**PREDICTING EMPLOYEE ATTRITION BY USING RANDOM FOREST**.

# INTRODUCTION

Employee attrition is “the movement of people into and out of employment within an organization”. A large number of departures in a relatively concentrated period of time will create difficulties for management . If the pace of recruitment and training cannot be maintained, important positions may remain vacant, resulting in daily work with that is not carried out, and a lack of advancement of the company’s development goals and plans. Thus, if we can predict which employees are likely to leave in the future, we can plan ahead, take steps to reduce the likelihood of departure. And the train and recruit new employees in advance, particularly, avoiding the loss of key positions and the leaking of core information. In this way, employee attrition is forecasting can play an important role in the steady development .

# OVERVIEW:-

Today, almost all companies are concerned about retaining their employees. However, they are not able to recognize the real factors that make them quit their jobs. Many factors could be responsible for that. Each company has its way Abstract. Today, almost all companies are concerned about retaining their employees. However ,they are not able to recognise the real factors that make them quit their jobs. Many factors could be responsible for that (for example: cultural, financial, etc.). Each company has its way to treat its employees and assure their happiness. But often no measures are taken of the satisfaction rate. As a result, in many cases, employees quit their employment suddenly without an apparent reason. In the last decades, Machine learning (ML) techniques have gained popularity among researchers. It can propose solutions to a wide range of problems. Then, ML learning has the potential to make predictions to anticipate employee attrition. In this paper, the authors compare state-of-the-art solutions for the proposed machine learning algorithms using a real data set sample size of 1469. The results could be used to warn managers in order to change their strategies or behaviour. It could also be used to make recommendations to the managers to add some policies in order to retain their employees in the company. This study aims to present a comparison of different

machine learning methods to give a prediction of employees who are likely to leave their

company. The data set includes information about the current employees and the employees who

had already quit their job with almost 50 valuable information units. This last combines many

factors: social, cultural, financial, professional, and relational factors. Six different ML algorithms

were used in this paper. Experimental results show that the Random Forest algorithm

demonstrated the best capabilities to predict the employees’ attrition. The best prediction accuracy

was 85.12, that is considered as good accuracy.

# PURPOSE:-

In this project, we make use of pandas, numpy, matplotlib and seaborn libraries. Employee attrition prediction as apart of human resource analytics could has been a research topic for a while. Research in employee attrition uses some of the classification models. Depending upon the data attributes and size. Layoffs, which occur due to a change or decrease of the business reduce employee morale and make it difficult to hire for other positions. When employee leave by choice, The company can decreases the costs while moving forward with other hiring decisions. Attrition is the important to understand because it can be decrease without incorporating staff departures. As the employee retire the company can perform hiring freeze. When the employee start to retire, the company doesn’t replace them.

# 2. LITREATURE SURVEY

# 2.1 EXISTING PROBLEM

## EXISTING APPROACHES OR METHOD TO SOLVE THIS PROBLEM

Research on employee prediction has been conducted for several decades. Many useful models have been proposed in both theory and practice. In recent years, some statistical analysis techniques have been used to predict or analyze employee intention. The two-phase cluster analysis method to predict employee predicting intention. The study added a self-organized mapping graph into the cluster analysis to find clusters of employee attrition intentional features. Normally we use a decision tree to predict employee intention.

# 2.2 PROPOSED SOLUTION:-

## RANDOM FOREST ALGORITHM

Employee Attrition Estimation Using Random Forest Algorithm

minimization of the penalized criterion defined for any sub-tree T of 𝑇��𝑥, noted

T ≤ 𝑇��𝑥, and for all α ϵ R

+by

Ʀα(T) = Ʀ(T, 𝐷�)+ α |T̃|, (4)

where T̃ denotes the number of leaves of the tree T and Ʀ(T, 𝐷�) corresponds to the empirical

error of the model T estimated from the data of the sample 𝐷�

The proposed algorithm has several parameters:

- The number of trees in the forest. Its default value is 500. Note that this

parameter is not really a parameter to calibrate in the sense that a larger value of

this parameter will always lead to more stable predictions than a smaller value

of this parameter.

- The number mtry of variables chosen for the division of each node. Its default

value is mtry = pd in classification. This is arguably the most important

parameter to calibrate as it can greatly influence the performance of the forest.

- The minimum number of nodesize observations below which a node is no

longer split. The default value for this parameter is nodesize = 1 for

classification. Usually, this setting is left at its default value.

- The number of observations per year in each bootstrap sample. By default, each

bootstrap sample contains an = n observations drawn with replacement in the

initial sample 𝐷𝑛.

-

Several authors have been interested in the choice and influence of these parameters

(Biau and Scornet, 2016; Genuer and Poggi, 2017). In general, the default values of the

parameters work well. Indeed, there are few theoretical results available for the Breiman

random forests. We can nevertheless cite a major result recently established, focusing on

the convergence of random forests in the additive model (Scornet et al., 2015).84) This is an approach based on random forest algorithm that consists of aggregating a

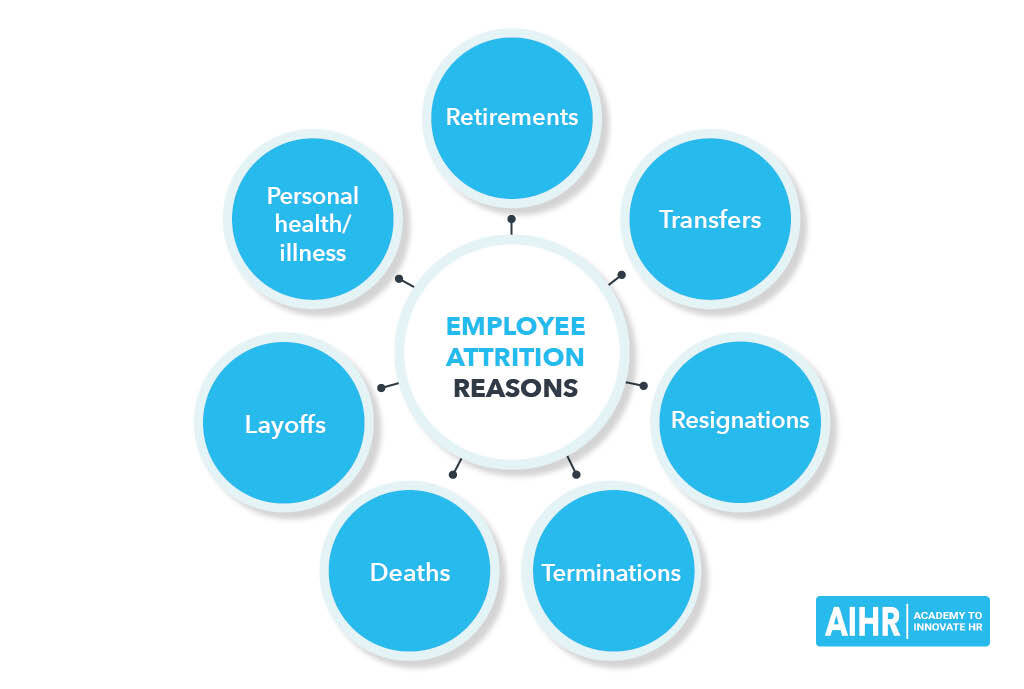
collection of estimators constructed from bootstrap samples (Fig. 8). A random forest is

an aggregation of random trees.

# 3.THEORITICAL ANALYSIS

The analysis has been carried to find out the most critical factor that effects an employee’s attrition.

The opposite attrition is retention. An employee’s attrition is negatively affect companies growth , every year companies come with different retention schemes for employee, so employee don’t leave the company. Those methods include retention bonuses, better training, promotion and stock option which mature after some definite period. So finding the essential factor for attrition will help organization to plan their retention schemes more targeted



Globally competitive organizations will depend on the uniqueness of their human resources and the systems for managing effectively to become successful. Outsourcing is a management process in the business context that has been well understood, tired and tested by the world. The issue of employee attrition has been found to be making huge economic impact on the business process outsourcing.

Employee attrition in BPO using factor analysis. Also multiple regression analysis was applied to analyze the collected data. It is found that the attrition factor employee’s salary has emerged as the critical factor affecting high attrition.

# BLOCK DIAGRAM

# 

# https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcSoW6uiQlktnnzPenEAD008pX3NTAc7-88Z1kRItCSiNPIAZ0dwDkKmpd06Ulg&s https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcTSqFBVJaKc2BTmUlwFHUGgJYuj9PKJVZTVHhnym81OdCAPbWBi2hG_BwIk2X4&s

# HARDWARE AND SOFTWARE DESIGNING REQUIREMENTS OF THE PROJECT

* Jupyter Notebook Environment
* Spyder Ide
* Random forest algorithms
* Python,(Pandas, numpy, matplotib, seaborn,sklearn)
* HTML
* Flask

“ These imported libraries that will be used throughout the program” in the project.

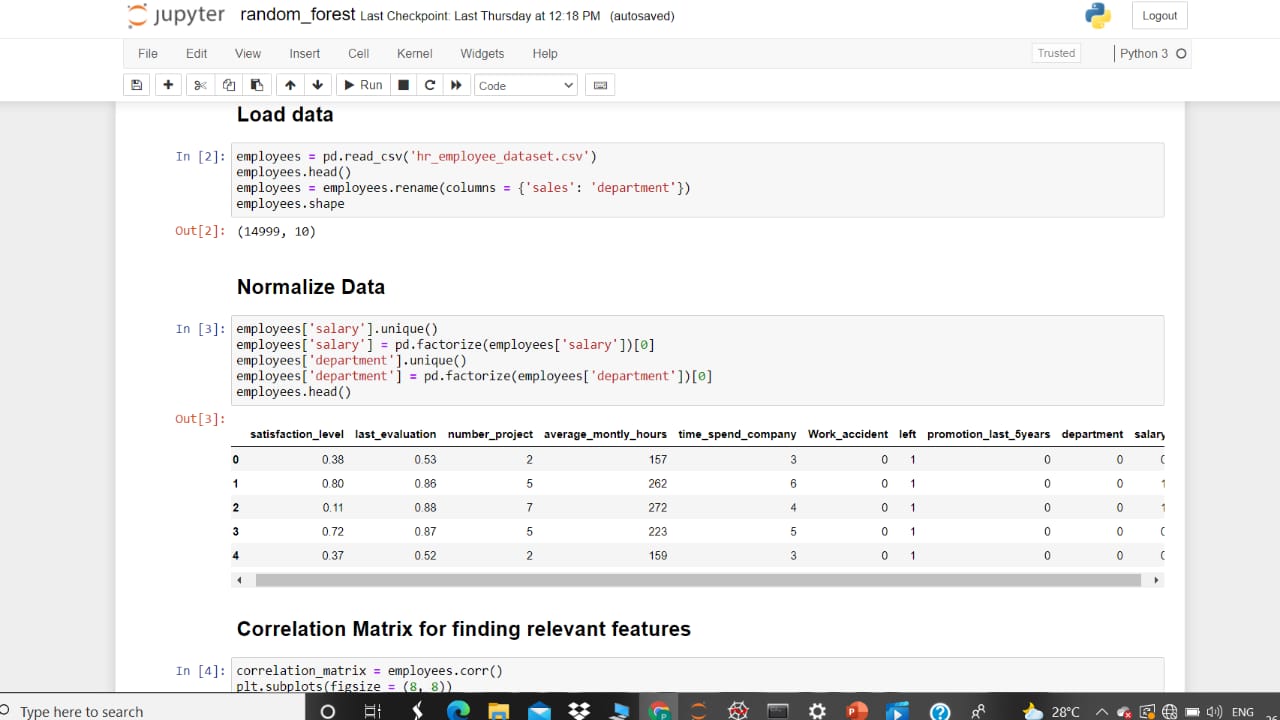
Requirements nof the project to develop we analyzed employee attrition using hardware and software designing methodologies.

* Data Collection
* Data pre processing
* Dividing the data into two parts “Training” and “Testing”.
* Build up the model using “TRANING DATA SET”.
* Accuracy Test using “TESTING DATA SET”.

# 4.EXPIREMENTAL INVESTIGATION

# ANALYSIS OR THE INVESTIGATION WHILE WORKING ON THE SOLUTION

We can predict the employee attrition by using random forest and by using random forest algorithms. And some of the few experimental investigations have been done on this work on the solution , predicting employee attrition is also done by using tree-based models, Decision tree, Logistic regression, python, and Machine Learning etc.



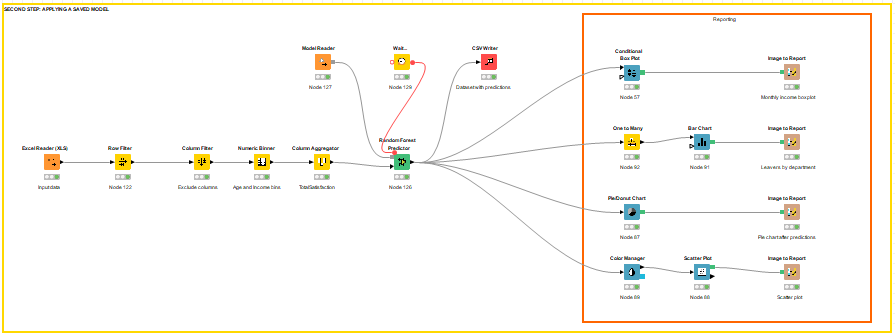
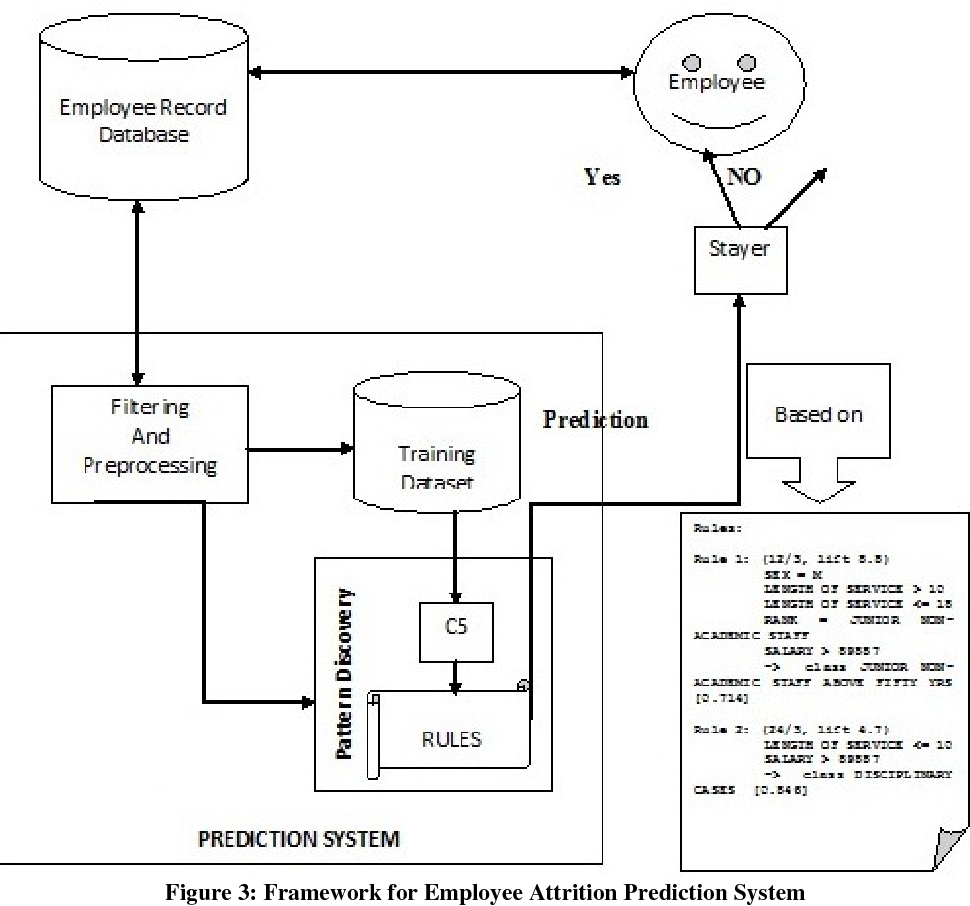
Here, the Load data statements read Rows from a text file into a table at a very high speed. The file can be read from the sever host or the client host depending on whether the LOCAL Modifier is given LOCAL also effects Data interpretation and Error Handling.

And coming to Normalized Data , It is the process of structuring Databases. Usually, in accordance with a series in order to reduce data redundancy and improves Data Integrity. Data dependencies are logical, all are related data items all stored together. There is no redundancy of data, all Data is stored in one place. Here, the employee data is Normalized data stored the employee’s information like salary, Department and the syntax of Normalized data finds the correlation matrix for relevant features.

# 

# 5. FLOWCHART

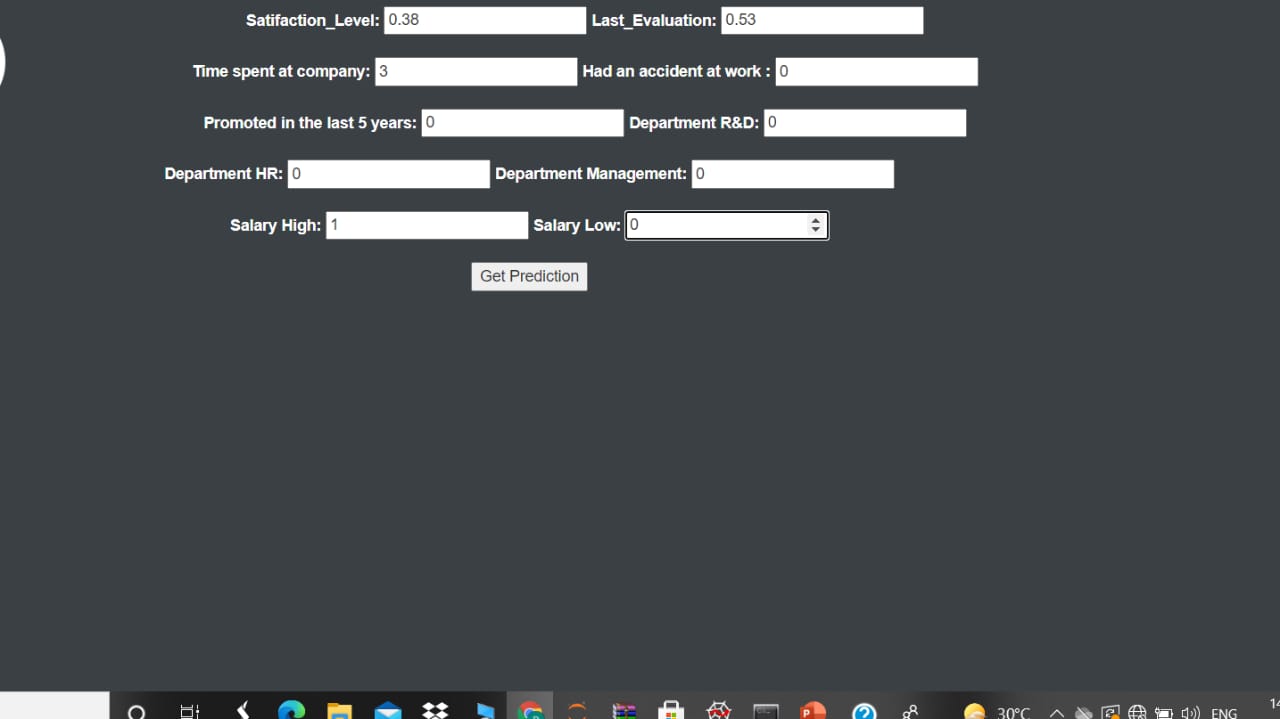
# DIAGRAM SHOWING THE CONTROL FLOW OF THE SOLUTION

Here, the control flow of the solution describes through the data set. Data set is a collection of data. The data of the solution describes the Employee Attrition, Job Level, Job Role, Total Working Years and Accuracy rate. 

Efficiency of the control floe of the solution depends on the employee’s attrition. Any company is primarily defined its employees, It takes the attrition rate or the employees intention of a company is Leaning towards a higher value. The nominal attrition of the company has been the major Reasons for Employee attrition.

# 6.RESULT

## FINAL OUTPUT OF THE PROJECT



# What is flask?

Flask is a Web Framework. This means flask provides you with tools, libraries and technologies that allow you to build a web application. This web application can be some web pages, a blog, a wiki or go as big as a web-based calendar application or a commercial website.

Flask is a part of the categories of the micro-framework. Micro-framework are normally framework with little to no dependencies to external libraries. This has pros and cons. Pros would be that the framework is light, there are little dependency to update and watch for security bugs, cons is that some time you will have to do more work by yourself or increase yourself the list of dependency by adding plugins. In this case of Flask, its dependencies are:

* Werkzeug a WSGI Library
* Jinja 2 which is its template engine

# RANDOM FOREST

We used Random forest for predicting employee attrition. Random forest is a classification algorithm. The Random forest model is far more advanced when compared and one would expect better results. The RF models were analyzed by varying the tuning parameter, the optimal model was obtained AUC was 0.8405 with an optimal threshold of 0.175. This is the major technique used in modeling predictions and behavior analysis.

# 7.ADVANTAGES AND DISADVANTAGES OF THE PROPOSED SOLUTION

Proposed solution means the combination of Hardware, Software , other products or equipment and they all services necessary to implement the solution described in its proposal.

Advantages:-

* No entry cost to your organization.
* No upfront costs such as conversion fee or setup fees.
* Increase in recruiting staff, recruiters will work with organizations to assist with any recruiting needs.
* Fast implementation which can be staged in groups, individuals or a combination if needed.
* Increases overall productivity and better experience.
* Decreases employee liability such as compensation claims, unemployment insurance claims etc.
* We can assess the value delivered by each proposed solution.

Disadvantages:-

* Information overloaded, companies have access to more information than they accustomed to managing.
* Organizational misunderstanding and poor co-ordination of information with decision-making needs.
* Longer time to complete than a bankrupty.
* Unable to detect with specific sources.
* Detection is not guarantee and required data may not available.

# 8.APPLICATIONS

# Where the solution can be applied:

Application of the proposed solution for the absolute to the pose problem. The number of trees in the forest. Its default value is 500. Note that this

parameter is not really a parameter to calibrate in the sense that a larger value of

this parameter will always lead to more stable predictions than a smaller value

of this parameter. According to the proposed solution, the most important features are Monthly Income,

Age, Daily Rate, Total Working Years and Monthly Rate. The proposed model repeatedly is proven the best fit for calculating employee

Attrition. The higher number of respondents have increased the

accuracy of the algorithm, and the broader range of featured included have added more

insight into the problem of attrition. With using five algorithms and more substantial

data set of employee attrition in proposed solution.

# 9.CONCLUSION

The word is changing, and so is our work setting. The development of technologies gives

us an opportunity to improve our data analysis and aim for more precise predictions.

Motivating and satisfying employees in the new work environment is challenging, and

so is keeping attrition rate low. Having loyal and long-lasting employees are essential to

most of the organizations. In this paper, the authors have presented the experiment of

finding the best-fit algorithm for employee behaviour measures within the given context.

The study is based on the author's previous research and is improved with a higher

amount of data. It supplements the author's previous research and conclusion that the

proposed model is the most appropriate algorithm for data set describing employee

satisfaction and attrition with an accuracy of 85,12%.

After data training and the use of the algorithms, the results were improved, and

several different features were highlighted. The most important features are Monthly

Income, Age, Daily Rate, Total Working Years and Monthly Rate. These results are

different from previous research analysis which didn't suggest age or salary of being

high importance factors, rather motivational factors and management being the highest

importance. These new results are a valuable addition to existing data and will be

researched more in detail in future.

# 10.FUTURE SCOPE

# Enhancements that can be made in the future:-

Predictive employee attrition model helps in not only taking preventing measures but also into making better hiring decisions. Deriving trends in the candidate’s performance out of past data is important in order to predict the future trends, as well as to board new employees. Employee attrition is predictable under stable circumstances, where in a set pattern can be deduced from certain parameters influencing the organizations at all times. Predictive model in order to predict the future trends and provide good understanding of workforce supply and demand. Minimize cost of new talent acquisition based on the employee profiling and company requirements.

# 11.BIBILOGRAPHY

## REFRENCES :-

Alao, D., Adeyemo, A. B. (2013). Analyzing employee attrition using decision tree algorithms.

Computing, Information Systems, Development Informatics and Allied Research Journal,

4(1), 17–28.

Biau, G., Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197–227.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

## Breiman, L., Friedman, J., Stone, C. J., Olshen, R. A. (1984). Classification and regression trees. [Scholarly articles for References for predicting employee attrition using random forest](https://www.google.com/url?q=http://scholar.google.co.in/scholar%3Fq%3DReferences%2Bfor%2Bpredicting%2Bemployee%2Battrition%2Busing%2Brandom%2Bforest%26hl%3Den%26as_sdt%3D0%26as_vis%3D1%26oi%3Dscholart&sa=U&ved=2ahUKEwjisKvxtYvyAhWvzDgGHc4VB80QgQN6BAgLEAE&usg=AOvVaw3BxGZVHgXkYu-9FIvftWkW)

scholar.google.com › citations

### [l/Predicting-Employee-Attrition: Using machine ... - GitHub](https://www.google.com/url?q=https://github.com/nitishghosal/Predicting-Employee-Attrition&sa=U&ved=2ahUKEwjCwbCztovyAhUEzTgGHZsaDc4QFjAAegQIAhAB&usg=AOvVaw0t9G6LDuDFSYX6GBAyRIVi)

[github.com › nitishghosal › Predicting-Employee-Attrition](https://www.google.com/url?q=https://github.com/nitishghosal/Predicting-Employee-Attrition&sa=U&ved=2ahUKEwjCwbCztovyAhUEzTgGHZsaDc4QFjAAegQIAhAB&usg=AOvVaw0t9G6LDuDFSYX6GBAyRIVi)

# APPENDIX

# A.SOURCE CODE:

# AAP.PY

from flask import Flask, render\_template, request, redirect, url\_for,session

import pandas as pd

import numpy as np

import pickle

from werkzeug.utils import secure\_filename

import os

from flask import jsonify

def prep\_data(df):

'''

:Assumption: Assuming the dataframe contains the required columns

: required columns : ['satisfaction\_level','last\_evaluation','time\_spend\_company','Work\_accident','promotion\_last\_5years','department\_RandD','department\_hr','department\_management',salary\_high', 'salary\_low']

:input: pandas dataframe

:output: pre-processed dataframe with selected columns

'''

num\_df = df[['satisfaction\_level','last\_evaluation','time\_spend\_company','Work\_accident','promotion\_last\_5years','department\_RandD','department\_hr','department\_management','salary\_high', 'salary\_low']]

new\_df = pd.concat([num\_df,cat\_df], axis=1)

return new\_df

def allowed\_file(filename):

return '.' in filename and filename.rsplit('.', 1)[1].lower() in ALLOWED\_EXTENSIONS

UPLOAD\_FOLDER = './templates'

ALLOWED\_EXTENSIONS = set(['csv'])

## Initialize the app

app = Flask(\_name\_)

app.config['UPLOAD\_FOLDER'] = UPLOAD\_FOLDER

@app.route('/home')

def analysis\_page():

# render a static template

return render\_template('home.html')

@app.route('/')

def index():

# redirect to home

return redirect(url\_for('analysis\_page'))

@app.route('/prediction', methods=['GET','POST'])

def prediction\_page():

if request.method == 'POST':

#check if post request has the file type

if 'file' not in request.files:

return render\_template('home.html', error='No File part',retJson ='No file part')

file = request.files['file']

# if user the did not select file

if file.filename == '':

return render\_template ('home.html',error='No file Selected', retJson='No File Selected')

#check for allowed extension

if file and allowed\_file(file.filename):

filename = secure\_filename(file.filename)

file.save(os.path.join(app.config['UPLOAD\_FOLDER'], filename))

# load the model from disk

loaded\_model = pickle.load(open('rf.pkl', 'rb'))

# read csv

data = pd.read\_csv(filename)

prediction = loaded\_model.predict\_proba(prep\_data(data))

# get percentage proba

retJson = []

count = 0

for prob in prediction:

count+=1

retJson.append("The probability of Employee Attrition with index {} : {} % ".format(count,prob[0] \* 100))

#retJson =jsonify({'retJson' :retJson})

return render\_template('home.html',error=None, retJson= retJson )

# render a static template

return render\_template('home.html')

@app.route('/attrition', methods=['GET','POST'])

def single\_prediction\_page():

if request.method == 'POST':

SatisfactionLevel = request.form['satifaction\_level']

Last\_Evaluation = request.form['last\_evaluation']

TimeSpentatCompany = request.form['time\_spend\_company']

Work\_Accident = request.form['Work\_accident']

Promotion\_in\_last\_5\_years = request.form['promotion\_last\_5years']

#DepartmentR&D =

request.form['department\_RandD']

#DeprtmentHR = request.form['department\_hr']

#Department\_Management = request.form['department\_management']

SalaryHigh = request.form['salary\_high']

SalaryLow = request.form['salary\_low']

# if len(Age) <= 0 or len(HourlyRate) <= 0 or len(OverTime) <= 0 or len(DailyRate) <= 0 or len(MonthlyIncome)<= 0 or len(TotalWorkingYears) <= 0 or len(YearsAtCompany) <= 0 or len(NumCompaniesWorked) <= 0:

# return render\_template('home.html', retJson= 'All filed is required to make prediction' )

#if OverTime == 'Yes':

# OverTime\_Yes = 1

#else:

# OverTime\_Yes = 0

#create a pandas dataframe last\_evaluation,'time\_spend\_company':time\_spend\_company, 'Work\_accident': Work\_accident,

'promotion\_last\_5years':promotion\_last\_5years, 'department\_RandD': department\_RandD, 'department\_hr':department\_hr,

'department\_management':department\_management, 'salary\_high': salary\_high, 'salary\_low':salary\_low }])

loaded\_model = pickle.load(open('rf.pkl', 'rb'))

#print(df.head())

df = pd.DataFrame([{'satisfaction\_level': 'satisfaction\_level', 'last\_evaluation':

#temp = [ Age, HourlyRate, DailyRate, MonthlyIncome,TotalWorkingYears, YearsAtCompany, NumCompaniesWorked,DistanceFromHome, OverTime\_Yes]

#temp = np.reshape(1,-1)

prediction = loaded\_model.predict\_proba(df)

retJson = []

for prob in prediction:

retJson.append("The probability is : {} % ".format(prob[0] \* 100))

return render\_template('prob.html',error=None, retJson= retJson )

# render a static template

return render\_template('home.html')

if \_name\_ =='\_main\_':

app.run(debug=true)

# BASE.HTML

<!DOCTYPE html>

<html lang="en">

<head>bug=True)

<meta charset="utf-8">

<meta name="viewport" content="width=device-width, initial-scale=1">

<link rel="stylesheet

Href=<https://maxcdn.bootstrapcdn.com/bootstrap/3.4.0/css/bootstrap/jquery/3.3.1/jquery.min.js>></script>

Src=<https://ajax.googleapis.com/ajax/libs/jquery/3.3.1/jquery.min.js>></script><script>

Src=<https://maxcdn.bootstrapcdn.com/bootstrap/3.4.0/js/bootstrap.min.js>></script>

<style>

/\* Remove the navbar's default margin-bottom and rounded borders \*/

.navbar {

margin-bottom: 0;

border-radius: 0;

}

/\* Set height of the grid so .sidenav can be 100% (adjust as needed) \*/

.row.content {height: 450px

}

/\* Set gray background color and 100% height \*/

.sidenav {

padding-top: 20px;

background-color: #3A4043;

height: 100%;

}

/\* Set black background color, white text and some padding \*/

footer {

background-color: #3A4043;

color: white;

padding: 15px;

}

/\* On small screens, set height to 'auto' for sidenav and grid \*/

@media screen and (max-width: 767px) {

.sidenav {

height: auto;

padding: 15px;

}

.row.content {height:auto;}

}

</style>

</head>

<body, style="background-color:#3A4043;text-align:center"><nav class="navbar navbar-inverse">

<div class="container-fluid">

<div class="navbar-header">

<button type="button" class="navbar-toggle" data-toggle="collapse" data-target="#myNavbar">

<span class="icon-bar"></span>

<span class="icon-bar"></span>

<span class="icon-bar"></span>

</button>

</div>

<div class="collapse navbar-collapse" id="myNavbar">

<ul class="nav navbar-nav">

<li class="active"><a>

</ul>

</div>

</div>

</nav>

<main role="main" class="container">

<body style="background-color:#3A4043;"></body>

{% block content %}

{% endblock %}

</main>

</body>

</html>

# HOME.HTML

{% extends "base.html" %}

{% block content %}

</br>

<h3><font color=white>Teamwork Predicting Employee Attrition using Random forest</font></h3>

<form method=post action="/attrition">

<b><font color=white> Satifaction\_Level:</font> </b>

<input type="number" name="satifaction\_level">

<b> <font color=white>Last\_Evaluation:</font> </b>

<input type="number" name="last\_evaluation">

<br><br>

<b><font color=white>Time spent at company:</font> </b>

<input type="number" name="time\_spend\_company">

<b><font color =white>Had an accident at work :</font> </b>

<input type="number" name="Work\_accident">

<br><br>

<b><font color=white>Promoted in the last 5 years:</font> </b>

<input type="number" name="promotion\_last\_5years">

<b><font color=white>Department R&D:</font></b>

<input type="number" name="department\_RandD">

<br><br>

<b><font color=white>Department HR:</font> </b>

<input type="number" name="department\_hr">

<b><font color=white>Department Management:</font> </b>

<input type="number" name="department\_management">

<br><br>

<b><font color=white>Salary High:</font> </b>

<input type="number" name="salary\_high">

<b><font color=white>Salary Low:</font> </b>

<input type="number" name="salary\_low">

<br><br>

<input type=submit value='Get Prediction' >

</form>

<h6> <font color=red> {{ retJson }} </font></h6>

</br>

{% endblock %}