Analysis and Visualization of performance indicators in university admission tests

Michela Natilli[1,2], Daniele Fadda[1,2], Salvatore Rinzivillo[2], Dino Pedreschi[1], and Federica Licari[3]

[1] Computer Science Department - University of Pisa, Italy
name.surname@di.unipi.it
[2] KDDLab, ISTI-CNR Pisa, Italy name.surname@isti.cnr.it
[3] CISIA - *Consorzio Interuniversitario Sistemi Integrati per l'Accesso*, Pisa, Italy
name.surname@cisiaonline.it

Abstract. This paper presents an analytical platform for evaluation of the performance and anomaly detection of tests for admission to public universities in Italy. Each test is personalized for each student and is composed of a series of questions, classified on different domains (e.g. maths, science, logic, etc.). Since each test is unique for composition, it is crucial to guarantee a similar level of difficulty for all the tests in a session. For this reason, to each question, it is assigned a level of difficulty from a domain expert. Thus, the general difficultness of a test depends on the correct classification of each item. We propose two approaches to detect outliers. A visualization-based approach using dynamic filter and responsive visual widgets. A data mining approach to evaluate the performance of the different questions for five years. We used clustering to group the questions according to a set of performance indicators to provide labeling of the data-driven level of difficulty. The measured level is compared with the *a priori* assigned by experts. The misclassifications are then highlighted to the expert, who will be able to refine the question or the classification. Sequential pattern mining is used to check if biases are present in the composition of the tests and their performance. This analysis is meant to exclude overlaps or direct dependencies among questions. Analyzing co-occurrences we are able to state that the composition of each test is fair and uniform for all the students, even on several sessions. The analytical results are presented to the expert through a visual web application that loads the analytical data and indicators and composes an interactive dashboard. The user may explore the patterns and models extracted by filtering and changing thresholds and analytical parameters.

Keywords: Performance evaluation · University entrance tests · Cluster analysis.

1 Introduction

In this paper we present an analytical process to explore the performances of questions included in the tests submitted to the students applying to several Italian Universities.

We evaluate the performance of each question based on the outcomes of the answers it received within the tests. From these performances we want to highlight outliers and anomalies. We followed two approaches:

- **Visualization-based approach**:
  - Analysis of the distributions of the proportion of right answers for each question in relation to the level of difficulty provided by the domain experts.
  - Analysis of the joint distributions of the proportions of correct, wrong and not given answers in relation to the corresponding difficulty level.
- **Data-mining approach**:
  - Cluster analysis on performance indicators, compared with the rule-based approach.
  - Market basket analysis on co-occurrences of questions within the tests

The analytical tasks listed above were implemented and integrated within a system that supports the users in the exploration of the performance of each question and the detection of anomalous performances of questions. We have designed a process that, starting from every single answer to each question in each test, evaluate a series of indicators (described in section 4.2), performs unsupervised analysis on such aggregations, and visualizes the results on a user-friendly web-based dashboard. The analyst can browse the analytical results by filtering on different dimensions: year, period of the year, topic of the test, discipline of the test. Items classified as anomalies are highlighted and flagged, and they can also be downloaded as *.csv* file for external analysis.

The data is provided by CISIA[4] (Consorzio Interuniversitario Sistemi Integrati per l'Accesso), a non-profit consortium formed by public universities. Currently, CISIA consortium counts 45 Universities and the Conferences of Engineering, Architecture and Sciences, CUIA - the Italian University Architecture Conference, the CopI - Conference for Engineering and Con.Scienze - National Conference of Presidents and Structure Directors University of Science and Technology.

The Consortium is open to the participation of all Italian universities; among the different statutory purposes, the main is to organize and coordinate the orientation activities for the access to the universities. CISIA organizes and provides access to admittance entry tests for students in many universities of the Consortium. For those faculties with a restricted number of admitted students, these tests are used as selection and ranking tools. These tests have two main purposes:

[4] http://www.cisiaonline.it

- for students enrolling the test, they provide a self-assessment of their preparation and aptitude to undertake the chosen discipline of studies;
- for the faculties and departments, the tests give a view of the actual skills and preparation of the students, allowing the management to prepare specific orientation and integrative training activities.

CISIA tests are currently available for six areas: Engineering, Economics, Pharmacy, Sciences, Humanities, and Agriculture.

2 Related work

The measurement and assessment of individual or collective performances are the starting steps to improve the quality of offered services, to enhance professional skills, to assess responsibility for results, integrity and transparency of the actions carried out.

Proper assessment requires a variety of methods; no single approach can test the whole of the performance. Designing assessment programs and selecting the best instruments for each purpose is not easy [1].

Many approaches can be used to design methods for evaluating performance and detecting anomalies [2], starting from *a-priori* defined indicators or using a completely data-driven approach, or a combination of the two. The advantage of the latter is that using one or more indicators it is possible to:

- Overcome personal judgment on measuring the performance
- Create a system that allows confrontation over time
- Construct a system that scales on large numbers

The measures and the approaches to measure performance are, obviously, strictly related to the field of evaluation: when evaluating scientic productivity the focus is on the metric $h - index$ also with all the limitation that this index has [3] or when measuring performance in sports (like soccer) measures like Pass Shot Value (PSV) or PlayeRank [4] have been used.

In the field of performance evaluation using tests, the focus has mainly been on the results of the test (students for an exam, Student Test Scores to Measure Teacher Performance [5]), but for those who build the tools there is the need to evaluate how the test performs or better how the single items that compose the test perform.

In this work, we propose two approaches for identifying anomalies in the behaviors of the different items composing a test.

The first method (the visualization-based approach) has been used as a starting point using simple proportions, and to give also to a non-expert audience the possibility to immediately understand the results. As stated in [6] "The basic idea of visual data exploration is to present the data in some visual form, allowing the human to get insight into the data, draw conclusions, and directly interact with the data". The advantage of this technique is to create a meaningful abstraction of the data, rather than trying to visualize it all at once [7].

The second method, a data-driven approach, using clustering analysis has been chosen to group data into classes with very similar characteristics (i.e. performances), with the scope of identifying groups of questions with anomalous behavior. An implementation of the k-means algorithm (optimized for one-dimensional space) has been chosen given its ability to group items with the same performance in homogeneous groups [8,9,16]. At the same time, the possibility of having combinations of questions with the same outcome was tested using a market basket analysis algorithm. The intuition behind this choice is that if two or more questions compare together and have the same result (right, wrong or not answered question), they probably measure the same "skill": the extraction of these rules can also help in identifying strange behaviors in the questions. Generally, this algorithm is mainly used for transactional data (i.e. the supermarket register) to identify set(s) of items purchased together [10], but it can be successfully used also on different kind of data (i.e. crash data [11]).

The results deriving from these analyzes have all been reported on a visual dashboard, to obtain an exhaustive and quick overview of the results obtained, simplify the interpretative work by parts of the domain experts and allow comparisons between different areas and different years. Through visualization, in fact, the results of data processing are made more accessible, straightforward, and user-friendly [12]. The choice of a dashboard is supported by the fact that, as stated in [13], "compared to visualization modalities for presentation and exploration, dashboards bring together challenges of at-a-glance reading, coordinated views, tracking data and both private and shared awareness". Furthermore, the integration of data mining and information visualization techniques has received a lot of attention, given its ability to filter and extract valuable patterns and to provide a better understanding of the final results [14].

3 Problem statement

CISIA Online Test (acronym TOLC) is a tool for orientation and assessment of the knowledge required for access to the Study Programs of Italian Universities, which can be used to select students for access. TOLC is an individual test, which is different from student to student, automatically composed for each student by a software. The software follows a set of *rules* (defined *a priori* by CISIA experts) to guarantee that all the tests generated are equivalent in terms of the level of difficulty. This means that in each TOLC there are a series of questions on different subjects with different level of difficulty. Thus it is crucial to have tests with comparable difficulties. CISIA has developed a methodology to provide a human-based classification of difficulty levels for each question and they exploit such labeling to compose equivalent individual tests.

The objective of our system is to provide an inspection platform where analysts may evaluate the labeling and the behavior of each question. The performance of a question is the outcomes of the answers of the students in terms of the number of correct or wrong answers. When a student has doubts for a specific question, she can decide to provide no solution: a missing answer has a

small penalty in the final grade, but there is a higher penalty in case of a wrong answer.

The basic strategy consists in the exploitation of the *a priori* level of difficulty of a question to define an expected performance: questions classified as "easy" should have a high proportion of right answers, while questions labeled as "difficult" should have a higher proportion of wrong answers.

The analytical system should automatically ingest the answers of the students and evaluate the classification of the level of difficulty of each question. The results of this analysis are made available with a visual interface to explore the performances of every single question during the time.

4 Analytical process

The analytical process is organized in two macro steps (see Figure 1): first, data is collected, aggregated and analyzed; secondly, the results are organized and optimized for fast interaction and visualization.

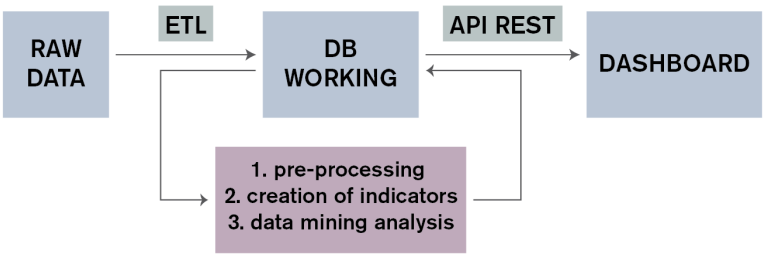

Fig. 1. The schema of the process.

4.1 Data loading and indicators extraction

The ETL (Extraction, Transformation, Load) phase is designed to incrementally update the performance indicators described below. Starting from the raw data (first box in Figure 1), the answers to each question in each test are collected and saved in a "working database". These data do not arrive in real-time since CISIA performs internal checks and assessment. Regularly, we can consider an update every week. The results of the tests are saved into a working area within a DBMS , where the analytical process is executed . At the moment of writing, data are related to the last six years (2014-2018). Table 1 reports the number of tests taken by students in different disciplines.

Table 1. Number of tests administered online.

	2014	2015	2016	2017	2018
biology					7.259
economics	5.144	10.382	14.365	21.463	33.184
pharmacy				3.871	6.706
engineering	16.526	30.048	35.981	51.013	55.449
science					13.748

4.2 Performance indicators

To have a data-driven criterion to measure the performance of every single question, we defined a series of indicators that summarized the performance of the questions in terms of correct, incorrect and not given answers. To represent the three possible outcomes for each question, we defined a new attribute, namely $R3$, which get values -1, 1, or 0, respectively for a wrong answer, a right answer, a not-given answer. From this attribute, we derive three new indicators: PR, the proportion of correct answers[5]; PW, the proportion of wrong answers[6]; PNA, the proportion of not answered questions[7]. The attributes have been calculated for each year and for each type of TOLC (e.g. engineering, economics, etc.).

We also define a series of derivative indicator, computed based on the previous ones. The first indicator *Perf1* provides a measure of the performance of the answers given, ignoring the cases when no answer was given.

$$Perf1 = \frac{sum(R3)}{count(R3 = -1) + count(R3 = 1)}$$

The second performance indicator *Perf2*, instead, takes into account the answers not given. This value is always less than or equal to $Perf1$.

$$Perf2 = \frac{sum(R3)}{count(R3)}$$

By introducing a simplification of the R3 attribute into two levels (naming it R2, where $R2 = 1$ if the answer is correct, while $R2 = 0$ if the answer is wrong or not given) it is possible to obtain an additional performance indicator.

$$PerfR2 = \frac{sum(R2)}{count(R2)}$$

The last performance indicator gives equal weight both to wrong and to not given answer.

[5] $PR = count(R3 = 1)/count(R3)$
[6] $PW = count(R3 = -1)/count(R3)$
[7] $PNA = count(R3 = 0)/count(R3)$

5 Performance evaluation through anomaly detection

Two different methods have been developed to highlight anomaly performance behaviors. The first method exploits visualization technique to compare outliers with the expected performance of questions on the basis of the level of difficulty: easy questions should have a more significant proportion of right answers. The second method uses data mining methods to identify groups of questions with similar performance and then compare these with the classification applied by the experts. In both approaches, the objective is to highlight those question whose classification needs to be revised.

5.1 Visualization-based anomaly detection

The visual approach we propose is based on the visualization of the expected behavior of each question based on the level of difficulty. In collaboration with the domain experts, we have identified a set of intervals for each level of difficulty. Figure 2 shows the values of PR for which an anomalous behavior should be highlighted. For example, an easy question with a low value of PR should be inspected to check if it should be classified as more difficult.

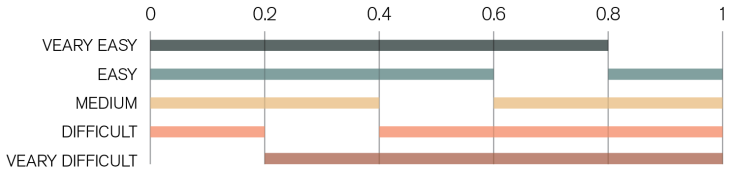

Fig. 2. Anomaly detection using the values of the PR indicator

This approach has a significant limitation: it considers mainly the PR indicator. However, difficult questions may produce two different behaviors: a high proportion of wrong answers and not-answers. To overcome this problem, we introduce a visualization based on *Ternary Plots*. This visualization allows a very effective representation of the behavior of each question, with a concurrent comparison of three indicators. This is a visual chart used mainly in geology to present proportions of soils or terrains.

This visualization is based on visual space determined by three axes: we map our indicators (PW, PR, PNA) to each axis. The triangle that is defined by these axes contains those points whose sum of values is constant (in our case 1). Each attribute interval is represented on one side of the triangle. Figure 3(a) shows an example of a visual representation of a set of questions. Each point is located accordingly to its indicator values and its color represents the level of difficulty assigned *a priori*. The vertexes of the triangle are annotated with the label of the three indicators: those points closer to one vertex have a high value of the corresponding indicator. In the example in figure 3 (a), we can notice a dark red

point (meaning a difficult question) with a very high proportion of right answers (it is close to vertex PR): this question should be checked to verify the correct classification.

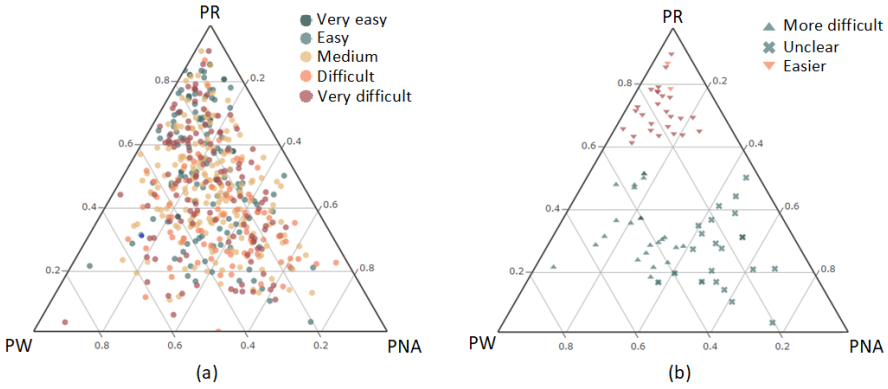

Fig. 3. Joint distribution of PR, PW and PNA (a) and anomaly detection (b) using a ternary plot

Each vertex, therefore, corresponds to the value 1 of a variable and the 0 of the other variables. To know the values of "Correct", "Wrong" and "Not given" relating to any point of the ternary diagram, it is necessary to draw from this point 3 lines that are parallel to the 3 sides of the triangle: the intersections of these lines with the sides of the triangle provide the values sought for each of the three variables.

When the number of questions is high, it may be difficult to identify anomalous points based on the color and position. Thus, we have developed a filter interface to visually highlight relevant points on the basis of a set of rules. These rules take into account the distribution of the three indicators for all the questions. For example, Figure 3(b) shows an example of detection that follows the following rules:

1. If the question was classified as "very easy" or "easy" and the value of the PW (wrong) is greater than the value corresponding to 75^{th} percentile of the distribution of the PW (right-tail of the distribution) then the question could be more difficult;
2. If the question was classified as "very easy" or "easy" and the value of the PNA is greater than the value corresponding to the75^{th} percentile of the distribution of the PNAs (right-tail of distribution) then the question may be unclear;
3. If the application was classified as "very difficult" or "difficult" and the PR value is greater than the value corresponding to the 75^{th} percentile of the PR distribution (right-tail distribution) then the question could be easier

The visual interface allows to dynamically change the value of the percentile threshold (as described in details in Section 7). The result of the filter is represented visually with the same color schema (mapping the level of difficulty) and with a new set of symbols:

↑ up-arrow: the question probably should be classified as more difficult;
↓ down-arrow: the question probably should be classified as more easier;
× cross: the question is not very clear.

5.2 Data mining anomaly detection

In this section, we present a data-driven approach to explore the whole dataset of answers, without relying on the definition of thresholds from the analysts or domain experts.

Cluster analysis We exploit cluster analysis to group questions with similar performance into clusters. We adopt *k-means*[15], a partitioning cluster algorithm that allows subdividing a set of objects into k groups based on their attributes. A centroid or midpoint identify each cluster. The algorithm follows an iterative procedure. Initially, it creates k partitions and assigns the entry points to each partition either randomly or using some heuristic information. Then calculate the centroid of each group. It then constructs a new partition by associating each entry point with the cluster whose centroid is closest to it. Then the centroids for the new clusters are recalculated and so on until the algorithm converges. Each question within the clustering is represented as a combination of the *Perf* indicators defined in Section 4.2. Given the possibility to the analyst to focus on one of the indicators at a time, we used an optimized implementation of *k-means* for uni-dimensional points: *ck-means* [16]. This algorithm performs better in the case in which each object has a single attribute[8]. We tested both algorithms and we stated that their performances are comparable for our case study. In our final implementation we adopted the *ck-means* algorithm. The k was chosen using the elbow method: a series of clustering runs on the dataset for a range of values of k (k from 1 to 20), and for each value of k the sum of squared errors (SSE) was calculated. According to the SSE distribution, we set $k = 5$. This number of clustering also allows an indirect comparison with the level of difficulties of the questions (see Section 7 for an example).

Pattern mining analysis Pattern mining analysis is a technique of analysis used primarily in marketing that analyzes the buying habits of customers in retail sales, finding associations on different products purchased, to obtain rules of association between products purchased together. In our domain, we use frequent items analysis [10] to verify how frequently questions with similar behavior in

[8] We used the Python implementation published in https://github.com/llimllib/ckmeans

terms of answers occur together. We want to check if in the composition of each test there is bias and two different questions repeatedly occur in many tests.

In the proposed application, a test can be seen as a transaction (a basket of goods) composed of many items (questions) and it would be analyzed to search if a particular combination tends to co-exist. We are interested only in the extraction of the frequent itemsets and in the verification that their support is below a statistical expected probability of co-occurrence. From the analysis, we assessed that there no itemset overly represented. Thus the construction of the tests (in terms of the composition of questions) is done in a fair manner.

6 Visual dashboard

All the analytical processes were organized into a visual dashboard, where the domain experts can formulate a hypothesis and dynamically explore the dataset through a set of filters, to be able to identify anomalies in the performance of the questions. The filter allows selecting specific subdimensions according to year, the period of the year, disciplines, topics. Figure 4 shows a schematic organization of the section with a description of the actual web application.

Fig. 4. The web-page schema.

The filters are organized in the top of the window and they are always visible to show the current active selection. A first section presents the various distribution of the data (performance indicators, distributions over year, trends of indicators, etc.)

The second part of the dashboard presents the interface for the outlier analysis, using both the visualization-based approach or the data-driven approach. The data-driven approach contains a cross-table to compare the results from the clustering analysis with the labels assigned *a priori* to the questions: in the diagonal of the matrix there are the questions that have a behavior similar to the expected one, while in the corners of the matrix the anomalies are present. It is possible to select, from a drop-down menu, the performance indicator to be used for the cluster analysis. For both sections, a selection of a set of outliers also produces an analytical table with all the attributes of the selected questions, with the possibility to download a *.csv* table for further investigations.

7 Test case

We describe here a typical analytical task that can be performed on the platform. We omit the discussion of the visual exploration of the distributions and trends and focus on the outlier detection task.

The second section of the dashboard has two tools dedicated to this analysis: a visual-based approach and a cluster-based method. We present here a case study using both methods. Figure 5 shows the resulting ternary plots after the commit of a filter. On the right, we selected three percentile thresholds for PR, PW, and PNA: the value of the 75th percentile of PR is used as a threshold to select those on the right side of the distribution. Since these questions are classified as easy (green color of the markers) and they have a large proportion of correct answers, the system suggests to check these questions to increase their level of difficulty. In the example, the question with ID 1234 (the id as been obfuscated to protect the original data, the indicators are real) has a PR value very low: this question should be classified as difficult, for example.

We repeat a similar analysis using cluster analysis. Figure 5 (left) shows the result of the selection of the cell in the cross table corresponding to cluster *0* and level of difficulty *very easy*. The selection highlight six questions. Among these six questions, there is the same question with ID 1234 that we discovered before.

It worth noting how the two approaches yield to different (but comparable) results. By comparing the two groups of points in the two charts, there is a subset of questions (4 questions in the central part of each ternary plot) that are in common in the two selections, but there are different questions in the remaining parts of the two charts. This is due to the fact that the first method (on the right) takes into account only the PR indicator, while the other method (on the left) exploit the *Perf2* indicator.

The choice between the two methods depends on the specific need of the experts: using the visual approach the performance of a question is seen through

an index at a time (PR, PW or PNA) while, using cluster analysis, we work on a composite indicator that takes into account the three proportions together.

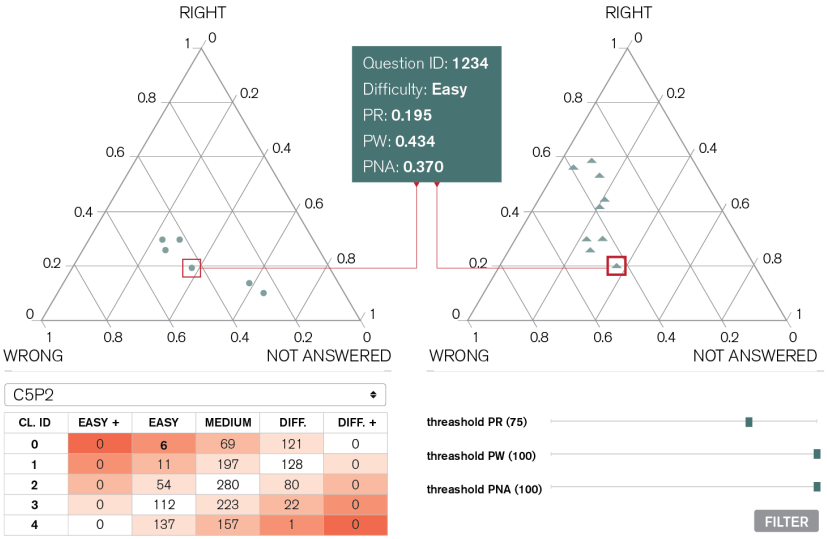

Fig. 5. An example of outlier detection using cluster analysis (left) and visualization-based approach (right).

8 Conclusion

In this paper we presented an analytical platform to evaluate the performance and anomaly detection of tests for admission to public universities in Italy. The process of analysis followed two different approaches: a visualization-based approach, where a set of rules provided by the domain experts are represented to create a visual highlight of candidate outliers; a data-driven approach where a clustering-based method is used to partition the set of questions into groups to be compared with the *a priori* classification of the level of difficulties.

The analytical results are made available to the users through a dynamic dashboard, where the user may set a filter to explore subdimension of the data, accordingly to the values of the year, the period of the year, the discipline and the topic.

The analytical tool is already deployed within CISIA consortium and it improved the inspection on the questions by enabling new detection mechanism, both the visual-based and the data-driven one. The system is being extended with specific analysis on the subtopics (for example considering "physics of fluid" rather than the general topic "physics").

References

1. Schuwirth, L. W., van der Vleuten, C. P.: How to design a useful test: the principles of assessment. Understanding Medical Education: Evidence, Theory, and Practice, 275-289 (2018).
2. Agrawal, S., Agrawal, J.: Survey on anomaly detection using data mining techniques. Procedia Computer Science, 60, 708-713, (2015).
3. Dorogovtsev, S., Mendes, J.F.: Ranking scientists. Nature Physics. 11. 10.1038 (2015).
4. Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., Giannotti, F.: PlayeRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach (2018).
5. Ballou, D., Springer, M. G.: Using student test scores to measure teacher performance: Some problems in the design and implementation of evaluation systems. Educational Researcher, 44(2), 77-86 (2015).
6. Keim, D. A. (2002). Information visualization and visual data mining. IEEE transactions on Visualization and Computer Graphics, 8(1), 1-8.
7. Paul, C. L., Rohrer, R., Nebesh, B.:. A "Design First" Approach to Visualization Innovation. IEEE computer graphics and applications, 35(1), 12-18, (2015).
8. Sarker, A., Shamim, S. M., Zama, M. S., Rahman, M. M.: Employees Performance Analysis and Prediction using K-Means Clustering & Decision Tree Algorithm. Global Journal of Computer Science and Technology (2018).
9. Lakshmi, T. M., Martin, A., Begum, R. M., Venkatesan, V. P.: An analysis on performance of decision tree algorithms using student's qualitative data. International Journal of Modern Education and Computer Science, 5(5), 18 (2013).
10. Agrawal, R., Imieliski, T., Swami, A.: Mining association rules between sets of items in large databases. In Acm sigmod record (Vol. 22, No. 2, pp. 207-216). ACM (1993).
11. Pande, A., Abdel-Aty, M.: Market basket analysis of crash data from large jurisdictions and its potential as a decision support tool. Safety science, 47(1), 145-154 (2009).
12. Tao, F., Qi, Q., Liu, A., Kusiak, A.: Data-driven smart manufacturing. Journal of Manufacturing Systems **48**, 157–169 (2018)
13. Sarikaya, A., Correll, M., Bartram, L., Tory, M., Fisher, D.: What Do We Talk About When We Talk About Dashboards?. IEEE transactions on visualization and computer graphics, 25(1), 682-692 (2019).
14. Shneiderman, B.: Inventing discovery tools: combining information visualization with data mining. Information visualization, 1(1), 5-12 (2002).
15. Rokach, L., Maimon, O.: Clustering methods. In Data mining and knowledge discovery handbook (pp. 321-352). Springer, Boston, MA (2005).
16. Wang, H., Song, M.: Ckmeans. 1d. dp: optimal k-means clustering in one dimension by dynamic programming. The R journal, 3(2), 29 (2011).