Statistical Data Mining Methods in Predicting Happiness and Habits

Sazan Kamal Sulaiman[1*], *Yousif* Sufyan Jghef[1], and *Abdulqadir* Ismail Abdullah[1], *Saadaldeen* Rashid Ahmed[2,3]

[1]College of Engineering, Department of Computer engineering, Knowledge university, Erbil, Iraq
[2]Computer Science Department, Bayan University, Erbil, Kurdistan, Iraq
[3]Artificial Intelligence Engineering Department, College of Engineering, Al-Ayen University, Thi-Qar, Iraq

Abstract. The objective of this study is to employ statistical data mining methods and con-duct a survey among young individuals to construct a model capable of forecasting overall happiness. This model will consider over a hundred characteristics, including lifestyle choices and musical tastes. We utilized boosting trees, subset se-lection, and GAM (Generalized Additive Models) techniques. In addition, we created actual test data to validate the model. All available approaches have found many lifestyle variables, including as energy levels, loneliness, desire to alter the past, eating properly, and spending time with friends, as significant determinants of happiness. We generated authentic test data to verify the model, utilizing rigorous testing protocols to evaluate its predicted precision and applicability across various demographics. Based on our investigation, the use of the gradient boost technique resulted in improved picture projections. The evaluation of the technique using a confusion matrix revealed an accuracy of 97.1% for training and a perfect accuracy of 100% for validation. The training phase achieved an accuracy of 62.5%, as shown by the confusion matrix, while the overall confusion matrix demonstrated a 92.0% accuracy in predicting happiness. The support vector machine, trained incrementally, demonstrated encouraging prospects for future investigation.

Keywords: Data Mining, Happiness, GAM, Boosting-Tree, Subset Selection, Lifestyle

1 Introduction

Exploring the application of statistical data mining technology gives a valuable opportunity to understand the secrets of human enjoyment and behaviour. The rise of big data and advanced analytics enables academics to explore the intricate cause-and-effect linkages among numerous elements that contribute to overall well-being.

* Corresponding author: sazan.sulaiman@knu.edu.iq

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

The fundamental goal of this research is to apply cutting-edge statistical data min-ing methodologies to construct prediction models capable of anticipating the general happiness of young people. By combining big data, powerful analytical approaches, and comprehensive survey data, we hope to determine the primary factors influencing individuals' happiness levels, including lifestyle choices, social behaviour, and per-sonal preferences.

The purpose of this research is anchored in the growing perspective of people's well-being, emphasizing the need to develop effective techniques to promote happiness and positive behaviours, especially among young people. In today's increasingly digitalized world, with a wealth of available data, there is a unique opportunity to employ statistical data mining technologies to get deeper insights into the elements contributing to happiness and behaviours.

The study contributes in numerous ways. Firstly, it presents a thorough picture of the different elements of happiness prediction and habit building, drawing on various variables and metrics. Moreover, it demonstrates the efficacy of modern statistical approaches, such as boosting trees and generalized additive models, in constructing models capable of accurate forecasting. Additionally, it gives practical advice for educators, legislators, and other stakeholders to promote well-being activities and encourage youth to develop positive behaviours.

2 Literature Review

2.1 Data collection

This survey aims to assess and analyze various humanistic indicators such as habits, emotions, and personality traits. The survey utilized the Kaggle database to access a dataset derived from research that examined the pastimes, behaviors, perspectives, and concerns of young individuals. As stated in [1], a total of 1,001 young individuals, aged 15 to 30, participated in this research. According to reference [2], the survey encompassed several subjects and was primarily structured into the following categories: demographics, personality characteristics, expenditure patterns, spending preferences, hobbies and interests, phobias, and music and film preferences. All items, excluding height and weight, were assessed based on self-reported replies using a linear scale ranging from 1 to 5. A score of 1 indicated low or strong disagreement, while a score of 5 indicated high or strong agreement. Queries such as "To what extent do you experience fear in darkness?" exemplify this category of investigation. For instance, "I experience difficulty in rousing myself in the morning." For instance, "I believe that all of my personality traits are favorable." After selecting happiness as the main component, the study examined how the other elements either contributed to or predicted happiness. The researcher in [3] quantified happiness by evaluating the spectrum of replies to the sentence "I am 100% happy with my life." The primary objective of the survey was to identify the variables that may potentially forecast this particular outcome. This challenge was approached by employing three widely used classification techniques: best subset selection, generalized additive models, and boosting trees. The process began by identifying a subset of predictors that were linearly connected and had the highest predictive power for happiness. This was accomplished through the utilization of forward, backward, and best subset selection techniques. The process of dividing the predictor space into sample areas was accomplished by a tree-based method and recursive binary splitting, as outlined in [4]. The boosting tree technique was chosen due to its high-ranking status as a tree-based model. After reviewing the initial tree, the weights of the outcomes were adjusted, making

them significant. These results are clearly discernible but challenging to pinpoint. As stated in reference [5], this data that has been assigned weights serves as the foundation for the second tree. The objective here is to improve the original picture projections. The primary obstacle encountered throughout the study was the arduous task of preserving the dataset, mostly owing to its limited scale and the absence of publicly accessible medical information that was utilized. Improved organization may be achieved by identifying the hyperplane that effectively separates the classes in data mining. A comprehensive implementation of the concept may be found in [6], which includes further mathematical and non-linear explanations. The approach is a deterministic, binary classification technique commonly employed for control purposes. These issues were successfully addressed, and outstanding outcomes were achieved in data mining methodologies and tree-based structured data mining. Additionally, four case studies were successfully resolved.

2.2 Subset Selection Approach

Generalized additive models, boosting trees, and best subset selection are three frequently employed data mining techniques for classification that are applied to address the problem in our study. The original strategy involved selecting a subset of predictors that were linearly connected and had the most predictive power for enjoyment. This was done using forward, backward, and best subset selection methods. In order to divide the predictor space into distinct areas, a tree-based approach was employed. Our alternative approach involved the utilization of recursive binary splitting, as documented in [8]. The boosting tree technique was chosen due to its high-ranking status as a tree-based model. If the distance value of a respondent was more than three standard deviations away from the mean, they were classified as potentially having completed the survey at random, as shown in Fig 1.

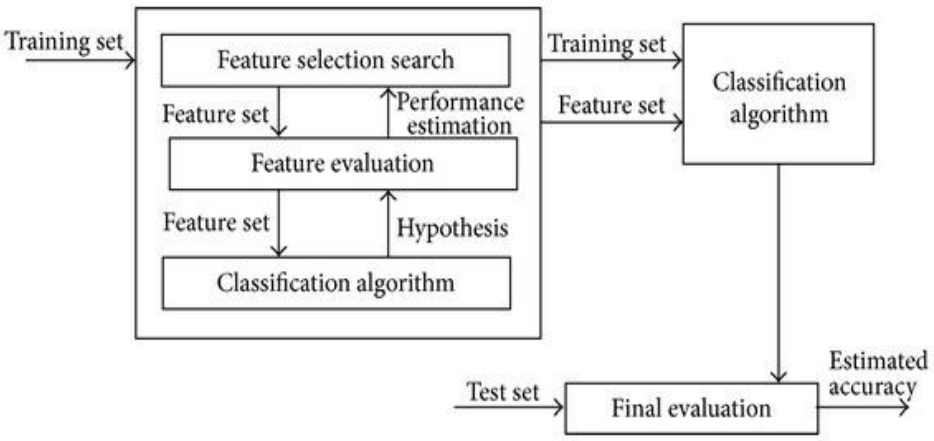

Fig. 1. The feature-based subset selection method [8].

2.3 Boosting Tree Method

To identify the key determinants of happiness, an effective approach would be to employ the boosting tree technique, which is a methodology based on decision trees. This study focuses on analyzing the gradient-boosting machine, which is utilized for predicting happiness levels. The boosting tree algorithm described in reference [9] was implemented to achieve this task. In the booster phase, several trees are included, with each tree enhancing the model by

utilizing data from the preceding one. The tuning parameters include the tree count, shrinkage parameter, and the number of splits within each tree, as shown in Fig 2.

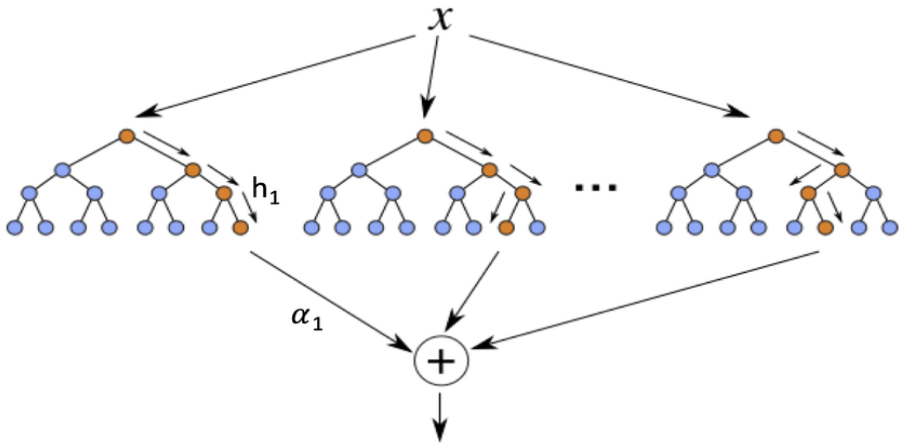

Fig. 2. The boosting tree cascading with number of trees [10].

Prior to using a boosting tree to analyze the impact of various variables on happiness, it was necessary to convert the happiness data from a 1–5 scale to a binary scale. In this binary scale, a value of 0 denoted not being happy, while a value of 1 indicated being happy. This was accomplished by categorizing any Happiness value that exceeded the average as Happy (1) and any Happiness value that fell below the average as Unhappy (0). To initiate the boosting model, the tree function was initialized, and all residuals were set to zero, ensuring that they corresponded to the values in the training set. A newly obtained tree with d divisions, as described in reference [11], was fitted by iteratively following these subsequent phases. Subsequently, it enhances our existing tree model by including a tree representation of the new tree into it. Finally, it deducts the count of tree values from the initial residuals to revise the residuals of the model.

2.4 Generalized Additive Model

Researchers can compare many generalized additive models in [12] by employing either the output variables obtained by subset selection or the variables with the highest relative effect in the boosting trees, or a combination of both. The forecast accuracy of each best model was directly compared by measuring them. The approach involved fitting Generalized Additive Models (GAMs) using different combinations of splines, second-degree polynomials, and linear predictor variables. This resulted in a reduction in the number of connections between individual predictors and the outcome. utilized additional polynomials and splines to model predictors that exhibited non-linear associations with our response variable. The degrees of freedom were determined using a mix of visual data analysis and cross-validation (K-Fold). In order to fit Generalized Additive Models (GAMs) to the six factors that were identified as crucial in subset selection, researchers in [13] employed a total of eleven distinct basis functions. This study examined the relationship between the component-target variable link and the confusion and difficulty of data mining for linear support vector machine training [14]. Consequently, the occurrence of flanging linear models, which are restructured reappearances, is more probable in the field of data mining. Data mining has several advantages, yet it also presents a few drawbacks, as referenced in [15]. The issue with data

mining for fitting purposes is that it results in the creation of an excessively intricate tree structure, which exhibits limited predicting ability when confronted with new data. Ultimately, the ideal tree cannot be replicated since data mining algorithms are insatiable. Although there are clear advantages to substituting null values with means, the most efficient approach was just to eliminate them. Due to the unavailability and insufficient size of the medical dataset in our nation, managing the dataset posed the most difficult component of the study project (as stated in [16]). These issues were successfully tackled, and outstanding outcomes were achieved in data mining methodologies and tree-based structured data mining. Additionally, four case studies were successfully resolved. Data mining aims to address the mathematical framework employed in resolving long-term value when confronted with challenges in supervised learning as shown in Fig 3.

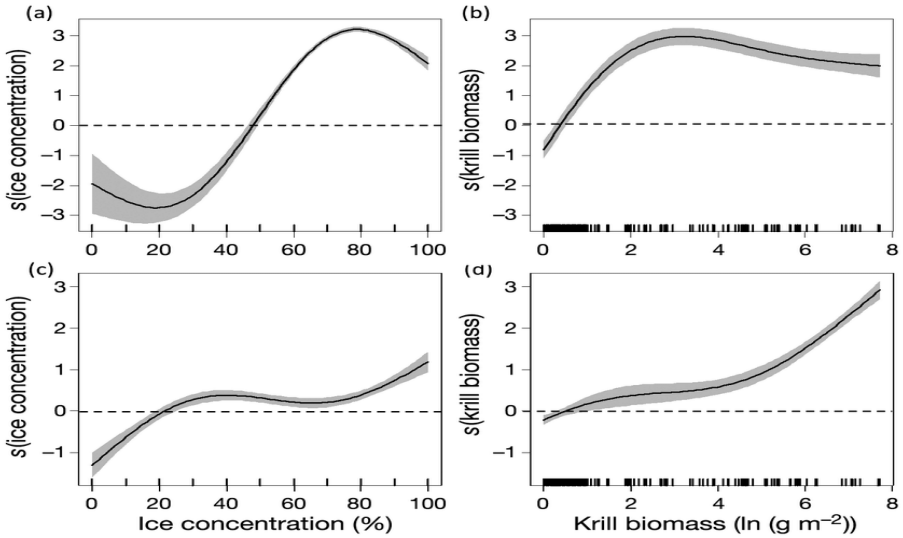

Fig. 3. The generalized additive model in predicting the functional relationship [16].

After deleting all missing data, our dataset was reduced by almost 30% using reference [17]. The lack of limits on variables or predictors that produce missing data is the root cause. Our data collection is small, thus used the column average instead of eliminating non-categorical data points. This method would save a lot of initial data. As seen in [18], averages have little influence on statistics due to little data volatility. The predictors are distributed to prevent filling missing data for variables with too many average values. The researchers' sole focus on pure statistical models throughout their joint research on chronic linear support vector machines [19] left gaps in machine-learning data mining advancement. This paper reviews the literature and proposes updated data mining technologies using the Support Vector Machine (SVM) algorithm. The suggested method evaluates four pre-processing circumstances using important F-score features. According to [20], used clinical data and machine learning to predict chronic renal disease. Data Mining Regression, Ada-Boost, Gradient Boosting Tree, Logistic Repressor (LR), and Support Vector Regression are the primary instructional methods being studied. As indicated in [21], this study employed UCI linear support vector machine dataset components. By comparing numerous regression models, the most accurate one is chosen for prediction. According to reference [22], data is continually grouped until each leaf has one instance. This prediction model was created using the Extreme Support Vector Machine [23]. In data mining, an Associate optimizes the Linear Support Vector Machine. Data mining in nursing refers to an upgraded Linear Support Vector

Machine. LSVM requires many call trees for its fundamental functionality. This idea shows that while individual call plaits may yield surprising outcomes, their combination improves performance and accuracy [24]. In case of quality loss, the additive technique can be used to slowly add additional trees [25]. This is because learning tree building is more complicated than gradient optimization.

The work solved scheduling and rebound difficulties using data mining techniques including controlled regression and support vector regression [26]. Each neuron repeats this calculation as signals propagate through the network layers, resulting in progressively abstract representations of the initial input data. The last network layer generates the prediction output, a class probability distribution. Randomly initializing all network settings is common. Thus, the network's first predictions are constrained. As mentioned in [27], this method is mostly used for structural concerns. This computational framework represents each data item as a coordinate in an n-dimensional space, where n is the number of highlights. The value assigned to each element estimates its importance. According to [28], creating order begins with a hyper-plane that differentiates the two groups. For the double straight classification non-probabilistic approach [29], non-mathematical explanations and a strong software are typically needed. By focusing on an event in a region, the data mining model may be depicted as a linear structure with divisions and a defined contour. It is therefore decided that recent occurrences would remain in the indistinguishable zone. According to [30], data mining's usefulness over broad domains is most exciting. It's hard to tell if comments were intentionally false, but some were. The models in [31] show that loneliness, energy levels, BMI, number of friends, and changing the past in polynomial two all affect happiness. Our model's remarkable congruence with student test results showed its reliability in extrapolating the original study's features to predict happiness. This study analyzed and predicted using data mining, a machine learning approach. Partitioning data into numerous groups allows factor calculations. According to [33], anomalies do not affect performance, hence pre-processing is unnecessary. split the dataset in half for testing and training. The free ensemble model, or data mining, uses numerous regression and classification trees. Our goal is to train the coaching dataset with the objective operation using data mining [34]. GAMs were trained on the original dataset. The predict function was applied to the new dataset to calculate Mean Squared Error using the same method as [35]. In this study, researchers categorize and identify visual items of the same class [36]. Instance segmentation combines object training, which predicts the bounding box and class of multiple image objects without segmenting them, with semantic segmentation, which labels each pixel with a semantic category without distinguishing between instances of the same category. According to reference [37], semantic segmentation may reliably identify agricultural areas even when surrounded by nearby regions of various image classes.

Thus, data mining can corrupt linear models with rearrangement reappearances [38]. Despite its many benefits, data mining has downsides. A complex tree structure created by data mining for fitting has poor prediction skills when presented with additional data. Data mining algorithms are insatiable; therefore the ideal tree cannot be recreated. According to [39], deleting null values yielded the best results, despite the reasonableness of replacing them with mean values. Given the expanding amounts of geospatial data and earth observation photographs, it seems appropriate for numerous remote sensing operations. This study [40] uses data mining-based boosting to strengthen weak learners. To train a linear support vector machine, data mining boosting adds each new picture to the starting data set.

3 METHODOLOGY

Upon completing the data preparation, it became evident that our dataset had many uniformly distributed missing components. Out of the four preprocessing scenarios, the most sensible decision was to choose relevant features and fill in missing data using column means. Training using supplementary data can be conducted without incurring any disadvantages as a result. Dividing the data into two or more groups enables the computation of factor computations.

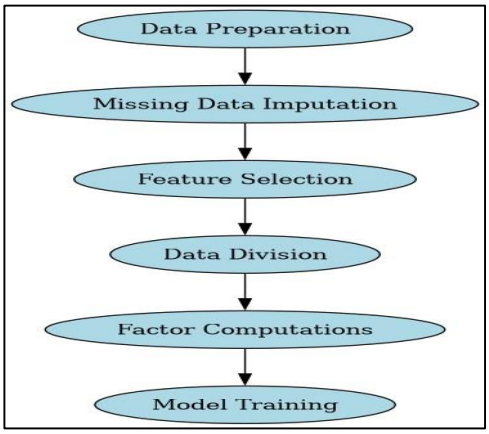

Fig. 4. Illustration of the Methodology Flowchart

One can see the most polarized regions by employing either an approximating function or a conflicting framework. This is utilized to enhance the data mining classification process by intelligently adjusting the data mining rule based on a partner degree. Data mining is characterized by more transparency and comprehensibility compared to several scheme calculations. Furthermore, this flowchart (Figure-4) outlines the step-by-step process followed in the methodology described in the review. It begins with data preparation, including handling missing data, proceeds to feature selection, data division, factor computations, and concludes with model training and pre-processing is superfluous as anomalies have little effect on performance as shown in Fig 5.

The data that we have acquired from numerous countries and areas and the primary elements that determine the happiness scores of their inhabitants are our main focuses. Such factors are as follows: GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, and views of corruption and we have taken 8 features. Each aspect is in itself a different part of societal well-being and is consequently vital for the study of the determinants of the happiness of people across different locations. The selection of the data splitting percentage is critical for guaranteeing the durability of the model. In our study, we decided for an 80/20 split, with 80% of the data assigned for training and 20% for testing. This decision was made to create equilibrium between retaining enough data to train the model properly and having sufficient unobserved data to evaluate its performance

appropriately.

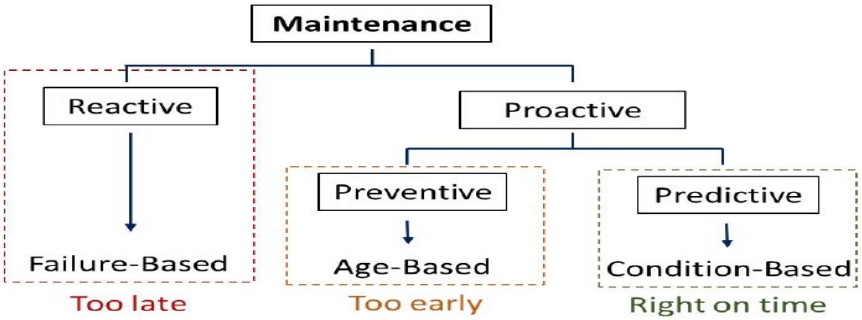

Fig. 5. The relationship between the maintenance of failure based and condition based predictive model for happiness.

Data mining is specifically designed to address the difficulties encountered in supervised learning by emphasizing the mathematical framework employed to extract enduring value. The dataset is divided in half to facilitate testing and training. The free ensemble model, commonly referred to as data mining, consists of many regression and classification trees. The purpose of our project is to utilize data mining techniques to train the coaching dataset as shown in Fig 6 and Fig 7.

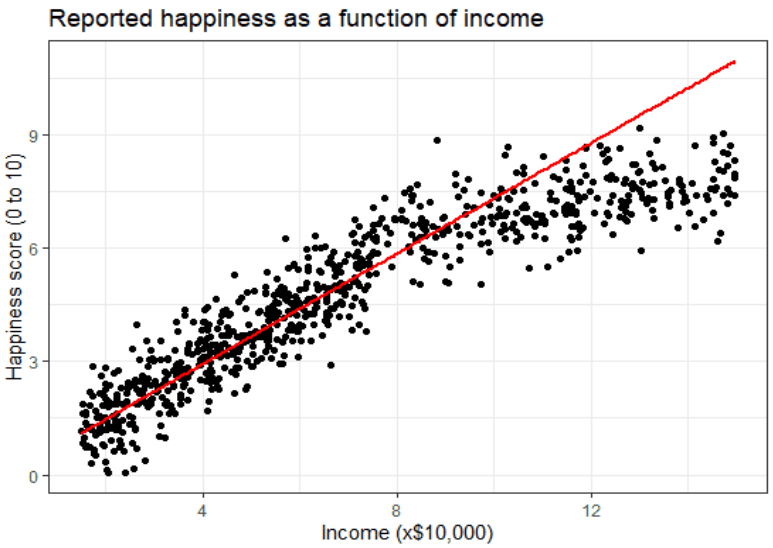

Fig. 6. The graph represents the happiness as a function of income with happiness score relevantly high

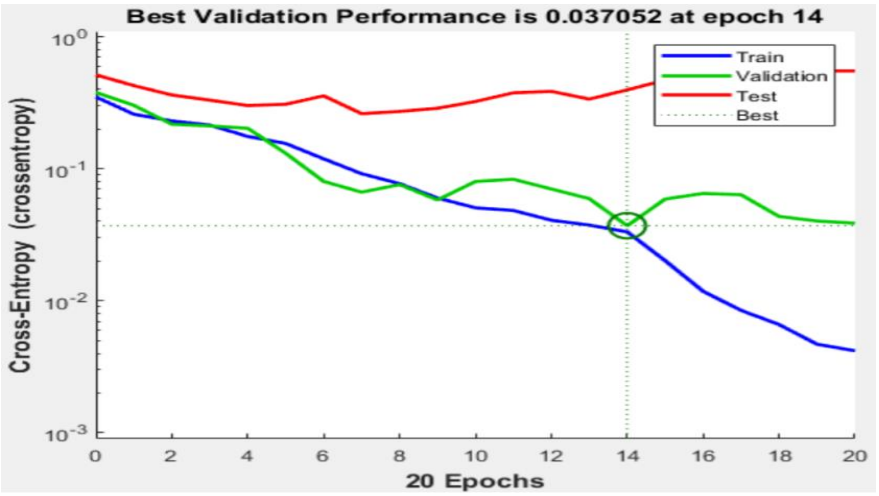

Fig.7. The graph represents the training, testing and validation of happiness prediction.

The models used to identify factors for predicting happiness analysis were highly robust. It constructed several generalized additive models (GAMs) by merging variables identified by boosting tree analysis and optimum subset selection. After analyzing the initial tree, we adjust the weights of the outcomes, which may be apparent yet challenging to see as shown in Fig 8.

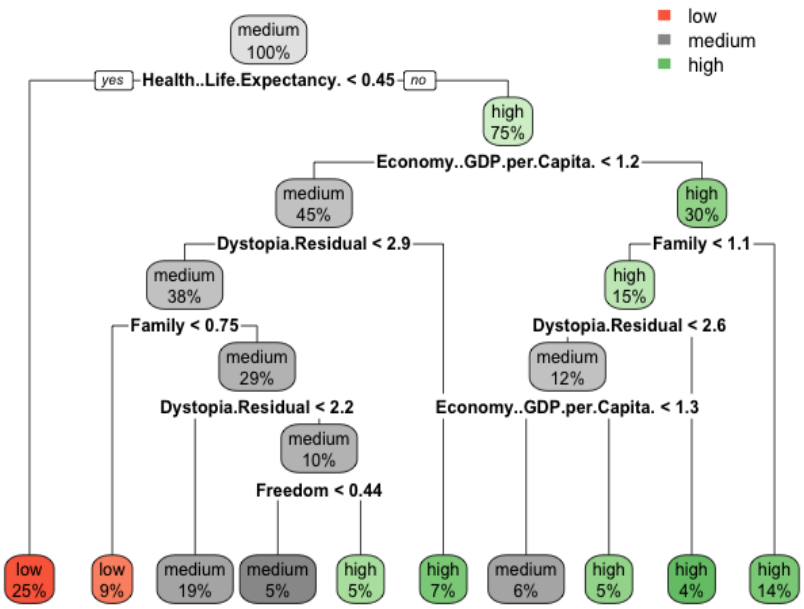

Fig. 8. The main factors that affect the performance of happiness prediction as per category.

The data with assigned weights is utilized to construct the second tree. The objective here is to improve the quality of the first picture projections. Consequently, our novel model is constructed using advanced data mining techniques. Analysts will systematically compute

the worth of each learning item by giving a precise numerical value to each component in the n-dimensional space. Identifying the hyperplane that effectively divides the two groups might simplify the task of handling them. One can identify the most polarized regions by employing either an approximating function or a conflicting basis. The data is divided into groups repeatedly until there is just one instance in each leaf. This method enhances the data mining classification process by intelligently adjusting the data mining rule based on a partner degree. Data mining is characterized by more transparency and comprehensibility compared to several scheme calculations. Furthermore, the Euclidean distance is not emphasized. Scaling the highlights is no longer necessary. Scaling also heightens the possibility of implicit misconceptions as a result of the modified characteristics. Data mining is specifically designed to address the difficulties encountered in supervised learning by emphasizing the mathematical framework employed to extract enduring value. If the sequential addition of new trees would result in a decrease in quality, we can choose for the additive strategy instead, as the process of learning tree construction is far more intricate than basic gradient optimization.

4 RESULTS

Additional data, collected from individuals in the same age range as the original dataset, was obtained after the survey was created with questions specifically targeting the key traits described earlier. Data cleaning was not required, except for the body mass index (BMI) variable. All questions that used the same scale included response possibilities ranging from 1 to 5. However, objects in fields frequently have contiguous boundaries and are positioned near one another. The outcome of semantic segmentation would be polygon items that are grouped together. Item segmentation enables us to distinguish between these comparable or overlapping occurrences of the identical category. Data mining has been adopted at a sluggish pace in the remote sensing business. This study examined the potential application of data mining techniques on aerial and satellite photographs to identify pictures as shown in Fig 9 and Fig 10.

Fig. 9. The confusion matrix of training with 97.1% for happiness prediction.

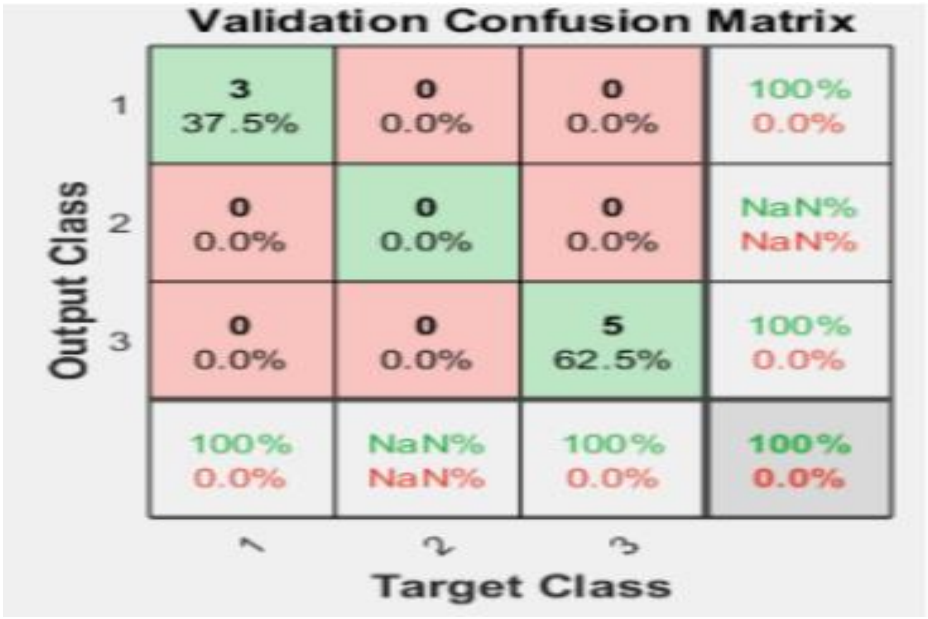

Fig. 10. The confusion matrix of validation with 100% for happiness prediction.

The data mining strategy may be elucidated more effectively in the first scenario by employing the gradient boost technique, which is rooted in data mining. The data mining approach commences by utilizing a decision tree and assigns equal weight to each conclusion based on every observation. After reviewing the initial tree, we adjust the weights of the outcomes, making them more or less significant. These results are clearly discernible but challenging to pinpoint. The second tree is built from this weighted information. The objective here is to improve the original picture projections. The primary obstacle in our research endeavor was managing the dataset, since it was limited in size and not accessible to the general public in our nation. Please refer to Fig 11 and 12 for further details. For our investigation, we utilized two main models: linear support vector machines (SVM) and linear regression. The linear SVM model was subjected to 100 iterations, each consisting of 20 epochs. This process yielded an accuracy rate of 92.0%. Unfortunately, there was a lack of particular information regarding the iteration and epoch details for the linear model. The results indicate that the linear SVM technique is highly effective in accurately forecasting happiness using the derived features from the dataset.

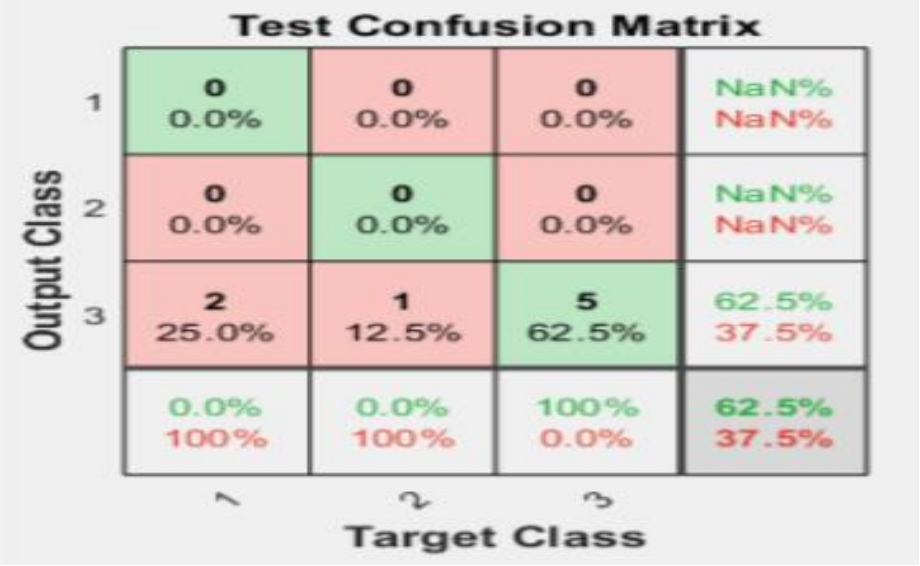

Fig. 11. The confusion matrix of training with 62.5% for happiness prediction.

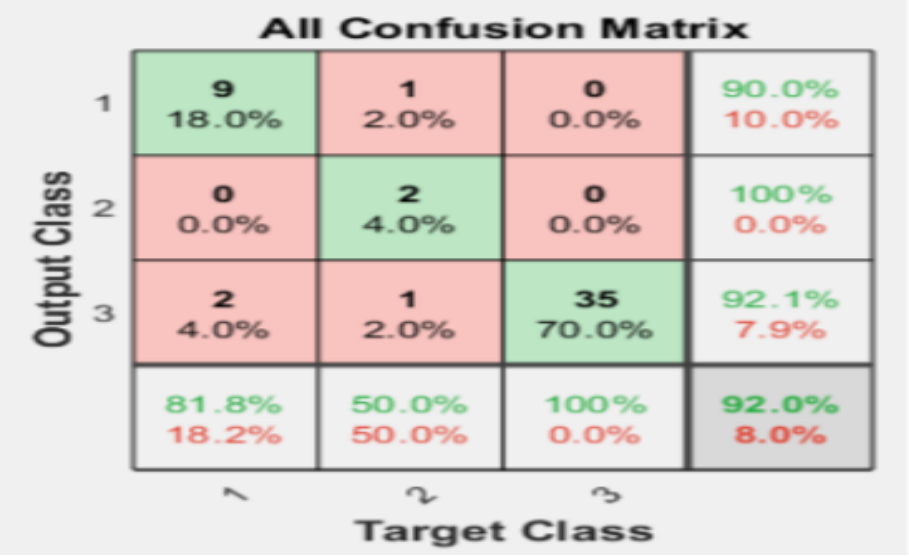

Fig. 12. The overall confusion matrix with 92.0% for happiness prediction.

By employing tree-based structured data mining methodologies, we successfully addressed the constraints and achieved optimal outcomes in four distinct case studies. Our novel solution is built upon data mining techniques. Forecasting the changing components may be achieved by utilizing a third tree that arises from establishing the identification coefficient of this distinct data mining set model. We repeat these processes a specific number of times. Subsequent trees facilitate the identification of inaccurately classified observations. The subsequent trees encompass the ultimate prognostications of the final model, which are a weighted amalgamation of the prognostications given by the tree models in the preceding ensemble model. Support Vector Machine trains several models in a sequential manner,

adding and progressing as it goes. Support vector machines train many models in a linear, cooperative, and progressive manner.

5 DISCUSSION

The dependability of the resultant model is contingent upon the comprehensiveness of the survey, given that our data is obtained from self-reported answers. It is appropriate to identify individuals who may have provided inaccurate information in the survey and exclude their replies. For analysis, two forms of deceit were selected: responding to the survey randomly and consistently providing identical responses. To achieve a response rate of 70%, the researchers in [41] followed the below processes to choose six individuals whom they felt replied in a random manner, and identified 10 individuals who provided identical responses. These people were excluded from the data set for the following segments. The researchers emphasized that data mining systems with a substantial number of layers or neurons are solely capable of acquiring linear mappings. Once the activation function is implemented, the resulting signal is sent to the neurons in the subsequent layer. As the signals propagate through the layers of the network, each neuron does this computation once again, leading to increasingly abstract representations of the initial input data. The prediction output, which represents a distribution of class probabilities, is created by the last layer of the network. As stated by reference [42], the default setting is to randomly establish all network configurations. Previously, it was mentioned that the network must ascertain the extent to which the parameters were altered during training. After each forward pass, a comparison is conducted between the anticipated label and the forecasted output. In order to accomplish this, employ a loss or cost function, such as cross entropy or mean squared error, to precisely measure the quality of the output. This is utilized to enhance the data mining categorization by intelligently adjusting the data mining rule with a degree of partnership. Data mining is characterized by more transparency and a higher level of comprehensibility compared to numerous scheme calculations. Performance is not affected by anomalies, hence pre-processing is unnecessary, as stated in [43]. Furthermore, the Euclidean distance is not emphasized. Scaling the highlights is no longer necessary. Scaling also heightens the possibility of implicit misconceptions as a result of the modified characteristics. Data mining considers all numerical and unmixed variables as significant information that deserves equal importance in the overall model, owing to the many types of data being collected as shown in Table 1.

Table 1. The comparison with other techniques.

Article	Technique	Accuracy
[44]	Gradient Boosting Tree	86.68%
[45]	Decision Tree & Random Forest	88.59%
Proposed	Linear Support Vector Machine	90.47%

Through a comparison of the shortcomings of the two data mining strategies in meeting the requirements of disadvantaged learners, the researchers in this study [45] identified significant differences between them. Regarding loss feature gradients, the increase in gradients operates in a comparable manner to how the data mining technique identifies vulnerabilities through the amplification of information points with higher weights. The loss function demonstrates a strong alignment between the model's coefficients and the base data, indicating a good fit. Our purpose is to optimize the loss function in a sensible manner. In a sales price projection based on regression analysis, the discrepancy between the actual and expected house prices would be used as the basis. The function for loss. To effectively

categorize nonperforming loans according to their default rates, the loss function would evaluate the accuracy of our predictive model in doing so. The popularity of gradient boosters stems from their ability to optimize a user-specified cost function, which is more suitable with real-world applications and provides greater flexibility compared to a loss function. Replacing null values with mean values was the most rational choice of the three options, while eliminating null values had the most significant outcomes. Our investigation involved a thorough and precise preprocessing of the dataset to ensure its high quality and relevance. Subsequently, we proceeded to train and optimize our linear support vector machine (SVM) model. In order to evaluate its efficacy, we utilized a range of performance indicators, including precision, recall, and F1-score. In addition, we employed cross-validation approaches to guarantee the strength and dependability of our model.

6 CONCLUSION

This research delves into the realm of happiness prediction by examining the intricate interplay of factors that contribute to it using statistical data mining techniques. We employed a meticulous data preparation technique to ensure the high quality of the data used. Subsequently, applied boosting trees and linear support vector machines (SVM) as our methods of choice. By the way, carried on refining our models to acquire the correct prediction of the happiness levels of young people through the cycles and the evaluation. The application of gradient boosting trees and linear SVM was able to provide us with a highly trustworthy model, and our linear SVM model obtained an exceptional accuracy of 92%. The fact that our system has an error rate of 0% demonstrates that our algorithm is successful in forecasting happiness and, thus, is a wonderful source of information for researchers and practitioners in the future. Nevertheless, our research has significant flaws. The huge and reachable dataset provided by us also caused issues in the construction and testing of our model. Consequently, the dependency on up-to-date survey data that is based on self-reporting elevates the risk of biases and mistakes that should be taken into account. The future work in this sphere should be focused on the wider range of data collection that encompasses a great number of people from diverse age groups and cultures. Besides, the incorporation of new machine learning techniques and data sources could enhance the predictive possibilities of our models. Finally, the results of our study demonstrate considerable progress in the field of study and prediction of happiness through the use of statistical data mining methods, which have the potential to be a useful tool in the development of strategies and interventions that will contribute to the promotion of positive health and behaviors in society.

References

1. Ritu, V. Gross national happiness: Meaning, measure and degrowth in a living development alternative. J. Political Ecol.(2017), 24, 476–490.
2. Jad, C.; Irani, A.; Khoury, A. The composite global well-being index (CGWBI): A new multi-dimensional measure of human development. Soc. Indic. Res. (2016), 129, 465–487.
3. Balestra, C.; Boarini, R.; Tosetto, E. What matters most to people? Evidence from the OECD better life index users' responses. Soc. Indic. Res. (2018), 136, 907–930.
4. Chen, Tianqi & Guestrin, Carlos. (2016). XGBoost: A Scalable Tree Boosting System. 785-794. 10.1145/2939672.2939785.
5. Energy-Efficient Classification for Resource-Constrained Biomedical Applications - Scientific Figure on ResearchGate. Available from:

https://www.researchgate.net/figure/Schematic-diagram-of-a-boosted-ensemble-of-decision-trees_fig2_325632132 [accessed 22 Nov, 2020]

6. Hatem, M.N.; Sarhan, S.S.; Rashwan, M.A.A. Enhancing recurrent neural network-based language models by word tokenization. Human Centric Comput. Inform. Sci. (2018), 8, 1–13
7. Aydadenta, H. A clustering approach for feature selection in microarray data classification using random forest. J. Inform. Process. Syst. (2018), 14, 1167–1175.
8. Pregibon, D.; Hastie, T.J. Generalized linear models. In Statistical Models in S; Momirovic, K., Mildner, V., Eds.; Routledge: London, UK, (2017); pp. 195–247.
9. Reiss, Christian & Cossio, Anthony & Santora, Jarrod & Dietrich, Kimberly & Murray, Alison & Mitchell, B & Walsh, Jennifer & Weiss, Elliot & Gimpel, Carla & Jones, Christopher & Watters, George. (2017). Overwinter habitat selection by Antarctic krill under varying sea-ice conditions: Implications for top predators and fishery management. Marine Ecology Progress Series. 568. 1-16. 10.3354/meps12099.
10. Estoque, R.C.; Togawa, T.; Ooba, M.; Gomi, K.; Nakamura, S.; Hijioka, Y.; Kameyama, Y. A review of quality of life (QOL) assessments and indicators: Towards a “QOL-Climate” assessment framework. Ambio (2018), 1–20.
11. Zhang, Z.; Lai, Z.; Xu, Y.; Shao, L.; Wu, J.; Xie, G. Discriminative elastic-net regularized linear regression. IEEE Trans. Image Process. (2017), 26, 1466–1481.
12. Moore, S.M.; Diener, E.; Tan, K. Using multiple methods to more fully understand causal relations: Positive affect enhances social relationships. In Handbook of Well-Being Diener; Oishi, S., Tay, L., Eds.; DEF Publishers: Salt Lake City, UT, USA, (2018).
13. M. H. B. A. Alkareem, F. Q. Nasif, S. R. Ahmed, L. D. Miran, S. Algburi, and M. T. ALmashhadany, “Linguistics for Crimes in the World by AI-Based Cyber Security,” 2023 7th International Symposium on Innovative Approaches in Smart Technologies (ISAS), Nov. (2023).
14. S. R. Ahmed AHMED, I. Ahmed Najm, A. Talib Abdulqader, and K. Basem Fadhil, “Energy improvement using Massive MIMO for soft cell in cellular communication,” IOP Conference Series: Materials Science and Engineering, vol. 928, no. 3, p. 032009, Nov. (2020).
15. Consoli, S.; Recupero, D. Using FRED for named entity resolution, linking and typing for knowledge base population. Commun. Comput. Inf. Sci. (2015), 548, 40–50.
16. Dridi, A.; Reforgiato Recupero, D. Leveraging semantics for sentiment polarity detection in social media. Int. J. Mach. Learn. Cybern. (2019), 10, 2045–2055.
17. Carta, S.; Corriga, A.; Ferreira, A.; Podda, A.S.; Recupero, D.R. A multi-layer and multi-ensemble stock trader using data mining and deep reinforcement learning. Appl. Intell. 2020, 1–17.
18. Barra, S.; Carta, S.M.; Corriga, A.; Podda, A.S.; Recupero, D.R. Data mining and time series-to-image encoding for financial forecasting. IEEE/CAA J. Autom. Sin. (2020), 7, 683–692.
19. Carta, S.; Ferreira, A.; Podda, A.S.; Recupero, D.R.; Sanna, A. Multi-DQN: An Ensemble of Deep Q-Learning Agents for Stock Market Forecasting. Expert Syst. Appl. (2020), 164, 113820.
20. Meena, K.S.; Suriya, S. A Survey on Supervised and Unsupervised Learning Techniques. In International Conference on Artificial Intelligence, Smart Grid and Smart City Applications; Springer: Berlin, Germany, (2019); pp. 627–644.

21. Van Engelen, J.E.; Hoos, H.H. A survey on semi-supervised learning. Mach. Learn. (2020), 109, 373–440.
22. Tehrani, A.F.; Ahrens, D. Supervised regression clustering: A case study for fashion products. Int. J. Bus. Anal. (IJBAN) (2016), 3, 21–40
23. Pes, B. Ensemble feature selection for high-dimensional data: A stability analysis across multiple domains. Neural Comput. Appl. (2020), 32, 5951–5973.
24. Jena, P.C.; Kuhoo; Mishra, D.; Pani, S.K. A novel approach for regularization of ensemble learning in classification and regression analysis. Indian J. Public Health Res. Dev. (2018), 9, 1406–1411.
25. Gayberi, M.; Gunduz Oguducu, S. Popularity Prediction of Posts in Social Networks Based on User, Post and Image Features. In Proceedings of the 11th International Conference on Management of Digital EcoSystems, Limassol, Cyprus, 12–14 November (2019); pp. 9–15.
26. De, S.; Maity, A.; Goel, V.; Shitole, S.; Bhattacharya, A. Predicting the Popularity of Instagram Posts for a Lifestyle Magazine Using Data mining. In 2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA), Mumbai, India, 7–8 April (2017).
27. Hong, L.; Dan, O.; Davison, B.D. Predicting popular messages in twitter. In Proceedings of the 20th International Conference Companion on World Wide Web, Hyderabad, India, 28 March–1 April (2011); pp. 57–58
28. Hoang, T.B.N.; Mothe, J. Predicting information diffusion on Twitter–Analysis of predictive features. J. Comput. Sci.(2018), 28, 257–264.
29. Rao, P.G.; Venkatesha, M.; Kanavalli, A.; Shenoy, P.D.; Venugopal, K. A micromodel to predict message propagation for twitter users. In Proceedings of the 2018 International Conference on Data Science and Engineering (ICDSE), Kochi, India, 7–9 August (2018); pp. 1–5.
30. Naseri, M.; Zamani, H. Analyzing and predicting news popularity in an instant messaging service. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July (2019); pp. 1053–1056.
31. S. R. Ahmed, A. K. Ahmed, and S. J. Jwmaa, “Analyzing The Employee Turnover by Using Decision Tree Algorithm,” 2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Jun. (2023).
32. N. Z. Mahmood, S. R. Ahmed, A. F. Al-Hayaly, S. Algburi and J. Rasheed, "The Evolution of Administrative Information Systems: Assessing the Revolutionary Impact of Artificial Intelligence," 2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkiye, (2023), pp. 1-7.
33. Peláez, J.I.; Martínez, E.A.; Vargas, L.G. Products and services valuation through unsolicited information from social media. Soft Comput. (2020), 24, 1775–1788.
34. Alduaiji, N.; Datta, A.; Li, J. Influence propagation model for clique-based community detection in social networks. IEEE Trans. Comput. Soc. Syst. (2018), 5, 563–575.
35. Boratto, L.; Carta, S. The rating prediction task in a group recommender system that automatically detects groups: Architectures, algorithms, and performance evaluation. J. Intell. Inf. Syst. (2015), 45, 221–245.
36. L. I. Khalaf, S. A. Aswad, S. R. Ahmed, B. Makki, and M. R. Ahmed, “Survey On Recognition Hand Gesture By Using Data Mining Algorithms,” 2022 International

Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Jun. (2022).

37. L. I. Khalaf, S. A. Aswad, S. R. Ahmed, B. Makki, and M. R. Ahmed, “Survey On Recognition Hand Gesture By Using Data Mining Algorithms,” 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Jun. (2022).
38. B. T. Yaseen, S. Kurnaz, and S. R. Ahmed, “Detecting and Classifying Drug Interaction using Data mining Techniques,” 2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Oct. (2022).
39. S. Rashid, A. Ahmed, I. Al Barazanchi, A. Mhana, and H. Rasheed, "Lung cancer classification using data mining and supervised learning algorithms on multidimensional data set," vol. 7, no. 2, pp. 438–447, (2019).
40. Awad, O. F., Sulaiman, S. K., & ALSHMEEL, G. H. A. (2023, October). Anomaly Detection and Security Enhancement Through Machine Learning in Administrative Information Systems. In 2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) (pp. 1-8). IEEE.
41. Guron, A. T., Anwer, M. A., Sulaiman, S. K., & AbdulSamad, S. J. (2023). Classification of the cause of eye impairment using different kinds of machine learning algorithms. Passer Journal of Basic and Applied Sciences, 5(2), 410-416.
42. Sulaiman, S. K., & Ismael, Y. S. (2022). Machine Learning-Based Personal Signature Recognition Model. NeuroQuantology, 20(10), 2170.
43. Zhou, Y.; Wu, Z.; Zhou, Y.; Hu, M.; Yang, C.; Qin, J. Exploring Popularity Predictability of Online Videos With Fourier Transform. IEEE Access (2019), 7, 41823–41834
44. Chaipornkaew, Piyanuch & Prexawanprasut, Takorn. (2019). A Prediction Model for Human Happiness Using Machine Learning Techniques. 33-37. 10.1109/ICSITech46713.2019.8987513.
45. Aich, Satyabrata & Choi, Ki-Won. (2018). A Machine Learning Approach to Predict Happiness Based on Sentiment Analysis of Twitter Data