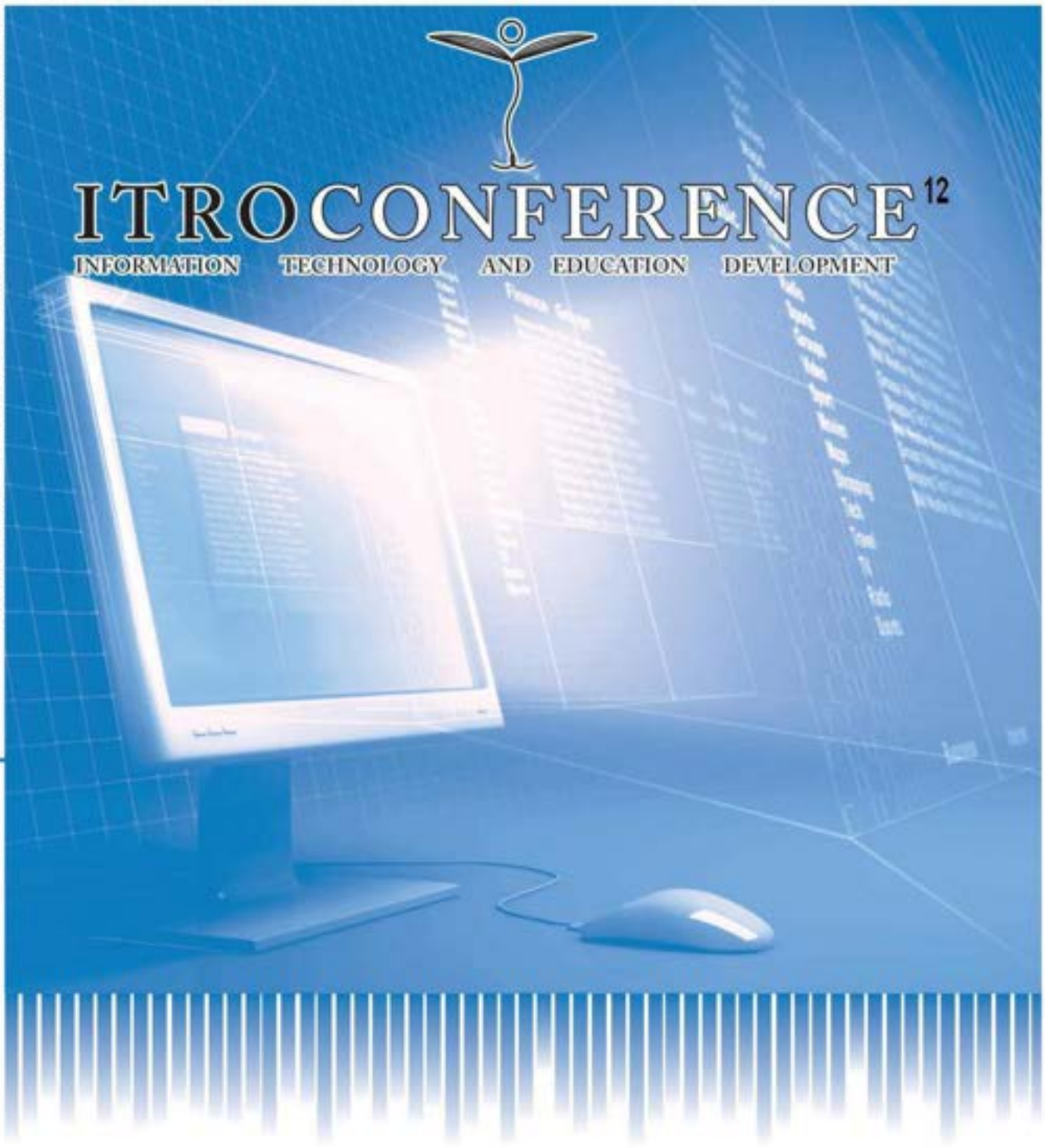
UNIVERSITY OF NOVI SAD
TECHNICAL FACULTY
"MIHAJLO PUPIN"
ZRENJANIN



ITROCONFERENCE 12
INFORMATION TECHNOLOGY AND EDUCATION DEVELOPMENT

PROCEEDINGS

ZRENJANIN, November 2021

UNIVERSITY OF NOVI SAD
TECHNICAL FACULTY ”MIHAJLO PUPIN”
ZRENJANIN
REPUBLIC OF SERBIA

XII INTERNATIONAL CONFERENCE OF

INFORMATION TECHNOLOGY AND DEVELOPMENT OF EDUCATION ITRO 2021

PROCEEDINGS OF PAPERS

XII MEĐUNARODNA KONFERENCIJA

INFORMACIONE TEHNOLOGIJE I RAZVOJ OBRAZOVANJA ITRO 2021

ZBORNIK RADOVA

ZRENJANIN, NOVEMBER 2021

Publisher and Organiser of the Conference:
University of Novi Sad, Technical faculty „Mihajlo Pupin", Zrenjanin, Republic of Serbia

For publisher:
Dragica Radosav, Ph. D, Professor,
Dean of the Technical faculty „Mihajlo Pupin", Zrenjanin, Republic of Serbia

Editor in Cheaf - President of OC ITRO 2021:
Snežana Jokić, Ph. D, Assistant Professor

Proceedings editor:
Snežana Jokić, Ph. D, Assistant Professor

Technical design:
Maja Gaborov MSc, Assistant
Aleksandra Stojkov MSc, Assistant
Nemanja Tasić MSc, Assistant

Circulation: **50**

ISBN: 978-86-7672-351-5

CIP - Каталогизација у публикацији
Библиотеке Матице српске, Нови Сад

37.01:004(082)(0.034.2)
37.02(082)(0.034.2)

INTERNATIONAL Conference of Information Technology and Development of Education ITRO (12 ; 2021 ; Zrenjanin)
Proceedings of papers [Elektronski izvor] / XII International Conference of Information Technology and Development of Education ITRO 2021 = Zbornik radova / XII međunarodna konferencija Informacione tehnologije i razvoj obrazovanja ITRO 2021, Zrenjanin, November 2021. - Zrenjanin : Technical Faculty "Mihajlo Pupin", 2022. - 1 elektronski optiČki disk (CD-ROM) : tekst, slika ; 12 cm

Sistemski zahtevi: Nisu navedeni. - Nasl. sa naslovnog ekrana. - Elektronska publikacija u formatu pdf opsega IX, 238 str. - Tiraž 50. - Bibliografija uz svaki rad

ISBN 978-86-7672-351-5

a) Информационе технологије -- Образовање -- Зборници б) Образовна технологија -- Зборници

COBISS.SR-ID 66049033

PARTNERS INTERNATIONAL CONFERENCE

South-West University „Neofit Rilski”
Faculty of Education, Blagoevgrad,
Republic of Bulgaria

SOUTH WEST UNIVERSITY
"NEOFIT RILSKI"

Technical University of Košice
Faculty of Electrical Engineering and Informatics
Slovak Republic

University Goce Delcev Stip
Republic of Macedonia

THE SCIENCE COMMITTEE:

Marina Čičin Šain, Ph.D, Professor, University of Rijeka, Croatia
Sashko Plachkov, Ph.D, Professor, South-West University "Neofit Rilski" /Department of Education, Blagoevgrad, Republic of Bulgaria
Sulejman Meta, Ph.D, Professor, Faculty of Applied Sciences, Tetovo, Macedonia
Márta Takács, Ph.D, Professor, Óbuda University, John von Neumann Faculty of Informatics, Budapest, Hungary
Nina Bijedić, Ph.D, Professor, Applied mathematics, Bosnia and Herzegovina
Mirjana Segedinac, Ph.D, Professor, Faculty of Science, Novi Sad, Serbia
Milka Oljača, Ph.D, Professor, Faculty of Philosophy, Novi Sad, Serbia
Dušan Starčević, Ph.D, Professor, Faculty of Organizational Sciences, Belgrade, Serbia
Josip Ivanović, PhD, Professor, Hungarian Language Teacher Training Faculty, Subotica, Serbia
Ivanka Georgieva, Ph.D, South-West University "Neofit Rilski", Faculty of Engineering, Blagoevgrad, Republic of Bulgaria
Miodrag Ivković, Ph.D, Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Momčilo Bjelica, Ph.D, Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Dragica Radosav, Ph.D, Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Dragana Glušac, Ph.D, Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Dijana Karuović, Ph.D, Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Ivan Tasić, Ph.D, Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Vesna Makitan, Ph.D, Assistant Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Marjana Pardanjac, Ph.D, Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Snežana Babić Kekez, Ph.D, Professor,Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Stojanov Željko, Ph.D, Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Brtka Vladimir, Ph.D, Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Kazi Ljubica, Ph.D, Assistant Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Berković Ivana, Ph.D, Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Nikolić Milan, Ph.D, Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Dalibor Dobrilović, Ph.D, Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Anja Žnidaršič, Ph.D Professor, Faculty of Organizational Sciences, Kranj, Slovenia
Janja Jerebic, Ph.D Professor, Faculty of Organizational Sciences, Kranj, Slovenia
Mirjana Kocaleva, Ph.D Professor, Faculty of Infomatics, University “Goce Delčev”, Štip, Nort Macedonia
Tatjana Grbić, Ph.D Professor, Faculty of Technical Sciences, Novi Sad, Serbia
Slavica Medić, Ph.D Professor, Faculty of Technical Sciences, Novi Sad, Serbia
Gordana Jotanović, Ph.D Professor, Faculty of Transport and Traffic Engineering, Doboj, BIH
Đurđa Grijak, Ph.D Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Bojana Perić Prkosovački, Ph.D Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Snežana Jokić, Ph.D, Assistant Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Stojanov Jelena, Ph.D, Assistant Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia

THE ORGANIZING COMMITTEE:

Snežana Jokić, Ph.D, Ass. Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, R. of Serbia - Chairman of the Conference ITRO 2020
Dragica Radosav, Ph.D, Professor, Technical Faculty “Mihajlo Pupin” Zrenjanin, R. of Serbia
Dragana Glušac, Ph.D, Professor, Technical Faculty “Mihajlo Pupin” Zrenjanin, R. of Serbia
Dijana Karuović, Ph.D, Professor, Technical Faculty “Mihajlo Pupin” Zrenjanin, R. of Serbia
Marjana Pardanjac, Ph.D, Professor, Technical Faculty “Mihajlo Pupin” Zrenjanin, R. of Serbia
Jelena Stojanov, Ph.D, Ass. Professor, Technical Faculty “Mihajlo Pupin” Zrenjanin, R. of Serbia
Vesna Makitan, Ph.D, Ass. Professor, Technical Faculty “Mihajlo Pupin” Zrenjanin, R. of Serbia
Đurđa Grijak, Ph.D Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Bojana Perić Prkosovački, Ph.D Professor, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Maja Gaborov, MSc, Assistant, Technical Faculty “Mihajlo Pupin” Zrenjanin, R. of Serbia
Nemanja Tasić, MSc, Assistant, Technical Faculty “Mihajlo Pupin” Zrenjanin, R. of Serbia
Aleksandra Stojkov MSc, Assistant, Technical Faculty “Mihajlo Pupin” Zrenjanin, R. of Serbia

With this publication, the CD with all papers from the International Conference on Information Technology and Development of Education, ITRO 2020 is also published.

INTRODUCTION

This Proceedings of papers consists from full papers from the International conference "Information technology and development of education" - ITRO 2021, that was held at the Technical Faculty "Mihajlo Pupin" in Zrenjanin on November 26th 2021.

The International conference on Information technology and development of education has had a goal to contribute to the development of education in Serbia and the Region, as well as, to gather experts from natural and technical sciences' teaching fields.

The expected scientific-skilled analysis of the accomplishment in the field of the contemporary information and communication technologies, as well as analysis of state, needs and tendencies in education all around the world and in our country has been realized.

The authors and the participants of the Conference have dealt with the following thematic areas:

- Education in crisis situations
- Educational challenges
- Theoretic and methodology questions of contemporary pedagogy
- Digital didactics of media
- Modern communication in teaching
- Curriculum of contemporary teaching
- E-learning
- Education management
- Methodic questions of natural and technical sciences subject teaching
- Information and communication technologies

All submitted papers have been reviewed by at least two independent members of the Science Committee. There were total of 94 authors that took part at the Conference from 12 countries, 3 continents: 52 from the Republic of Serbia and 42 from foreign countries such as: Macedonia, Bosnia and Herzegovina, Hungary, Slovakia, India, Bulgaria, Rumania, Albania, USA, Canada, Malaysia. They were presented 49 scientific papers.

The papers presented at the Conference and published in Proceedings can be useful for teachers while learning and teaching in the fields of informatics, technics and other teaching subjects and activities. Contribution to the science and teaching development in this Region and wider has been achieved in this way.

The ITRO Organizing Committee would like to thank the authors of papers, reviewers and participants in the Conference who have contributed to its tradition and successful realization.

Chairman of the Organizing Committee

Snežana Jokić, Ph.D, Ass. Professor

CONTENTS

INVITED LECTURE

D. Sladić, A. Radulović, M.Zarić, B. Markoski
IMPORTANCE OF LEARNING SOA IN MODERN GIS LECTURES……....2

SCIENTIFIC PAPERS

Ž. Namestovski, A. Buda, G. Molnár, Z. Szűts
SOCIAL ASPECTS OF DISTANCE LEARNING DURING THE COVID-19 PANDEMIC…...9

M. Gaborov, D. Karuović, M. Kavalić, D. Milosavljev, S. Stanisavljev, J. Bushvati
COVID 19 AND ONLINE LEARNING PLATFORMS……...13

M. Majstorović, D. Radosav
DISTANCE LEARNING FROM THE PERSPECTIVE
OF STUDENTS DURING THE COVID-19 PANDEMIC……16

A. Mamić, M. Blagojević, T. Đuričić
ANALYSIS OF LMS USED IN THE PROCESS OF DISTANCE LEARNING
IN PRIMARY EDUCATION, DURING THE COVID 19 PANDEMIC…….20

R. Zamurović, D. Radosav
VIDEO GAMES AS A PROMISING EDUCATIONAL OPTION FOR ALL AGES…….27

E. Karamazova, M. Kocaleva
CASE STUDY: WHICH MATH TOPICS STUDENTS HAVE A
PROBLEM WITH WHEN THEY START UNIVERSITY STUDYING…….34

D. Bikov, B. Shterjev, D. Siracheski
USE OF EDUCATIONAL HARDWARE AND SOFTWARE
TO ENCOURAGE CHILDREN TO CODE……..38

M. Kavalić, M. Pečujlija, S. Stanisavljev, D. Milosavljev, M. Gaborov, M. Bakator
LOCUS OF CONTROL IN THE FUNCTION OF STUDENTS' ACADEMIC SUCCESS…….43

B. Saliu
DISCUSSION THREAD ON GOOGLE CLASSROOM AND GROUP
COMMUNICATION: A CASE STUDY OF LANGUAGE CENTER STUDENTS……48

D. Kreculj, M. Gaborov, N. Ratkovic Kovacevic, V. Nikolic, S. Minic, N. Cvorovic
IMPLEMENTATION OF DRONES IN TEACHING……...53

E. Pavlova Tosheva
THE EVOLUTION OF WEB BASED LEARNING PLATFORMS……60

M. Kocaleva, E. Karamazova, B. Zlatanovska, D. Karuović
MOBILE TEACHING AND LEARNING – BENEFITS,
PERSPECTIVE AND CHALLENGES…….64

G. Škondrić, I. Hamulić, E. Junuz
LMS CONCEPTUAL MODEL THAT RECOGNIZE
ALL FORMS OF LEARNING OUTCOMES……...67

S. Šević, D. Glušac
PEDAGOGICAL DIMENSION OF TEACHING INFORMATICS AND COMPUTING……..70

S. Jokić, V. Srdić, I. Kostovski
THE INFLUENCE OF ETOS ON THE QUALITY OF SCHOOL WORK……..75

C.M. Bande, A.Stojanova, N.Stojkovikj, M.Kocaleva, L.K.Lazarova, B. Zlatanovska
LEARNING DATA MINING COURSE USING LANGUAGE R……...79

N.Stojkovikj, A. Stojanova , L. K. Lazarova, M. Miteva
AGENT-BASED MODELLING AND SIMULATION……87

M. Kocaleva, B. Zlatanovska, E. Karamazova, N. Stojkovikj, A. Stojanova
USING WEKA FOR FINDING OUTPUT FOR GIVEN FUNCTION…….93

A. Mamić, M. Blagojević, T. Đuričić
ANALYSIS OF THE REPRESENTATION OF OBJECT-ORIENTED
PROGRAMMING LANGUAGES IN PRIMARY EDUCATION……97

D. Krstev, A. Krstev, S. Dimitrov
DATA PROCESSING USING ANALYTICAL
HIERARCHICAL PROCESS IN REAL CIRCUMSTANCES……...104

S.Mrđen, E. Brtka, V. Makitan
COMPARISON OF C ++ AND PYTHON
PROGRAMMING LANGUAGES IN TEACHING……108

I. Borjanovic
THE VIRTUAL PHYSICS LABORATORY……..112

S. Jokic, A. Ilic, M. Hadzic, V. Srdić
METHODOLOGICAL APPROACH TO ELECTRICITY
PRODUCTION WITHIN THE FIELD ’RESOURCES AND
PRODUCTION’ IN 8TH GRADE OF PRIMARY SCHOOL……..115

L.K. Lazarova, M.Miteva , A.Stojanova
MODERNIZATION OF MATHEMATICS EDUCATION
BY USING EDUCATIONAL E-PLATFORMS……..121

I. Hamulić, G. Škondrić, E. Junuz
DYNAMIC SOCIAL NETWORK ANALYSIS
VISUALIZATION SOFTWARE: A COMPARATIVE REVIEW……..126

Lj. Kazi, D. Radosav, N. Chotaliya
USABILITY EVALUATION FRAMEWORK FOR WEB PORTALS
OF TECHNICAL SCIENCES HIGHER EDUCATION INSTITUTIONS:
A CASE STUDY WITH SERBIAN STATE UNIVERSITIES……...129

S. Mrđen, E. Brtka, V. Makitan, M. Sisak
EXAMPLE OF AN APPLICATION IN THE PYTHON PROGRAMMING LANGUAGE….135

M. Živić, M. Pardanjac, J. Barbarić
APPLICATION OF 3D PRINTING IN EDUCATION……...139

N. Koceska, S. Koceski
VIRTUAL LABORATORY AS PROGRESSIVE WEB APPLICATION…….142

S. Dimitrov, D. Krstev, A. Krstev
IMPROVEMENT OF THE STATIC CHARACTERISTICS
OF PILOT OPERATED PRESSURE RELIEF VALVES……...147

M. Kocaleva, B. Petrovska, N. Stojkovikj, A. Stojanova, B. Zlatanovska
REVIEW OF SENTINEL-2 APPLICATIONS……155

S. Arsovski, B. Markoski, V. Premcevski, P. Vasiljevic, A. Sofic
REVIEW ON DEEP LEARNING ARCHITECTURES……..160

M. Bakator, D. Radosav. N. Đalić, S. Stanisavljev, D. Milosavljev, E. Terek Stojanović
THE ROLE OF ADVANCED ICTS IN EFFECTIVE CRM……...168

D. Banović, Z. Kazi
ELECTRONIC APPLICATION OF CHILDREN FOR
ENROLLMENT IN PRESCHOOL INSTITUTION……173

T. Milić, I. Berković, E. Brtka, I.Vecštejn, K. Ivanović
THE USE OF WEB TOOLS 2.0 IN EDUCATION…….178

B. Sobota, P. Lovas, Š. Korečko, M. Mattová
VIRTUAL REALITY TECHNOLOGIES USAGE IN THE AREA OF
MANAGEMENT AND THERAPY OF PHOBIAS AND COGNI-TIVE ABILITIES……..182

B. Sobota, M. Mattová, J. Boguščiak, M. Hudák, Š. Korečko
WHEELCHAIR SIMULATOR IN WEB VIRTUAL REALITY……187

S. Stanisavljev, D. Radosav, Z. Košut, S. Jokić, J. Vukajlović, S. Zec
IMPORTANCE OF EMPLOYEE TRAINING FOR INDUSTRY 4.0……192

A. Krstev, A. Velkova Krstev
THE IMPACT OF AUGMENTED REALITY IN ARCHITECTURAL DESIGN USING
COMBINED METHOD OF DATA AGGREGATION AND SEGREGATION……196

D. Krstev, S. Dimitrov, A. Krstev*
VEHICLE ROUTING PROBLEM WITH DISTANCE
CONSTRAINTS AND CLUSTERING USING MATLAB……200

A. Velinov, N. Koceska, S. Koceski
APPLICATION OF THE MQTT PROTOCOL IN TELEPRESENCE ROBOTS……..205

R. Timovski, S. Koceski, N. Koceska
CREATING 3D OBJECTS USING PHOTOGRAMMETRY…….210

M. Gaborov, S. Popov, D. Karuović, D. Radosav, D. Milosavljev, E. Terek-Stojanović
THE APPLICATION OF SCRUM IN COMPANIES:
A SYSTEMATIC LITERATURE REVIEW……...216

M. Knežević, N. Bobinac
DIGITAL MARKETING OF AGRICULTURAL HOLDING IN REPUBLIC OF SERBIA….221

M. Knežević, N. Bobinac
TESTING USING SELENIUM……...224

M. Kavalić, M. Pečujlija, Ž. Stojanović, S. Stanisavljev, M. Bakator
THE EFFECTS OF LOCUS OF CONTROL ON ENTREPRENEURIAL BEHAVIOR……...228

M. Knežević
APPLICATION FOR GRAPE SALES……233

M. Knežević, N. Bobinac
GRAPE VINE PROTECTION RECORD OF AN
AGRICULTURAL HOLDING IN REPUBLIC OF SERBIA…….236

Learning Data Mining Course Using Language R

C.M. Bande*, A.Stojanova*, N.Stojkovikj*, M.Kocaleva*, L.K.Lazarova*, B. Zlatanovska*

*Faculty of computer science, "Goce Delcev" University, Stip, Republic of Macedonia
cveta.martinovska@ugd.edu.mk, aleksandra.stojanova@ugd.edu.mk natasa.stojkovik@ugd.edu.mk,
mirjana.kocaleva@ugd.edu.mk, limonka.lazarova@ugd.edu.mk, biljana.zltanovska@ugd.edu.mk

Abstract - Data mining, also known as knowledge discovery, is a process of discovering patterns and knowledge from large amounts of data, turning raw data into useful information. The data sources can be databases, data warehouses, the web, and other information repositories or data that are streamed into the system dynamically. Data mining course is crucial subject in computer science education. Finding proper tool for learning data mining is important in process in education. In this paper is considered programming language R, as a helping tool in process of learning data mining course. Some examples of machine learning algorithms implemented in R, are given.

I. INTRODUCTION

Data mining is a process which finds useful patterns from large amount of data. Today, we live in a world where vast amounts of data are collected every day. Analyzing such data is an important need. Data mining is a process that turns a large collection of data into knowledge. It makes use of complex mathematical algorithms to study data and then evaluate the possibility of events happening in the future based on the findings. Data mining is also referred to as knowledge discovery of data or KDD [1,2].

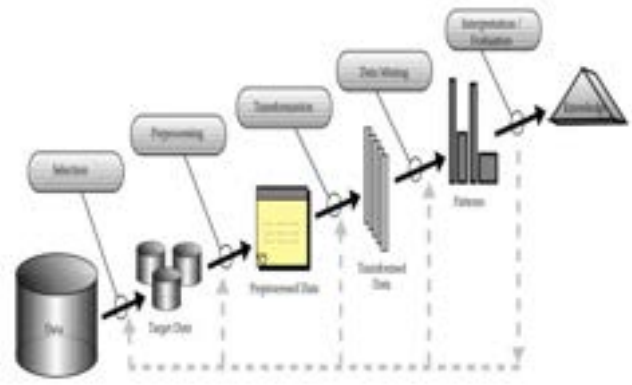


Figure 1. Knwolage discovering process

Data mining is usually used by businesses to draw out specific information from large volumes of data to find solutions to their business problems. It has the capability of transforming raw data into information that can help businesses grow by taking better decisions. Data mining has several types, including pictorial data mining, text mining, social media mining, web mining, and audio and video mining amongst others [1, 3].

Types of data that can be minded are, data stored in the database, data warehouse, transactional data, and other types of data like, data streams, engineering design data, sequence data, graph data, spatial data, multimedia data, and more.

Data mining process can be useful in many areas like, healthcare, market analysis, customer relationship management (CRM), manufacturing engineering, finance and banking, education, etc.

Data mining can be with a great potential to transform the healthcare system completely. It can be used to identify best practices based on data and analytics. Therefore, it can help healthcare facilities to reduce costs and improve patient outcomes. Data mining, along with machine learning, data visualization, statistics, and other techniques can be used to make a difference. Data mining together with different techniques can be used for forecasting patients of different categories. This can help patients to receive intensive care when and where they want it. Data mining can also help healthcare insurers to identify fraudulent activities [4-8].

When data mining is used in area of market analysis it represents modelling technique that uses hypothesis as a basis. Retailers can use this technique to understand the buying habits of their customers. They also can use this information to make changes in the layout of their store and to make shopping a lot easier and less time consuming for customers [1,8].

CRM involves acquiring and keeping customers, improving loyalty, and employing customer-centric strategies. Every business need customer data to analyze it and use the findings in a way that they can build a long-lasting relationship with their customers. Data mining can be useful for them to do that.

A manufacturing companies depend on the data or information available to it. Data mining can help these companies in identifying patterns in processes that are too complex for a human mind to understand. They can identify the relationships that exist between different system-level designing elements, including customer data needs, architecture, and portfolio of products. Data mining can also prove useful in forecasting the overall time required for product development, the cost involved in the process, and the expectations companies can have from the final product.

Bankers can use data mining techniques to solve the baking and financial problems that businesses face by finding out correlations and trends in market costs and business information. This task can be very difficult without data mining as the volume of data that they are dealing with is too large. Managers in the banking and financial sectors can use this information to acquire, retain, and maintain a customer.

Use of data mining in education is still in its nascent phase. Data mining aims to develop techniques that can use data coming out of education environments for knowledge exploration. The purposes that these techniques are expected to serve include studying how educational support impacts students, supporting the future-leaning needs of students, and promoting the science of learning amongst others. Educational institutions can use these techniques to not only predict how students are going to do in examinations but also make accurate decisions. With this knowledge, these institutions can focus more on their teaching pedagogy.

Considering all this mention earlier, data mining is useful process in many areas and its learning is crucial for helping many aspects of life. In this paper we first introduce data mining algorithms and techniques and later we concentrate on learning data mining using language R thru some useful examples [1,8].

II. DATA MINING PROCESS, TECHNIQUES, AND ALGORITHMS

Three steps are involved in process of data mining, these steps are: Exploration, Pattern identification and Deployment. The first step is data exploration. In this step data is cleaned and transformed into another form, and important variables and therefore nature of data based on the problem are determined. When data is explored, refined and defined for the specific variables the second step is performed. Second step is forming pattern identification. Identifying and choosing the patterns is crucial for making the best prediction. And at the end, patterns are deployed for desired outcome.

Before the actual data mining process could occur, there are several processes involved in data mining implementation [1,2 3].

First step is complete research of objectives, available resources, and requirements. That would help create a detailed data mining plan that effectively reaches the set goals. Next step is data quality checks. As the data gets collected from various sources, it needs to be checked and matched to ensure no bottlenecks in the data integration process. The quality assurance helps spot any underlying anomalies in the data, such as missing data interpolation, keeping the data in top-shape before it undergoes mining. The next step is data cleaning. It is believed that 90% of the time gets taken in the selecting, cleaning, formatting, and anonymizing data before mining. The next step is data transformation that can be divided in to 5 sub-stages: data smoothing, data summary, data generalization, data normalization and data attribute construction. In data smoothing process, noise is removed from the data. In data summary process the aggregation od data sets is applied. In data generalization process, the data gets generalized by replacing any low-level data with higher-level conceptualizations. In data normalization process data is defined in set ranges and in the data attribute construction process, the data sets are required to be in the set of attributes before data mining. The next step is data modeling process. Data modeling process is used for better identification of data patterns, several mathematical models are implemented in the dataset, based on several conditions [1,8].

Various algorithms and techniques like Classification, Clustering, Regression, Association, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

Association is one of the most used data mining techniques. In this technique, a transaction and the relationship between its items are used to identify a pattern. This is the reason this technique is also referred to as a relation technique. Association and correlation is usually to find frequent item set findings among large data sets. It can be used to conduct market analysis, which is done to find out

all those products that customers buy together on a regular basis. This technique is very helpful for retailers who can use it to study the buying habits of different customers. Retailers can study sales data of the past and then lookout for products that customers buy together. Then they can put those products in close proximity of each other in their retail stores to help customers save their time and to increase their sales. Association Rule algorithms need to be able to generate rules with confidence values less than one. However, the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little value. Types of association rule are: Multilevel association rule; Multidimensional association rule; Quantitative association rule.

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit-risk applications are particularly well suited to this type of analysis. Classification finds its origins in machine learning. It classifies items or variables in a data set into predefined groups or classes. It uses linear programming, statistics, decision trees, and artificial neural network in data mining, among other techniques. Classification is used to develop software that can be modelled in a way that it becomes capable of classifying items in a data set into different classes. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

There are different types of classification models, the most used ones are: Classification by decision tree induction; Bayesian Classification; Neural Networks; Support Vector Machines (SVM) and Classification Based on Associations.

Clustering can be said as identification of similar classes of objects. This technique creates meaningful object clusters that share the same characteristics. By using clustering techniques, it can be identified dense and sparse regions in object space and can be discovered overall distribution pattern and correlations among data attributes. Clustering is often confused with classification, but if those two techniques are properly understood, it won't have any confuse. Unlike classification that puts objects into predefined classes, clustering puts objects in classes that are defined by it. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification.

Some of types of clustering method are: Partitioning Methods; Hierarchical Agglomerative (divisive) methods; Density based methods; Grid-based methods; Model-based methods.

Prediction as data mining technique predicts the relationship that exists between independent and dependent variables as well as independent variables alone. It can be used to predict future profit depending on the sale, foe example. Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what has to be predicted. Many real-world problems are not simply prediction. For example, sales volumes, stock prices, and product failure rates can be very difficult to predict because they can depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models [5,6,7].

Types of regression methods that usually used are: Linear Regression; Multivariate Linear Regression; Nonlinear Regression; Multivariate Nonlinear Regression.

Neural network presents a set of connected input/output units, and each connection has a weight that corresponds with it. During the learning phase, network learns by adjusting weights, therefore the

network can predict the correct class labels of the input tuples. It represents the connection of a particular machine learning model to an AI-based learning technique. Since it is inspired by the neural multi-layer system found in human anatomy, it represents the working of machine learning models in precision. It can be increasingly complex and therefore needs to be dealt with extreme care. Neural networks could derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs. The most used neural network is Back Propagation.

III. R LANGUAGE

R is a system for statistical analyses, graphics representation and reporting created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team. R is both a software and a language considered as a dialect of the S language created by the AT&T Bell Laboratories. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac [9, 10].

The core of R is an interpreted computer language which allows branching and looping as well as modular programming using functions. R allows integration with the procedures written in the C, C++, .Net, Python or FORTRAN languages for efficiency.

R is available in several forms: the sources (written mainly in C and some routines in Fortran), essentially for Unix and Linux machines, or some pre-compiled binaries for Windows, Linux, and Macintosh. The files needed to install R, either from the sources or from the pre-compiled binaries, are distributed from the internet site of the Comprehensive R Archive Network (CRAN) where the instructions for the installation are also available.

R has many functions for statistical analyses and graphics. The data can be also visualized immediately in their own window and can be saved in various formats (jpg, png, bmp, ps, pdf, emf, pictex, xfig; the available formats may depend on the operating system). The results from a statistical analysis are displayed on the screen, some intermediate results (P-values, regression coefficients, residuals, . . .) can be saved, written in a file, or used in subsequent analyses.

The R language allows the user, for instance, to program loops to successively analyze several data sets. It is also possible to combine in a single program different statistical functions to perform more complex analyses.

R users may benefit from a large number of programs written for S and available on the internet, most of these programs can be used directly with R. R could seem too complex for a non-specialist because of its complexity. A prominent feature of R is its flexibility. Whereas a classical software displays immediately the results of an analysis, R stores these results in an “object”, so that an analysis can be done with no result displayed. The user may be surprised by this, but such a feature is very useful. Indeed, the user can extract only the part of the results which is of interest. For example, if one runs a series of 20 regressions and wants to compare the different regression coefficients, R can display only the estimated coefficients: therefore, the results may take a single line, whereas a classical software could well open 20 results windows.

The following are the important features of R:

- R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.
- R has an effective data handling and storage facility,
- R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
- R provides a large, coherent and integrated collection of tools for data analysis.
- R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.

R is world’s most widely used statistics programming language. It's the number one choice of data scientists and supported by a vibrant and talented community of contributors. R is taught in universities and deployed in mission critical business applications [9,10].

In this paper we are using R language in the process of learning data mining course.

IV. USING LANGUAGE R IN THE PROCESS OF LEARNING DATA MINING COURSE

We will consider only the decision tree algorithms for classification and regression because it has a number of advantages over classical approaches such as linear regression, logistic regression, or linear discriminant analysis.

Advantages of decision trees:

- Decision trees can be very easily explained to people.
- Decisions from decision tree are much closer to human decisions, compared to other classification and regression approaches.
- Trees can be displayed graphically, and easily interpreted by non-experts.
- Trees can easily handle quality predictors without creating dummy variables
- Disadvantages of decision trees:
- Trees usually do not have the same level of accurate prediction as other classification and regression approaches [11-13].

V. BUILDING A REGRESSION TREE

First step is importing data in language R. The "read.csv" function is used to import data into R. After performing the function, the data will be given to the variable. (*movie* in the example, Figure 2)

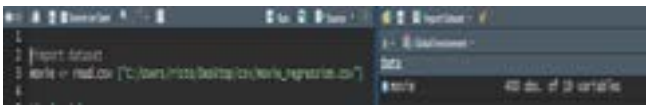


Figure 2.

The "View (name_of_variable)" function is used to view the data. After performing this function, a database will open in a new tab. The next step is data processing.

This step is also called lack of values. Which means if a value is missing in the database, that value must be replaced. This step is critical, because if there is a lack of value in the database, the models cannot be trained, i.e., R will not be able to perform the training functions of the model. This step is a mandatory step.

The summary (name_of_variable) command is performed, this function is used in order to make a summary of the data, i.e., to see if values are missing in any of the variables.

In the summary there are Min, 1rdQu, Meadian, Mean, 3rdQu, Max values. When a variable is missing a value, the NA's value appears.

In the Figure 3 is shown the summary of the variable *movie*. In the variable Twitter_hastgs, NA's with a value of 8 appears, which means that in the variable Twitter_hastgs there are 8 empty fields / values that need to be filled.

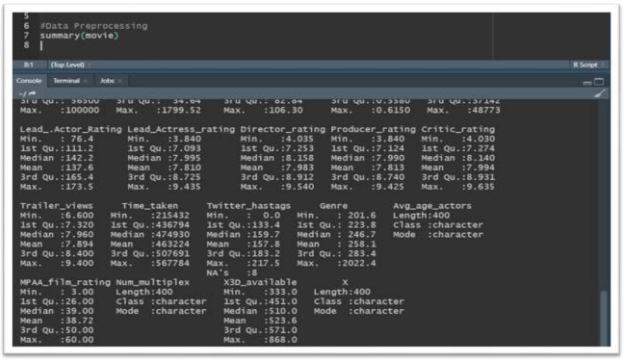


Figure 3.

NA's or empty fields need to be changed to a harmless value, such as Mean or Meadian. Which means the average value of the variable Twitter_hastgs should be found, and that value should be inserted in the empty fields. The following code is used to perform that:

movie$Twitter_hastags[is.na(movie$Twitter_hastags)] <- mean(movie$Twitter_hastags , na.rm = TRUE)

The complete code for importing data set and data preprocessing is given below:

```
1. #import dataset
2. movie <- read.csv ( "C:/Users/risto/Desktop/csv/movie_regression.csv" )
3. View ( movie )
4. #Data Preprocessing
5. summary ( movie )
6. movie$Twitter_hastags[is.na(movie$Twitter_hastags)] <- mean(movie$Twitter_hastags , na.rm = TRUE)
7. summary(movie)
```

The next step is dividing data set into test and training set. This division is done in order to be seen the performance of the previously unknown data model. Usually, the data is shared between 80-20, which means that 80% of the data will be used to train the model, and 20% of the data will be used to test the model. A package called *caTools* is required to perform the division. First the package should be installed with *install.packages('caTools')* command,

and then started with *library(caTools)* command. After that, *set.seed(0)* command is used. By performing the *set.seed(0)* function, it is ensured that everyone will get the same data sharing.

Next, a new variable called split is created. This variable will be created based on the *movie* dataset, which means that it will has the same number of views as the movie dataset. The command SplitRatio = 0.8 sets that 80% of the split variable will have a TRUE value, and the remaining 20% will have a FALSE value.

Whenever the value of the *split* variable is TRUE, which is almost 80% of the time, that value will be placed in the training set, and new variable *train* is obtained, by using the command:

train=subset(movie, split==TRUE)

And when the value of the *split* variable is FALSE, that value will be placed in the test set and new variable *test* is obtained, and new variable *train* is obtained, by using the command:

test=subset(movie, split==FALSE)

The complete code for obtaining training and test set is given below:

```
#Test-Train Split
install.packages ( 'caTools' )
library(caTools)
set.seed(0)
split = sample.split (movie , SplitRatio = 0.8)
train = subset(movie, split == TRUE)
test = subset(movie , split == FALSE)
```

For performing regression in R, first step is building a regression tree. In order to build a regression tree, some packages must first be installed. The *rpart* package is required to create a decision tree. The *rpart.plot* package is required to perform the plot.

The complete code for building regression tree is given below:

```
#install required packages
install.packages( 'rpart' )
install.packages( 'rpart.plot' )
library(rpart)
library(rpart.plot)
#Run regression tree model on train set.
regtree <- rpart ( formula = Collection~. , data = train , control = rpart.control(maxdepth = 3 ))
#Press F1 on rpart for help on this function
#Plot the decision Tree
rpart.plot(regtree, box.palette = "RdBu", digits = -3)
#Predict values at any point
test$pred <- predict(regtree , test , type = "vector")
MSE2 <- mean (( test$pred - test$Collection )^ 2)
```

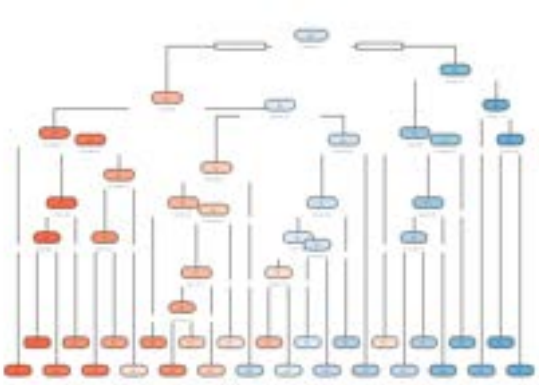
After bulding and ploting a regretion tree, values for other movies or for future whatching are predicted (line 12).

In the next step, the Mean Square Error is calculated. In the next code line is the difference between the predicted values and the actual values, then the difference is squared and the mean value of all those values is found. This value is assigned to the variable MSE2 (line 13).

Large decision trees with many nodes and many divisions are difficult to interpret and overcrowd the training set, leading to poor test set performance. In order not to have such problems, the growth of the tree should be controlled. Therefore, it can be decided before devision to omit certain divisions, and this solution is called tree pruning. When working with pruning, many large trees are used, which are then shaped. The parts of the tree that are not needed are cut. The complete code of tree pruning in R is given below:

```
#Tree Pruning
fulltree <- rpart(formula = Collection~. , data = train , control = rpart.control(cp = 0))
rpart.plot(fulltree, box.palette = "RdBu" , digits = -3)
printcp(fulltree)
plotcp(regtree)
mincp <- regtree$cptable [which.min(regtree$cptable [ , "xerror"]) , "CP" ]
prunedtree <- prune(fulltree, cp = mincp)
rpart.plot(prunedtree, box.palette = "RdBu" , digits = -3)
test$fulltree <- predict(fulltree, test , type = "vector" )
MSE2full <- mean((test$fulltree - test$Collection)^2)
```

The only difference from the bulding a regression tree is the control parameter cp. Since cp = 0, the tree will grow like a normal tree, no pruning will be performed on this tree. And due to this control parameter, the maximum length of the tree will be obtained. Which means that the fulltree variable has total regression without constraints.



The relative error in the tree is changed with the value of cp. The *printcp(fulltree)* and *plotcp(regtree)* commands (line and line 5) need to be performed in order to find the minimum relative error.

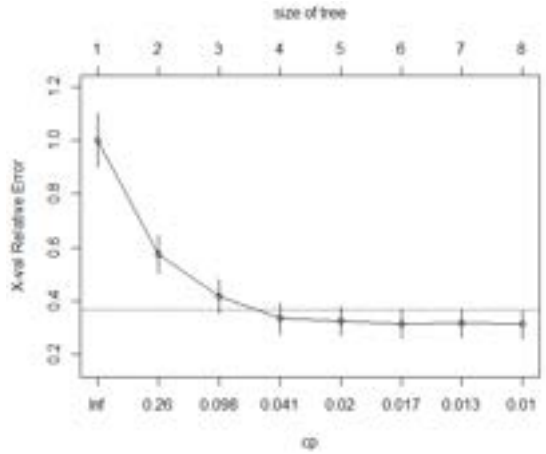
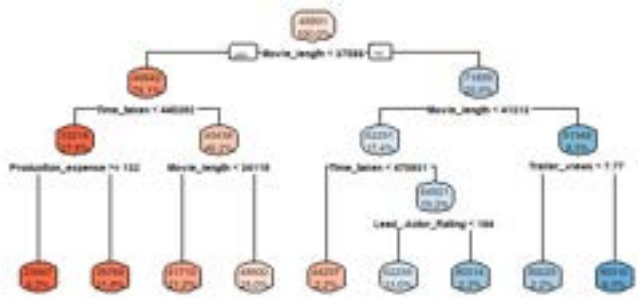


Figure 4. Finding the miniumum relative error in regression tree

In order to find the specific value of cp, which cross-validated error is the minimum, line 6 comand is performed. In this way a new variable mincp is created which has a value of almost 0.01. Using this value,the tree will be shaped (by performing pruning on the fulltree, line 7). When prunedtree is ploted (line 8) a small easy-to-read tree is obtained.



Next step is finding the value of MSE for fulltree, that will be done by using the predicted values (line 9 and line 10).

After obtaining the new variable *MSE2full*, the difference between MSE2, (Mean Square Error of regresion tree) and *MSE2full* (Mean square error of whole tree) can be noticed. A decrease in MSE values can be observed.

The classification tree provides categorical variables, such as whether the film will win an Oscar or not. The output tree in the classification is similar to the tree in the regression, only there are a few small differences in the model and the backend.

In classification trees, *mode* is used to predict the model.

VI. BUILDING A CLASIIFICATION TREE

The same code for regression can be used for classification. Only a few modifications need to be done. Code for Importing data set is the same (only new varieable *moviec* is created and new csv file is used). The next steps are similar, Finding and filling in the missing values, Database division into test and training set, Installing the necessary packages and Building a classification tree.

The function of building a tree is *rpart*. The only change that needs to be made is to mention the *method* (method='class'). All other parameters are the same. A new variable *classtree* is obtained, which has all the necessary information about the classification tree. The *rpart.plot* function is used to plot the tree.

Next step is to predict the values for each position. The difference in this prediction function is the type parameter. Here, instead of the formula being equal to a 'vector' (regression), the formula will be equal to a 'class'. Later a comparison of the performance of the classification decision tree with the predicted values need to be done, and also the overall accuracy need to be checked.

The complete code for makinkg calsiification in R is given below:

```
1. #import dataset
2. moviec <- read.csv("C:/Users/risto/Desktop/csv/classification file .csv", header = TRUE)
3. View(moviec)
4. #Data Preprocessing
5. summary(moviec)
6. moviec$Twitter_hastags[is.na(moviec$Twitter_hastags)] <- mean(moviec$Twitter_hastags,na.rm= TRUE)
7. summary(moviec)
8. #Test-Train Split
```

```
install.packages('caTools')
library(caTools)
set.seed(0)
split = sample.split (moviec , SplitRatio = 0.8)
trainc = subset(moviec, split == TRUE)
testc = subset(moviec , split == FALSE)
#install required packages
install.packages('rpart')
install.packages('rpart.plot')
library(rpart)
library(rpart.plot)
#Run classification tree model on train set.
classtree<- rpart(formula=Start_Tech_Oscar~.,data=trainc,method='class',control=rpart.control(maxdepth = 3))
#Plot the decision Tree
rpart.plot(classtree, box.palette = "RdBu", digits = -3)
#Predict values at any point
testc$pred <- predict(classtree , testc , type = "class")
```

VII. CONCLUSION

Data mining is very useful technique in many areas of life which need processing of large amount of data. Data mining has wide application domain almost in every industry where the data is generated that's why data mining is considered one of the most important frontiers in database and information systems and one of the most promising interdisciplinary developments in Information Technology. Therefore, learning and understanding data mining is crucial to deal with new challenges. In this paper we presented language R as language that could be used for easily learning and understanding data mining course, with giving some examples that represent the basis of data mining.

REFERENCES

[1] Bharati, M., & Ramageri, M. (2010). Data mining techniques and applications.

[2] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[3] Lei-da Chen, T. S., & Frolick, M. N. (2000). Data mining methods, applications, and tools. *Information systems management*, *17*(1), 67-68.

[4] Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.

[5] Tan, P. N., Steinbach, M., & Kumar, V. (2013). Data mining cluster analysis: basic concepts and algorithms. *Introduction to data mining*, 487-533.

[6] Weiss, S. M., & Indurkhya, N. (1998). *Predictive data mining: a practical guide*. Morgan Kaufmann.

[7] Rokach, L., & Maimon, O. (2005). Decision trees. In *Data mining and knowledge discovery handbook* (pp. 165-192). Springer, Boston, MA.

[8] Rohit Sharma (2021) Data Mining Techniques: Types of Data, Methods, Applications (https://www.upgrad.com/blog/data-mining-techniques/#18_Neural_Networks)

[9] Paradis, E. (2005). *R for Beginners* (pp. 37-71). Institut des Sciences de l'Evolution. Université Montpellier II.

[10] R Programming Language – Tutorialspoint https://www.tutorialspoint.com/r/r_tutorial.pdf

[11] Alison : Machine Learning and Decision Trees *https://alison.com/topic/learn/114685/simple-classification-tree#course-plan*

[12] JIGSAW ACADEMY *https://www.jigsawacademy.com/blogs/data-science/decision-tree-in-machine-learning/*

[13] Rstudio *https://www.rstudio.com/products/rstudio/*