Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.199

IJCSMC, Vol. 9, Issue. 1, January 2020, pg.1 – 11

A NOVEL APPROACH FOR DATA ANALYSIS BASED ON SIMULTANEOUS DATA TYPES TEXT, IMAGE, AND NUMBER

Udhayakumarapandian[1]; Srinethe S[2]
Data Scientist, Alphind Software Solution, Ekkatutangal, Chennai, India[1]
Bachelor of Engineering, Department of Computer Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India[2]

ABSTRACT: ***The amount of data present in the world increases by terabytes every second. The need for a proper approach to analyse them plays a vital role in the betterment of the data analytics field. With the appropriate approach, even untrained data-savvy people will be able to exploit the availability of data and gain useful information from them. The combined knowledge extracted from the one dimensional data i.e. textual and numerical data as well the two dimensional i.e. images will aid us to do valuable manipulations on the data for hidden patterns to interpret better decision making. Though there are many approaches for numerical analysis of data, there are no specific approaches to combine image/textual analysis of data with its mathematical counterpart. In this study, we address this challenge in order to get better accuracy and performance ultimately. Finally we stand on the point of application of any type and we enumerate few examples in our endeavour to explain the process.***

KEYWORDS: ***Data analytics, Employee attrition, Sentiment analysis, Regression, Classifier, ensemble***

I. INTRODUCTION

Data mining is ubiquitous everywhere. From multinational companies to Governments, everybody uses it to reap maximum benefits from the data available. The application of data analytics ranges from marketing, loan prediction to employee attrition, health care applications ranging from drug discovery, disease diagnosis, prediction of prognosis, health insurance, etc. It is present in all fields from managerial science to biological sciences using artificial intelligence. As the collection and storage of data became more natural, the need for approaches for analysis of big data became pervasive. The existing approach perform analysis on numerical data effectively, but when it comes to textual data or image data, they play little to no role in analysing them. We, as humans, tend to store information in the text form for better understanding and usage. But due to the large volume of data available, we are not able to analyse the data manually. So we tend to ignore the valuable data in the text format during

analysis because of the lack of available approaches to do so. Therefore in this study, we propose a data analytic approach model which combines the knowledge from numerical and text data to provide the best results to the end user. Now consider one of the worst problems faced by the companies worldwide, employee attrition or turnover. The companies use their available data set, to predict the employees who are likely to leave or quit the company in the present year and the methods that must be taken to prevent them. Through numerical data, we can easily predict at-risk employees. But we also need to predict the reason for that prediction, which is possible only through text analysis. The valuable information collected from the exit interviews must be analysed to know the cause of their decision and what the company has to do to prevent it. This scenario also contains the micro expressions images of the employees and our approach analyses all the vital information gives rise to effective employment.

II. RELATED WORKS

Alao D. & Adeyemo A. B.[2] analysed the role of Decision tree algorithms in their research on employee attrition. The decision tree used normalised information gain of each attribute as a node. After implementing many classifiers on the data, they found that SeeRule and Boost Seetree classifiers gave the highest accuracy of 0.74. The dataset contained 309 complete records of employees, of one of the higher institutions in Nigeria who worked between 1978 and 2006. With useful analyses, the better feature impacting classification at 100% was the duration of work and salary.

Aasheesh Barvey et al.[3] used an ensemble classification methodology to classify between employees who had left the organisation and those who are still working in the organisation. They used linear regression to predict the total tenure of each employee in the organisation, this way the "lead time" available with the management for employees who can leave the organisation in the future is calculated.

Anil Kumar Dubey et al.[5] analysed various regression data and found that logistic regression gives the highest accuracy of 97% among other techniques. They also experimented with different preprocessing methods like one hot encoding and label encoding to obtain this increased accuracy.

Boosting involves fitting a sequence of weak learners over modified data. The data modification at each step involves assigning higher weights to the misclassified training samples in the previous iteration. As iterations proceed, samples that are difficult to predict are given ever-increasing influence. This process forces the weak learner to concentrate on the cases that are missed by its predecessor. Rohit Punnoose et al.[1] used XGBoost, a boosted tree algorithm in predicting the employee turnover of an organisation.

In the study, a firm analysis of employee attrition, Nathan Bennett et al.[4] used hierarchical regression to analyse the data. The authors found out that higher the benefits provided by the company, lower the attrition rate of the company.

K. M. Suceendran et al.[14] in the study of applying classifier algorithms to organisational memory to build an attrition predictor model, focused mainly on how the HR must collect exit interview data from all the employees who are exiting the company. The valuable information gained from the above process can be used to predict at-risk employees and help the organisation mitigate further employee attrition.

Amir Mohammad Esmaieeli Sikaroudi et al.[13] analysed different approaches to predict employee turnover. They concluded the study by saying that random forest classifier is better than Support Vector Machine(SVM), K-Nearest Neighbours(KNN) and many other classifiers. They used grid search and k-fold cross validation for increased accuracy along with CN2 algorithm. Random forest classifier gave an accuracy of 0.906.

In the study, Employee Attrition: What Makes an Employee Quit? Alex Frye et al.[6] used the model coefficients to find the best features of the data. They also said that among logistic regression, random forest and K-nearest neighbour classifiers, logistic regression produces the most trained results.

III. PROPOSED SOLUTION

Our solution deploys the python components fort the following process:

1. Text processing to classify the opinion the text by the packages Numpy, Scikit-learn, Bs IV
2. Image processing for detecting the object using scipy.spatial, distance, imutils perspective, contours, Numpy, argparse, cv2
3. Measuring object using SciPy, cv2,scikit-image, PyTorch

After implementing successfully the test were carried out for validating the equivalence classes of inputs for the given application such as health care, customer relationship etc.

As mentioned above, this paper proposes a method to integrate textual and numerical analyses in a data analytics approach using employee attrition as an example.

Instead of just predicting at-risk turnover employee, we can find the following:

- What was the reason for the employee leaving the company?
- How to prevent further attrition?
- How many employees may leave the company for the same reason?
- Which employees are likely to turnover?

Pseudo Code for Logistic Regression:

Pseudocode: given α, X, y where X is the group of the class used for prediction and y is the predicted class

Step 1: Initialize a = <1, .., 1> ^{T}

Step 2: Normalize X

Step 3: Repeat until convergence

$$a = a + \alpha \, m \, XT \, (y - g(Xa))$$

Step 4: Output a

IV. DATASET AND IT'S DESCRIPTION

IV.I DATA UNDERSTANDING

Data understanding includes the tasks of gathering and understanding the data that will be used in the data analysis. For this research, we used IBMWatson dataset with 35 attributes and 1470 observations. We selected the data from the dataset after going through different publically available datasets on the open sources. We couldn't use real-time data because it was essential to maintain data privacy of the organisations. But keeping this in mind, the IBM data was chosen as it has the attributes which can be related to the real-time scenarios of any organisation. The IBMWatson dataset is crafted in such a way that it is easily relatable and gives complete insight with the HR department of any organisation. The site[16] provided with the required dataset. The features are as follows:

Age	MonthlyRate	BusinessTravel	DailyRate
DistanceFromHome	Education	EducationField	EmployeeCount
EnvironmentSatisfaction	Gender	HourlyRate	JobInvolvement
JobRole	JobSatisfaction	MaritalStatus	MonthlyIncome
NumCompaniesWorked	Over18	OverTime	PercentSalaryHike
RelationshipSatisfaction	StandardHours	StockOptionLevel	TotalWorkingYears
WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion
Department	YearsWithCurrManager	EmployeeNumber	Attrition
TrainingTimesLastYear	PerformanceRating	JobLevel	

The various attributes that were selected for predicting employee attrition in different research papers are as follows in Table 1.

	Authors and Year	**Attributes Given/Accuracy**
Technical Summary	Alao and Adeyemo (2013)	Demographics, Salary, Job-length Using these attributes, we achieved an accuracy of 0.74.
	Nagadevara and Srinivasan (2008)	Absenteeism, Late-coming Using these attributes, we achieved an accuracy of 0.79
	Rombaut and Guerry (2017)	Work-specific factors
	Sengupta (2011) Harter et al. (2002)	Employee Satisfaction
	Cahyani and Budiharto (2017)	Age, Tenure, Department, Gender, Health Status, Skills
Management Summary	Pande and Chung (2017)	Working hours, Salary, Employment years, Family and Health problems
	Mihajlovic et al. (2008)	Work-environment, Job satisfaction, Job involvement, Work pressure, Career growth, Education and Training, Relationship with the manager, Performance rating, Relationship with other teammates, Work life balance
	Allen et al. (2010)	Role Clarity and Conflicts, Promotions Opportunities, Job Scope, Tenure, Age, Marital Status, Gender, Organisational Commitment, Race, Work Stress, Job Previews
	Avey et al. (2009)	Work Stress, Work Overload, Increase business travel
	Batt and Valcour (2003)	Work-Family Balance, Overtime, Bonuses, Job security

Table 1: Attributes responsible for Employee Attrition

IV.II. DATA PREPARATION

Data comes from various raw sources and contains noise as well as irrelevant information. It is essential to refine this data so that it can be suitable for building useful models and generate better results.

Once the data is selected, the third phase is to prepare this data. This phase includes tasks like cleaning, transformation, and removing unwanted data. We performed the following functions to prepare the data:

•In this study, the data for the employee attrition had various attributes which added no value to the model, i.e. it did not aid in any way to increase the accuracy and the precision of the model. Hence these attributes were removed in the process of data cleaning. Table 2 shows all the final attributes of this dataset, which were used to build the model. The vital features of the dataset were found using the model coefficients of all the features. This process is called feature selection. So, further analyses are done using these eight features.

Age	Business Travel	Daily rate	Department
Distance from Home	Education	Education field	Years at the company

Table 2: Attributes that were selected using feature selection

V. MATERIALS AND METHODS

V.I DATA MINING

Data mining is an essential step of the knowledge discovery process which involves analyses and methods for finding patterns, trends and relationships in data. The target is to acquire the knowledge required for future decision-making policy. Data mining analyses contain association analysis, clustering and prediction modelling. Association analysis is nothing but finding implicit correlation among items, simultaneous events or frequent patterns. It has a wide range of usages in market management, telecommunication networks etc. Prediction and classification models contain various model types such as decision trees, artificial neural networks and regressions. Clustering, in contrast, is an unsupervised learning algorithm. Unsupervised methods don't have a class or target feature, and all data features seen as distances.

V.II LOGISTIC REGRESSION

Based on various analytic studies, we concluded that logistic regression is the best way to predict employee attrition. Logistic regression is a predictive modelling regression algorithm which gives a binary categorical prediction. Logistic Regression is a supervised machine learning algorithm.

V.III META CLASSIFIERS

Meta classifiers are ensemble machine learning techniques used to reduce bias and variance. It runs an algorithm inside another algorithm which enables fine-tuning of the model. This fine tuning results in converting the weak learners to strong learners with increased accuracy.

V.III.I ADABOOST DECISION TREE

AdaBoosting is a general ensemble method that creates a robust classifier from several weak classifiers. This algorithm transforms a weak classifier to a strong one by retraining all the data classified wrongly by the classifier before. This retraining makes sure that the classifier after, will take that data into extra consideration and make it perfect. AdaBoost was the first successful boosting algorithm developed for binary classification.

V.III.II BAGGING

Bagging is an ensemble ML technique that combines multiple machine learning algorithms predictions to produce increased accuracy than any individual model. Bootstrap Aggregation or Bagging is a process used to reduce the variance for algorithms that have high variance while retaining the bias. This reduction in variance happens when we average the predictions in different spaces of the input feature space.

V.IV METRICS USED FOR MEASURING PERFORMANCE:

A)Accuracy:

Accuracy is the ratio of the dataset correctly classified by the method to the total number of samples in the dataset.

$$Accuracy = \frac{Number\ of\ samples\ classified\ correctly}{Total\ number\ of\ samples\ in\ the\ dataset} \quad (1)$$

B)Recall:

Recall is the ratio of errors correctly predicted to all the errors that occurred.

$$Recall = \frac{True\ Positive}{True\ positive+False\ Negative} \quad (2)$$

C)Precision:

Precision is the ratio of the actual errors among all the predictions to all that were classified as errors.

$$Precision = \frac{True\ Positive}{True\ Positive+False\ Positive} \quad (3)$$

D)F-Measure:

F-Measure is the measure of the test's accuracy and is defined as a weighted average of precision and recall.

$$F - Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

VI. ARCHITECTURE DIAGRAM

The architecture diagram gives an overview of all the process involved in mining the data. First, the dataset must be carefully understood to perform feature selection. Here feature selection is done by finding the model coefficients of all the attributes in the dataset. This feature selection aids in finding the crucial attributes of the dataset. Following feature selection is model selection, which helps in finding the best classifier algorithm, which gives the highest performance for the given dataset. Model selection is followed by numerical and textual analysis, which offers combined knowledge and required prediction to the end user.

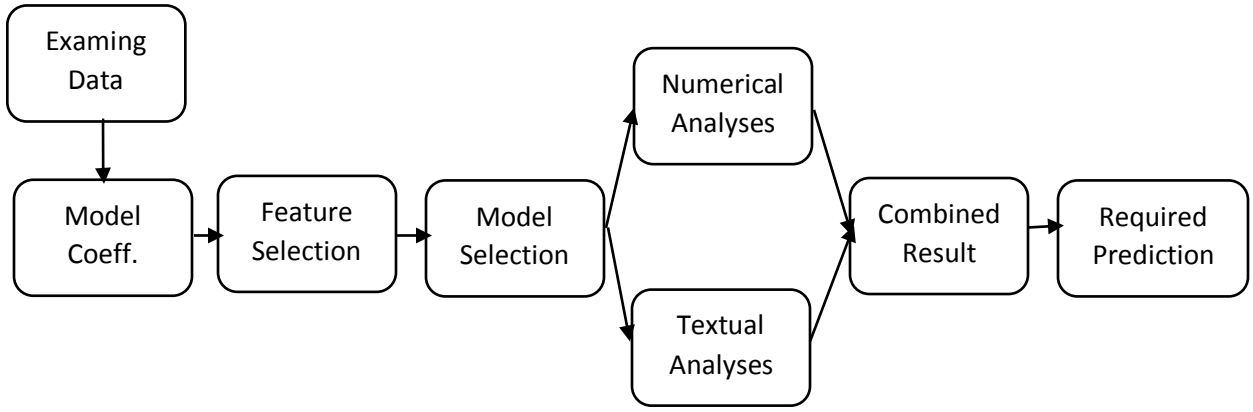

Figure 1: Architecture Diagram for the processes involved in mining

VI.I CORRELATION

A statistical test was done with the correlation matrix to check if the attributes in the dataset have any correlation among each other. A dataset with good attributes should not correlate among themselves. A correlation matrix was used to find the highly correlated attributes, as these variables can harm the models by carrying the same information resulting in overfitting. Hence, it is crucial to handle the problem of correlation. After going through the correlation matrix in Figure 2, we observed that there are several attributes with high correlation. So under the pretext of the attributes having the same information, we eliminate them.

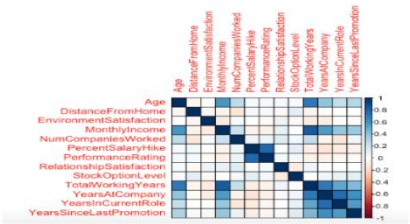

Figure 2: Correlation matrix for the attributes in the dataset

VII. RESULT AND ANALYSIS

The summary of the results obtained by the analysis of the dataset is as follows. We used logistic regression, AdaBoost and Bagging classifiers to know the best way to move forward. The accuracy of the testing data for the different classifiers is summarised below. The train to test split ratio is 80:20 in the dataset.

Classifier Model	Accuracy
Logistic Regression	0.84317
AdaBoost	0.84285
Bagging	0.83673

Table 3: Accuracy of various classifier models

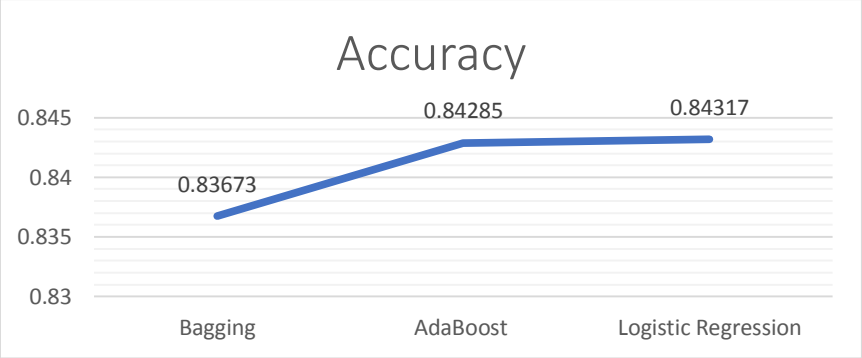

Figure 3: Accuracy of different classifier models

From the above Table 3 and Figure 3, we understood that logistic regression is the best way forward. So we experimented with the different tuning methods in logistic regression to obtain an accurate model. We built regression models with varying sizes of the dataset and achieved the following results.

Name of the model	Number of samples	Accuracy
Model0	250	0.84399
Model1	500	0.84799
Model2	700	0.84571
Model3	1000	0.83499
Model4	1200	0.83833
Model	1470	0.84317

Table 4: Relation between number of samples and accuracy

After this, we duplicated the dataset to increase the available dataset and recorded their accuracy. We applied cross-validation with ten folds to the following models.

Number of samples [times]	Accuracy
1470[1]	0.84082
2940[2]	0.840819
4410[3]	0.84127
5880[4]	0.84183

Table 5: Comparision of accuracy with the increased duplicated dataset

From the above Table 4 and Table 5, we concluded that the number of samples in the dataset and cross-validation plays little to no role in increasing the accuracy. We then experimented with the different train to test split ratios in the dataset to find the best split ratio.

Train test split	Accuracy
0.6 : 0.4	0.84149
0.7 : 0.3	0.84217
0.8 : 0.2	0.84217
0.9 : 0.1	0.84217

Table 6: Different train to test split ratio and its accuracy

So, from the above observations, we concluded that complete dataset with 1470 entries and 80:20 split ratio with no cross-validation is the best way to move forward. The performance metrics for the final model is as follows:

	Precision	Recall	F1-score	Support
0	0.84	1.00	0.91	371
1	0.50	0.01	0.03	70
Avg/total	0.79	0.84	0.77	441

Table 7: Performance Metrics

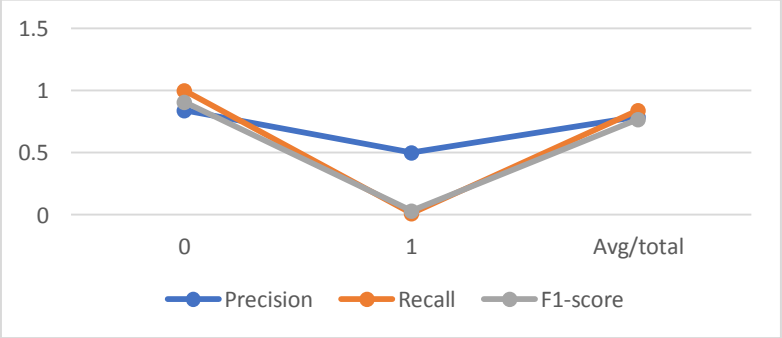

Figure 4: Precision, Recall and F1 score for 0-no, 1-yes of Attrition class

We then calculated the confusion matrix for the model and obtained the following results. Confusion matrix, also known as an error matrix, is a specific table layout that allows visualisation of the performance of an algorithm based on true positive, true negative, false positive and false negative values.

a	b	<-- classified as
100	137	**a = Yes**
39	1194	**b = No**

Table 8: Confusion Matrix

The accuracy metrics of the model are tabulated in Table 9. It is followed by the piecharts which show the distribution of employees who are at-risk of attrition based on their education field and the department that they work. All these visualisations give a better representation of the data.

TP Rate	FP Rate	MCC	ROC Area	PRC Area	Class
0.422	0.032	0.491	0.835	0.624	**Yes**
0.968	0.578	0.491	0.835	0.953	**No**
0.880	0.490	0.491	0.835	0.900	

Table 9: Accuracy metrics of the model

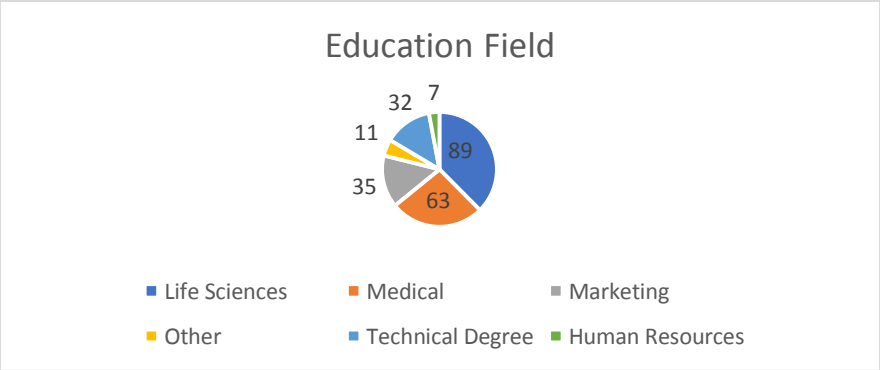

Figure 5: Distribution of at-risk attrition employees based on their education field

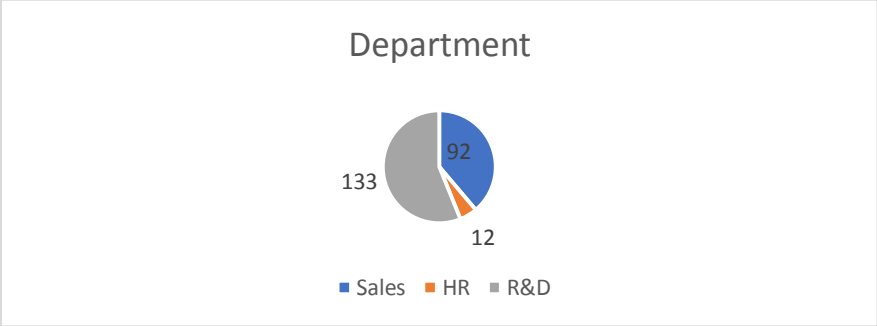

Figure 6: Distribution of at-risk attrition employees based on Departments they work

VIII. TEXTUAL ANALYSIS

The exit interview data collected from the employees leaving the company and the data from the feedback form filled by the employees every three months were used to gain helpful information in the analyses. We applied naive Bayes multinomial analysis for analysing the different emotions expressed by the employee when leaving and while working. The data in the text form is converted into analysable data for proper manipulation. The output helped in understanding the needs and requirements of the employees and the reason for their turnover. The result helped the company in mitigating further attrition as they were able to predict at-risk employees and the reason for their turnover.

Fig 7: Expressions for employees various emotions

IX. IMPLEMENTATION STRATEGY

The models were implemented using Python programming and run in Spyder IDE. A user-friendly GUI was provided to the companies through which they fed their data into the approach to get a proper prediction. The approach enables the companies to change the number of samples, classifiers used etc. according to their needs.

The Sklearn python package proved very useful by providing functions to perform regression, clustering, cross-validation etc. Sklearn is a free ML library based on Python programming language. The package also has data preprocessing approaches like normalisation, which proved to be helpful in data cleaning.

X. POTENTIAL BENEFITS

The Organisations around the world may use this approach to analyse textual data along with numerical data. This analysis enables companies to get new and unexplored information in all fields. This research will help the higher management to take better decisions and create a preliminary plan of actions to avoid the sudden attrition of resources. Here, we also target to provide the administration with the accurate prediction of employees who have the highest chances of leaving, thus focusing the research towards the true positive accuracy and F1 score.

XI. FURTHER WORK

Our work deals the combined formats of data to support the attributes with in themselves as well as the target output. This approach encourages us to go for further datasets from the domain where attributes are inherently structured differently. One such domain health care, is a potential candidate involving patients' disease profiles in terms of Imaging reports, physicians' prescription text notes, and direct numerical observational readings. This can be augmented by dynamic sequence of images in video format and made support for the prediction results. By this proposal of foreseeing one may travel more distances in health care industry for flawless diagnosis and fruitful treatment.

REFERENCES

[1] Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms. *algorithms*, *4*(5), C5.

[2] Alao, D. A. B. A., & Adeyemo, A. B. (2013). Analyzing employee attrition using decision tree algorithms. *Computing, Information Systems, Development Informatics and Allied Research Journal*, *4*.

[3] Barvey, A., Kapila, J., & Pathak, K. (2018). Proactive Intervention to Downtrend Employee Attrition using Artificial Intelligence Techniques. *arXiv preprint arXiv:1807.04081*.

[4] Bennett, N., Blum, T. C., Long, R. G., & Roman, P. M. (1993). A firm-level analysis of employee attrition. *Group & Organization Management*, *18*(4), 482-499.

[5] Dubey, A. K., Maheshwari, I., & Mishra, A. Predict Employee Retention Using Data Science.

[6] Frye, A., Boomhower, C., Smith, M., Vitovsky, L., & Fabricant, S. (2018). Employee Attrition: What Makes an Employee Quit?. *SMU Data Science Review*, *1*(1), 9.

[7] Kane-Sellers, M. L. (2007). *Predictive models of employee voluntary turnover in a North American professional sales force using data-mining analysis*. Texas A&M University.

[8] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, 3-24.

[9] Kumar, R. K., Poonkuzhali, G., & Sudhakar, P. (2012, March). Comparative study on email spam classifier using data mining techniques. In *Proceedings of the International MultiConference of Engineers and Computer Scientists* (Vol. 1, pp. 14-16).

[10] Lau, S., & Albright, L. (2011). ADVICE AND ANALYSIS SOLUTIONS Positive turnover, disability awareness, employee selection guidelines. *HRMagazine*, *56*(1), 20.

[11] O'Byrne, T. (2013). History of Employee Engagement–from Satisfaction to Sustainability. *HR Zone*.

[12] SEBT, M. V., & YOUSEFI, H. (2015). Comparing data mining approach and regression method in determining factors affecting the selection of human resources. *Fen Bilimleri Dergisi (CFD)*, *36*(4).

[13] Sikaroudi, E., Mohammad, A., Ghousi, R., & Sikaroudi, A. (2015). A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing). *Journal of Industrial and Systems Engineering*, *8*(4), 106-121.
[14] SUCEENDRAN, K., SARAVANAN, R., DIVYA ANANTHRAM, D. S., KUMAR, R. K., & SARUKESI, K. Applying Classifier Algorithms to Organizational Memory to Build an Attrition Predictor Model.
[15] Tamizharasi, K., & Rani, U. (2014). Employee Turnover Analysis with Application of Data Mining Methods. *International Journal of Computer Science and Information Technologies*, *5*(1), 562-566.
[16] https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

