

***A PROJECT ON***  
**“EMPLOYEE ATTRITION USING MACHINE LEARNING”**

SUBMITTED IN  
PARTIAL FULFILLMENT OF THE REQUIREMENT  
FOR THE COURSE OF  
DIPLOMA IN BIG DATA ANALYSIS



***SUNBEAM INSTITUTE OF INFORMATION TECHNOLOGY***

‘Plot no R/2’, Market yard road,  
Behind hotel Fulera, Gultekdi  
Pune – 411037.  
MH-INDIA

**SUBMITTED BY:**

**Vijay Y Wartale (75802)**

**Shivam Panchlinge(75182)**

**UNDER THE GUIDENCE OF:**

Mrs. Pradnya Dindorkar  
Faculty Member  
Sunbeam Institute of Information Technology, PUNE.



## **CERTIFICATE**

This is to certify that the project work under the title 'Employee Attrition Using Machine Learning' is done by Vijay Y Wartale & Shivam Pnchlinge in partial fulfillment of the requirement for award of Diploma in Big Data Analysis Course.

**Mrs. Pradnya Dindorkar**  
**Project Guide**

**Mrs. Pradnya Dindorkar**  
**Course Co-Ordinator**

Date:

## **ACKNOWLEDGEMENT**

A project usually falls short of its expectation unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of gratitude towards Mr. Nitin Kudale (Center Coordinator, SIIT, Pune) and Mrs. Pradnya Dindorkar (Course Coordinator, SIIT ,Pune) and Project Guide Mrs. Pradnya Dindorkar.

We are deeply indebted and grateful to them for their guidance, encouragement and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form.

Last but not the least we thank the entire faculty and the staff members of Sunbeam Institute of Information Technology, Pune for their support.

Vijay Y Wartale  
DBDA March 2023  
Batch,SIIT Pune

Shivam Panchlinge  
DBDA March 2023  
Batch,SIIT Pune

# **TABLE OF CONTENTS**

## **1. Introduction**

1. Introduction And Objectives
2. Why this problem needs To be Solved?
3. Dataset Information

## **2. Problem Definition and Algorithm**

1. Problem Definition
2. Algorithm Definition

## **3. Experimental Evaluation**

1. Methodology/Model
2. Exploratory Data Analysis

## **4. Results And Discussion**

## **5. GUI**

## **6. Future Work And Conclusion**

1. Future Work
2. Conclusion

## **1. Introduction**

### **1. Introduction And Objectives:**

Employee attrition is project based on data of employee's in the company. Here we analyze the data and clean the data to use it's all potential. In the data set we have total 10 columns that we analyze the data and help company to better understand their employee's. before doing that , let me point out the objective of this analysis. The employee may take a new job or retire. An attrition policy takes advantage of tis inevitable changeover to reduce overall staff. Our Main Objective is to predict the status of employee that it will leave or not.

### **2. Why this problem needs To be Solved?**

Employee attrition helps the company to predict the employee will leave the company or not . This helps the company to know their employee very well so company help them to stay in the company by giving offer or resign them by checking their status.

We have very two helpful columns that help the company to know their employee perception about their company and they are happy with company work culture or not. The columns are last year satisfaction level and current year satisfaction level this is the some question about company culture this help company to add new features and remove also. This columns are very helpful for our projects.

### **3.Dataset Information.**

#### **HR\_Employee\_Data.csv**

It has 10 columns.

Emp\_Id: specific employee id of all employee

Satisfaction Level: satisfaction level range between 0 to 100%.

Last Evaluation: Last evaluation range between 0 to 100%.

Number Project: How many project done.

Average Monthly Hours: monthly average hours of company in a year.

Time Spend : how many year spend in company.

Left: have categorical value 1 and 0.

Promotion In Last 5 year: have categorical value 1 and 0.

Department: Department number between 1 to 10.

Salary: Salary between three parameter low, medium and high.

## **2. Problem Definition and Algorithm:**

### **1. Problem Definition**

When considering attrition, many leaders tend to focus on the problem of high turnover with good reason. Companies also suffer productivity losses and lots of profit when there is a large amount of continuous churn in the workforce. Top talent, in particular, can be very difficult to find and expensive. In fact, our metric of interest will be the accuracy, precision, recall and f1 score. The metric is not very complicated. The further away from the actual outcome our forecast is, it is very helpful for HR team. Optimally, we predict the employee will leave or not.

### **2.1 Algorithm Definition:**

**SVM(Support Vector Machine):** It is a supervised machine learning algorithm that can be used for both classification and regression. However, it is mostly used in classification problems. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional. SVM separates the classes using hyperplane. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane.

**KNN(K nearest neighbours):** The k-nearest neighbours (KNN) algorithm is a simple, easy-to-implement supervised machine learning. Algorithm that can be used to solve both classification and regression problems. However, it is more widely used in classification problems in the industry. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection. The KNN algorithm assumes

**Random forest:** is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

**Decision Tree:** algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data). In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

**Ada Boost:** AdaBoost can be used to boost the performance of any machine learning algorithm it is best used with weak learners these are model that achieve accuracy just above random chance on a classification problem the most suited and therefore most common algorithm used with AdaBoost are decision trees with one level.

**Gradient Boost:** Gradient boosting algorithm can be used for predicting not only continuous target variable but also categorical target variable. When it is used as regressor, the cost function is mean square error and when it is used as classifier then the cost function is log loss. In gradient boosting we train the multiple models sequentially and for each new model the model gradually minimize the loss function using the gradient descent method.

**XGBoost:** or extreme gradient boosting is one of the well-known gradient boosting techniques (ensemble) having enhanced performance and speed in tree-based (sequential decision trees) machine learning algorithms. XGBoost was created by Tianqi Chen and initially maintained by the Distributed (Deep) Machine Learning Community (DMLC) group. It is the most common algorithm used for applied machine learning in competitions and has gained popularity through winning solutions in structured and tabular data. It is open- source software. Earlier only python and R packages were built for XGBoost but now it has extended to Java, Scala, Julia and other languages as well.

### **3.Experimental Evaluation:**

#### **1. Methodology:**

The objective of this project is to predict Employee Attrition in the Company. The data set is contained from Kaggle and has 10 columns. First off all we analyzed data then cleaned it to use its all potential.

#### **Loading in raw data**

```
df = pd.read_csv('./HR_Employee_Data.csv')
```

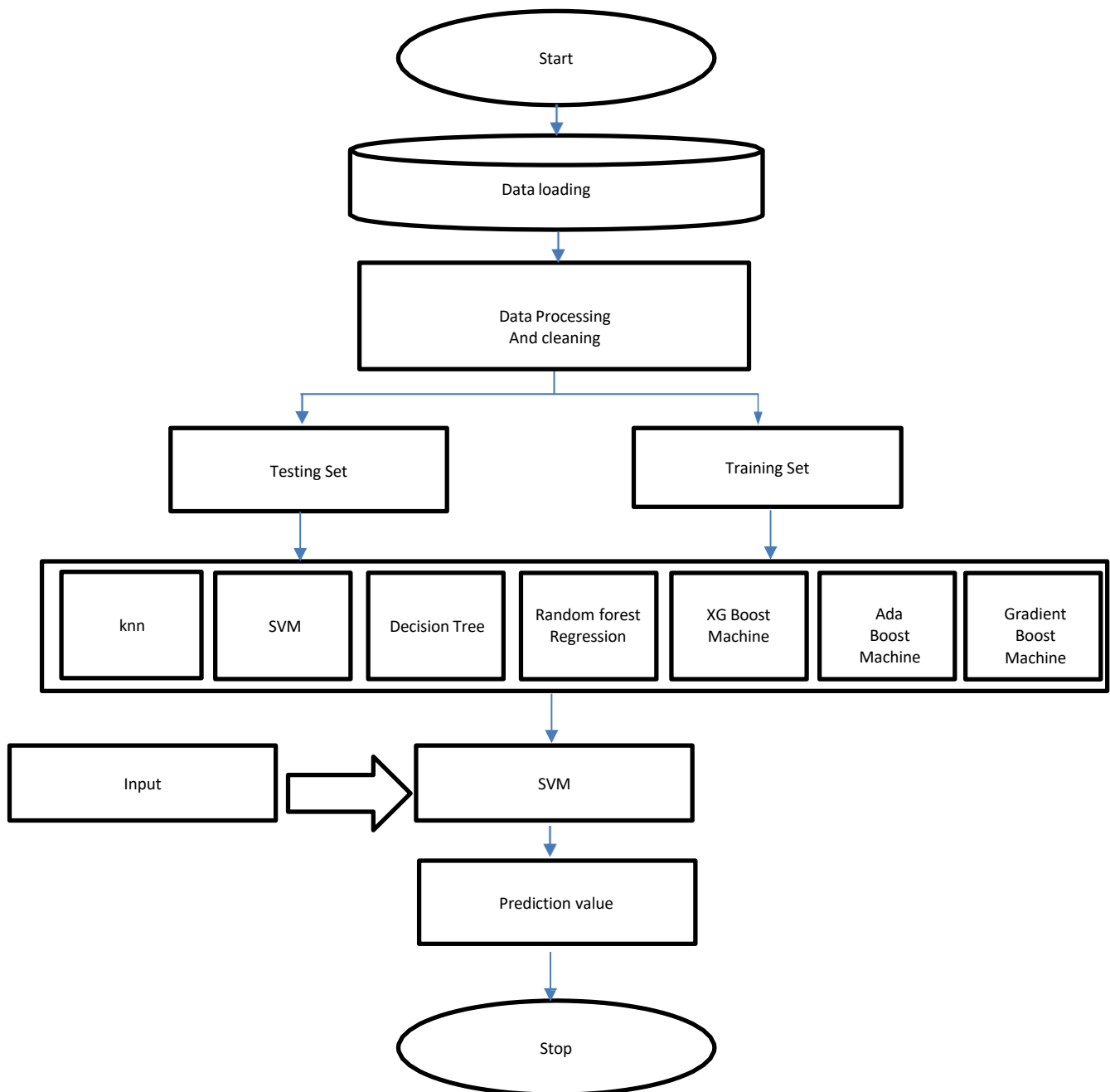
#### **Preprocessing:**

After loading the data, we use the info command to get information about data after we use describe command to get count, mean, std deviation, min, max of our data.

Then we use isna command to find missing value in the data. we got missing value and we perform some operation on that to clean the data we remove the all null value we put average of values of that specific column to missing value and also change the data type of project.

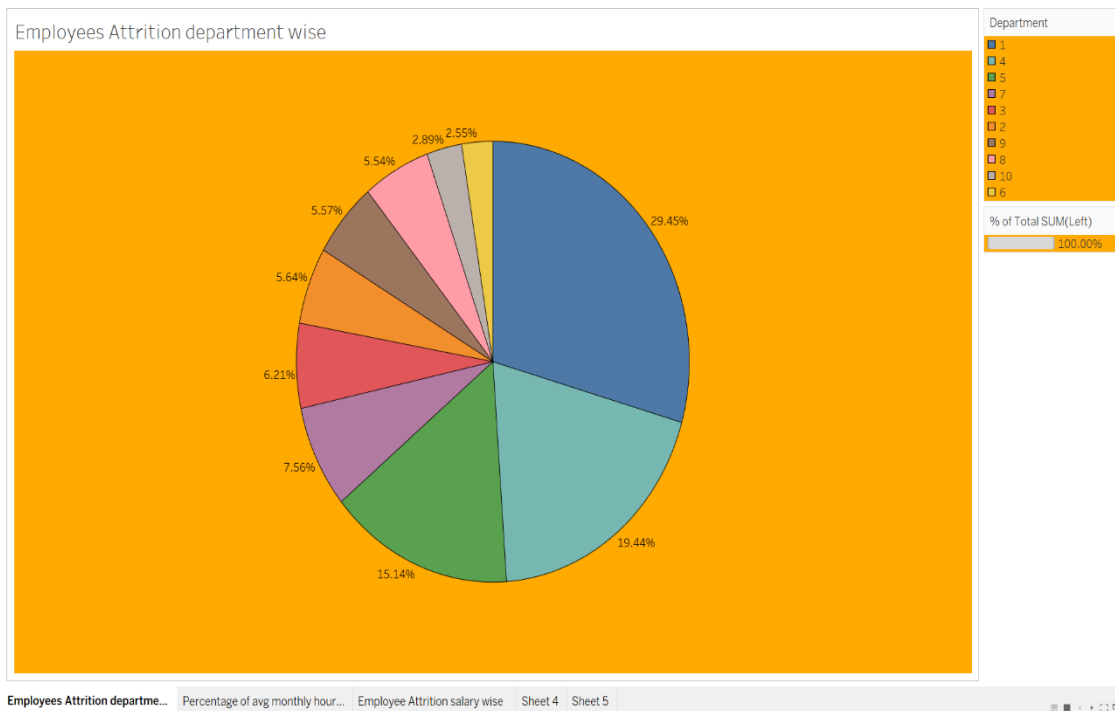


## Flow Diagram :

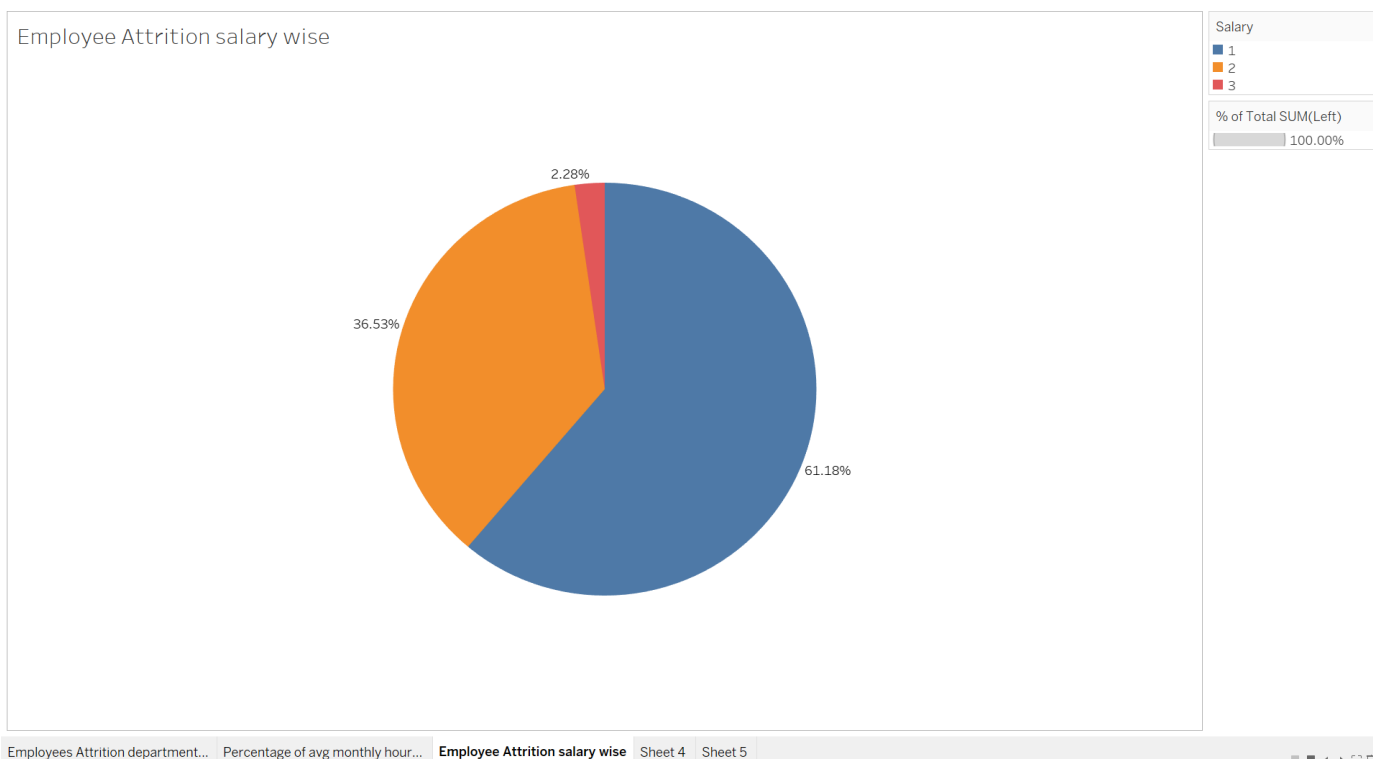


### 3.2 Exploratory Data Analysis

Here we plot the all employee left from each department to understand which department faces most changeover of employees.



In the fig3 it shows the Employee attrition salary wise by using salary.



In the another visualization we take the average value of important parameter and done the analysis.

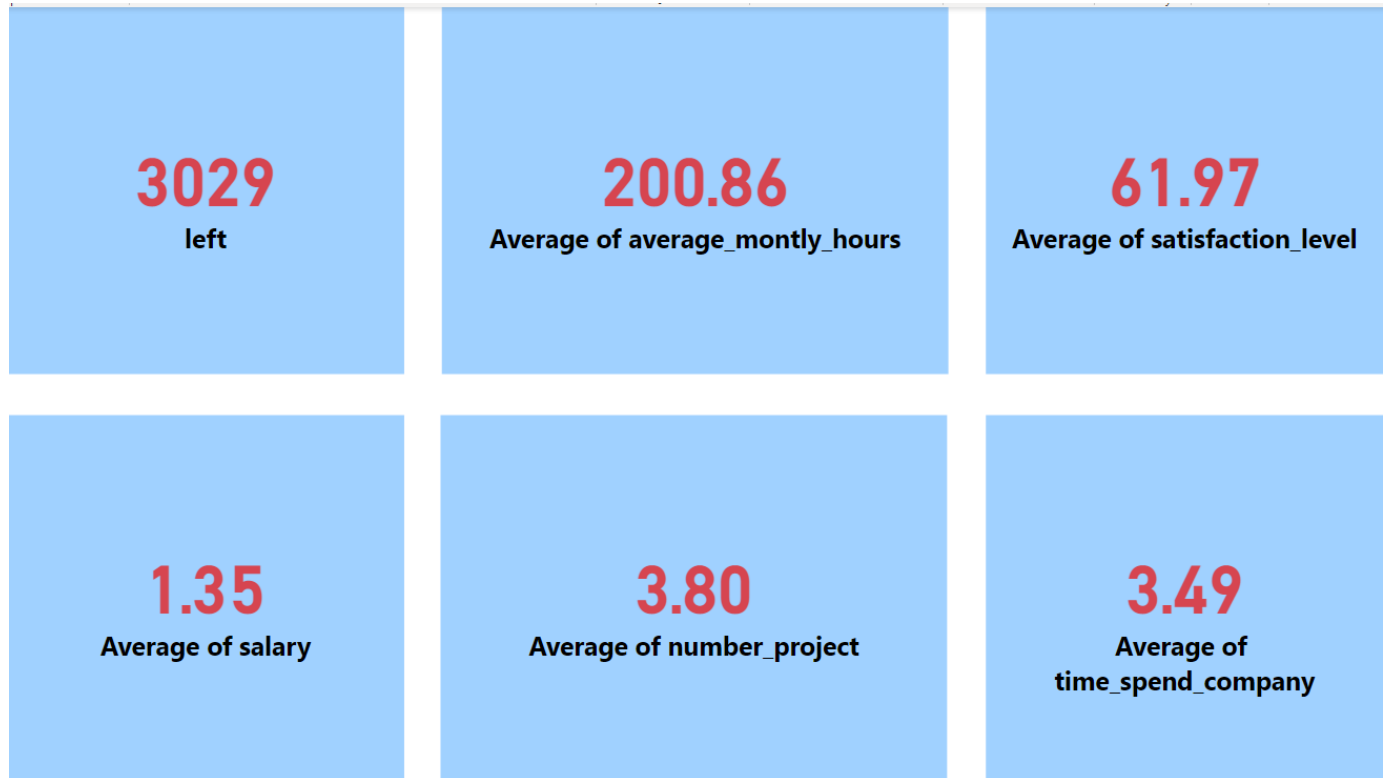


Fig 4: Average of parameters

#### 4. Results and discussion:

Knn(K nearest neighbor), SVM(support vector machine), random forest, decision tree, XG boost, Ada Boost, Gradient boost this algorithm we used in our project in all of this algorithm most of it overfitting so to avoid that we use SVM algorithm with 0.89 accuracy and it gives us very accurate result.

Below we write the function for the all function and accuracy table

#knn model to check accuracy

```
def knn():
```

```
    from sklearn.neighbors import KNeighborsClassifier
```

```
    # create the model
```

```
    model = KNeighborsClassifier()
```

```
    # train the model
```

```
    model.fit(x_train, y_train)
```

```
    return model
```

```
#support vector machines model to check accuracy
def svm():
    from sklearn.svm import SVC

    # create the model
    model = SVC(C=3.0)

    # train the model
    model.fit(x_train, y_train)

    return model

#decision tree model to check accuracy
def decision_tree():
    from sklearn.tree import DecisionTreeClassifier

    # create the model
    model = DecisionTreeClassifier()

    # train the model
    model.fit(x_train, y_train)

    return model

#Random forest model to check accuracy
def random_forest():
    from sklearn.ensemble import RandomForestClassifier

    # create the model
    model = RandomForestClassifier()

    # train the model
    model.fit(x_train, y_train)

    return model
```

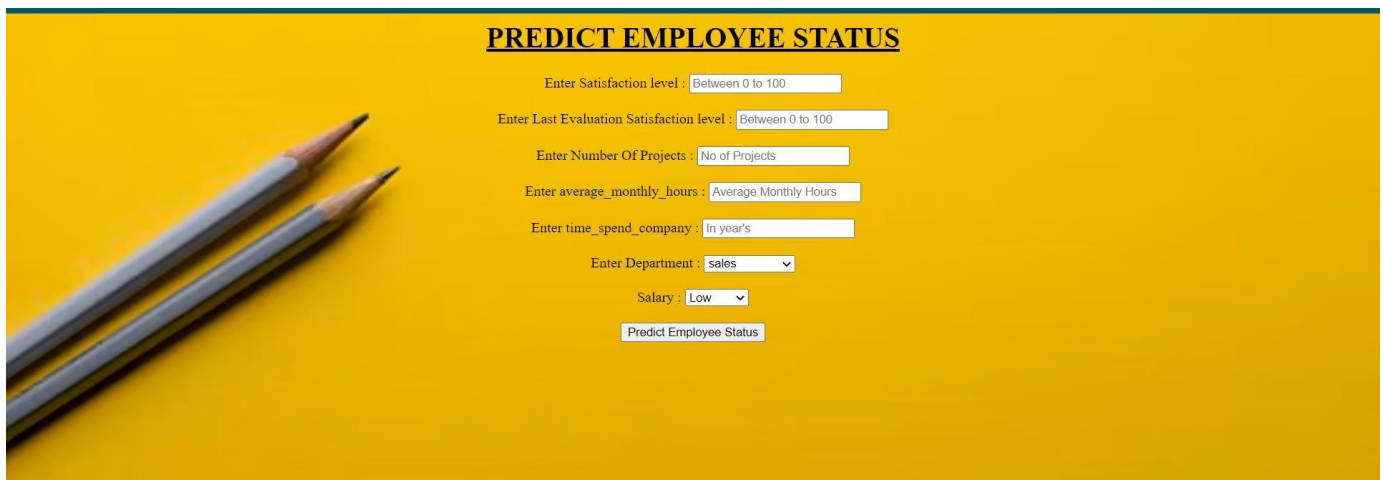
In this way we implemented all remaining algorithms.

Accuracy table with some results:

name	accuracy	precision	recall	f1
knn	0.95	0.87	0.89	0.88
svm	0.89	0.84	0.63	0.72
decision_tree	0.98	0.96	0.97	0.96
random forest	0.99	1.00	0.97	0.98
xgboost	0.99	0.99	0.97	0.98
Ada Boost	0.97	0.92	0.92	0.92
Gradient Boost	0.98	0.96	0.94	0.95

## 5. GUI:

GUI is made using Flask framework. **Flask** is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools



**PREDICT EMPLOYEE STATUS**

Enter Satisfaction level :

Enter Last Evaluation Satisfaction level :

Enter Number Of Projects :

Enter average\_monthly\_hours :

Enter time\_spend\_company :

Enter Department :

Salary :

## 6. Future work And Conclusion

### 6.1Future Work:

In future we have to collect more data about employee so we understand the employee more. So it helps us to keep talented employee in the firms and help them to grow with us.

## **2. Conclusion:**

- Satisfaction level plays most important role in the analysis and the status.
- Work pressure of employee also we have understand by how many hours they spend the company so we implement some features for stress relief.
- SVM model gives us the very accurate value with most accuracy other model are overfit the data.

