Employee Attrition Prediction Using Machine Learning Tool PYTHON JUPITER

1.Giriraj H,2.Kashinath V Vernekar
Co Author: MAHENDRA KUMAR B

1,2 PG SCHOLORS, DEPT. OF MCA, DSCE
CG-ASST.PROF, DEPT. OF MCA, DSCE

Abstract - ***This mini project developed under a group of team with the guidance of experts which predicts the employee attrition using the machine learning tools jupyter notebook using python .One of the core objectives of machine learning is to instruct computers to use data or past experience to solve a given problem. A good number of successful applications of machine learning exist already, including classifier to be trained on email messages to learn in order to distinguish between spam and non-spam messages, systems that analyze past sales data to predict customer buying behavior, fraud detection etc. Machine learning can be applied as association analysis through Supervised learning, Unsupervised learning and Reinforcement Learning but in this study we will focus on strength and weakness of supervised learning classification algorithms.. We are optimistic that this study will help new researchers to guiding new research areas and to compare the effectiveness and impuissance of supervised learning algorithms.[8]***

Keywords **– Data Science, ,Logistic Regression, PreProcessing Tecniques, Prediction**

I. INTRODUCTION

Attrition, in Human Resource terminology,refers to the phenomenon of employees leaving the company. Attrition in a company is usually measured with a metric called attrition rate,which is simply measures the number of employees moving out of the company(voluntary resigning or lay off by the company). Attrition rate is also referred as churn rate or turnover.[9]

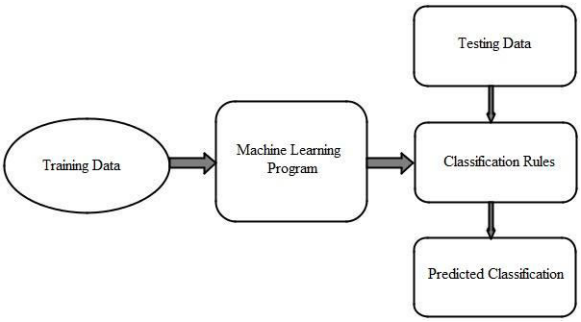
"the rate of shrinkage in size or number" in the best of worlds,employees would love their jobs,like their co-workers, work hard for their employers,get paid well for their work, have ample chances for advancement, and flexible schedules so they could attend to personal or family needs when necessary. And never leave. But then there's the real world, employees, do leave,either because they want more money, hate the working conditions, hate their co-workers, want a change, or because their spouse gets a dream job in another state.[10]

Data Science is a mixture of various tools, various algorithms, and machine learning principles with the objective to discover hidden patterns from the raw data.A Data Analyst as a rule clarifies what is happening by handling history of the information. Then again, Data Scientist not exclusively does the exploratory investigation to find bits of knowledge from it, yet in addition utilizes different propelled machine learning calculations to recognize the event of a specific occasion later on. A Data Scientist will take a gander at the information from numerous edges, at times edges not known before. Along these lines, Data Science is essentially used to settle on choices and forecasts making utilization of prescient causal examination, prescriptive investigation (prescient in addition to choice science) and machine learning.

II. MACHINE LEARNING

Machine learning is the process of making the machine tolearn itself through patterns and training data sets. Training data sets are data which is given to machine for understanding the hidden patterns within data and make relations for own understanding. It helps in working of machines efficiently by making them processed like a human brain. Pattern recognition is the most challenging task for developers to use such algorithms that allows different machines to work according to the requirement.
This paper emphasizes on making prediction of retention of an employee within an organization such that whether the employee will leave the company or continue with it. It uses the data of previous employees which have worked for the company and by finding pattern it predicts the retention in the form of yes or no. It uses various parameters of employees such as salary, number of years spent in the company, promotions, number of hours, work accident, financial background etc.

Considering new processing innovations, machine adapting today isn't care for machine learning of the past. It was conceived from design acknowledgment and the hypothesis that PCs can learn without being customized to perform assignments; specialists intrigued by manmade brainpower et.al [6] needed to check whether PCs could gain from information. The iterative part of machine learning is essential claiming as models are presented to new information, they can. freely adjust. They gain from past calculations to deliver solid, repeatable choices and results. It's a science that is not new – but rather one that is increasing crisp energy. While numerous machine learning calculations have been around for quite a while, the capacity to naturally apply complex scientific computations to huge information again and again, quicker and speedier is a current advancement.[17]



Machine learning algorithms are differentiated as supervised or unsupervised.

III TECHNOLOGY

We have utilized Python programming dialect, which is a translated, progressively written dialect and least difficult in grammar. Python is utilized for every one of the applications like in IOT advancement, information science field, web improvement, scripting reason and so forth. Consequently, now it is being utilized generally over the globe.

Python contains various number of libraries accessible in it, this makes it simple to use for each application like for web rejecting delightful cleanser, for GUI improvement TKinter, for web network urlib2, for machine learning sklearn et.al [8], numpy, pandas and so on. Python is one of the for the most part utilized dialect for Data Science applications since it gives libraries, for example, Pandas, nltk which can oversee substantial number of datasets into fitting way, it gives representation libraries like Matplotlib, Bokeh, Seaborn and so on that are exceedingly expressive regarding charts and plots portrayals.

The sklearn library is one which gives bigger number of machine learning calculations, for example, direct and various relapse, polynomial relapse, choice tree characterization and so on., to make expectations, bunching and grouping of information in number of billions.Machine learning is a branch in software engineering that reviews the outline of calculations that can learn. Run of the mill errands are idea learning, work learning or "prescient demonstrating", bunching and finding prescient examples. These undertakings are found out through accessible information that were seen through encounters or directions, for instance. The expectation that accompanies this teach is that including the experience into its assignments will in the end enhance the learning. However, this change needs to occur such that the learning itself ends up programmed with the goal that people like ourselves don't have to meddle any longer is a definitive objective.

Scikit-learn is the most helpful library for machine learning in Python. It is on NumPy, SciPy and matplotlib, this library contains a great deal of efficient devices for machine learning and factual displaying including arrangement, relapse, bunching and dimensionality lessening.[11] Scikit-learn gives a scope of directed and unsupervised learning calculations through a reliable interface in Python. It is authorized under a lenient disentangled BSD permit and is circulated under numerous Linux appropriations, empowering scholastic AND BUSINESS UTILIZE [12]

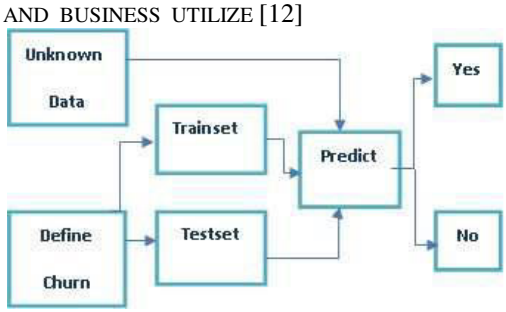


FIG. 2. PREDICTION METHODOLOGY

IV PREPROCESSING TECHNIQUES

In straightforward words, pre-preparing et.al [9] alludes to the changes connected to the information before nourishing it to the calculation. In python, scikit-learn library has a pre-assembled usefulness under sklearn. pre-processing. The information we get from client is as crude information, so it needs to get perfect, change and decrease to make it proper for applying strategies on it, this procedure is known as preprocessing. Require scientific sandbox in which you can perform examination for the whole term of the task. You have to investigate, preprocess and condition information preceding demonstrating. Further, you will perform ETLT (remove, change, stack and change) to get information into the sandbox. It enhances the general nature of the information and effectiveness of the model to deliver comes about. There are numerous more alternatives for pre-preparing as –

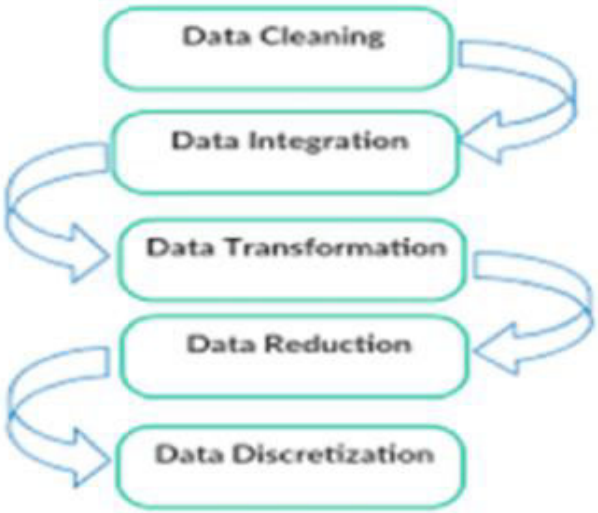


Fig. 3. Preprocessing Techniques

V. METHODOLOGY USED FOR PREDICTION

Utilizing this expectation demonstrate, which intends to foresee whether a representative will proceed or leave the association based upon the investigation of the information of past workers. The expectation factors incorporate fulfillment level, last assessment, normal month to month

hours, compensation, work mischance, advancement, time spent at the organization and division, in view of these parameters, diverse machine learning models like calculated relapse, choice tree order and so forth are connected to foresee which worker will leave straightaway and the variables that are most huge in this choice.

A. Linear Regression:

Coordinate backslide is the path toward finding the association between two ward factors using a straight condition. It is the most principal kind of making figures using backslide that is known as coordinated learning, in it a planning dataset is used to set up the machine with the objective that when we ask for to impact desires it to will can make comes to fruition using the association between the components. It can be used for most prominent two elements for various variable conjectures polynomial backslide is used. It produces data as some motivating force after associated distinctive preprocessing methods. It is the most broadly perceived system used for fitting a backslide line. It figures the best-fit line for the watched data by constraining the aggregate of the squares of the vertical deviations from each datum point to the line. Since the deviations are first squared, when included, there is no counterbalancing among positive and negative regards.[13]

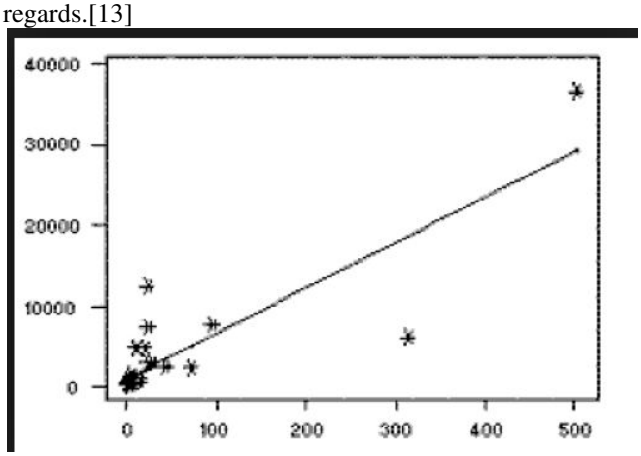


Fig: linear regression

b. Logistic Regression:

Backslide is the route toward making desire the association state of two ward factors. the minimum complex kind of the backslide condition with one dependent and one free factor is portrayed by the condition et.al [10]

$y = m + c*x$

where y = assessed subordinate variable score,

m = enduring, c =regression coefficient, and x = score on the self-sufficient variable.[14]

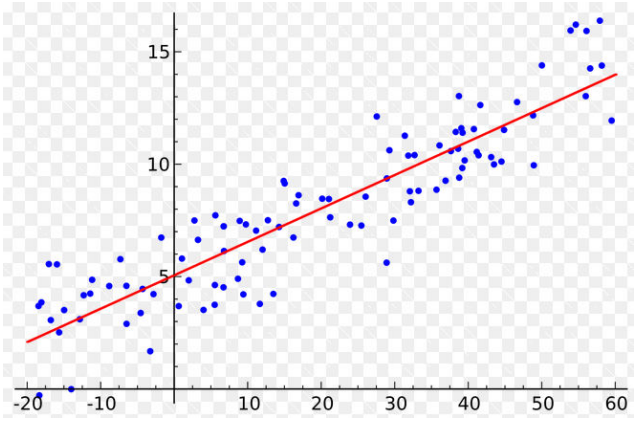
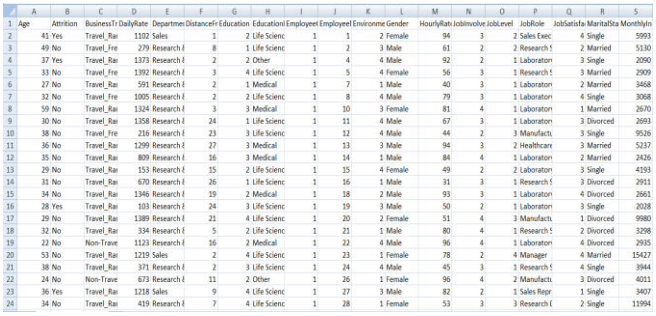


Fig Logistic Regression

VI. RESULT AND DISCUSSION

This report expects to foresee whether a worker will proceed or leave the association in view of the examination of the information of past representatives.

The expectation factors incorporate fulfillment level, last assessment, normal month to month hours, pay, work mischance, advancement, time spent at the organization and division, in view of these parameters, distinctive machine learning models like strategic relapse, choice tree characterization and so on are connected to anticipate which worker will leave straightaway and the components that are most critical in this choice. Through this paper an organization can choose its strategies to keep great representatives from leaving the organization. Information science part that utilized as a part of this report is to take crude information from csv document and then apply diverse handling component to settle on information helpful in settling on choices from it like arrangement of dataset, Label Encoding, Onehot Encoding and include scaling. It at that point applies diverse relapse models to anticipate whether the worker will leave the organization or not as 0 and 1. If 0 comes in the outcome that implies that the worker will proceed with the organization, however if 1 comes then the representative will leave the organization. Here is given the example information that we utilized for making expectations, it is in an unthinkable frame which contains segments as fulfillment level, last assessment, number of undertakings, normal month to month hours, years spent in the organization, work mischance, advancement, office and pay

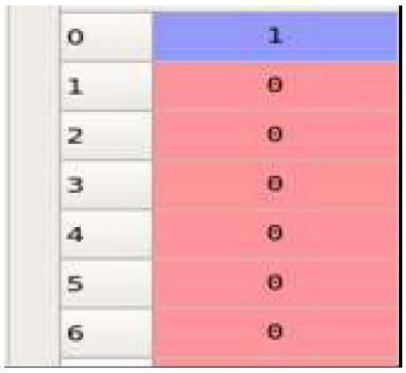


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Age	Attrition	BusinessTr	DailyRate	Departme	DistanceF	Education	Educationl	EmployeeI	EmployeeI	Environme	Gender	HourlyRat	JobInvolve	JobLevel	JobRole	JobSatisfa	MaritalSta	MonthlyIn
2	41	Yes	Travel_Rai	1102	Sales	1	2	Life Scienc	1	1	2	Female	94	3	2	Sales Exec	4	Single	5993
3	49	No	Travel_Fre	279	Research &	8	1	Life Scienc	1	2	3	Male	61	2	2	Research S	2	Married	5130
4	37	Yes	Travel_Rai	1373	Research &	2	2	Other	1	4	4	Male	92	2	1	Laboratori	3	Single	2090
5	33	No	Travel_Fre	1392	Research &	3	4	Life Scienc	1	5	4	Female	56	3	1	Research S	3	Married	2909
6	27	No	Travel_Rai	591	Research &	2	1	Medical	1	7	1	Male	40	3	1	Laboratori	2	Married	3468
7	32	No	Travel_Fre	1005	Research &	2	2	Life Scienc	1	8	4	Male	79	3	1	Laboratori	4	Single	3068
8	59	No	Travel_Rai	1324	Research &	3	3	Medical	1	10	3	Female	81	4	1	Laboratori	1	Married	2670
9	30	No	Travel_Rai	1358	Research &	24	1	Life Scienc	1	11	4	Male	67	3	1	Laboratori	3	Divorced	2693
10	38	No	Travel_Fre	216	Research &	23	3	Life Scienc	1	12	4	Male	44	2	3	Manufactu	3	Single	9526
11	36	No	Travel_Rai	1299	Research &	27	3	Medical	1	13	3	Male	94	3	2	Healthcare	3	Married	5237
12	35	No	Travel_Rai	809	Research &	16	3	Medical	1	14	1	Male	84	4	1	Laboratori	2	Married	2426
13	29	No	Travel_Rai	153	Research &	15	2	Life Scienc	1	15	4	Female	49	2	2	Laboratori	3	Single	4193
14	31	No	Travel_Rai	670	Research &	26	1	Life Scienc	1	16	1	Male	31	3	1	Research S	3	Divorced	2911
15	34	No	Travel_Rai	1346	Research &	19	2	Medical	1	18	2	Male	93	3	1	Laboratori	4	Divorced	2661
16	28	Yes	Travel_Rai	103	Research &	24	3	Life Scienc	1	19	3	Male	50	2	1	Laboratori	3	Single	2028
17	29	No	Travel_Rai	1389	Research &	21	4	Life Scienc	1	20	2	Female	51	4	3	Manufactu	1	Divorced	9980
18	32	No	Travel_Rai	334	Research &	5	2	Life Scienc	1	21	1	Male	80	4	1	Research S	2	Divorced	3298
19	22	No	Non-Trave	1123	Research &	16	2	Medical	1	22	4	Male	96	4	1	Laboratori	4	Divorced	2935
20	53	No	Travel_Rai	1219	Sales	2	4	Life Scienc	1	23	1	Female	78	2	4	Manager	4	Married	15427
21	38	No	Travel_Rai	371	Research &	2	3	Life Scienc	1	24	4	Male	45	3	1	Research S	4	Single	3944
22	24	No	Non-Trave	673	Research &	11	2	Other	1	26	1	Female	96	4	2	Manufactu	3	Divorced	4011
23	36	Yes	Travel_Rai	1218	Sales	9	4	Life Scienc	1	27	3	Male	82	2	1	Sales Repr	1	Single	3407
24	34	No	Travel_Rai	419	Research &	7	4	Life Scienc	1	28	1	Female	53	3	3	Research &	2	Single	11994

fig Dataset for prediction

When the accuracy of the result is being calculated from the previous analysed data with the help of confusion matrix and the accuracy score, this result is being

compared with the available data to find the result accuracy and 97% of the predictions are made correct.



0	1
1	0
2	0
3	0
4	0
5	0
6	0

Fig Result

The figure contains the result in the form of 0 or 1 as o representing the employee who will not leave the company and 1 representing as employee who will going to leave the company.

VII. DATA VISUALIZATION

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.[15]

COUNT PLOT

A count plot can be thought of as a histogram across a categorical, instead of quantitative, variable. The basic API and options are identical to those for barplot(), so you can compare counts across nested variables[16]

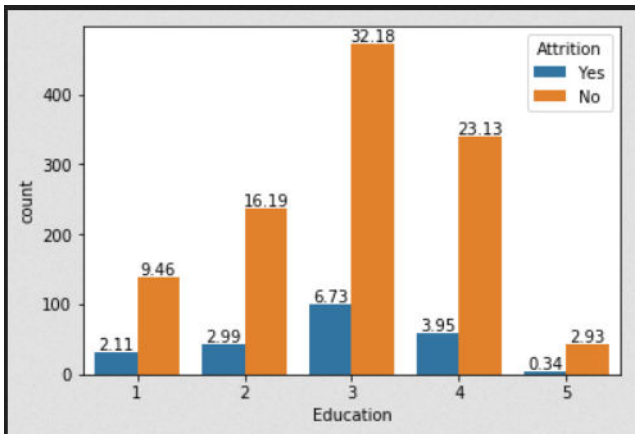
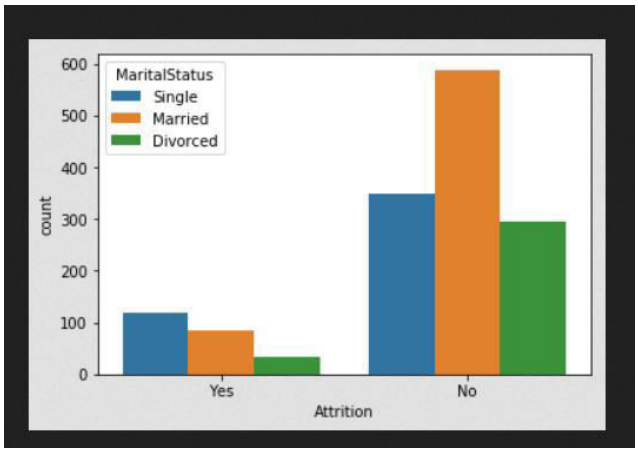


Fig: count plot

Above graph showcases the percentages of the employees across Education levels and corresponding attrition

Observations from graphical representations:

Attrition is higher in employees who are single



VIII CONCLUSIONS

In this investigation, we become more acquainted with that maintenance of a representative inside an association can be discover utilizing strategic relapse procedure, which delivers an outcome with 97% exactness. It can likewise help in discovering the components that are influencing the representatives in the association like pay level, work stack, advancements and so forth.

The future extent of information science is brilliant; consequently, this procedure can be utilized as a part of any association for better worker administration and for their fulfillment. This paper can be additionally reached out as it requires information as .csv records just, so this impediment can be expelled

REFERENCES

[1] Piotr Płoński (MLJAR), "Human-first Machine Learning Platform," Human Resource Analytics Predict Employee Attrition.
[2] Le Zhang and Graham Williams (Data Scientist, Microsoft), "Employee Retention with R based Data Science Accelarator".
[3] Ashish Mishra (Data Scientist, Experfy), "Using Machine Learning to Predict and explain Employee Attrition".
[4] Rupesh Khare, Dimple Kaloya and Gauri Gupta, "Employee Attrition Risk Assessment using Logistic Regression Analysis," from 2nd IIMA International Conference on Advanced Data Analysis, Business Analytics and Intelligence.
[5] Randy Lao, "Predicting Employee Kernelover," Kaggle.
[6] Sandra W. Pyke & Peter M. Sheridan, "Logistic Regression Analysis of Graduate Student Retention," from The Canadian Journal of Higher Education, Vol. XXIII-2, 1993.
[7] Prof. Dr. Vjollca Hasani and Prof. Dr. Alba Dumi, "Application of Logistic Regression in the Study

of Students' Performance Level," Journal of Educational and Social Research Italy.
[8]Web Source scholar.in
[9] Referred paper from Milton
[10] Referred paper from ABES
[11] Referred paper from Northumbria
[12] Referred paper from jaypee university
[13] Referred paper from Unitec
[14] Referred paper from HUS&T
[15] Web Source www.tableu .com
[16]Web Source standford.edu
[17] Referred paper from C Q University