Journal of Theoretical and Applied Information Technology
15th May 2023. Vol.101. No 9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ANALYTIC APPROACH OF PREDICTING EMPLOYEE ATTRITION USING DATA SCIENCE TECHNIQUES

R. VINSTON RAJA[1,*], A. DEEPAK KUMAR[2,+], DR. I. THAMARAI[3,+] ,S. NOOR MOHAMMED[4,+], R. RAJESH KANNA[5,+]

[1,*]Assistant Professor, Department of Information Technology, Panimalar Engineering College, Chennai

[2,+] Assistant Professor, Department of Computer Science and Engineering , St. Joseph's Institute of Technology, Chennai

.[3,+]Associate Professor, Department of Computer Science and Engineering, Panimalar Engineering College Chennai City Campus, Chennai.

[4,+]Assistant Professor, Department of Computer Science and Engineering, Loyola Institute of Technology, Chennai

[5,+]Assistant Professor of Department of Computer Science and Engineering at Agni College of Technology, Thalambur, Chennai.

[1,*]rvinstonraja@gmail.com, [2,+]deepakkumar@stjosephstechnology.ac.in, [3,+] thamarai.panimalar@gmail.com, [4,+]noormugam786@gmail.com, [5,+]rkanushaquafarms@gmail.com

ABSTRACT

Employee turnover has turn out to be a large venture for data technological know-how companies. The departure of key software program builders would possibly reason large loss an IT business enterprise in view that they additionally leave with essential commercial enterprise understanding and integral technical skills. It is fundamental for IT companies to apprehend developer turnover in order to keep certified builders and reduce injury due to developer exit. In this research, monthly self-report of the software developers includes developer's activities, working hours, no of projects they have been assigned etc. will been taken into account for analysis for doing the prediction with the help of data science algorithm.By the usage of NB algorithm, KNN algorithm and SVM algorithm, prediction mannequin hasbeen in contrast on the experimental groundwork and supply the end result of which algorithmis performing better. Then, this fantastic mannequin will be given to HR managers to predict whether or not the worker will depart the corporation or not.

Keywords : *Employee Attrition, Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Turnover Rate*

1. INTRODUCTION:

Employees are the treasured property of any organization. However, if you quit your job unexpectedly, the company will cost a lot of money. Not only are new employees wasting money and time, but new employees are also spending time making profits for their companies. Employee turnover is a diversity of existing staff and is replaced by new staff for a period of time. Staff turnover (Lingfeng Bao et all.,2017) is one of the most urgent troubles in the enterprise's body of workers management. Previous lookup on the difficulty of workforce turnover failed to be successful in managing, supervising, and stopping employee turnover on manufacturing traces with excessive turnover rates.

Employee churn (Andry Alamsyah et all., 2018) has a range of bad penalties for a business, such as unequal workload distribution, huge monetary losses, and the extra time required to recruit a replacement, all of which can lead to an expand in consumer unhappiness. Tangible fees encompass coaching costs and the time it takes from when a worker begins to when they begin contributing intangible expenses contain what is closing when an environment friendly worker quits: new product ideas, high-quality challenge management, or client relationships.

A survey of 1,000 full-time employees

performed by way of the on-line recruitment company (Tara Safavi et all., 2018) Headhunter.net reviews that 78% would take a new function if the proper possibility comes alongside and 48% these that are employed are on the lookout for sparkling opportunities. Because software program engineers may go away with a lot of crucial data and expertise, if developer turnover (Mingfei Teng et all.,2019) is no longer properly managed, it can undermine the success of a software program venture and end result in big losses for the firm. As a result, being in a position to pick out who will go away the organization early will permit the corporation to preserve brilliant software program builders whilst minimizing the loss when they leave.

Data science is the most precious and broadly used technique for the exploration and evaluation of giant extent of records to accumulate valid, novel, probably useful and clever patterns hidden in data. Data science techniques are extensively used for modeling real-world problems. Specifically, for classification, regression and clustering, data science has been used extensively in recent times. The key tendency in statistics technological know-how is to extracttreasured data from the large quantities of facts saved in files, databases, and different repositories by using creating tremendous methods of examining and deciphering such data.

Many data scientist communities are working on this problem for the earlier prediction of employee attrition (Dilip Singh Sisodia et all., 2017).

2. PROBLEM DEFINITION

In today's world many industries and specially IT are experiencing high attrition rate. Some common causes of attrition in their organization are known to managers and HR departments.This difficulty is triggered by using disappointment with various elements of a job, such as profession aspirations, work location, salary, overall performance management, job satisfaction, and managers, amongst others. Employee attrition (Francesca Fallucchi et all.,2020) manipulate is integral to the long-term fitness and success of any organization. Employees who leave on their own accord have a negative influence (James M. Vardaman et all., 2015) on the organization or project in which they are employed. Any industry's HR and senior management, as well as policymakers, are collaborating to reduce voluntary exits.

3. SCOPE

Retaining the high-quality personnel ensures patron satisfaction. Increased revenues and comfortable colleagues and staff (Xiao-Li Qu et all., 2015). To hold employees, organizations spend a lot of cash on training, presenting onsite opportunities, and paying above-market salaries. Build a data-driven turnover (Thomas Hugh Feeley et all., 2019) prediction model to predict future turnover (Zubin R et all., 2013), each at combination degrees as properly as for figuring out people with excessive threat of attrition. Using data science classification algorithm, employee turnover can be predicted accurately. Data science is an aggregate of records and AI. Data Science is the series of historic and latest trends in statistics, AI, and machine learning. Data science is a process to do analysis such as frequent pattern mining andassociation rule mining, classification, clustering and prediction to find the intelligent pattern hidden in the huge amount of dataset.

This research predicts whether developers will leave the company after a positive period of time, based on monthly reports. For prediction, Naive Bayes algorithms, KNN and SVM computing device algorithms that used to classify and predict the developer turnover the usage of factors,such as pleasure level, remaining evaluation, wide variety projects, average_monthly_hours, time_spent_company, etc. The overall performance of all the algorithms has in contrast in phrases of accuracy. then the great mannequin will predict the worker turnover. The insights, alongside with data-driven predictive models (Mohammad Nayeem Hasan et all., 2019), can be used to sketch positive plans for lowering attrition, enhancing retention, lowering attrition expenses and mitigating attrition effects.

4. SYSTEM ANALYSIS:

4.1. Existing System

Existing system includes only few attributes for analysis and also deals with qualitative observations and simple statistical analysis. The qualitative observations deal with the data that can be observed through human senses (Huang Xu, Zhiwen Yu et all., 2015). They do not involve measurements or number. The simple statistical analysis includes Mean, Standard deviation, median, finding the size of data, variance etc., the results produced by this technique are not precise. With the current increase in IOT and connected device, we now have access to so much of data and along with it an increase needs to manage and understand data.

4.2. Proposed System

Employee turnover is an issue that has been the focus of research over the last few decades. Some industries, such as call centers, tend to have higher turnover (Bin Lin et all., 2017) issues than others, but in general this affects all industries. Employers are also affected in other ways, such as the loss of expertise or knowledge that employees have. This research collects and analyzes a dataset of monthly employee reports to predict whether developers will leave the company after a period of time. Naive Bayesian algorithms, K-nearest neighbors, and support vector machine data science techniques were applied to the predictions.

Data Science is an interdisciplinary field that incorporates computer science, mathematics, statistics and domain knowledge. Data Science is a process to do the analysis such as frequent pattern mining, Association rule mining, classification, clustering, regression etc. to find the intelligent patterns hidden in the huge number of datasets. Using Data Science, more accurateprediction result will be obtained. Naive Bayes algorithm is used both for classification and predication. NB, KNN and SVM algorithms are applied to the employee monthly report data and the performance is compared. The efficient model is used for delivering the result of whether the employee will leave the company or not.

5. SYSTEM SPECIFICATION:

R is a data manipulation, statistical computing, and graphics programming language and environment. John Chambers and colleagues at Bell Laboratories invented it. Similar to S, but available on all platforms (Linux, Mac, and Windows). Regression, Clustering, Classification,Association Rule Mining, and more graphical and statistical (Vachik S. Dave et all., 2018) approaches are available. It compiles and operates on a range of UNIX and comparable platforms, as nicely as Windows and Mac OS X. The R software program environment's supply code is typically written in C, FORTRAN, and R.

The Integrated Development Environment (IDE) for R is RStudio. R is made less difficult to use with RStudio. It comes with a code editor, compiler or interpreter, debugger, and visualization tools (Jian Tang et all., 2015), as well as charting, history, syntax-highlighting editor, and workspace management tools. JJ Allaire, the developer of the ColdFusion programming language, launched RStudio. RStudio is accessible in each open supply and industrial forms, with two versions: RStudio Desktop (for Windows, macOS, and Linux) and RStudio Server (for gettingaccess to RStudio whilst it is jogging on a faraway Linux server). The C++ programming language is used to create RStudio. RStudio makes R more organized and provides new features.

6. SYSTEM ARCHITECTURE:

The figure below shows the architectural illustration of this design defines the inflow of data for Employee Attrition. There are more than a few steps for inspecting the worker attrition such as records importing, records pre-processing, prediction the usage of Naive Bayes algorithm, prediction the use of KNN algorithm, prediction the use of assist vector desktop and the overall performance assessment between Naive Bayesian, KNN and SVM algorithm.

The first process is data importing, Data has to be loaded in to the R surroundings for analysis. The

employee data set is loaded into environment for prediction operation. The packages necessary for Naive bayes, KNN and SVM have to be loaded into the R environment. For Naive bayes algorithm, Naive Bayes package, for KNN and for SVM, e1071 have to be loaded into R Environment.

In records pre-processing, attribute selection, standardization and normalization features will be applied. In standardization, raw statistics is modified into common, comprehensible format. In attribute selection, maintain solely the attributes which is affecting the evaluation and it is no longer critical to preserve all the attributes for doing the analysis. In Normalization, imply of the attribute will be zero and trendy deviation will be 1. The pre-processed information is given to sampling, in sampling, dataset is partition into two sets, training dataset and testing dataset with the chance of 80% and 20%.

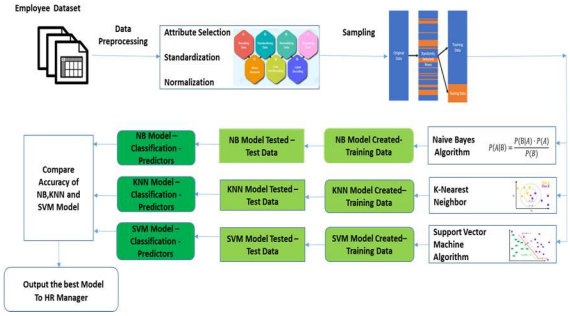


Figure-1. Architectural Diagram

The overall performance of NB, KNN and SVM is in contrast in terms of accuracy in which mentioned in Figure-1. The confusion matrix is created for discover the accuracy of the model. Therefore, exceptional mannequin will be given to the HR supervisor to generate the output of whether or not a developer will depart the organization or will proceed in the company.

7. MODULES:

There are five Modules

- Data Importing and Preprocessing
- Model Generation Using Naive Bayes Algorithm
- Model Generation Using K-NN Algorithm
- Model Generation Using SVM Algorithm
- Comparison of NB, KNN and SVM Algorithm

8. DATA IMPORTING AND PREPROCESS:

Data is available in any file format like .txt, .csv,.xlsx, spss etc. Data have to be loaded into R surroundings for analysis. Once data have been extracted from the file it should be stored in a data frame. Libraries necessary for classification algorithm-NB and SVM have to be installed into the R environment. For NB algorithm, naive Bayes package, for KNN, class package and for Support Vector Machine, e1071 package have to be installed and loaded into R environment.

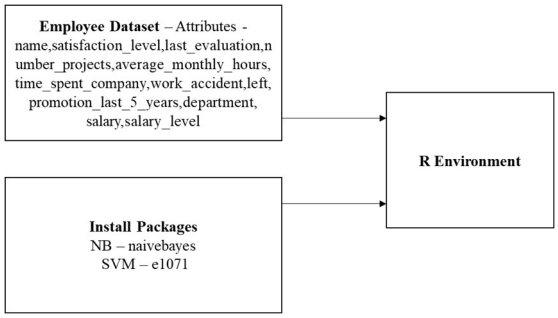


Figure-2. Data Importing and Preprocessing

Data preprocessing is the statistics mining method that includes reworking raw records into comprehensible format in which mentioned in Figure-2. The raw statistics is rather inclined to noise, missing values, and inconsistency. Real-world records are regularly incomplete, inconsistent, and inclined to inaccuracies. Simple statistical techniques must be used to overcome the missing values problem. To improve the quality of the records, the mining sequence is preprocessed from the

raw statistics. It is now not indispensable to preserve all the attributes for doing the analysis, we can preserve solely the attributes which is affecting the analysis.

This improves layout effectiveness, improves contrasting processes, and enables collaborative

Journal of Theoretical and Applied Information Technology
15th May 2023. Vol.101. No 9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

search and large-scale (Liangyue Li et all., 2019) analysis. Normalization is a scaling approach that shifts and rescales values to make them vary between zero and 1. It's occasionally referred to as Min-Max scaling.

Formula used for Normalization:
X-Xmin/Xmax-Xmin

9. MODEL GENERATION USING NAIVE BAYES ALGORITHM:

Structured and unstructured facts can both be classified. Data is labeled into a set of lessons the use of the classification procedure. A classification problem's primary cause is to decide which class or type new statistics will belong to. Structured and unstructured records can each be classified. Data is labeled into a set of instructions the use of the classification procedure. A classification problem's fundamental motive is to decide which class or classification new information will belong to. A classification challenge's essential intention is to parent out which category or type sparkling information will fall into.

$$P(B/C) = \frac{P(C/B)P(B)}{P(C)}$$

The Naive Bayes algorithm is primarily based on Bayes' theorem and assumes that each pair of elements is independent. This algorithm simply requires a little extent of coaching records to estimate the proper parameters. In comparison to more complicated algorithms, Naive Bayes classifiers are incredibly fast. Preprocessed data is given as a input to Naive Bayes algorithm and Naive Bayes model has been created, which is used for classification and prediction.

Naive Bayes Algorithm DFD Preprocessed data is given as a input to Naive Bayes algorithm and Naive Bayes model has been created, which is used for classification and prediction in which mentioned in Figure-3.

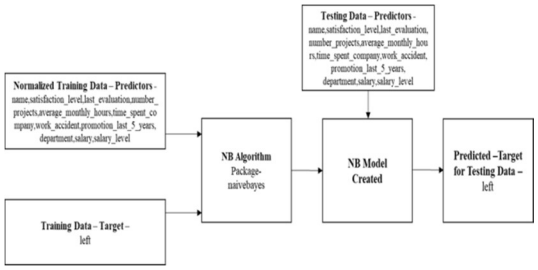


Figure-3. Naive Bayes Algorithm DFD

9.1. Algorithmic Steps for NavieBayes Classification

Step 1: Create a frequency table from the data set
Step 2: Create Likelihood table by finding the probabilities values of each attribute
Step 3: Now, compute the posterior probability for each class using the NavieBayes equation. Class with highest posterior probability in the outcome of prediction in which mentioned in Table-1,2,3.

Table-1. Data Sets

Day	Climate	Temperature	Humidity	Air Speed	Play
1.	Sunny	Hot	High	Weak	No
2.	Sunny	Hot	High	Strong	No
3.	Overcast	Hot	High	Weak	Yes
4.	Rain	Mild	High	Weak	Yes
5.	Rain	Cool	Normal	Weak	Yes
6.	Rain	Cool	Normal	Strong	No
7.	Overcast	Cool	Normal	Strong	Yes
8.	Sunny	Mild	High	Weak	No
9.	Sunny	Cool	Normal	Weak	Yes
10.	Rain	Mild	Normal	Weak	Yes
11.	Sunny	Mild	Normal	Strong	Yes
12.	Overcast	Hot	Normal	Weak	Yes
13.	Overcast	Hot	Normal	Weak	Yes
14.	Rain	Mild	High	Strong	No

Table-2. Frequency Table

Climate			Temperature			Humidity			Air Speed			Play	
	Y	N		Y	N		Y	N		Y	N	Y	N
Sunny	2	3	Hot	2	2	High	3	4	Weak	6	2	9	5
Overc	4	0	Mil	4	2	Norm	6	1	Stro	3	3		

ast			d			al			ng				
Rainy	3	2	cool	3	1								

Table-3. Likelihood Table

Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	Weak	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	Stg	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Sunny, Cool, High, Strong

P(Yes |S,C,H,S) = P(S,C,H,S | Yes) *P(Yes) / P(S,C,H,S)

P(No | S,C,H,S) = P(S,C,H,S | No) *P(No) / P(S,C,H,S)

P(S,C,H,S) = P(S)*P(C)*P(H)*P(S)

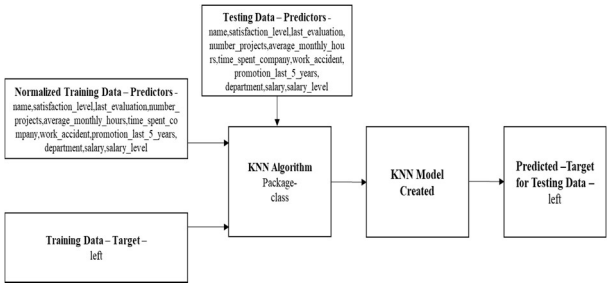
= 5/14 * 4/14 *7/14*6/14 = 0.0216

Likelihood for Yes = 2/9 * 3/9 * 3/9* 3/9 = 0.0082
Likelihood for No = 3/5 * 1/5 * 4/5* 3/5 = 0.0576
Posterior Probability of Yes = (0.0082* 9/14) / 0.0216 =0.245Posterior Probability of No = (0.0576 * 5/14) / 0.0216 =0.952
The posterior probability for the No is higher, so the probability of no is higher. Once Naive Bayes model have been created, by using testing data, model can be evaluated and by giving only predictor variables, target variable can be predicted.

10. MODEL GENERATION USING K-NEAREST NEIGHBOR:

KNN is used for each Classification and Regression. Classifying the facts factors primarily based totally on how its neighbor are classified. K-Nearest Neighbor is a Supervised Learning-primarily based totally Machine Learning set of rules this is one of the maximum basic (Bryan Perozzi et all., 2014). The KNN set of rules assumes that the brand-new case/facts and present instances are comparable and locations the brand-new case with inside the class this is maximum much like the present categories. The KNN set of rules may be used for each regression and type, however it's miles greater typically applied for type tasks in which mentioned in Figure-4. The KNN set of rules is a non-parametric set of rules, this means that it makes no assumptions approximately the facts.

Let (Xi, Ci) Xi denotes feature value, Ci denotes labels (Class) for each Xi.Let X be the data point for which label have to be find out using KNN.Preprocessed data is given as a input to KNN algorithm and KNN model has been created, which



is used for classification and prediction.

Figure-4. KNN Algorithm DFD

10.1 KNN Algorithm Steps:

Step1:Calculate D(X,Xi)i=1,2,3,…n; where d denotes Euclidean Distance between data points.

Step2: Arrange the calculated n Euclidean Distance in increasing order.

Step3: Let K be a +ve integer, take first K distance from the sorted list.

Step4: Let Ki denotes the number of points belonging to the i^{th} class among k points.

Step5: if Ki > Kj then put X in class i

How to choose the K value is Sqrt (no of data points)

Odd value of K is selected to avoid confusion between the 2 classes in which mentioned in Table-4.

.

Table-4. Height and Weight Data Sets

Weight	Height	Class
51	167	Underweight
62	182	Normal
69	176	Normal
64	173	Normal
65	172	Normal
56	174	Underweight
58	169	Normal
57	173	Normal
55	170	Normal

On the basis of given data classify the following datasetWeight = 57 kg
Height = 170cmClass=?
Calculate the Euclidean distance of new data point from all the points.

Euclidean distance between 2 points (x,y) and (a,b)
isdist(d) = sqrt((x-a) ^2 + (y-b) ^2)
Let K=3

ED for 1st data point
Sqrt ((57-51) ^2 + (170-167) ^2) = 6.7
Select K entries which are closest to the new sample. Find the most common classification of these entries. This is the classification of the new sample in which mentioned in Table-5.

.

Table-5. Euclidean Distance

Weight	Height	Class	Euclidean Distance
58	169	Normal	1.4
55	170	Normal	2
57	173	Normal	3
56	174	Underweight	4.1
51	167	Underweight	6.7
64	173	Normal	7.6
65	172	Normal	8.2
62	182	Normal	13
69	176	Normal	13.4

Majority of neighbors are pointing Normal, so the data point (57,170) belongs to Normal. Once KNN model have been created, by giving only predictor variables, target variable canbe predicted.

12. MODEL GENERATION USING SUPPORT VECTOR MACHINE:

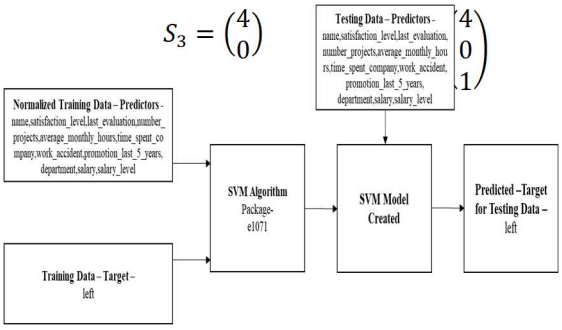
Preprocessed data is given as a input to SVM algorithm and SVM model has been created, which is used for classification and prediction.

Figure-5. Support Vector Machine Algorithm DFD

$$S_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \qquad \tilde{S}_1 = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}$$

$$S_2 = \begin{pmatrix} 2 \\ -1 \end{pmatrix} \qquad \tilde{S}_2 = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}$$

$$S_3 = \begin{pmatrix} 4 \\ 0 \end{pmatrix} \qquad \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix}$$



SVM is a time period that refers to a kind of computer that makes use of vectors to remedy for classification and regression analysis, it is a machine learning approach. Learning algorithms train supervised learning models. They look for patterns in the enormous volume of data. By creating two parallel lines, an SVM creates parallel divisions. It creates flat and linear partitions by separating the space in a single pass in which mentioned in Figure-5. A clear separation, as large as practicable, should be used to separate the two categories. Use the hyper plane to partition the data. In a high- dimensional space (Aditya Grover et all., 2016) an SVM produces hyper planes with the biggest margin to classify data.

Journal of Theoretical and Applied Information Technology
15th May 2023. Vol.101. No 9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The distance between the two classes' closest data points is represented by the margin betweenthem in which mentioned in Figure-6. The lower the classifier's generalization error, the larger the margin. Following the training map, new data is placed in the same space to forecast which category it belongs to. Byusing training data, categorize the new data into various divisions.

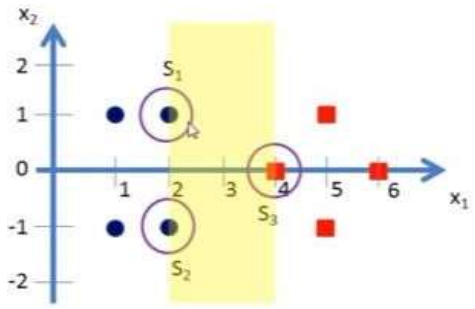


Figure-6. Hyper plane partitioning the Data

First step is to select support vectors to get the boundary between the 2 classes. The figure-6 shows the selection of 3 support vectors S1, S2 and S3. Use Vectors augmented with 1 as bias input and differentiate these with over-tilde symbol. The support vectors and modified support vectors are shown below.

Form the linear equations, and find the value of the parameters.

$$a_1 \tilde{S}_1 . \tilde{S}_1 + a_2 \tilde{S}_2 . \tilde{S}_1 + a_3 \tilde{S}_3 . \tilde{S}_1 = -1\ (-ve\ class)$$
$$a_1 \tilde{S}_1 . \tilde{S}_2 + a_2 \tilde{S}_2 . \tilde{S}_2 + a_3 \tilde{S}_3 . \tilde{S}_2 = -1\ (-ve\ class)$$
$$a_1 \tilde{S}_1 . \tilde{S}_3 + a_2 \tilde{S}_2 . \tilde{S}_3 + a_3 \tilde{S}_3 . \tilde{S}_3 = +1\ (+ve\ class)$$

Substitute the modified vector values in to the equation.

$$a_1 = \begin{pmatrix}2\\1\\1\end{pmatrix}.\begin{pmatrix}2\\1\\1\end{pmatrix} + a_2 = \begin{pmatrix}2\\-1\\1\end{pmatrix}.\begin{pmatrix}2\\1\\1\end{pmatrix} + a_3 = \begin{pmatrix}4\\0\\1\end{pmatrix}.\begin{pmatrix}2\\1\\1\end{pmatrix} = -1$$

$$a_1 = \begin{pmatrix}2\\1\\1\end{pmatrix}.\begin{pmatrix}2\\-1\\1\end{pmatrix} + a_2 = \begin{pmatrix}2\\-1\\1\end{pmatrix}.\begin{pmatrix}2\\-1\\1\end{pmatrix} + a_3 = \begin{pmatrix}4\\0\\1\end{pmatrix}.\begin{pmatrix}2\\-1\\1\end{pmatrix} = -1$$

$$a_1 = \begin{pmatrix}2\\1\\1\end{pmatrix}.\begin{pmatrix}4\\0\\1\end{pmatrix} + a_2 = \begin{pmatrix}2\\-1\\1\end{pmatrix}.\begin{pmatrix}4\\0\\1\end{pmatrix} + a_3 = \begin{pmatrix}4\\0\\1\end{pmatrix}.\begin{pmatrix}4\\0\\1\end{pmatrix} = +1$$

After simplification, the following equations are obtained.

$$6a_1 + 4a_2 + 9a_3 = -1$$
$$4a_1 + 6a_2 + 9a_3 = -1$$
$$9a_1 + 9a_2 + 17a_3 = +1$$

After further simplification, the parameters values will be obtained as follows:

$$a_1 = a_2 = -3.25\ and\ a_3 = 3.5$$

The hyperplane that discriminates the positive and negative class is given by

$$\tilde{w} = \sum_i a_i \tilde{s}_i$$

Apply all the parameters and support vector values and add them together, we will getthe following equation:

$$\tilde{w} = a_1\begin{pmatrix}2\\1\\1\end{pmatrix} + a_2\begin{pmatrix}2\\-1\\1\end{pmatrix} + a_3\begin{pmatrix}4\\0\\1\end{pmatrix}$$
$$\tilde{w} = (-3.25).\begin{pmatrix}2\\1\\1\end{pmatrix} + (-3.25).\begin{pmatrix}2\\-1\\1\end{pmatrix} + (-3.5).\begin{pmatrix}4\\0\\1\end{pmatrix} = \begin{pmatrix}1\\0\\-3\end{pmatrix}$$

Vectors are augmented with bias, therefore separate hyperplane from bias

$$y = wx + b$$

The hyperplane equation is:

$$w = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

The bias of offset value is:

$$b = -3$$

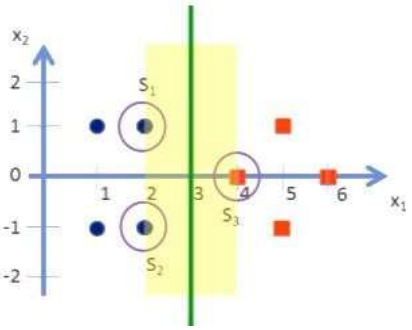


Figure-7. Hyperplane

When the value of w is (1 0), then the hyperplane is a vertical line. When the value of w is (0 1), then the hyperplane is a horizontal line. Since the offset value is -3 then hyperplane have to draw in the value 3 in which mentioned in Figure-7.

14. COMPARISON OF NAIVE BAYESIAN, K-NEAREST NEIGHBOR AND SUPPORT VECTOR MACHINE:

Once the classification mannequin is completed, it is imperative to consider its performance. The overall performance of the Naive Bayesian, KNN and SVM computing device classification fashions are in contrast in phrases of accuracy. The confusion matrix is created to calculate the accuracy of the classification model. The model's performance is described by the confusion matrix, which outputs a matrix/table. Error matrix is another name for it. The matrix contains a total number of correct and wrong predictions in a summary manner in which mentioned in Figure-8. The matrix appears in the table below.

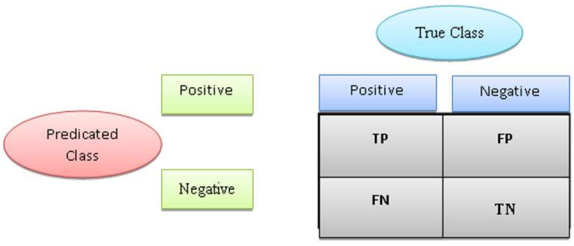


Figure-8. Accuracy Matrix

$$\boldsymbol{Accuracy} = (\boldsymbol{TP} + \boldsymbol{TN}) / (\boldsymbol{TP} + \boldsymbol{FP} + \boldsymbol{FN} + \boldsymbol{TN})$$

Naive Bayes Algorithm predicted the developer turnover with the accuracy of 76%, KNN with the accuracy of 94% and SVM with the accuracy of 96%. Performance of Support Vector Machine is more accurate when compared to Naive Bayes Algorithm and KNN for this dataset. So, SVM Algorithm mannequin is given to HR supervisor to l predicts and classify whether ornot the developer will go away the corporation or not.

15. THE PERFORMANCE OF NB, KNN AND SVM ALGORITHM

The performance of NB, KNN and SVM algorithm is compared using the accuracy of the classification model. Naive Bayes Algorithm predicted the developer turnover with the accuracy of 76%, K-Nearest Neighbor with the accuracy of 94% and Support Vector Machine with the accuracy of 96%. Performance of Support Vector Machine is more accurate when compared to Naive Bayes Algorithm and K- Nearest Neighbor algorithm for this employee dataset. The results of the experiment indicate that the accuracy of the Support vector machine algorithm is better when compared to Naive bayes algorithm and K- Nearest Neighbor in which mentioned in Figure-9.

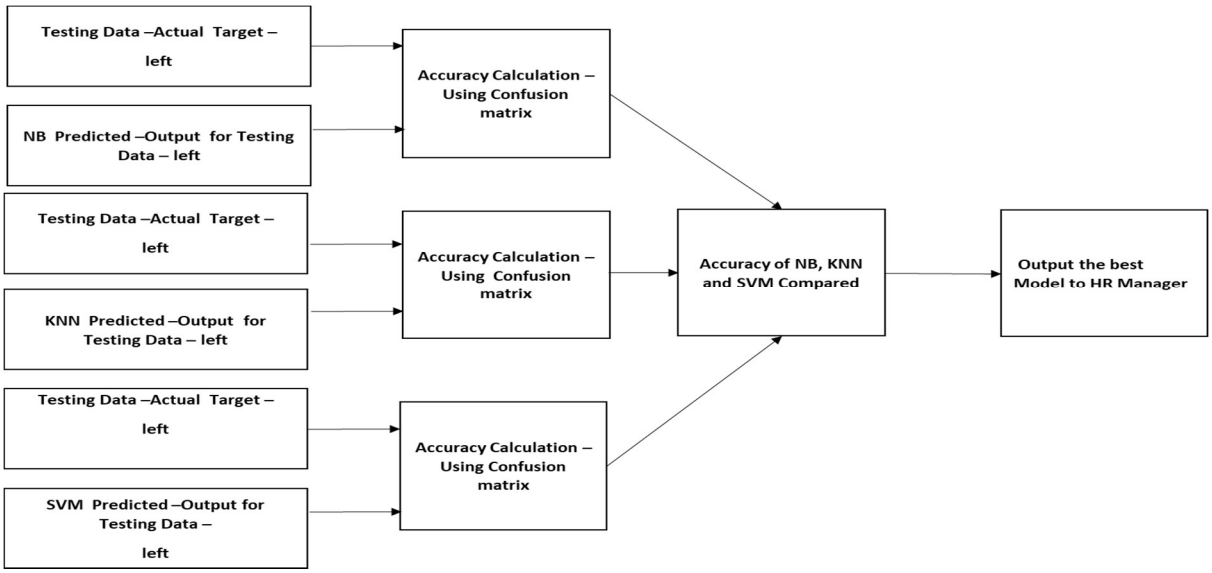
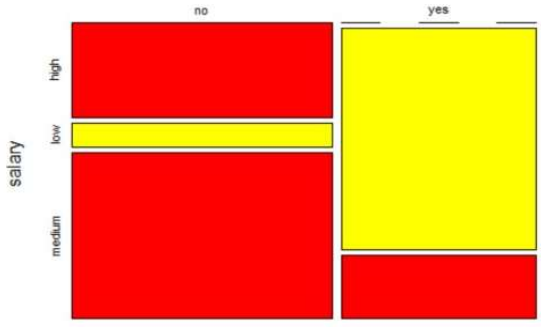


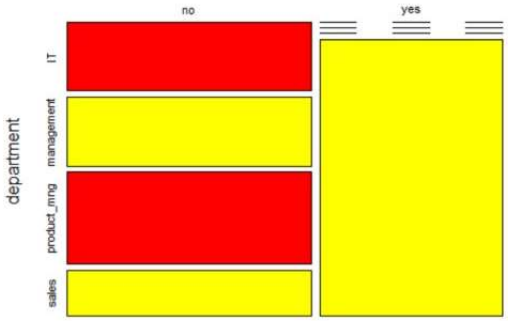
Figure 9 Performance Comparison of NB, K-NN and SVM DFD

16. PERFORMANCE OF NB ALGORITHM

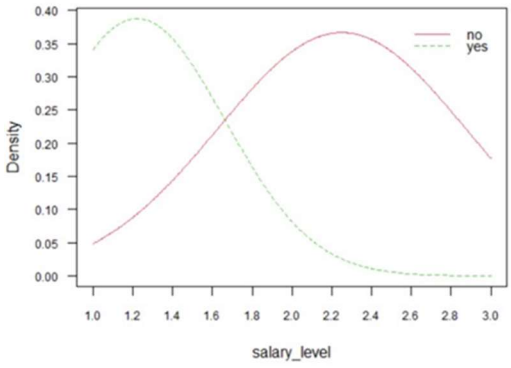
Naive Bayes Algorithm predicted the developer turnover with the accuracy of 76%, KNN with the accuracy of 94% and Support Vector Machine with the accuracy of 96%. Performance of Support Vector Machine is more accurate when compared to Naive Bayes Algorithm and KNN for this dataset in which mentioned in Figure-10. So, SVM Algorithm model is given to HR manager to l predicts and classify whether the developer will leave the company or not.



(a) Salary Analysis



(b) NB Model Inference – Department Vs Attrition



c) NB Model Inference – Salary Vs Attrition

Figure-10. Naive Bayes Algorithm predicted the employee turnover with the accuracy of 76%, KNN with the accuracy of 94% and SVM with the accuracy of 96%.

Journal of Theoretical and Applied Information Technology
15th May 2023. Vol.101. No 9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

17. IMPLEMENTATION

Implementing the Naive Bayes Algorithm predicted the employee attrition using r-programming which is mentioned in Figure-11, 12, 13. Algorithms have in contrast in phrases of accuracy. Naive Bayes Algorithm predicted the employee turnover with the accuracy of 76%, KNN with the accuracy of 94% and SVM with the accuracy of 96%. This task concludes, overall performance of SVM algorithm is greater correct over accuracy of Naive Bayesian algorithm and KNN

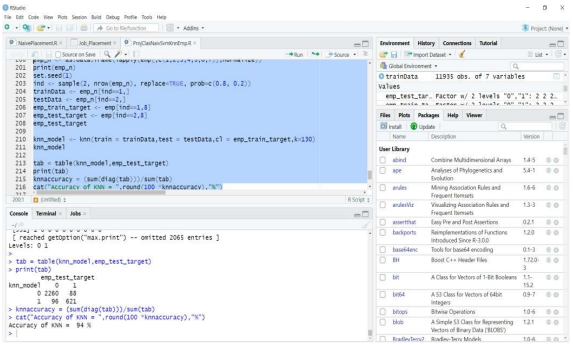


FIGURE-11: Performance of NB Algorithm

18. PERFORMANCE OF KNN ALGORITHM

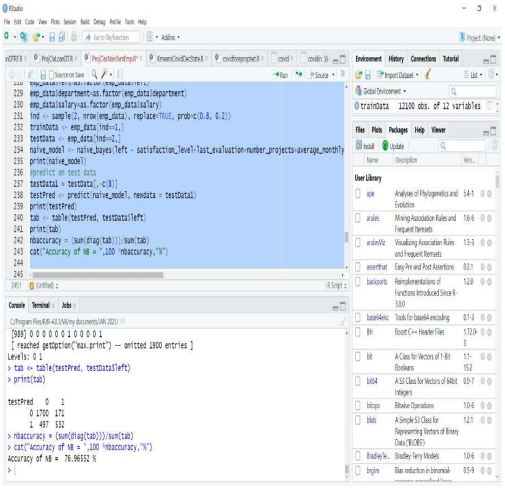


Figure-12. Performance of KNN Algorithm

19. PERFORMANCE OF SVM ALGORITHM

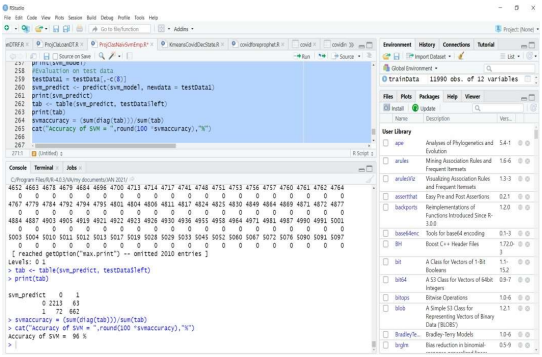


Figure-13. Performance of SVM Algorithm

20. CONCLUSION

This research addresses the Naive Bayes, KNN and SVM algorithm for Classifying and predicting, whether the employee will leave the company or not. The overall performance of each SVM, KNN and Naive Bayes Algorithms has in contrast in phrases of accuracy. Naive Bayes Algorithm predicted the employee turnover with the accuracy of 76%, KNN with the accuracy of 94% and SVM with the accuracy of 96%. This task concludes, overall performance of SVM algorithm is greater correct over accuracy of Naive Bayesian algorithm and KNN. This SVM mannequin can probably assist a business enterprise to predict the departure of key software program builders and they can be retained in the company, via taking proactive motion such as offering earnings hikes or bendy timing or through higher managing workload variance amongst mission contributors etc. to keep away from massive loss to company.

The future enhancement might include some more data science algorithms and deep learning algorithms. Along with Naïve Bayes, KNN and Support Vector Machine, the dataset can be examined with Deep Neural Network and overall performance of the algorithm is evaluated and as soon as the best algorithm has been found, that algorithm will be used to predict whether or not the precise worker will stay or go away the company.

REFERENCES

[1] Lingfeng Bao, Zhenchang Xing, "Who Will Leave the Company? A Large-Scale Industry Study of Developer Turnover by Mining Monthly Work Report", IEEE/ACM International Conference on Mining Software Repositories, 2017.

[2] Andry Alamsyah; Nisrina Salma, "A Comparative Study of Employee Churn Prediction Model", IEEE International Conference on Science and Technology, 2018.

[3] Dilip Singh Sisodia, " Evaluation of Machine Learning Models for Employee Churn Prediction ", IEEE International Conference on Inventive Computing and Informatics, 2017.

[4] Francesca Fallucchi, "Predicting Employee Attrition Using Machine Learning Techniques", Mdpi Journal of computers, 2020.

[5] Mohammad Nayeem Hasan, "A Comparison of Logistic Regression and Linear Discriminant Analysis in Predicting of Female Students Attrition from School in Bangladesh", International Conference on Electrical Information and Communication Technology, 2019.

[6] Xiao-Li Qu, " A decision tree applied to the grass-roots staffs' turnover problem", IEEE International Conference on Grey Systems and Intelligent Services, 2015

[7] L. Li, H. Jing, H. Tong, J. Yang, Q. He, and B.-C. Chen, ''Nemo: Next career move prediction with contextual embedding,'' in Proc. 26th Int. Conf. World Wide Web Companion,Apr. 2017, pp. 505– 513.

[8] M. Teng, H. Zhu, C. Liu, C. Zhu, and H. Xiong, ''Exploiting the contagious effect for employee turnover prediction,'' in Proc. 33rd AAAI Conf. Artif. Intell., Jul. 2019, pp. 1166– 1173.

[9] T. H. Feeley, J. Hwang, and G. A. Barnett, ''Predicting employee turnover from friendshipnetworks,'' J. Appl. Commun. Res., vol. 36, no. 1, pp. 56–73, Feb. 2008.

[10] T. H. Feeley, S.-I. Moon, R. S. Kozey, and A. S. Slowe, ''An erosion model of employee turnover based on network centrality,'' J. Appl. Commun. Res., vol. 38, no. 2, pp. 167–188, May 2010.

[11] J. M. Vardaman, S. G. Taylor, D. G. Allen, M. B. Gondo, and J. M. Amis, ''Translating intentions to behavior: The interaction of network structure and behavioral intentions in understanding employee turnover,'' Org. Sci., vol. 26, no. 4, pp. 1177–1191, 2015.

[12] B.Perozzi, R.Al-Rfou, and S. Skiena,''Deepwalk Online learning of social representations,'' in Proc. 20th ACM SIGKDD Int.Conf. Knowl. Discovery Data Mining, Aug. 2014, pp. 701–710.

[13] A. Grover and J. Leskovec, ''node2vec: Scalable feature learning for networks,'' in Proc.22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2016, pp. 855–864.

[14] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, ''Line: Largescale information network embedding,'' in Proc. 24th Int. Conf. World Wide Web, May 2015, pp. 1067–1077.

[15] H. Xu, Z. Yu, H. Xiong, B. Guo, and H. Zhu, ''Learning career mobility and human activity patterns for job change analysis,'' in Proc. IEEE Int. Conf. Data Mining, Nov. 2015, pp. 1057–1062.

[16] V. S. Dave, B. Zhang, M. Al Hasan, K. Aljadda, and M. Korayem, ''A combined representation learning approach for better job and skill recommendation,'' in Proc. 27th ACM Int. Conf. Inf. Knowl. Manage. (CIKM), Oct. 2018, pp. 1997–2005.

[17] Z. R. Mulla, K. Kelkar, M. Agarwal, S. Singh, and N. E. Sen, ''Engineers' voluntary turnover: Application of survival analysis,'' Indian J. Ind. Relations, vol. 2013, pp. 328–341, Oct. 2013.

[18] B. Lin, G. Robles, and A. Serebrenik, ''Developer turnover in global, industrial open source projects: Insights from applying survival analysis,'' in Proc. IEEE 12th Int. Conf. Global Softw. Eng. (ICGSE), May 2017, pp. 66–75.