trees are considered one of the most popular approaches for classifying data. Researchers from various scientific and technical disciplines such as statistics, machine learning, pattern recognition and data exploration have addressed the problem of growing a decision tree based on available data. The decision tree contains a root, that is, a node that has no input branches. All other nodes have exactly one input branch. The node with the output branches is called the inner node. All other nodes are called sheets (also known as terminal or decision nodes). In decision trees, each inner node divides the input data (examples) into 2 or more subspaces according to certain discrete attribute input functions. In the simplest and most common case, each data item considers one attribute, so the data space is divided by the value of the attribute under

consideration. Each leaf of a tree is associated with a class and represents the best matching target value. Assigning a class to certain data is done by running through the decision tree process starting from the root node and ending at the leaves.

Researchers opt for less complex decision trees, as they may be considered more understandable. The complexity of a tree has a key influence on its accuracy. The complexity of a tree is explicitly controlled by the stop criterion used and by the applied pruning method. The decision tree is a special method that portrays a particular, real problem as a tree with multiple branches, whose every branch is in fact a possible alternative solution to that particular, previously identified problem. The decision tree is employed in the decision-making process for finding optimal solutions, especially where there are certain uncertainties and risks, where it is obvious that there are several possible alternative solutions to the problem that the decision focused on.

From a graphical perspective, the decision tree is represented in such a way that the tree branches show possible alternative directions, the nodes of the places where the decision is made. The creation of a decision tree virtually leads to the creation of a chain of interrelated and dependent decisions that are involved in the final decision making. By making a final decision, the decision maker and the organization will achieve the planned and desired goals. In the decision-making process, the decision maker most commonly uses objective tools such as probability theory and the decision tree. The decision tree is a method that is primarily used for graphical representation in the decision-making process. This method is often used as a graphical tool in the decision-making process, it consists of a series of steps and is usually employed for evaluating specific decisions. Actually, the decision tree in the decision-making process is a map of possible solutions at different stages which is available to the manager, the decision maker.

The decision tree is a method that decision makers apply in situations where they are required to make several decisions in a row, and where each decision has a significant impact on a particular stage of the decision process that follows. What distinguishes this decision-making method from others is its ability to return to the initial stage of a situation in case the previous decision was wrongly made and does not yield the desired result.

5. The process of Preparing Ddata for Analysis

Before data can be converted into information that will be used by a certain person, it must be in a format that is tailored to the individual type of user. That is why it is important to implement an ETL (Extract, Transform, Load) procedure. ETL is a process of thorough data storage, so producing accurate and quality data is of utmost importance. The ETL process itself is not visible to end users, but it is based on a key part of building a data warehouse.

For this reason, Microsoft has developed S*QL Server Integration Services*, as a tool to be employed for the ETL process already mentioned, when it is necessary to connect to a data source and download that data to its own server where it will be processed in the future. SSIS allows data to be downloaded from a variety of sources, regardless of the format in which it was stored, no matterif it was Excel, a database, a plain text file or something else. This tool has the ability to access data, retrieve it, and then begin the process of data transformation. The last step consists in loading the data into a custom data warehouse.

The first step in demonstrating the employment of the decision tree model in the process of improving the business intelligence system consists in collecting, processing, transforming and storing the data that will be investigated. This paper uses traffic accident data for the territory of the city of Belgrade for 2016, which in taken from the open data portal and weather data for each day in 2016.

The development strategyof the Republic of Serbia is based on the comparison and monitoring of the best and most advanced countries in all fields, including traffic. For these reasons, it is necessary to act strategically in order to activate the traffic safety system in the Republic of Serbia. The Republic of Serbia has started to make its data available, and the competent institution responsible for the processing of open data is the Office for Information Technology and Electronic Governance within the Directorate for Electronic Governance.

Open data is digital data available to the public. Its technical and legal characteristics allow it to be used, reused and redistributed by anyone,

at any time and anywhere. Open data can help governments, citizens and organizations achieve better results in the field of public services.

Datasets in the open data lookup have been provided in .CSV format, which is extremely suitable for transformation, compression and analysis, and it also allows conversion for conversion to many other formats suitable for data analysis.

The published document on the number of traffic accidents that occurred on the territory of the City of Belgrade in 2016 contains eight variables (columns).

Column names are defined as follows:

ID_ACC - A traffic accident identification number, which is presented as a seven-digit number.

TIME_ACC - The date and time of the traffic accident is given in the format dd.mm.yyyy, hh: mm (day.month.year, hour:minute). From the given data, it can be concluded that the complete data is available in the database for all twelve months of 2016.

KIND_ACC - The type of traffic accidents is divided into: 1) traffic accidents with material damages; 2) traffic accidents involving injured persons; 3) traffic accidents with dead persons.

TYPE_ACC - The name and type of traffic accidents is divided into: 1) single-vehicle traffic accidents; 2) traffic accidents involving at least two vehicles (no turning); 3) traffic accidents with at least two vehicles (turning or crossing); 4) traffic accidents involving parked vehicles; 5) pedestrian accidents.

DESC_ACC - A more detailed description of traffic accidents which contains 68 events: 1) a single-vehicle traffic accident (11 events); 2) traffic accidents involving at least two vehicles (no turning) (9 events); 3) traffic accidents involving at least two vehicles (turning or crossing) (18 events); 4) traffic accidents with parked vehicles (5 events); 5) pedestrian accidents (25 events).

PLACE_ACC - Features the name of the 17 city municipalities or of the town of the accident.

Weather data was taken from the website of the Republic Hydrometeorological Institute of Serbia.

After collecting the data, it is transformed and enrolled in the data warehouse, which is the basis of business intelligence.

Dimensional modeling is a logical design technique aimed at presenting data in a standard format that ensures a high system performance. The result of dimensional modeling is a dimensional data model that involves the definition of dimensions, hierarchies, and relationships. In the dimensional model, data is organized to describe dimensions and variables of the star scheme. Dimensions define each transaction and are stored in tables that are linked to a fact table. Variables represent numerical data which is stored in a fact table. This creates a data model that uses multidimensional arrays in order for users to catch a glimpse of the respective data as it is viewed by executives, managers, analysts or planners. Users typically view and analyze data from dimension. A star schema was used for implementing the data warehouse.

The star schema stores all dimensional information in a single table. Each level of the hierarchy is represented by a column or a set of columns in a dimension table. A dimensional object can be used in order to define a hierarchical relationship between two columns (or a set of columns). Without dimensional objects, hierarchical relationships can onlybe defined by metadata. The attributes are featured in the columns of the dimension tables. Dimension tables contain descriptive text information. The attributes in the dimension table are used as constraints when querying. Each dimension table has its own primary key, and they all participate in the creation of the primary key of the fact table.

Fact tables are used for storing measures, they contain information that is usually of a numerical type and they can contain a large number of records. Fact tables contain a composite primary key, which consists of several foreign keys (one for each dimension table) and columns for each dimension. Based on this implementation, one can get different views of the same data.

For each variable, the dimensions that are linked through the fact table are defined and implemented.

Since a data warehouse is a specific database designed for supporting decision-making in a particular organization, it is rich in data but, on the other hand, poor in knowledge. Unlike business-oriented transaction systems, data warehouses are subject-oriented, which means they are focused on business process entities. Data integration in

data warehouses ensures that data is displayed in consistent formats based of certain conventions when specifying names, restrictions on domains, attributes, and dimensions. Data in data warehouses is time-dependent which means that any data contained in a data warehouse is related to a certain moment in time. Last, the data in the data warehouses is invariable, that is, as soon as certain data is entered into the data warehouse, it is only possible to access itbut not to change it.

As it was stated, the raw data required for business analyses is located in different locations, in different formats andis constantly collected and stored in systems designed for automating the operations that are performed on a daily basis. Yet all of these facts are beyond the reach of business decision makers. Since data is collected from different files, data analysis may lead to contradictory results. Also, data formats and semantics are different across databases. Data warehouses are designed for a large amount of read-only data, providing information that is used for decision making. Following a complex approach to collecting, processing, transforming, and loading data into a data warehouse, implementation of a decision tree model is performed in order to develop models that will help one identify, by accurately ranking the respective variables, data patterns and their interconnections in a way that is adaptable to the end users and understandable for him. The goal is to enable the users, using the application to select any attribute for which one would want to follow its future behavior, as well as an arbitrary choice of attributes whose predictive power should be measured.

Taking into account the achievements to date in this field, one of the ways of analysing available data on existing facts and of extracting the necessary information as a starting point for defining, developing and improving decision making will be outlined below. A particular emphasis is placed on the use of the decision tree model, on the ability to improve the performance of this model through optimization of parameters and the choice of an appropriate attribute selection method, as well as on monitoring and comparing of results achieved by the selected model for balanced and unbalanced data.

The paper uses Shiny, a package from the R software package that allows one to create interactive web applications directly through R in an extremely simple way.

Since the data is already prepared, memory space is not wasted on processing it, data is only loaded into R by performing simple queries, e.g. using the view from SQL Server Management Studio. The code used for importing library to Shiny application and creating Open Database Connectivity (ODBC) from R to SQL Server Management Studio is presented below. Also, the package rpart.plot was installed, in order to plot an rpart model, automatically tailoring the plot for the model's response type. Set.seed function is used to set the seed of R‘s random number generator, which is useful for creating simulations or random objects that can be reproduced. Shiny comes with a reactive programming library that one can use to structure one`s application logic. By using this library, changing input values will naturally cause the right parts of the R code to be re-executed, which will in turn cause any changed outputs to be updated.

```
library(shiny)
library(RODBC)
install.packages(“rpart.plot”)
library(“rpart”)
library(“rpart.plot”)
set.seed(678)
```

Code used for the connection, which uses the already created ODBC is presented below:

```
con<-odbcConnect(“data_traffic_accident_2016”,
uid=””, pwd=””)
```

After that the dataset from SQL Server Management Studio was loaded, from certain query and used it in R. Further on, the data cleanup was loaded. Since Shiny web apps are interactive, the input values can change at any time, and the output values need to be updated immediately to reflect those changes.

```
data_traffic_accident <-sqlQuery(con,”SELECT *
FROM [2016].[dbo].[View_ data_traffic_accident]”)
```

Essentially, the Shiny package gives one the ability to use only R for creating a graphical interface (in the form of a web application) for the analyses that should be performed through R. As such, the results of certain data analyses can be very striking.

Also, certain procedures are often repeated over and over in R, which may be automated in some way. This is handled using Shiny, since one can create a graphical interface for these procedures,

only by setting the necessary inputs, after which the output itself will appear on the screen.

6. Implementation of the Decision Tree Model

As it was mentioned, in the practical part of the work, the aforementioned problem will be solved on the basis of traffic accident data for the territory of the City of Belgrade in 2016, with the aim of predicting the potential occurence of future traffic accidents based on the application of the stable decision-making model data on the weather conditions that prevailed on a certain day. The available dataset consists of 51 attributes and 17275 instances. Therefore, since most of the time is spent preparing the data, it is necessary to harness the potential of some software tools to reduce the dimensions of the data set and identify the attributes, that is, the features that play the most important part in predicting traffic accident patterns, because otherwise, this huge investment in data preparation would lose its purpose since the respective data would provide little knowledge.

Shiny applications have two components, a user interface object and a server function, that are passed as arguments to the `shinyApp` function that creates a Shiny app object from this UI/server pair. The Shiny web framework is fundamentally about making it easy to wire up input values from a web page, making them easily available to the user in R, and having the results of one's R code written as output values back out to the web page. The source code for both of these components is listed below. The user interface of the application (i.e., how the app will be displayed to the user in a web browser) is described by a ui.R file. However, it also allows users to create widgets, which are interactive controls that affect the application. User interface is employed for designing appearance of the form peculiar to titles and positions of selection fields of variables which is presented below:

```
library(dplyr)
install.packages("magrittr")
df = subset(mydata, select = -c(x,z) )
ui <- fluidPage(
   titlePanel('Plotting Decision Tree'),
   sidebarLayout( sidebarPanel( h3(Saobracajne nezgode),
         uiOutput('choose_y'), uiOutput('choose_x'),
       actionButton('c50', label = 'Generate decision tree')),
 mainPanel(verbatimTextOutput('tree_summary'),
         plotOutput('tree_plot_c50'))))
```

Figure 1 presents implemented user interface in Shiny.

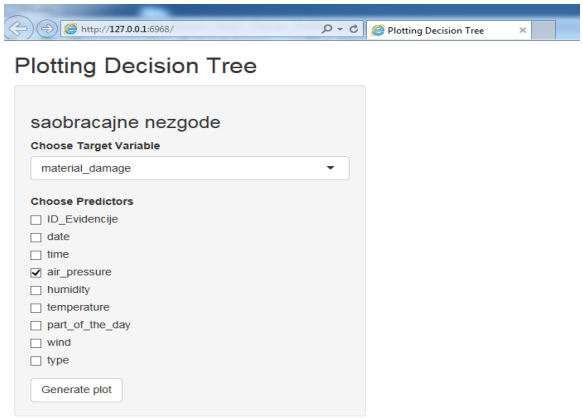

Figure 1. Implemented user website in Shiny

User interface (ui.R) is populated with information from an R session (which can be thought of as the server for this application). The Shiny architecture allows one to pass information back and forth between the user-interface and the server. Here is the corresponding server.R code for the UI above:

```
server <- function(input, output) {
   output$choose_y <- renderUI({
      is_factor <- sapply(data_traffic_accident, FUN =
is.factor)
      y_choices <- names(data_traffic_accident)[is_factor]
   selectInput('choose_y', label = 'Choose Target
Variable', choices = y_choices) })
      output$choose_x <- renderUI({x_choices <-
names(data_traffic_accident)[!names(data_traffic_
accident) %in%  input$choose_y]
      checkboxGroupInput('choose_x', label = 'Choose
Predictors', choices = x_choices)})
   fit <- rpart(as.formula(paste(isolate(input$choose_y),
'~',  paste(isolate(input$choose_x), collapse = +'))),
method="class", data=
  data_traffic_accident, control=rpart.control(minsplit =
50, cp=0.0000001, maxdepth = 7))
   output$summary(fit)
   output$rpart.plot(fit, type=4, extra=6)
   output$rpart.plot(fit)}
```

The above codesshows the procedure for implementation of the decision tree model in the R software package. This represent the first step to visualizing and exploring data by creating dynamic diagrams. Decision trees were used for supervised machine learning with this type of data sets because they are not sensitive to the scale of features and they can handle features that are both quantitative and qualitative. It can be seen that the decision tree model can be customized by adjusting and leveling the respective parameters in order to provide the

most accurate analytics and interactive parameter selection, as R can provide a wide variety of graphical representations. I used parameters:

Minsplit signifies the number of instances in the decision-making tree node needed to perform the branching.

The parameter *cp (complexity parameter)* judges the relative improvement in the decision-making tree's accuracy created through further branching.

The parameter *maxdepth* represents depth of the decision-making tree, i.e., the greatest number of characteristics of an instance that need to be checked to make a decision.

The decision tree model is suitable for this type of problems because it not only uses known conditionsto make predictions, but it also estimates the prediction probability which makes it suitable for interpreting the results.

The implementation of the decision tree model, carried out in the way described above, has created the opportunity to improve business intelligence by discovering relationships between arbitrarily selected variables, which can help decision makers reach a conclusion more easily and predict future situations, as well as describe, review and compare the decisions made so far throughout their business process, explain and compare them with the methods used to that point, and indicate the possibility of improving them. Figure 2 illustrates implemented decision tree in Shiny. The user can get the decision trees, based on the entered parameters, and predict future events, so one will make decisions based on them.

The process of developing a decision tree model based on the classification and regression tree model required developing a model with 3 to 8 levels of depth. The results indicate that a 7-tier C&RT tree model can be selected as a base model. Up to the seventh level, the values of the absolute average error are constantly decreasing, and the value of the linear regression coefficient is increasing, while in the 8th level of 8-tier modelsthe indicators stagnate.

So far, the emphasis is on visualizing the relationship between arbitrarily selected components and on a better understanding of data, from the perspective of an end-user who is a direct user of a software solution and who is actually targeting the business intelligence system, with the aim of obtaining the required information in the shortest possible time period, by using the relevant data contained in the data warehouse. Decision trees can also provide information about the "weight" and influence of certain variables on the behavior of an arbitrarily chosen variable. This is another parameter that the end user can consider when making a decision.

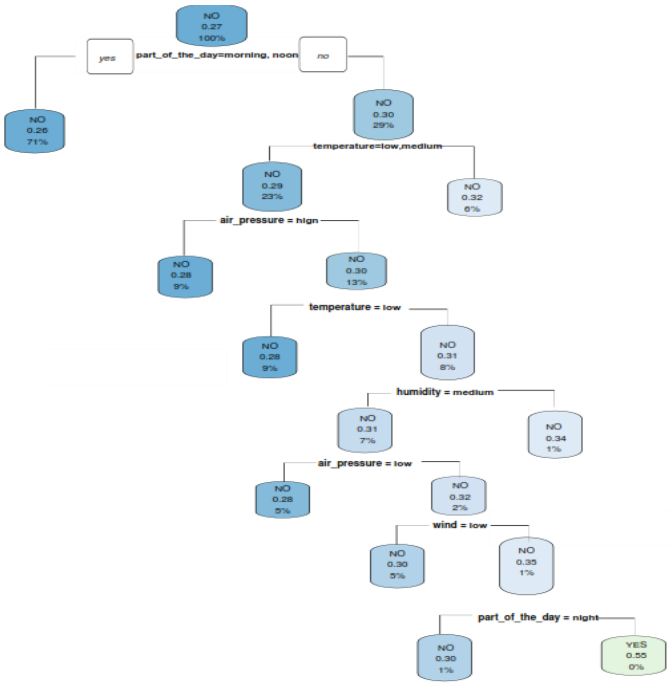

Figure 2. Implemented decision tree model in Shiny

7. Conclusion

This paper analyzes the current researchefforts to improve Business Intelligencebased on previously published papers that focus entirely or partially on the elements related to the use of the R software package.

The design and development of a business intelligence system should take into account whether it facilitates the work of end users, whether the data it uses is expensive and easily accessible, whether the information it generates is expensive, whether the respective data is readily available, etc.

This process should not neglect what is actually necessary for the end user. The end user wants to use "fresh" information, anytime, from any location, and to be able to integrate it into the existing system with a minimal effort. Also, the end user wants a unique and integrated data source, to generate reports, and to have a simple decision support tool.

One of the main causes of failure in implementing a business intelligence system is the human factor, i.e. the ability of end users to change and adapt to the changes brought about by a new technology. Although there are numerous methods for planning, analyzing, designing, building, using and maintaining business intelligence systems, they did not yieldany positive results yet. It is necessary to find ways and methods to change the default mode.

In order to solve this problem, it is necessary that the people from the management structures of the system have a positive attitude and that the end users accept new systems and professionalism, i.e. knowledge of the technology involved and of its application possibilities.

The paper proposes the application of software package R in order to improve the business intelligence system, in the case when the data involved was preloaded, transformed and stored in one of the Database Management Systems.

The expected impact is that by implementing a stable decision-making model based on software package R, one may discover new connections and regularities, which were unknown before, and may become able to make better and more prudent decisions, involving business models that will better fit the terms and requirements of users.

This work leads to a conclusion that data accumulated in a certain business sphere can be used to create new value that can be useful to the decision-makers. This allows decision-makers to uncover and utilize hidden potential of existing data.

The prediction method shows how, based on input data already stored and processed, a potential traffic accident can be anticipated, which might help prevent it. This reduces the time necessary for decision makers to respond to changes that may occur, by obtaining data that results from the prediction of the acceptable absolute error and the degree of linear correlation.

REFERENCES

Atzeni, P., Jensen, C. S., Orsi, G., Ram, S., Tanca, L. & Torlone, R. (2013). The relational model is dead, SQL is dead, and I don't feel so good myself, *ACM SIGMOD Record*, *42*(2), 64-68.

Blanco-Justicia, A., Domingo-Ferrer, J., Martínez, S. & Sánchez, D. (2020). Machine Learning Explainability Via Microaggregation and Shallow Decision Trees, *Knowledge-Based Systems*, *194*, 105532. DOI: 10.1016/j.knosys.2020.105532

Covington, D. (2016). *Analytics: Data Science, Data Analysis, and Predictive Analytics for Business*. CreateSpace Independent Publishing Platform.

Ferrari, A. & Russo, M. (2016). *Introducing Microsoft Power BI*. Microsoft Press.

Fotache, M. & Strimbei, C. (2015). SQL and Data Analysis. Some Implications for Data Analysits and Higher Education, *Procedia Economics and Finance*, *20*, 243-251. DOI: 10.1016/S2212-5671(15)00071-4

Grandhe, P., Damarla, V. P & Mohammad, S. (2019). Extensive data set analysis & prediction using R, *Journal of Physics: Conference Series*, *1228*(1), p. 012048. DOI: 10.1088/1742-6596/1228/1/012048

Grossman, T., Komargodski, I. & Naor, M. (2020). Instance Complexity and Unlabeled Certificates in the Decision Tree Model. In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*, Schloss Dagstuhl-Leibniz-Zentrum für Informatik (pp. 56:1–56:38).

Gupta, P. & Sagar, B. B. (2020). Decision Support System for Business Intelligence Using Data Mining Techniques: A Case Study*, Advances in Computational Intelligence*, 81-94. Springer, Singapore. DOI: 10.1007/978-981-13-8222-2_7

Höppner, S., Stripling, E., Baesens, B., vanden Broucke, S. & Verdonck, T. (2020). Profit Driven Decision Trees for Churn Prediction, *European Journal of Operational Research*, *284*(3), 920-933.

Inmon, W. H., Strauss, D. & Neushloss, G. (2010). *DW 2.0: The architecture for the next generation of data warehousing*. Elsevier, Morgan Kaufmann.

Jatain, A. & Ranja, A. (2017). A Review Study on Big Data Anaylsis Using R Studio, *International Journal of Computer Science and Mobile Computing*, *6*(6), 8-13.

Jishag, A. C., Athira, A. P., Shailaja, M. & Thara, S. (2020). Predicting the Stock Market Behavior Using Historic Data Analysis and News Sentiment Analysis in R. In *First International Conference on Sustainable Technologies for Computational Intelligence* (pp. 717-728). Springer, Singapore.

Keer, P. K. & Misra, S. (2020). HR Analytics-Transforming Traditional HR to New Age Process in Today's Digitalized ERA, *Studies in Indian Place Names*, *40*(71), 2912-2919.

Manohar, S., Mittal, A., Naik, S. & Ambre, A. (2015). A dynamic classifier using decision tree algorithm, *International Journal of Advanced Research in Computer Science and Software Engineering*, *5*(1), 628-631.

Nguyen, T. D., Kasmarik, K. E. & Abbass, H. A. (2020). Towards Interpretable Deep Neural Networks: An Exact Transformation to Multi-Class Multivariate Decision Trees, *arXiv:2003.04675v2*, *arXiv.org*.

Safavian, S. R. & Landgrebe, D. (1991). A Survey of Decision Tree Classifier Methodology, *IEEE Transactions on Systems, Man, and Cybernetics*, *21*(3), 660-674.

SolidIT (2018). *DB-Engines Ranking*. Available at: <https://dbengines.com/en/ranking, accessed 09th December 2019>.

Wang, H., Li, S., Tian, Y. & Aitouche, A. (2016). Intelligent Proportional Differential Neural Network Control for Unknown Nonlinear System, *Studies in Informatics and Control*, *25*(4), 445-452. DOI: 10.24846/v25i4y201605

Wickham, H. (2019). Data science: how is it different to statistics?, *IMS Bulletin*, *48*.