JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Exploring Machine Learning in Higher Education: Prediction of Student Performance

Dr. Kumud[1], Dr. Prashant Dixit[2], Prof. Sarvottam Dixit[3]

[1]Assistant Professor, D.S.College, Aligarh, Uttar Pradesh

[2]Assistant Professor, Mewar University, Chittorgarh, Rajasthan

[3]Vice-Chancellor, Mewar University, Chittorgarh, Rajasthan

Abstract:

Machine learning (ML) is used constantly in various industries because its possibility to give innovative and unique solutions to different partners of an organization. ML is utilized in higher education industry also, to provide insights and supporting activities of an educational organization. The higher educational organizations have commonly several data sources, which they can adapt in their activities. These systems provide the raw data, which can be used with machine learning algorithms.

The main aim of this study is to explore the ML in higher education organisations. Moreover, the objective is to provide an example of ML-based project and its implementation utilizing ML project management approach. The CRISP-DM was selected in this. CRISP-Dm is an approach to execute the development task and solve the research questions. Several unsupervised and supervised ML algorithms were used during the research process. The research exists about ML utilization in higher education, but each research is conducting a different type of grants, because of raw datasets and contexts. This study provides a ML-based results related to the VIRTA data and systems.

The use of ML project was a success in overall and the formation of the models were executed in this study. The results indicate that CRISP-DM approach can be adapted in higher education organizations in several course of action. ML provides worth in student performance prediction when the algos are developed based on the needs of an organization and its raw data. The results of this study can be used as well other higher education organizations. However, more research and raw data are needed to make the prediction of student performance more correct. This additional data could be collected from different areas, for instance, learning management, project management, student management, and reporting systems.

Introduction:

Machine learning (ML) is used continuously in different industries because its possibility to provide innovative solutions of an organization. ML is used in higher education institutions to give insights and supporting activities of an educational institution. This scientific field is known as an educational data mining

(EDM) and it is defined to be a field, which is" the area of scientific inquiry centered around the development of methods for making discoveries within the unique type of data that come from educational settings, and using those methods to better understand students and the settings which they learn in."

Research is also needed to develop a successful understanding of the industry's needs. In the light of the disturbing patterns in graduate unemployment we must start thinking, researching, planning, investigating, constructing and implementing some kind of tools to evaluate employability to help us correct shortcomings or improve academic results after graduation [1]. To create a well-organized and structured machine learning model, many professional elements are required. First of all, they need data scientists or expert who have special experience in the industry segment, especially on chosen the right algorithms and tunning the hyperparameters [2]. The main objective of this study is to explore the Machine Learning in higher education organization. Moreover, the aim is to provide a pragmatical example of ML-based project using ML project management approach to give an answer to research questions.

The research is about Machine Learning utilization in higher education organisation, but each analysis is conducting a different type of contribution, because of contexts and datasets. The analysis has three questions, which are answered based on practical implementation of Machine Learning based system:

1. Which ML (supervised) model performs best in student performance prediction?
2. How CRISP-DM model can transform in the higher education organization?
3. Which ML (unsupervised) model performs best in student categorization?

Theoretical Background:

This research provides the review to state-of-the art literature related to ML and EDM in educational factor. ML algos is an integral part of educational data mining and objective of this study is to explore the ML and its utilization in higher educational organization.

Educational Data Mining (EDM):

In the last decades, EDM has attained massive attention from researchers due to the existence of massive educational details which is accessible from many sources. The main aim of EDM is to make DM models more effectively in order to safeguard the numerous amounts of educational information and to develop a protective atmosphere for the student's learning. In this approach, diverse models have been deployed for DM and its analytics (Baker et al., 2014) [03]. Moreover, prediction models were used namely, Classification, Regression, and Latent factor evaluation technologies.

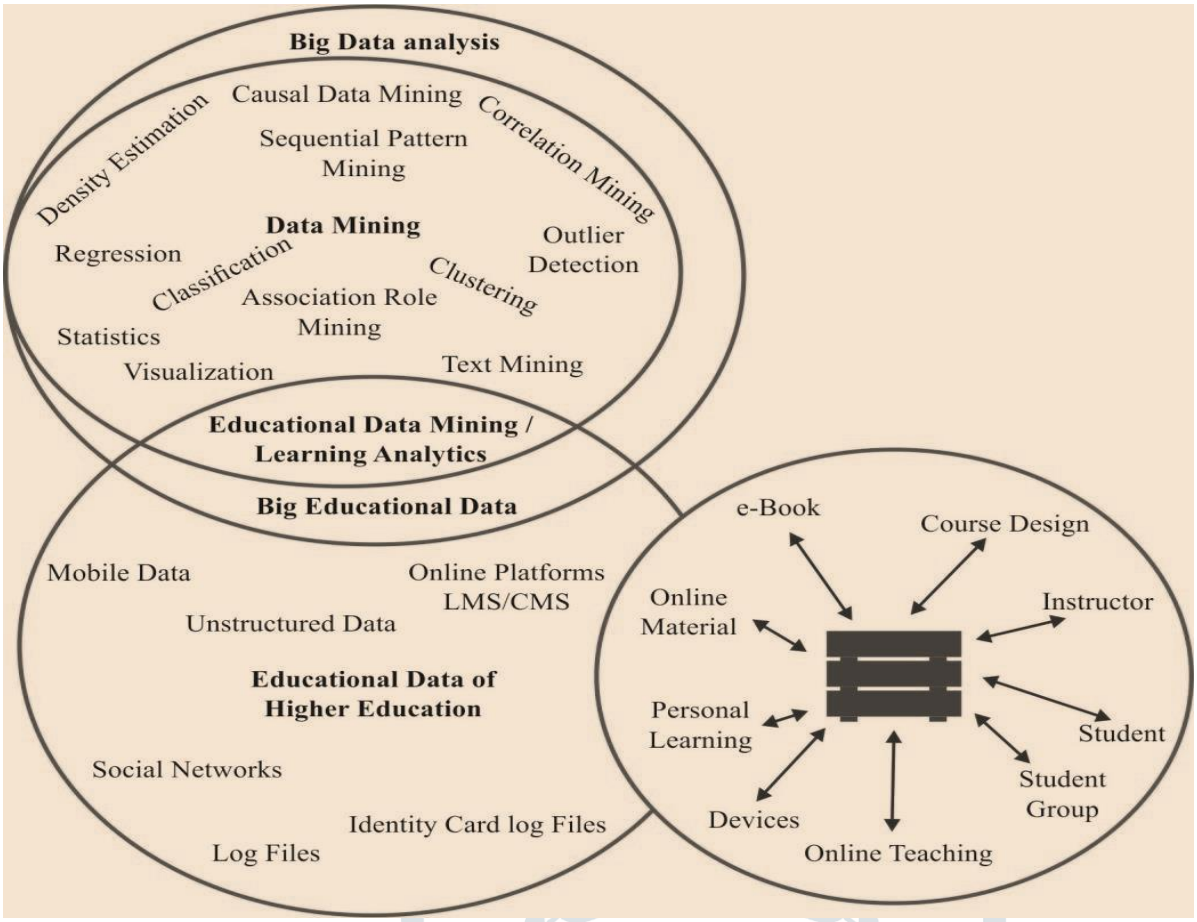

Figure-1: An illustration of Data Mining in Higher Education

The DM tools are collaborated with academics in enhancing the students' learning methodologies by exploring, filtering, and estimating the parameters relevant to student's features or behaviours (Baradwaj *et al.*, 2012) (Figure 1) [04]. The major challenges faced by any educational institutions lies in the number of placements it gets and the number of successful graduates it produces.

Life Cycle of Educational Data Mining:

The application of EDM models is composed of various phases (Figure 2). At the initial phase, the method is developed with the responsibility of identifying essential data. Then, data has been filtered from an accurate educational platform. Subsequently, data has to be pre-processed, as it is aroused from diverse sources with distinct templates and hierarchy levels. The identical patterns are attained while using EDM models which is interpreted. Finally, the simulation outcome recommends that using changes in the teaching process is not an appropriate solution and the analysis is carried out after changing the teaching process previously.

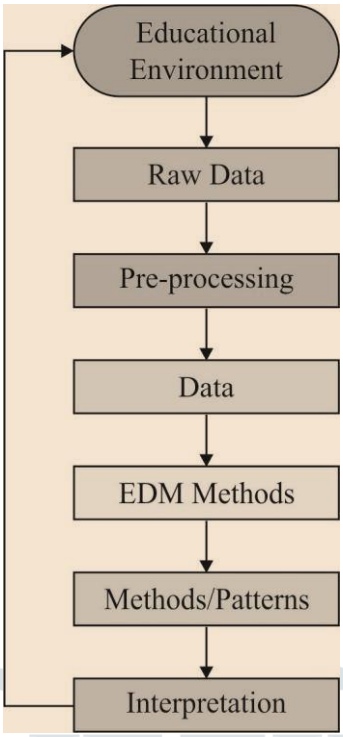

Figure 2: Workflow of Educational Data Mining

Every data point is divided into closest centroid relied on a distance measure. Diverse class labels of the points are represented in various colors. This categorization is not a novel clustering of points as cluster borders are slightly ineffective. Afterward, for all classes, a novel centroid has been re-determined. As the centroids were modified for students who signed for massive times; yet is not applicable in numerous progresses, perhaps the short attention spans. According to the interpretation of 4 clusters, a tutor designs unique learning paths for various students. (Siemens et al., 2012) [05] indicates that EDM is used for automatic identification. Then, EDM's method in classifying the problems into tiny portions as it resolves the problems using ITS. This illustration is extended to detect learning results. Finally, (Figure 3) depicts some of the typical approaches applied in EDM, they are Clustering and Visualization.

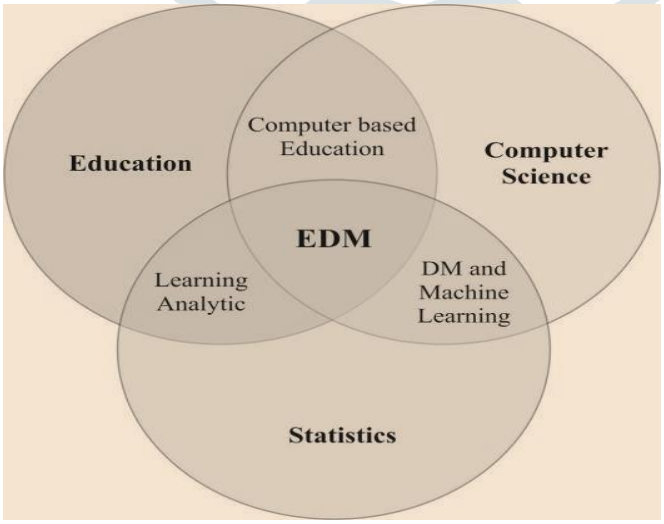

Figure 3: Main areas involved in Educational Data Mining

Machine Learning:

Machine Learning and data modelling require a process and this process commonly follows the same pattern. Everything begins from a decent issue and finish to final version of the Machine Learning model, which can be adapted by different systems and software's. Figure 4 is illustrating a common ML algorithm development process.

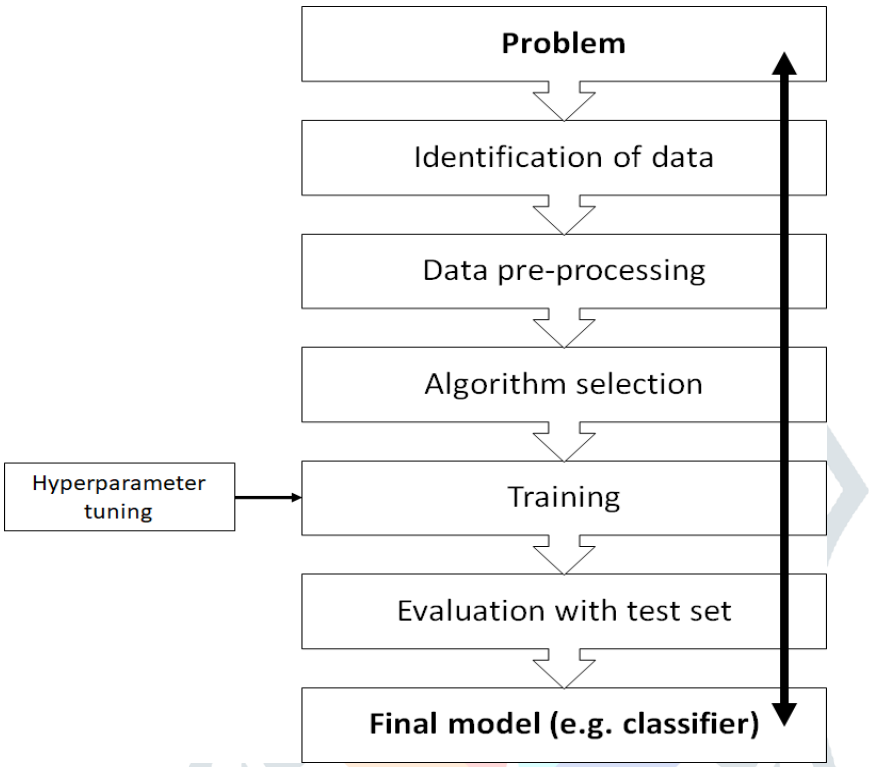

Figure 4: Machine learning algorithm development (Adapted from Ayodele, 2017) [06]

Machine Learning Algorithms:

An ML model is defined as a computer-intensive mechanism and applies re-sampling and iterative methodologies for classification approaches. ML approaches are considered with optimal subset selection and eliminate the issues of classical classifiers like over-fitting as well as distributional demands of parameters. ML technologies that have emerged in computer science with logic and basic mathematics, statistics as ML approaches do not estimate the group features rather it is initialized with an arbitrary group separator and tunes frequently till satisfying the classification groups. ML examines the tuning variables and individual ML functions became unstable, which makes a suitable process. As the non-statistical nature is embedded, these approaches can apply the data in various formats like nominal data that generates maximum classification accuracies.

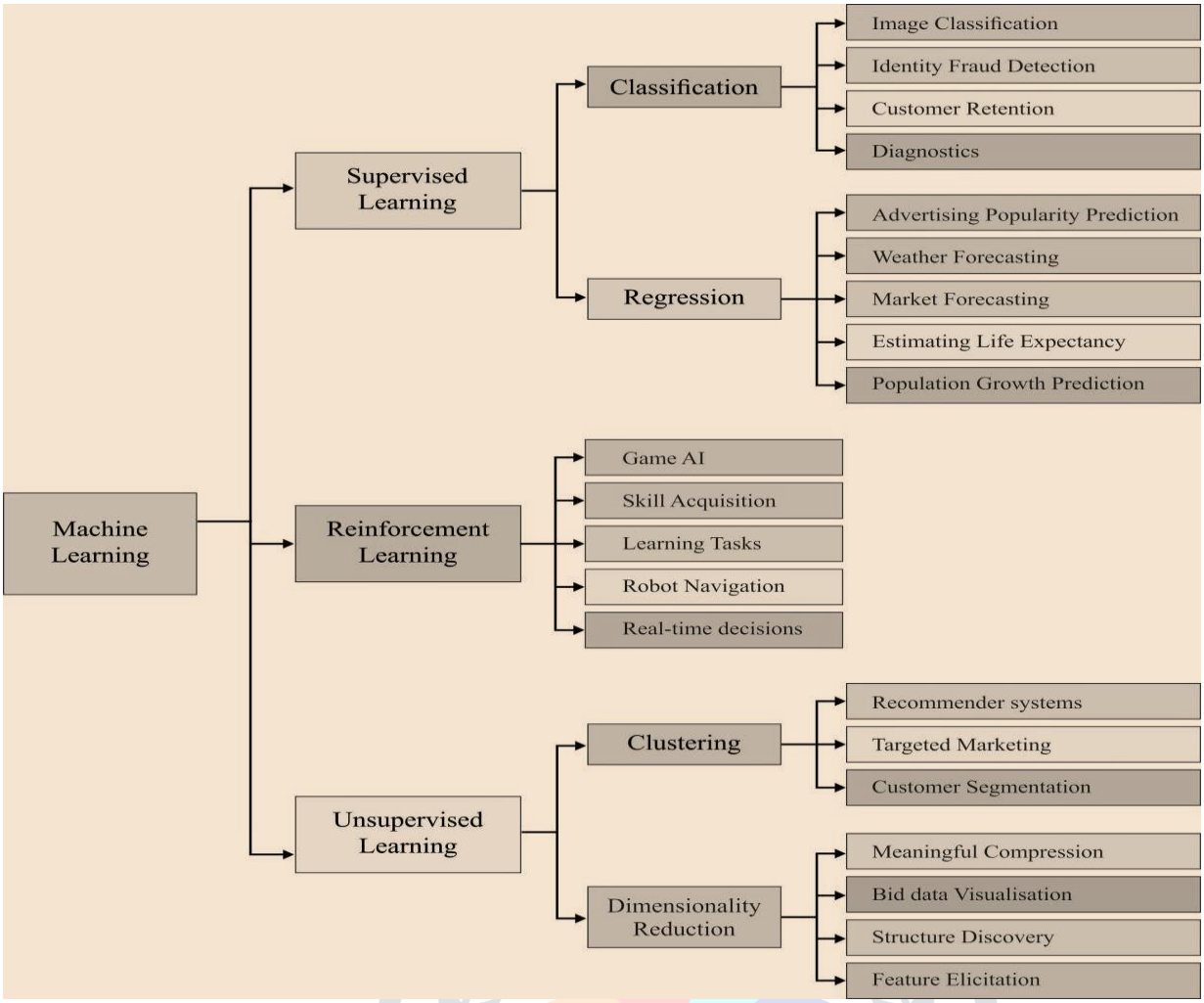

Figure 5: Summary of machine learning algorithms (Adapted from Prashant Dixit, 2022) [07]

Student Performance Prediction with Machine Learning:

ML utilization to predict student performance has been implemented with various systems and datasets by researchers acting in different fields of science. The main objective is to find patterns from the data, which define when student performance is good enough to proceed with the studies and not to dropout while studying. In my previous research I used Decision support system model for students' performance [08] and also used multi-agent system based educational data mining [09] and also used case-based reasoning knowledge-based system (CBR-KBS), this is also done with the help of data mining techniques. [01]. Thakar et al. (2015) have implemented a literature review related to ML utilization in different educational industries, which results table 1 illustrate as appropriate.

Methodology	Key Findings
Decision Tree	Success chances of curriculum estimation by implementing student profiling with storyboard system.
Classification and Clustering	The performance of the final year students.
K-means Clustering	Students' learning behavior analyzation to check the performance of students and predicted weak students.
Classification Tree Models	Enrolment attributes of pre-identify success of students.
Clustering, Association Rules and Decision Trees	Students' academic achievement, students drop out, and students' financial behavior.
Neural Network, Rough Set Theory	Dropout prediction of the course.
Decision Tree	Storyboard (e-learning system) success paths of students.
Decision Tree	High school student's evaluation and studying effectiveness
Decision Tree	Forecasting model for students' marks to identify negative learning habits or behaviors of students.
Association Rule	Student's achievement prediction systematically.
Rough Set Theory	Weak student prediction.

Association Rules, Apriori Algorithm	The success and failure factors of students.
Genetic Algorithm, Novel Spatial Mining Algorithm	Final grade prediction based on features extracted from log data in a web-based system.
J48 and farthest first algorithms	Managerial information on understanding, predicting, and preventing academic failure.
Statistical Approach, Association Rules,	Hidden patterns for students to avoid becoming low performer ones.
Grammar Trees, Regular Expressions	Concept Map Mining (CMM) done from essays written by students to judge their knowledge.
Decision Tree, Rough Set Theory, Naïve Bayes	The produced system generated rules to predict the final grade in a course under study.
Fuzzy Approaches	Fuzzy approaches to classify students' academic performance
Decision Tree	A roadmap for the application of data mining in higher education by pre-identifying weak students.
Genetic Algorithm	Final grade prediction of students based on features extracted from logged data in an education Web-based system.
Decision Tree, Naive Bayes Algorithm, NN	Student performance prediction before the final examination.
Decision Tree, Sota, Naïve Bayes	Comparison of three algorithms in terms of prediction of students result.
Cluster Analysis, NN, Logistic Regression and Decision Tree.	Three techniques comparison for understanding undergraduate's student Enrolment data.
Statistical Classifier, Decision Tree, Rule Induction, Fuzzy, NN	Classifier model for educational use in terms of accuracy and comprehensibility for decision making.
Co-relational Statistics, Multiple Regression	Relationships between the Participants' personality preferences and their employability attributes.
Bayesian Method and Decision Tree Method	Comparison of classification accuracy between two algorithms to evaluate employees' performance.
Decision Trees, NN, SVM, Logistic Regression	Student satisfaction determination for a particular course.
Classifiers, Random Tree, Random Forest	The faculty evaluation based on different parameters.
Decision Tree	Assisted in selecting students for enrollment in a course.
Decision Tree and Clustering	Learning disabilities prediction of school-age children.
Modified Apriori Algorithm	Evaluation index system and teaching index method based on data mining.
Association Rules	The patterns in matching organization and student interests, where they meet each other's requirements.
Data warehouse	Data warehouse to facilitate and provide a thorough analysis of department's data.
OLAP, Data warehouse	Curriculum's establishment analysis from several angles.
Bayesian Network, Decision Forest	Nbtree as the classifiers to predict student sequences for course registration planning
Conceptual	The data mining process in management education and focusing on academic aspects of admission and counseling process.

Table 1. Machine learning in educational industries (Adapted from Thakar et al., 2015)

This study is using CRISP-DM methodology as the main approach in development of Machine Learning based system. Moreover, a survey was implemented to provide an understanding of the requirements of the institution, which supports ML-based system development.

1. Method of Research:

The aim of this research is to provide an understanding of general methodological choices of the study. This research can be concluded as a case-based study: A case-based study is "an empirical inquiry that investigates a contemporary phenomenon in depth and within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident" [11]. This study is a single case study, because a single institution selected and it is acting as well the unit of analysis in this study. The study is utilizing CRISP-DM methodology as the main approach of implementation.

The Case Study:

Mewar University (MU) is located in Rajasthan, India. It has about 5000 students (about 10% students from outside of India) in total, 800 graduating students annually, 200 international degree students, 180 academic staff, 250 international mobility students and 160 other staff in total. It has four different

Departments, which are Business and Management, Engineering and Technology, Law and Agriculture. Vision is **"Reach the unreached."**

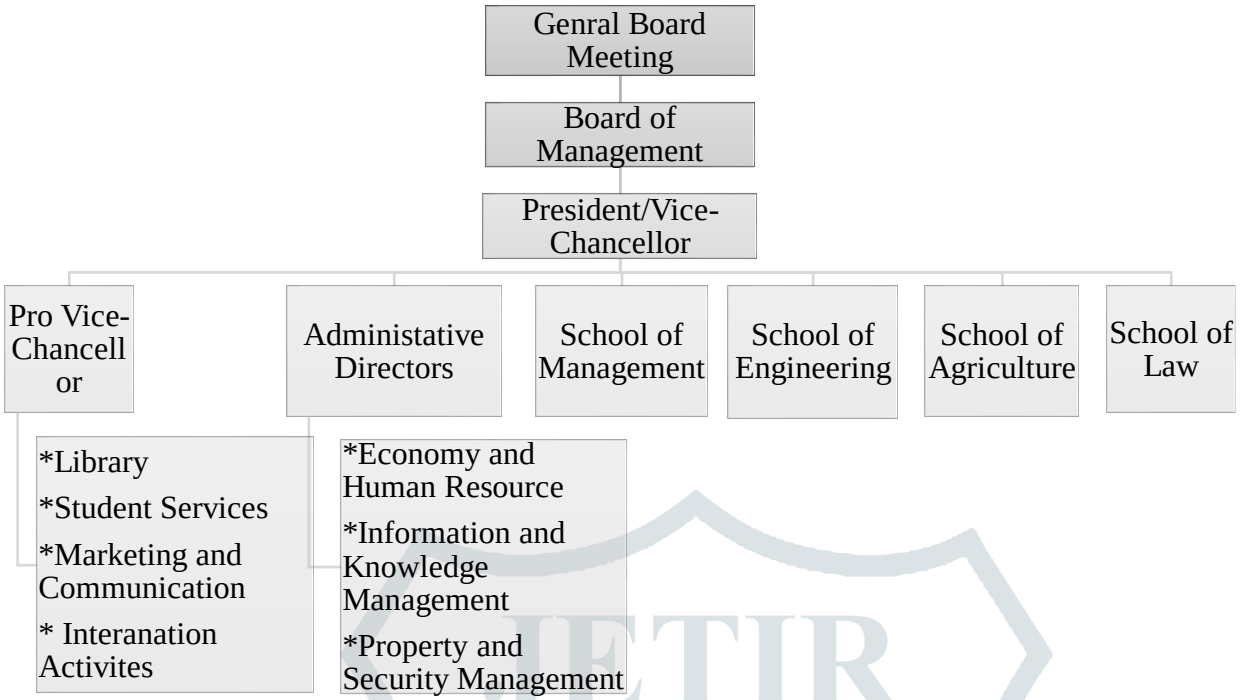

Figure 6: Organotin Chart

CRISP-DM (as an Approach):

The CRoss Industry Standard Process for Data Mining (CRISP-DM) is a process model that serves as the base for a data science process. It has six sequential phases:

1 Business understanding – What does the business requirements?
2 Data understanding – What raw data do we have / need? Is it ok?
3 Data preparation – How do we organize the raw data for modelling?
4 Modelling – What modelling techniques should we use?
5 Evaluation – Which model best meets the business needs?
6 Deployment – How do users access the results?

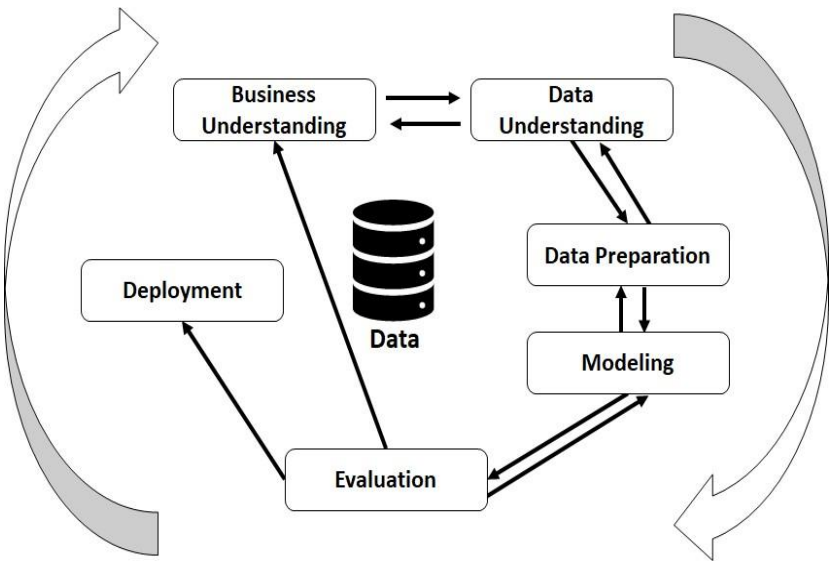

Figure 7: CRISP-DM Model

The CRISP-DM model is based on 6 different phases as figure 7 is illustrating. The process is highlighting the continuous and developing nature of data and Machine Learning. The solution is not completed even though the ML project or process. The solution needs constant renewal and iteration for updates. Moreover, raw data is not commonly static and, thus, requires constant observation for quality. Each phase has its own important role in the CRISP-DM process the following table is presented-

Table 2. CRISP-DM model phases

CRISP-DM model phase	Description of the phase
1. Business understanding	**Determine Business Objectives** • Background • Business objectives • Business success • Criteria **Assess Situation** • Inventory of resources • Requirements, assumptions, and constraints • Risks and contingencies • Terminology • Costs and benefits **Determine Data Mining Goals** • Data mining goals • Data mining success • Criteria **Produce Project Plan** • Project plan • Initial assessment of tools and techniques
2. Data understanding	**Collect Initial Data** • Initial data collection report **Describe Data** • Data description report **Explore Data** • Data exploration report **Verify Data Quality** • Data quality report

3. Data preparation	**Data Set** • Dataset description **Select Data** • Rationale for inclusion or exclusion **Clean Data** • Data cleaning report **Construct Data** • Derived attributes • Generated records **Integrate Data** • Merged data **Format Data** • Reformatted data
4. Modeling	**Select Modeling Technique** • Modeling technique • Modeling assumptions **Generate Test Design** • Test design **Build Model** • Parameter settings • Models • Model description
	Assess Model • Model assessment • Revised parameter settings
5. Evaluation	**Evaluate Results** • Assessment of data mining results (c.f. Business success criteria) • Approved models **Review Process** • Review of process **Determine Next Steps** • List of possible actions • Decision

6. Deployment	**Plan Deployment** • Deployment plan **Plan Monitoring and Maintenance** • Monitoring and maintenance plan **Produce Final Report** • Final report • Final presentation **Review Project** • Experience documentation

The Data Sources and Raw Data Description:

Different data sources and tools are used in higher education industry. Each function of an institution has commonly its own data sources, which is a challenge when developing different ML-based solutions. Thus, data aggregation is required to implement different ML models and analyses. However, some data origins already aggregate the raw data, which can be used in several ways in ML projects (e.g., VIRTA data). The several data sources are illustrated in figure 8-

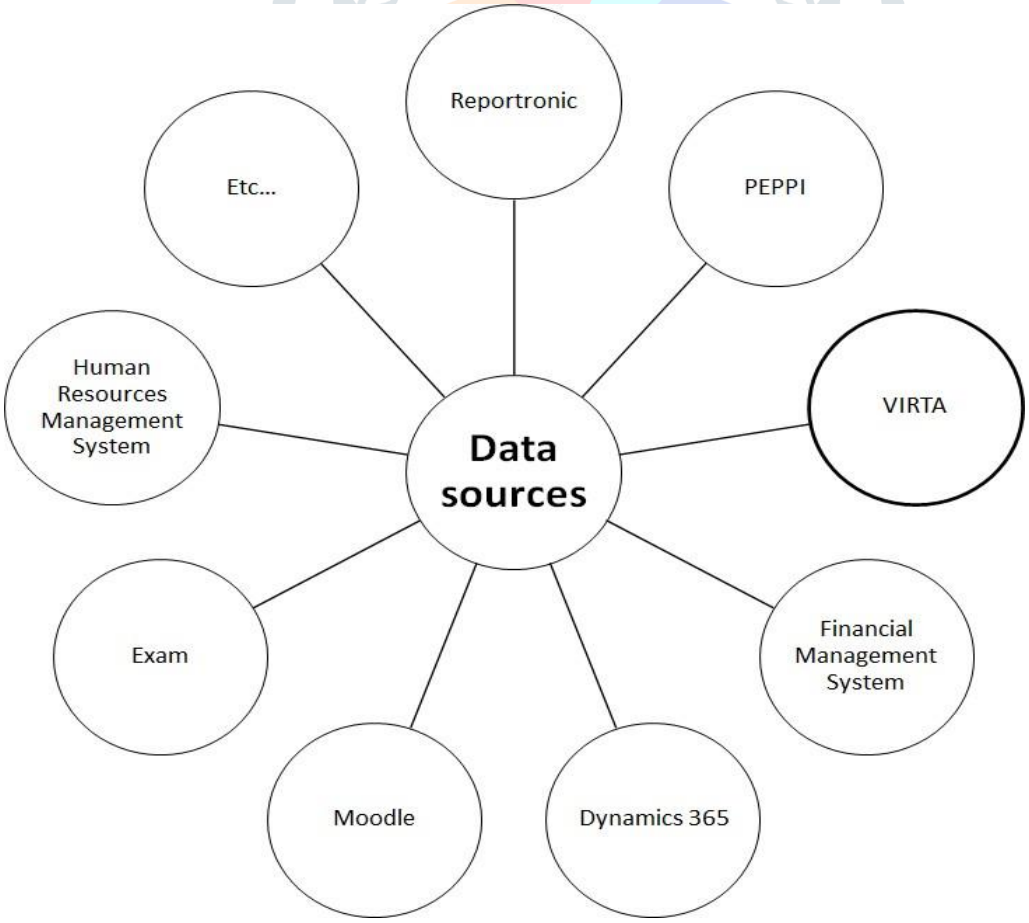

Figure 8: Different Data Sources

In Figure 8 illustrates different data sources in the case. The VIRTA data have an important role in the prediction modelling in Machine Learning in the case organization. However, VIRTA data is not complete, but VIRTA data was utilized in this study in several ways to create the ML models. VIRTA data was selected as the main data source of data in this study for Machine Learning algos since it includes enough relevant information to predict. Moreover, VIRTA data is relatively uncomplicated to aggregate with other datasets in the future if needed.

Process:

This study was carried out as a development process using CRISP-DM. This study process included all relevant steps and the process consist of 2 different main phases: pre-processing phase and Machine Learning development phase. The pre-processing phase is mainly for formatting the raw data for Machine Learning algorithms and Machine Learning development phase are mainly for modelling, evaluation, and deployment. Figure 10 is illustrating the pre-processing phase.

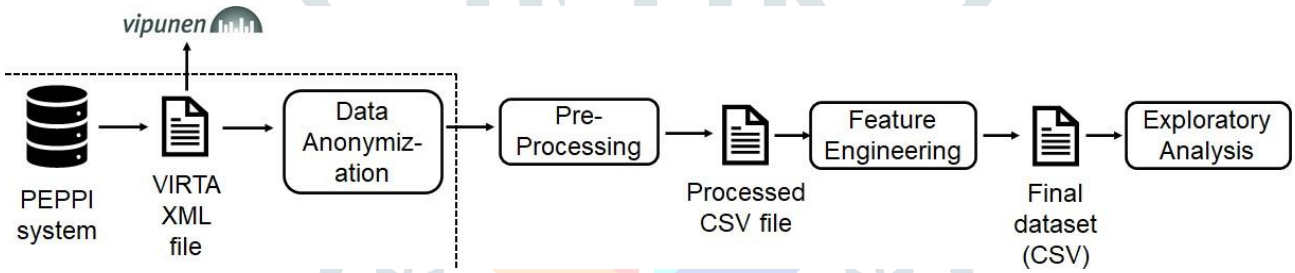

Figure 10. Pre-processing (Phase 1)

The VIRTA data is constructed from the data of PEPPI system (student administration system). The script, which constructs the VIRTA XML dataset is provided by PEPPI consortium and PEPPI student administration system institution are using the same script since the data should be aggregable with other institutions XML files in CSC systems. This same file can be used in data analytics and Machine Learning. The data is anonymized before pre-processing phase and anonymization is implemented by IT department.

VIRTA XML file requires pre-processing and feature engineering to provide the raw data in an appropriate format for Machine Learning algorithms. The pre-processing (Phase 1) was implemented as its own Python algo and feature engineering as a separate algorithm. The exploratory data analysis was implemented after the final feature engineering calculations were finished. There are 2 main data saving points (CSV files) in the process as figure 10 is illustrating. The next step was to proceed Machine Learning development phase after the final data was created Figure 11 is illustrating Machine Learning development phase.

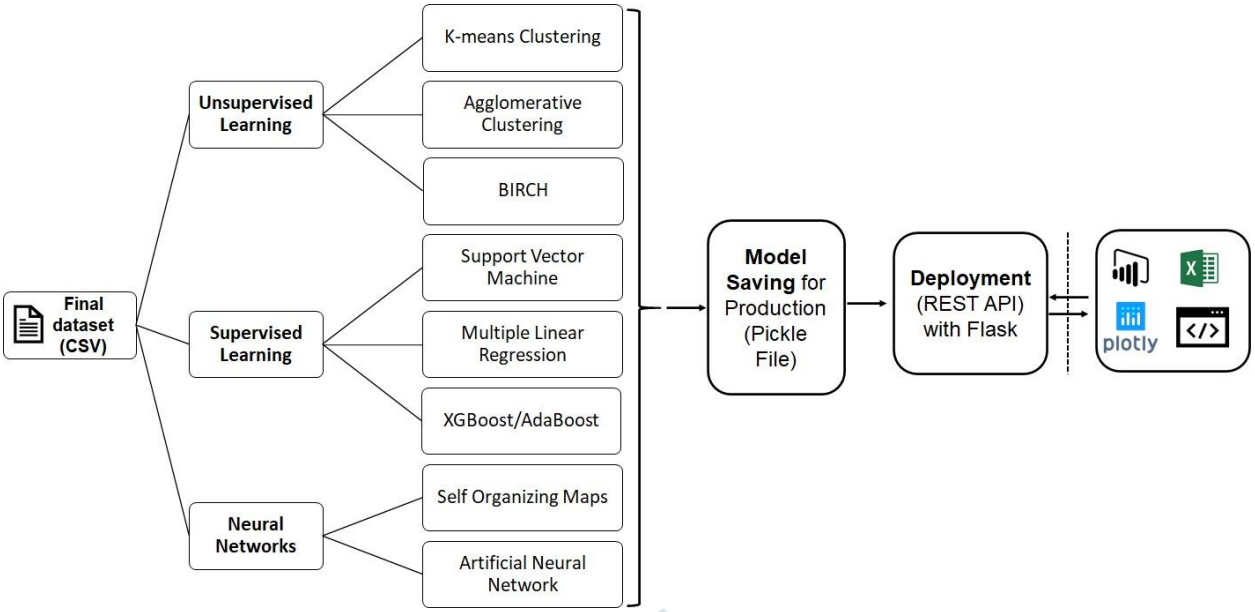

Figure 11: Machine Learning Development (Phase 2)

The end dataset, which was created in previous phase, is used in Machine Learning development phase. It is possible to adapt three different Machine Learning approaches in this study: supervised learning, unsupervised learning, and neural networks. Selected algorithms in the study are based on existing Educational Data Mining and Machine Learning literature. The main aim of unsupervised is to find out the different types of students based on features with clustering. The main objective of supervised learning and neural networks is to prediction of students, who will graduate on time (student performance). On the other hand, the same approach can be used identify students with lower performance.

Both are relevant Machine Learning problems in higher education organisations. Suitable models are picked after the testing and evaluation. Finally, the models are saved to Pickle files, which can be read by the REST API, which will be used as a deployment solution for the Machine Learning models. Different systems and software's (e.g., Power BI and Excel) can use the models in practice with certain parameters using the REST API.

Results of this study:

The implementation of Machine Learning based prediction system is following CRISP-DM approach and formatted on the basis of CRISP-DM. Moreover, this study is providing insights from the results of the survey, which was sent to key stakeholders at MU. The selected programming language in this study is Python (version 3.8.5) and its main data analytics and Machine Learning libraries (e.g., sklearn and pandas).

Weakness and Strength:

It has a good background of raw data and its processing in overall based on the results of the survey. This provides a good base structure for adapting the Machine Learning approaches. However, most of the systems are in their own institutional units are not working in the same manner, which is relatively common in different institutions. This requires combining different raw data sources and as well processes which can gather raw data from the different systems where access is impossible from one reason or another.

Resourcing to Machine Learning and data analytics development is needed in the near future in the this. Resources are needed to these activities now and even more in the future when the field of data analytics

and Machine Learning is developing in the processes. Data gathering and preparation processes needs development, which are related to the students. This development should be mainly fixed to data preparation, pre-processing, feature engineering, and collecting of proper data.

Conclusion:

The objective in this study was to explore Machine Learning in the higher education institution and answer three research questions. The objective of artificial intelligence in higher education organisation is to support activities. Research questions of this study were answered based on practical implementation of Machine Learning based system. In this study was dataset utilized was VIRTA data. Moreover, Python was selected as the programming language for implementing the Machine Learning based system. Selection was natural since several libraries exists to implement, for instance, data pre-processing and Machine Learning modelling, and Python tends to be an industry standard in development of Machine Learning based solutions.

The first question was "Which ML (supervised) model performs best in student performance prediction?". In this study the performance defines student's ability to receive enough ECTS points during a semester and, thus, on the time of graduate. The prediction can be implemented with several algorithms. This study was adapting SVM and XGBoost algos to predict the student performance based on the feature of graduate on time. The SVM algorithm performed much better based on accuracy data. Thus, SVM was selected to deployment of performance prediction of a student.

The second research question was "How CRISP-DM model can transform in the higher education organization?". The CRISP-DM model is a good fit for developing Machine Learning based solutions as well in higher education institution. Thus, the process of CRISP-DM gives all relevant steps needed to implement the Machine Learning based solution based on this study. The data preparation phase requires most of the computation power and time with the VIRTA data used in this study. On the other hand, multidimensional data always needs computation power and time in all environments. Moreover, this phase includes feature engineering, which needs the development of scripts as well to provide target feature based on the background features.

The third question was "Which ML (unsupervised) model performs best in student categorization?". The students can be categorized in different ways and two different unsupervised Machine Learning algorithms were used in this study. This study adapted k-means and agglomerative hierarchical clustering based on earlier studies implemented in the higher education institutions. The clustering was implemented separately for undergraduate and graduate students. The performance of k-means clustering was better based on the study and four different student clusters was found by the algorithm. The k-means clustering was chosen to development phase. K-means clustering algorithm can be used to find the different student types in the case organization.

This study was exploring Machine Learning in higher education institution. The research was a single case

study and, so it should be acknowledged that more research should be carried out to raise the validity of the research. Moreover, this study was adapting a single raw dataset (VIRTA), which does not necessarily include all relevant datasets to implement the prediction as precise as it should. This data comes from a single source and should aggregated with other useful data, for instance. Moreover, collecting data directly from the students using, for example, crowdsourcing could be beneficial (Sivula & Kantola, 2014; Sivula, 2016) [12] [13]. This type of a system could give valuable insights directly from the students, which can be combined with existing raw data from several data sources in higher education organisation.

References:

[1] Prashant Dixit, Dr. Harish Nagar, Dr. Sarvottam Dixit, "Student Performance Prediction Using Case Based Reasoning Knowledge Base System (CBR-KBS) Based Data Mining", International Journal of Information and Education Technology, Vol. 12, No. 1, January 202

[2] Dr. Prashant Dixit, Dr. Kumud, Ankit Upadhyay, "Analysis on Automated Machine Learning Applications", International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.9, Issue 8, page no. b490-b497, August-2022,

[3] Baker, R.S. and Inventado, P.S. (2014). Educational data mining and learning analytics. In Learning analytics Springer, New York, NY, Chap. 4. ISBN: 978-1- 4614-3304-0.

[04] Baradwaj, B.K. and Pal, S. (2012). Mining educational data to analyze students' performance. International Journal of Advanced Computer Science and Applications.,2(6):63-69.

[05] Siemens, G. and Baker, R.S.D. Learning analytics and educational data mining: towards communication and collaboration. International conference on learning analytics and knowledge, pp. 252-254, 2012.

[06] Ayodele, T.O. (2021). Types of Machine Learning Algorithms. In Zhang, Y. New Advances in Machine Learning. InTech: Croatia.

[07] Prashant Dixit, "Psychometric Analysis of Graduate Students using Machine Learning" (2022)

[08] Prashant Dixit, Dr. Harish Nagar, Prof. Sarvottam Dixit, "Decision Support System Model for Student Performance Detection using Machine Learning", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 10 Issue 05, May-2021.

[09] Prashant Dixit, Dr. Harish Nagar, Prof. Sarvottam Dixit, "The Role of Multi-Agent System Based Educational Data Mining Techniques for Student Performance Prediction", Design Engineering, ISSN: 0011-9342 | Year 2021, Issue: 7 | Pages: 4073- 4088.

[10] Thakar, P., Mehta, A., & Manisha. (2015). Performance analysis and prediction in educational data mining: A research travelogue. International Journal of Computer Applications, 110(15), pp. 60-68.

[11] Yin, R. (2009). Case study research: design and methods (4th edition). Sage Publications Inc: California.

[12] Sivula, A. (2016). Generic Crowdsourcing Model for Holistic Innovation Management. Acta Wasaensia, 355.

[13] Sivula, A., & Kantola, J. (2014). Crowdsourcing in a project lifecycle. International Conference on Knowledge Management in Organizations. Cham: Springer.