**EMPLOYEE ATTRITION USING RANDOM FOREST**

**1.0 Introduction**

**1.1. Overview**

There are several areas in which organisations can adopt technologies that will support

decision-making: artificial intelligence is one of the most innovative technologies that is widely used to assist organisations in business strategies, organisational aspects and people management. In recent years, attention has increasingly been paid to human resources (HR), since worker quality and skills represent a growth factor and a real competitive advantage for companies. After having been introduced to sales and marketing departments, artificial intelligence is also starting to guide employee-related decisions within HR management.

Data has become a strategic asset for most organisations across multiple sectors, including those linked to business processes. All types of organisations benefit from the adoption of new technologies and collection, management and analysis of data bring numerous benefits in terms of efficiency and competitive advantage. In fact, analysing large amounts of data can lead to improvements in decision-making processes, the achievement of pre-established corporate objectives and better business competitiveness. The application of artificial intelligence in the field of HR allows companies to transform data into knowledge by implementing predictive models: such models allow predictions on employees using data collected by the company over the previous years, thus reducing critical issues and optimising all HR activities.

**1.2 Purpose**

**Attrition** is a problem that impacts all businesses, irrespective of geography, industry and size of the company. It is a major problem to an organization, and predicting turnover is at the forefront of the needs of Human Resources (HR) in many organizations. Organizations face huge costs resulting from employee turnover. With advances in machine learning and data science, it’s possible to predict the employee attrition, and we will predict using Random Forest Classifier algorithm .The purpose is to support decisions that are based not on subjective aspects but on objective data analysis. The goal of this work is to analyse how objective factors influence employee attrition, in order to identify the main causes that contribute to a worker’s decision to leave a company, and to be able to predict whether a particular employee will leave the company.

**2.0 Literature Survey**

Research on employee attrition 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 turnover intention. We applied the two-phase cluster analysis method to predict employee attrtion intention. The study added a self-organized mapping graph into the cluster analysis to find clusters of employee attrition intention features. We used a decision-tree to predict employee attrition intention. The study calculated the relationship among job-satisfaction and work arrangements and employee turnover intention; it created a classification method to identify and group potential attrition employees.

**2.1 Existing Problem**

Companies invest a lot of time and resources in employee recruiting and training, according to

their strategic needs. Therefore, the employees represent a real investment for organisations. When an employee leaves the company, the organisation is not only losing a valuable employee, but also the resources, specifically money and HR staff effort, that were invested recruiting and selecting those employees and training them for their related tasks. Consequently, the organisation must continuously invest in recruiting, training and developing new staff to fill vacant job positions. Training a new employee is a long and costly process and it is of full interest of the company to control and decrease the employee attrition rate: attrition is defined as an employee resigning or retiring from a company. Moreover, satisfied, highly motivated and loyal employees form the core of a company and also have an impact on the productivity of an organisation. In the literature, some authors suggest retaining only happy and motivated employees as they tend to be more creative, productive and perform better, which in the end generates and sustains improved firm performance. As job dissatisfaction is shown in the economic literature as a good predictor of turnover intention job satisfaction data are powerful predictors of both separations and resignations, even controlling for wages, hours and standard demographic and job variables

**2.1 Proposed Solution**

In this project, we perform an analysis of the reasons or motivations that push an employee to leave the company and consequently allow the HR department to take timely appropriate countermeasures such as improving the work environment or production incentives. Starting from the dataset, we identify the main factors related to the employee’s attrition and we propose a real classification, based on the statistical evaluation of the data. The application of classification algorithms can support the HR management by allowing the adoption of staff management support tools in the company.

The proposed solution involves the following phases:

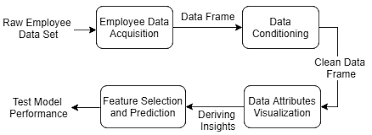
* Collect the employee dataset, which consists of current and past employee observations
* Apply various data cleaning techniques to prepare the dataset
* Start a descriptive analysis of data to detect the key factors and trends that contribute to attrition
* Elaborate the dataset for the training and testing phase and try several classification algorithms to process it
* Based on the results collected with test data, compare many performance metrics of machine learning models and select which model best fits and gives the most accurate results for the given problem
* Then build a flask application with the help of a pickle file integrate the application with your machine learning model .The final UI will predict whether the employee will stay or leave the company.

In general the below figure depicts the phases:



**3.0 Theoretical Analysis**

**3.1 Block Diagram**



**3.2 Hardware / Software Designing**

Hardware Requirements:

The minimum configuration required on computer:

* Processor: 1 gigahertz (GHz) or faster processor.
* RAM: 1 gigabyte (GB) for 32-bit or 2 GB for 64-bit.
* Hard disk space: 16 GB for 32-bit OS 20 GB for 64-bit OS

Software Design:

* Programming Language: Python
* Platform: Jupyter Notebook – version 3.7.1,Spyder
* Front End: HTML,CSS
* Backend:Python

**4.0 Experimental Investigations**

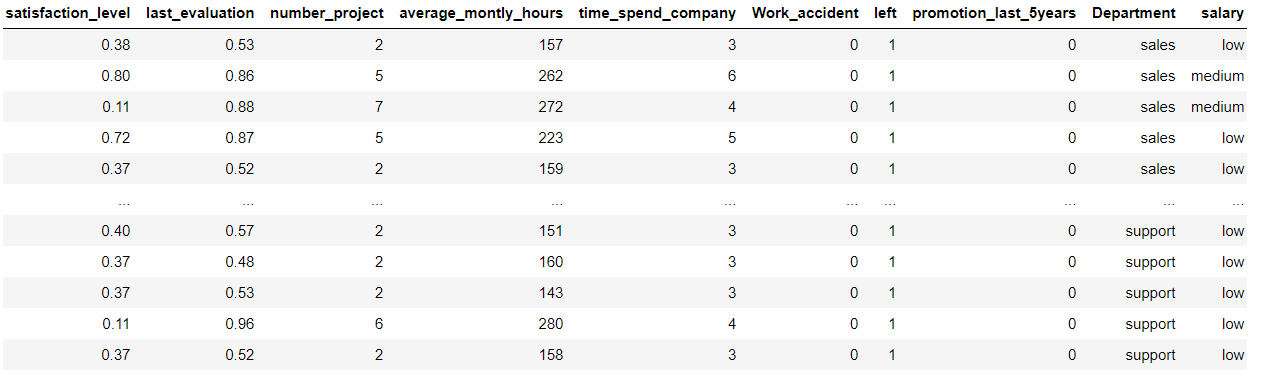
**4.1. Data Overview**

The dataset consists of 25491 obseravtions and 10 variables.  
Each row in dataset represents an employee; each column contains employee attributes:

satisfaction\_level (0–1)  
last\_evaluation (Time since last evaluation in years)  
number\_projects (Number of projects completed while at work)  
average\_monthly\_hours (Average monthly hours at workplace)  
time\_spend\_company (Time spent at the company in years)  
Work\_accident (Whether the employee had a workplace accident)  
left (Whether the employee left the workplace or not (1 or 0))  
promotion\_last\_5years (Whether the employee was promoted in the last five years)  
sales (Department in which they work for)  
salary (Relative level of salary)

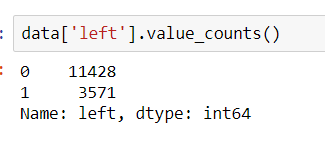
In the experiment, we use Python with NumPy, SciPy, SciKit-learn, Pandas, and Matplotlib and develop three tools: one tool is for feature ranking based on the RF, one is a visual tool for analysis of the feature variables and the target variable, and one is a modeling tool for the RF based on the weighted F-measure.

The below figure depicts the dataset

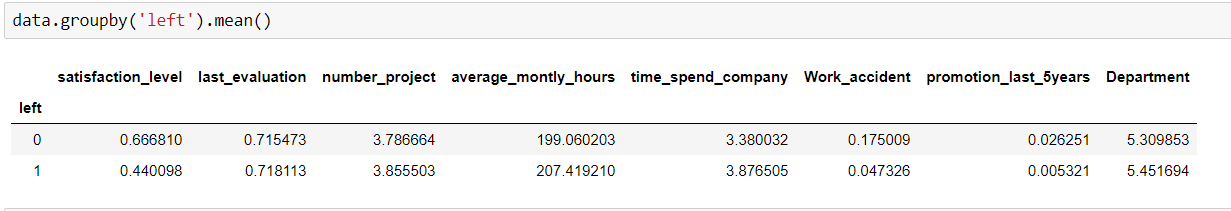


##### 4.2. Data Exploration

##### Let us find out the number of employees who left the company and those who didn’t!



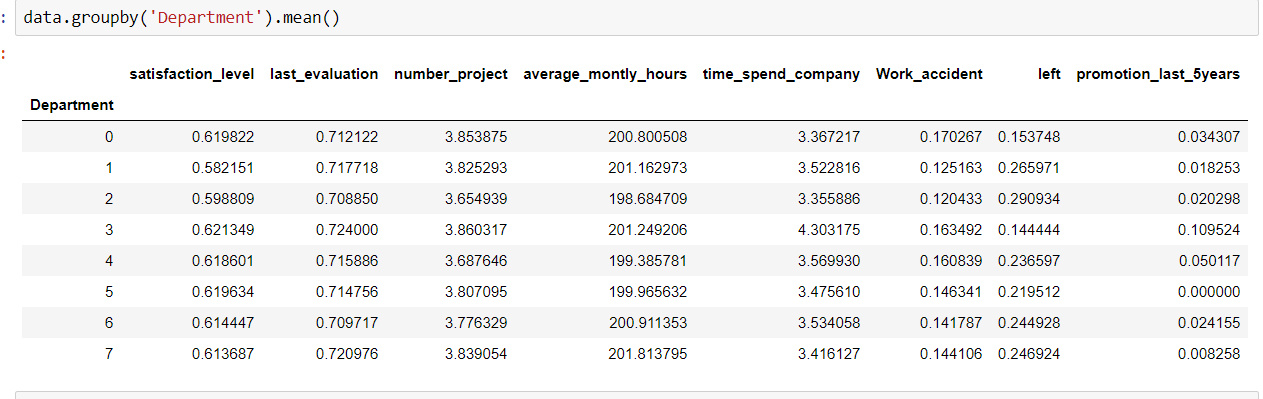
We observe that 5990 employees left the company, which is 23.4 percent of the total employees in the organization.

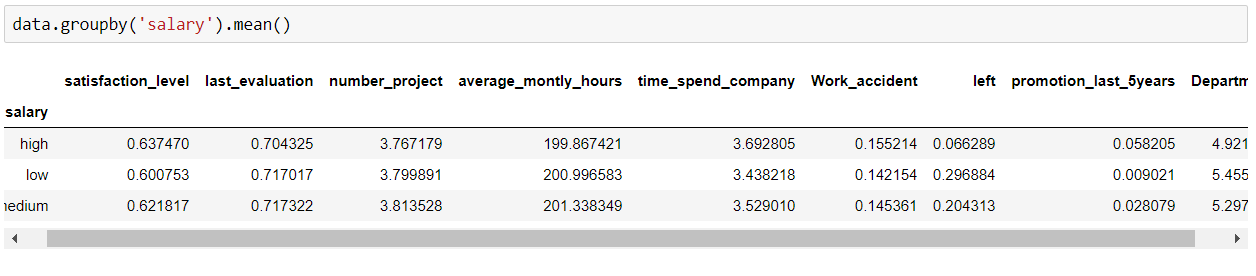


There are Several observations made:

* The average satisfaction level of employees who stayed with the company is higher than that of the employees who left.
* The average monthly work hours of employees who left the company is more than that of the employees who stayed.
* The employees who had workplace accidents are less likely to leave than that of the employee who did not have workplace accidents.
* The employees who were promoted in the last five years are less likely to leave than those who did not get a promotion in the last five years.

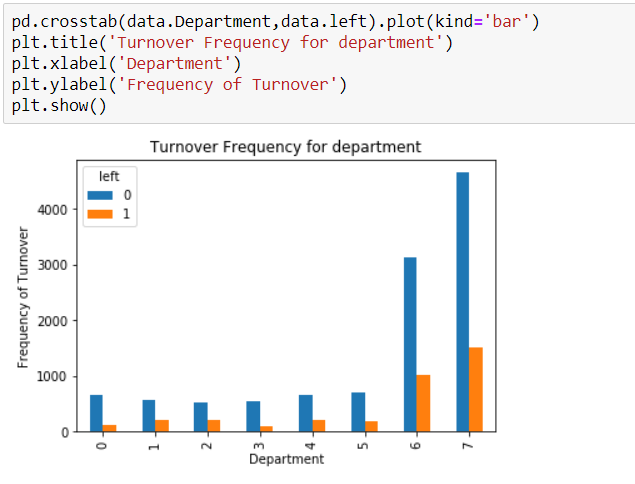
We can calculate categorical means for categorical variables such as department and salary to get a more detailed sense of our data.





**4.3 Data Visualization**

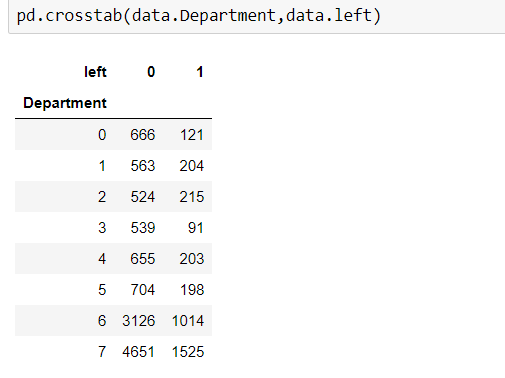
Let us visualize our data to get a much clearer picture of the data and the significant features.



It is evident from the above figure that the number of employees who left are the highest from department 7 which is the RandD Department.

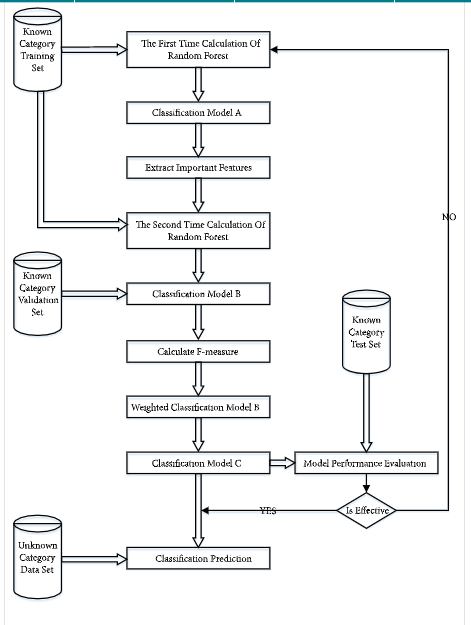


It is evident that the frequency of employee turnover depends a great deal on the department they work for. Thus, the department can be a good predictor of the outcome variable.



The proportion of employee turnover depends a great deal on their salary level; hence, the salary level can be a good predictor in predicting the outcome.

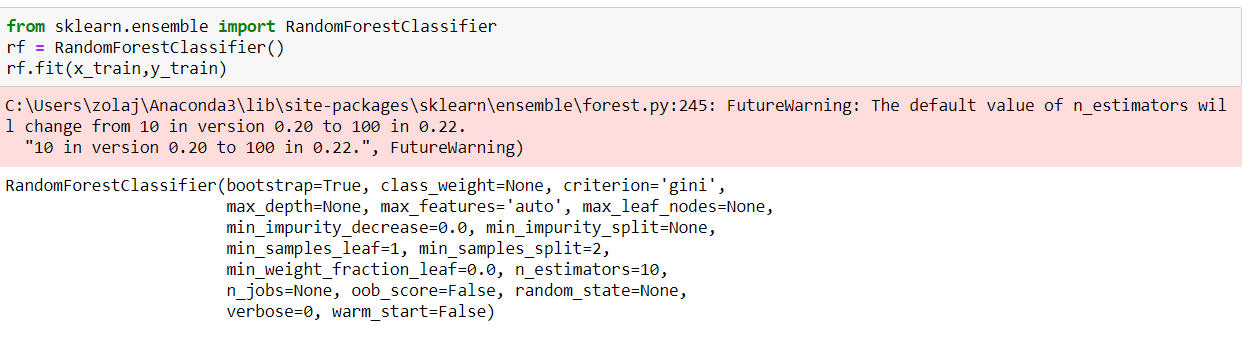
**5.0 Flowchart**



**6.0 Result**

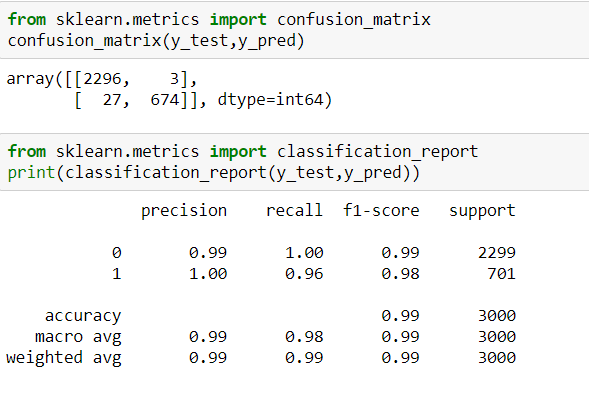
Once after splitting the data into train and test, the data should be fed to an algorithm to build a model.

The machine learning algorithms used for this model here is Random Forest Classifier. The Random Forest Classifier is a set of decision trees from a randomly selected subset of the training set. It aggregates the votes from different decision trees to decide the final class of the test object.



**Precision and Recall:**

We construct a confusion matrix and a classification report to visualize predictions made by a classifier and evaluate the accuracy of classification.



Hence we can observe that we are getting an f1 score of 0.99.

**7.0 Advantages and Disadvantages**

Advantages

This application of machine learning in the field of HR allows companies to transform data into knowledge by implementing predictive models: such models allow predictions on employees using data collected by the company over the previous years, thus reducing critical issues and optimising all HR activities.

Disadvantages

There are various other parameters that need to be considered for employee attrition apart from what we have considered. Like employees health, their behavioural and emotional aspects also need to be addressed.

**8.0 Applications**

There are several applications for this project:

* The application of classification algorithms can support the HR management by allowing the adoption of staff management support tools in the company.
* The results obtained from the data analysis demonstrate that the adoption of machine learning systems can support the HR department in the company staff management.
* It reduces the huge amount of time spent by the HR in assessing the employee.

**9.0 Conclusion**

Human Resource is the main pillar for any organization. The growth level as well as market penetration are duly depends on the strength of the employees. Now a day due to increased availability of technology resources and people with high competency makes great success for any firm. But the prime issues which are normally addressed in any organization are only the attrition. This is a great challenge as well as retention is also the prime task. In this project we have used Random Forest Classifier used by the various researchers for employee prediction strategy. In future research it is possible to improve the analysis by considering new employees’ opportunities as well as adverse working conditions (e.g., harm and hazard) and poor promotion prospects, discrimination and low social support, that are positively related to employees’ turnover intention

**10.0 Future Scope**

In terms of study limitations, adding the feature sorting and weight calculation processes does mean that the modeling has a higher cost in terms of time than other algorithms. In future research, improvement in operational efficiency and overall prediction accuracy could be analyzed. In addition, this algorithm is not suitable for industries with high turnover rate. How to increase the universality of the algorithm still needs to be studied further.

**11.0 Bibliography**

[1] S. Jahan, “Human Resources Information System (HRIS): A Theoretical Perspective”, Journal of Human Resource and Sustainability Studies, Vol.2 No.2, Article ID:46129, 2014.

[2] C. Cortes and V. Vapnik, Support-vector networks. Machine learning, 20(3), 273-297, 1995

[3] RenukaAgrawal, Jyoti Singh, and Zadgoankar .S, “Formative Assessment For Performance Evaluation Of Faculty Using Data Mining”, International Journal Of Advances In Electronics And Computer Science, ISSN: 2393-2835.

[4] HosseinAlizadeh, Buinzahra Branch and Islamic, 2016 ,“Introducing A Hybrid Data Mining Model ToEvaluate Customer Loyalty”, Engineering, Technology & Applied Science Research Volume. 6,No.6,1235-1240.

[5] Amir Mohammad EsmaieeliSikaroudi ,Rouzbehghousi and Ali Esmaieelisikaroudi, 2015 “A Data Mining Approach To Employee Turnover Prediction” (Case Study: Arak Automotive Parts Manufacturing), Journal Of Industrial And Systems Engineering Volume. 8, No. 4.

[6] Anjali A. Dudhe and SachinSakhare .R, January 2018, “Teacher Ranking System To Rank Of Teacher As Per Specific Domain” ,Journal On Soft Computing ICTACT, Volume: 08, Issue: 02, Issn: 2229-6956

**12.0 Appendix**

**12.1 Source Code**

Spyder code:

from flask import Flask,render\_template,request

import pickle

import numpy as np

app=Flask(\_\_name\_\_) #your application

rf=pickle.load(open('demo.pkl','rb'))

@app.route('/') # default route

def home():

return render\_template("proj.html")

@app.route('/predict',methods=['post'])

def predict():

satisfaction\_level=float(request.form['satisfaction\_level'])

last\_evaluation=float(request.form['last\_evaluation'])

number\_project=float(request.form['number\_project'])

average\_montly\_hours=float(request.form['average\_montly\_hours'])

time\_spend\_company=float(request.form['time\_spend\_company'])

Work\_accident=float(request.form['Work\_accident'])

promotion\_last\_5years=float(request.form['promotion\_last\_5years'])

salary=float(request.form['salary'])

Department=float(request.form['Department']) a=np.array([[satisfaction\_level,last\_evaluation,number\_project,average\_montly\_hours,time\_spend\_company,Work\_accident,promotion\_last\_5years,salary,

Department]])

print(a)

result=rf.predict(a)

print(result)

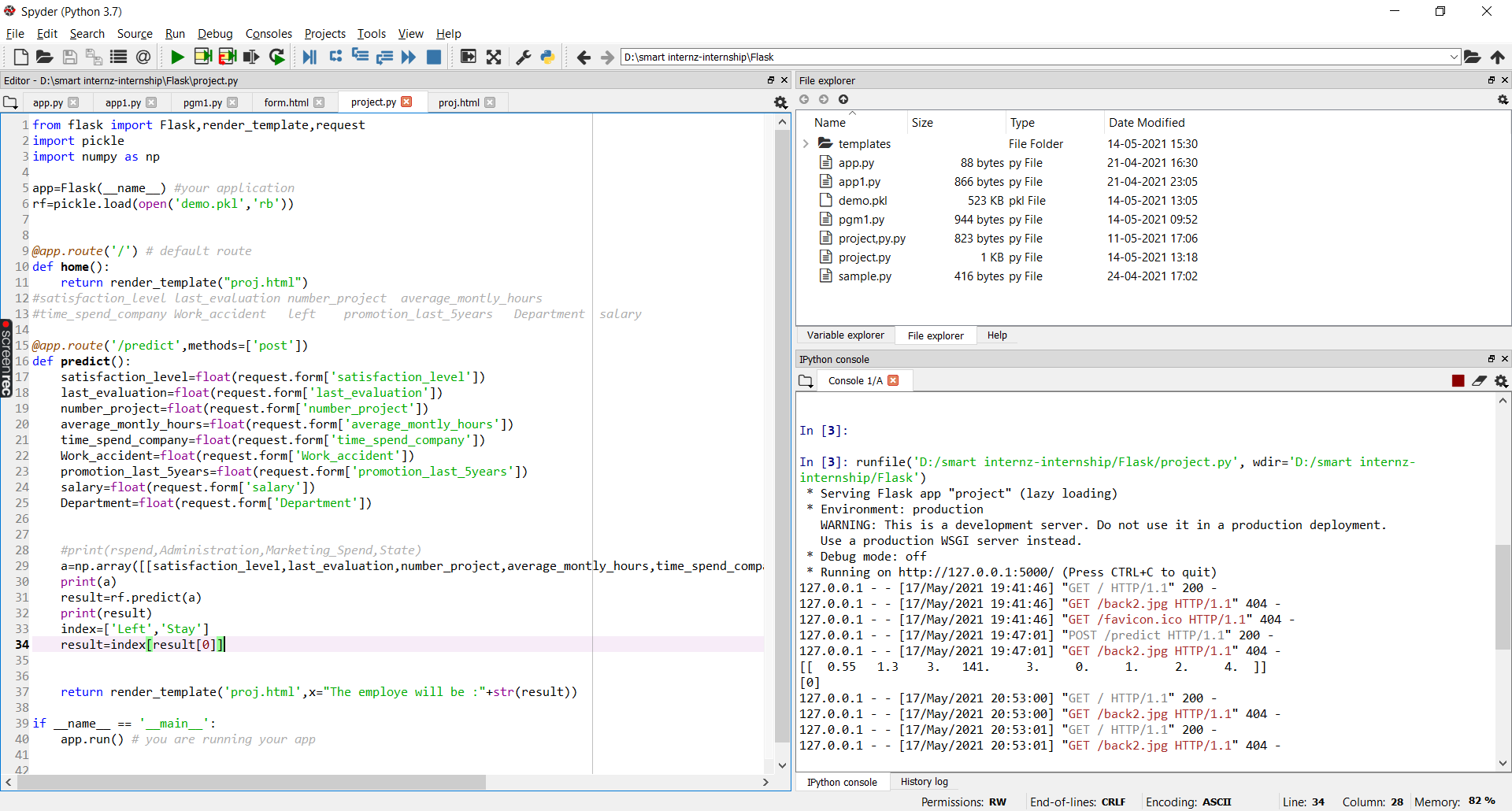
index=['Left','Stay']

result=index[result[0]]

return render\_template('proj.html',x="The employe will be :"+str(result))

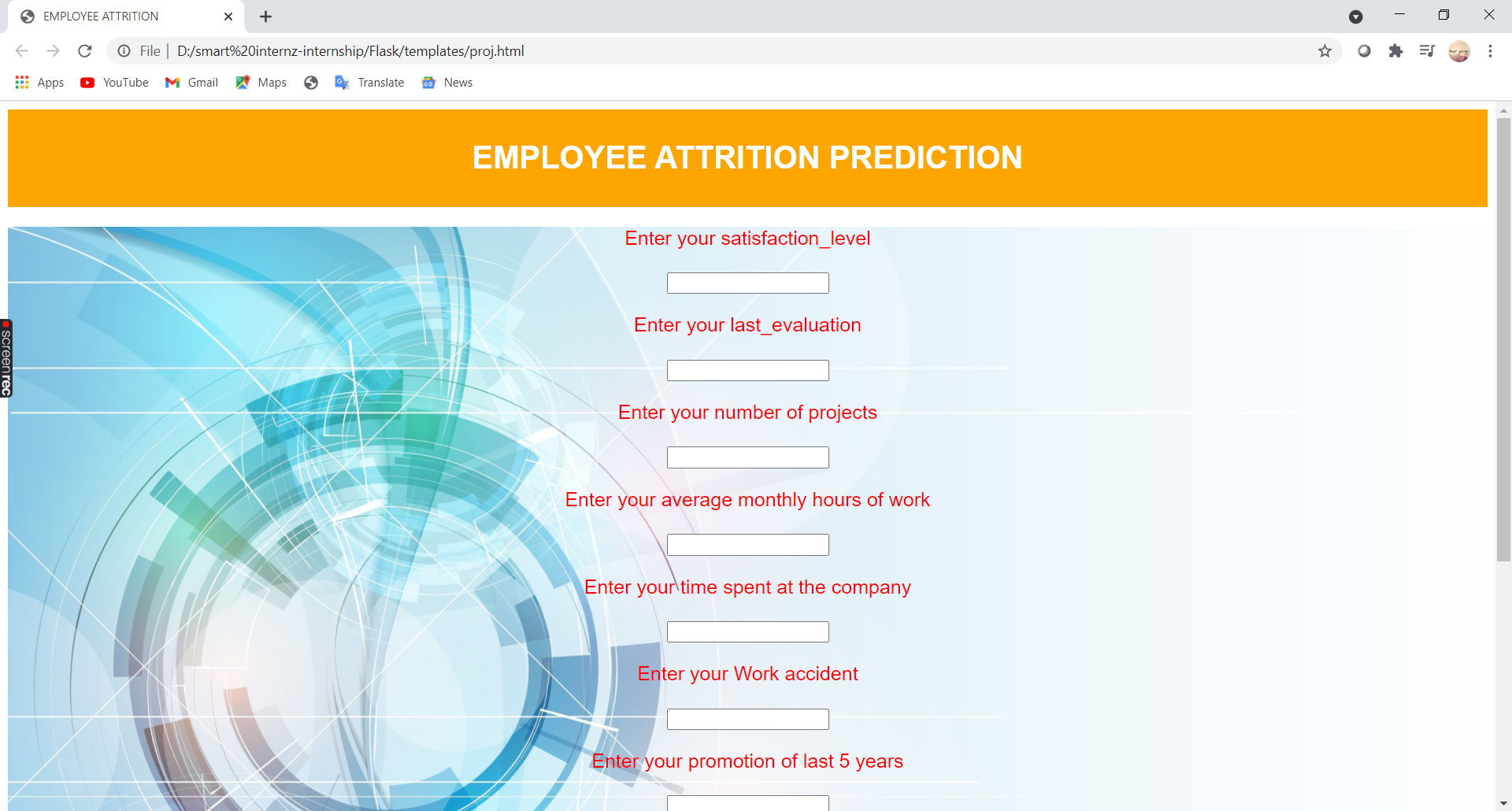
if \_\_name\_\_ == '\_\_main\_\_':

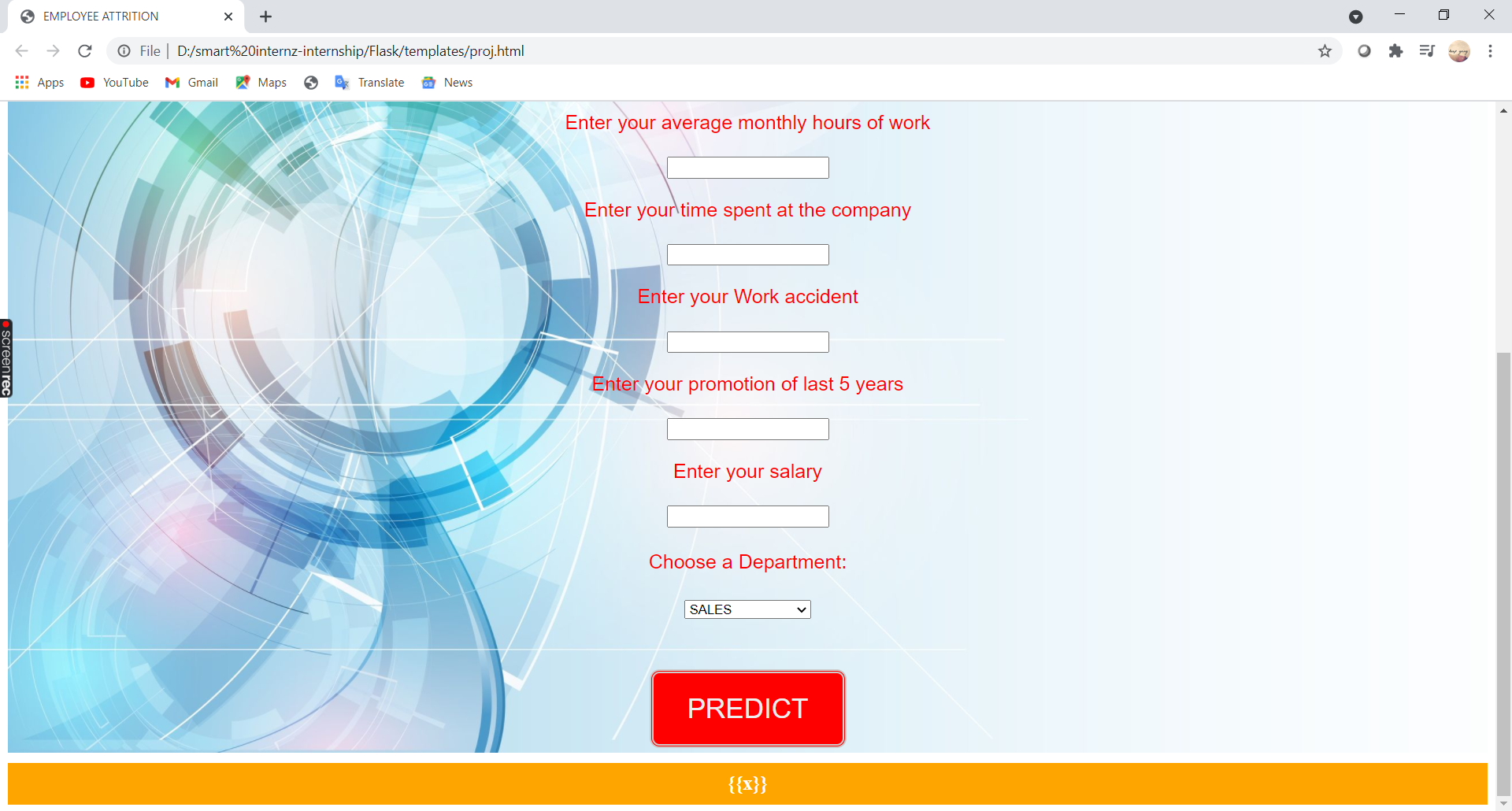
app.run() # you are running your app



**12.2 UI Output Screenshot**

Here we will run the application to get the prediction on the web application. We will deploy our model using flask. In this section, we have built a web application that is integrated into the model we built. A UI is provided for the uses where he has to enter the values for predictions. The enter values are given to the saved model and prediction is showcased on the UI.





By giving the values in all the fields we can predict whether the employee will stay in the company or leave the company.