**EMPLOYEE ATTRITION Prediction using Random Forest**

1. **INTRODUCTION**
   1. **Overview**

In today’s competitive economy and its growing technological specialisation, acquisition, study and analysis of data are giving rise to new knowledge, referred to as “knowledge economy”. Information technologies are not only a source of data but are, above all, an enabling factor for data analysis, making it possible to process large data collections and allow information to be extracted from them. 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 Within organisations there are several areas in which the adoption of artificial intelligence impacts on a company’s decision-making activities. In recent years, increasing attention has been focused on human resources (HR), since the quality and skills of employees constitute a growth factor and a real competitive advantage for companies . In fact, after becoming more adopted in sales and marketing areas, artificial intelligence is now also starting to guide company decisions regarding their employees, with the aim of basing HR management decisions on the analysis of objective data rather than subjective considerations. In general, companies try to maximise their profits. In companies where workers perform simple tasks, they can resort to on-call, occasional and temporary work (as in the gig economy), because they have fewer contractual obligations. However, for companies in which workers perform more specialised tasks, the specialisation and continuity of work of the employee becomes essential. The importance of skills, knowledge and continuous learning ability has proven to be fundamental for businesses. 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. Companies invest a lot of time and resources in employee recruiting and training, according to their strategic needs. Therefore, the employees (to a greater or lesser extent) 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.

* 1. **Purpose**

The reasons for employee turnover rate (attrition) are mainly related to their motivation to work and satisfaction measures . Employees who are satisfied will less likely decide to leave the company. Satisfaction measures are also related to performance. More satisfied employees show higher performance measures satisfaction is a result of intrinsic motivational factors such as recognition, professional growth opportunities and a good feeling about the organization. The factors contributing to dissatisfaction avoidance include effective senior management and supervisor, satisfaction with salary and benefits and good relationships with co-workers. According to the Two-factor theory - by fulfilling extrinsic factors, employees can feel neutral, but not extra satisfied . If the needs of extrinsic factors are met, then employees can get motivated and in turns satisfied by intrinsic factors

1. **LITERATURE SURVEY**
   1. **Existing Problem**

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. 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. After the training, the obtained model for the prediction of employees’ attrition is tested on a real dataset provided by IBM analytics, which includes 35 features and about 1500 samples. Results are expressed in terms of classical metrics and the algorithm that produced the best results for the available dataset is the Gaussian Naïve Bayes classifier. It reveals the best recall rate (0.54), since it measures the ability of a classifier to find all the positive instances and achieves an overall false negative rate equal to 4.5% of the total observations.

* 1. **Proposed System**

This part describes the dataset and contains block diagrams, flow diagrams, evaluation matrices, and the study’s procedure and methodology. The below figure(1) depicts the block diagram of the proposed system. The framework utilizes the EA prediction dataset. After preprocessing and feature selection, Random Forest Classifier have been used. All the components of this diagram have been discussed in the following sub sections.

1. **THEORITICAL ANALYSIS**
   1. **Block Diagram**

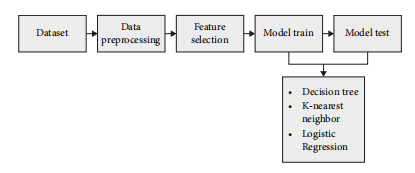
****

Figure-1 Block Diagram of the Proposed System

* 1. **Hardware/Software Designing**
* Google Co-Laboratory
* VS(Visual Studio) Code
* Python 3.10.4

1. **EXPERIMENTAL INVESTIGATIONS**

**4.1.1 Dataset**

The dataset contains target feature, identified by the variable Attrition: “No” represents an employee that did not leave the company and “Yes” represents an employee that left the company. This dataset will allow the machine learning system to learn from real data rather than through explicit programming. If this training process is repeated over time and conducted on relevant samples, the predictions generated in the output will be more accurate.

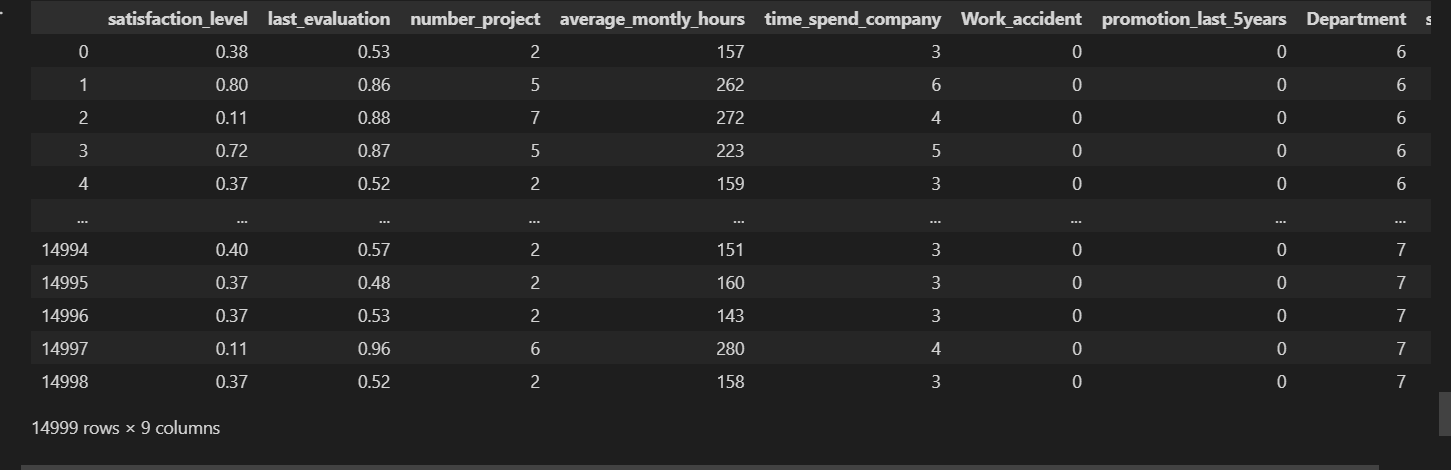


Figure-2 Data set

**4.1.2 Data Pre-Processing**

Data Preprocessing. Prior to model building, data preprocessing is required to remove unwanted noise and outliers from the dataset that might cause the model to diverge from the proper training set. This stage tackles anything that is impeding the model’s efficiency. After collecting the necessary data, it must be cleaned and prepared for model construction. The dataset is next searched for null values. However, this dataset contains no null values. Figure 3 shows there is no missing data available in this

dataset. Here, the output values “False” and “0” indicate the absence of null values. After completing data preparation and handling the unbalanced dataset, the next step is to build the model. To increase the accuracy and efficiency of this task, the data is split into training and testing segments, with an 80/20 ratio of training to testing. Following the

model’s splitting, it is trained using a number of classifification techniques. The classifification methods used in this research include the Random Forest classifification method,

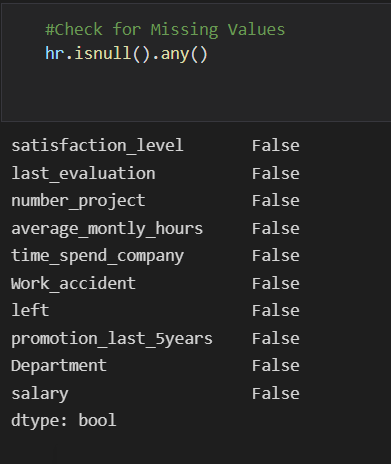
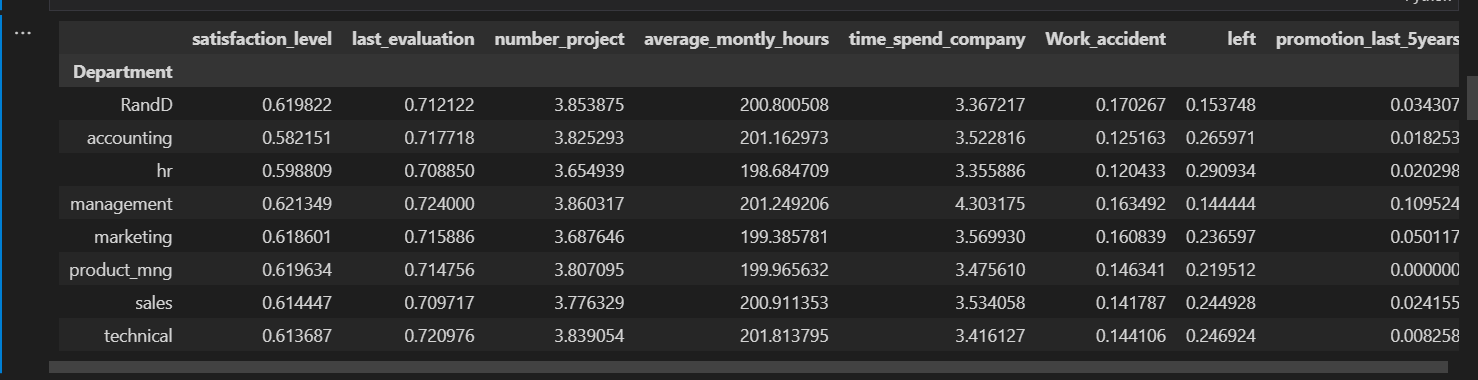


Figure-3 checking for Missing values

**4.1.3 Data Exploration**

At this point we generated the descriptive statistics of the dataset in order to observe the characteristics of all variables. We considered the following variables: count, unique, top, frequency, mean, standard deviation (std), minimum and maximum values (min/max), 25%/50%/75% percentile. An extract of the overall dataset is reported in Figure



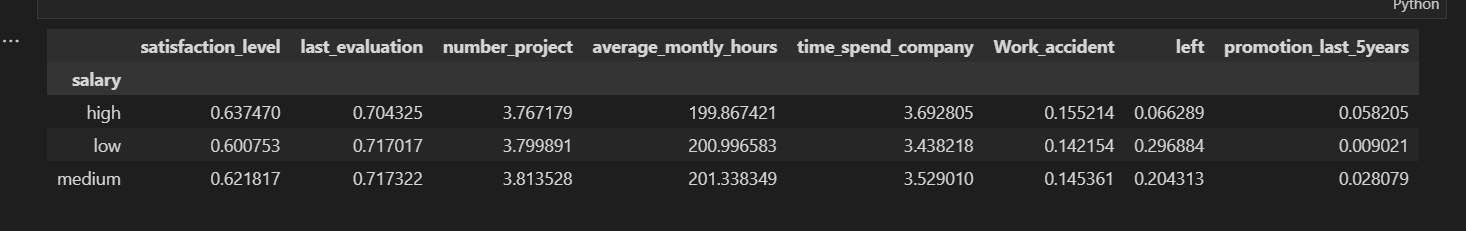


Figure-4 Dataset descriptive statistics.

Data exploration phases were complete, it was necessary to proceed with Categorical Encoding before moving to the descriptive analysis. In this phase we transform data to the correct format to perform the analysis. It is necessary to change the format of some variables to allow greater readability and a comparison with other analysed quantities, in order to avoid ambiguous results. The format modification involved arranging values into uniform clusters to make values comparable, or mapping for each numeric value with a corresponding categorical variable.

**5. Algorithms**

The following machine learning algorithms have been used to predict Employee Attrition

1. Random Forest Classifier

**5.1 Random Forest Classifier**

 “Random Forest Classifier” refers to the classification algorithm made up of several decision trees. **The algorithm uses randomness to build each individual tree to promote uncorrelated forests, which then uses the forest's predictive powers to make accurate decisions**.

****

1. Figure-5 RF Working Procedure

RF classifier is an ensemble method that trains several decision trees in parallel with bootstrapping followed by aggregation, jointly referred as bagging Bootstrapping indicates that several individual decision trees are trained in parallel on various subsets of the training dataset using different subsets of available features. Bootstrapping ensures that each individual decision tree in the random forest is unique, which reduces the overall variance of the RF classifier. For the final decision, RF classifier aggregates the decisions of individual trees; consequently, RF classifier exhibits good generalization. RF classifier tends to outperform most other classification methods in terms of accuracy without issues of overfitting. Like DT classifier, RF classifier does not need feature scaling. Unlike DT classifier, RF classifier is more robust to the selection of training samples and noise in training dataset. RF classifier is harder to interpret but easier to tune the hyperparameter as compared with DT classifier

1. **RESULTS AND DISCUSSION, PERFORMANCE ANALYSIS**

**6.1 Random Forest Classsifier**

The higher the AUC, the better is the performance of the model at distiniguishing between the positive and negative classes.Random Forest model(Area=50.4%) To conclude,Random Forest did a better job in classifying the positive class in the dataset and gives the most accuracy in result.

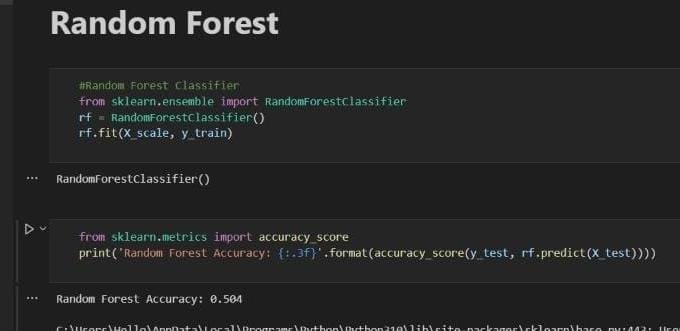


Figure-7 Accuracy of Random Forest

**6.2 10 Fold Cross Validation**

10-fold cross validation would perform the fitting procedure a total of ten times, with each fit being performed on a training set consisting of 90% of the total training set selected at random, with the remaining 10% used as a hold out set for validation.

10 Fold Cross Validation acuuracy is 99.0%

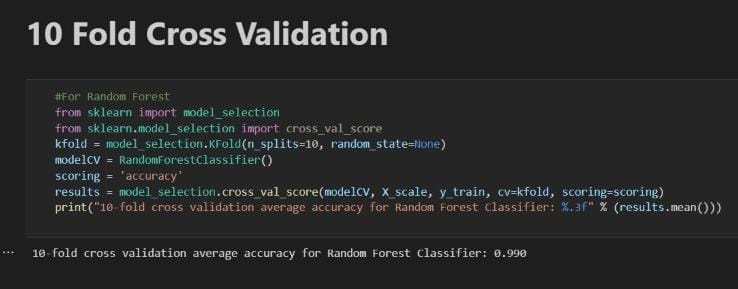


Figure-8 Accuracy of 10 Fold Cross Validation

**6.3 Precision And Recall**

Precision and recall are performance metrics used for pattern recognition and classification in machine learning. These concepts are essential **to build a perfect machine learning model which gives more precise and accurate results**.

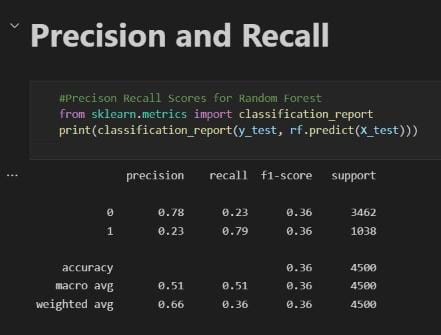


Figure-9 Precision and Recall

**Confusion Matrix**

Figure-(8) shows the confusion matrix. The confusion matrix rates machine learning classification models’ performance. All models were evaluated using the confusion matrix. The confusion matrix illustrates how often our models guess correctly and incorrectly. Poorly predicted values received false positives and negatives, whereas

properly predicted values received genuine positives and negatives. The model’s accuracy, precision-recall trade-off, and AUC were assessed after grouping all predicted values in the matrix.

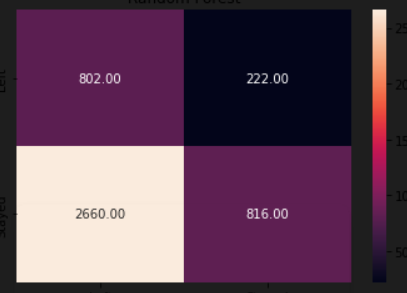


Figure6-Confusion Matrix Diagram

**Chapter – 5, Summary and Conclusions**

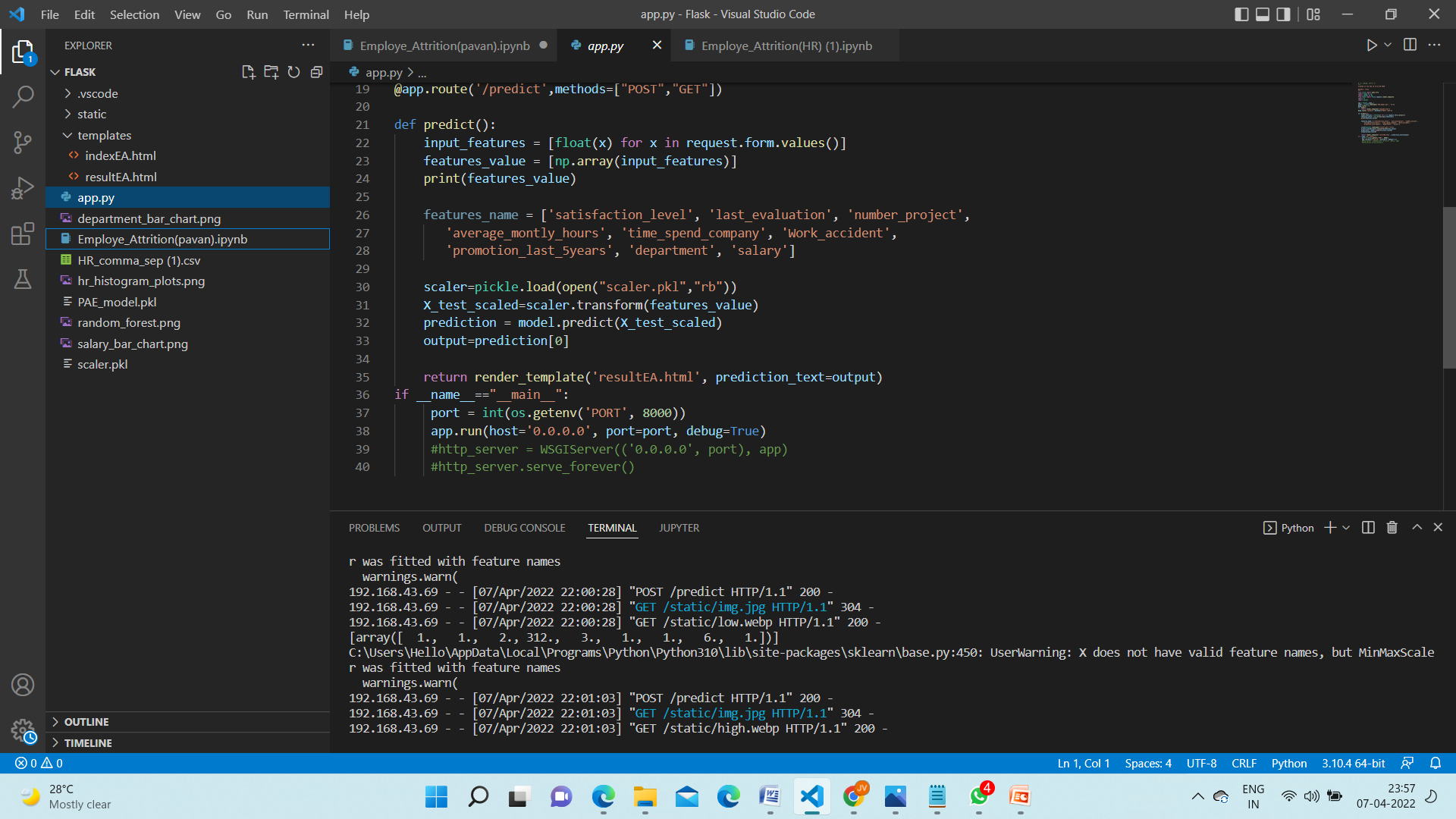
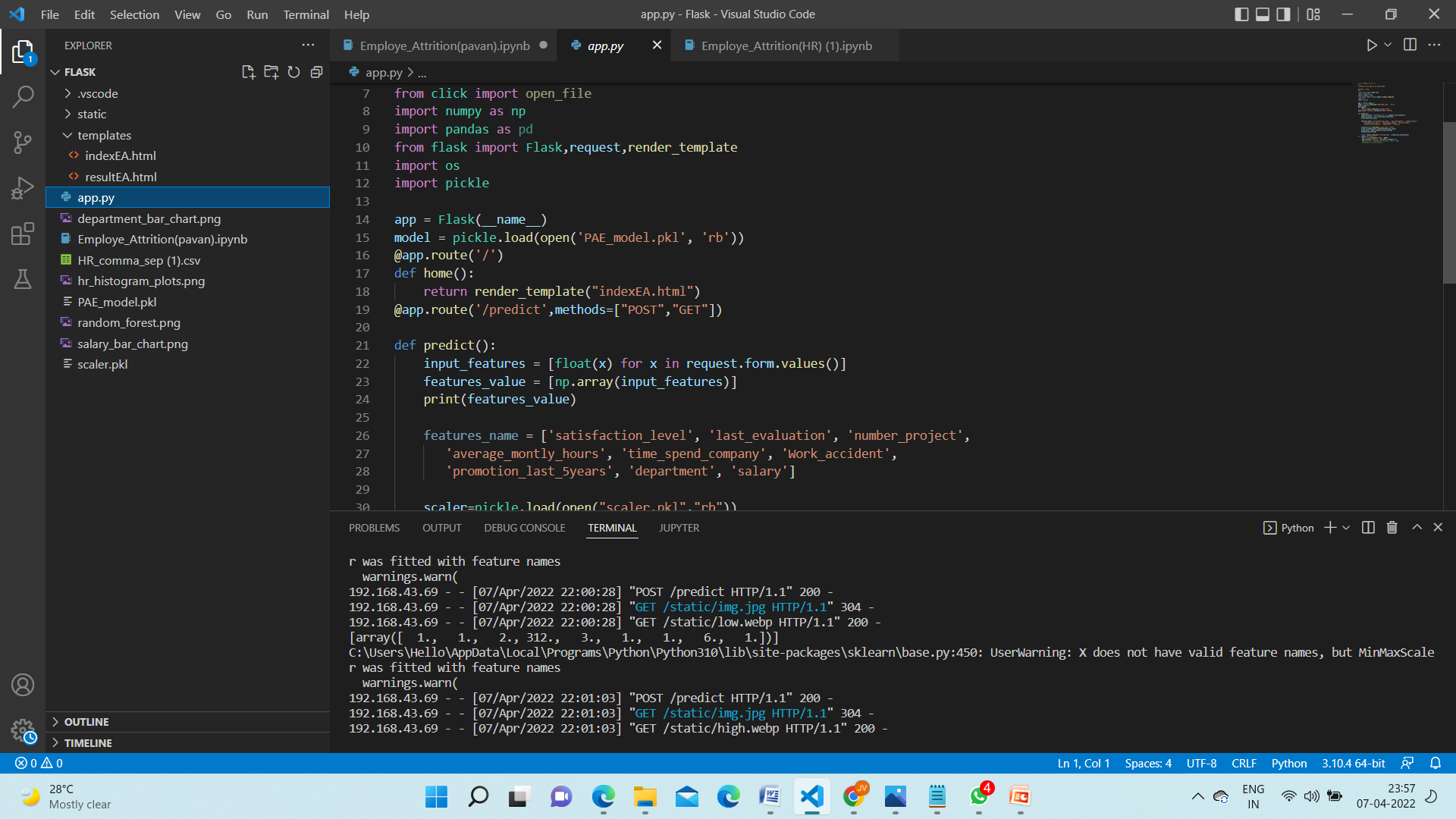
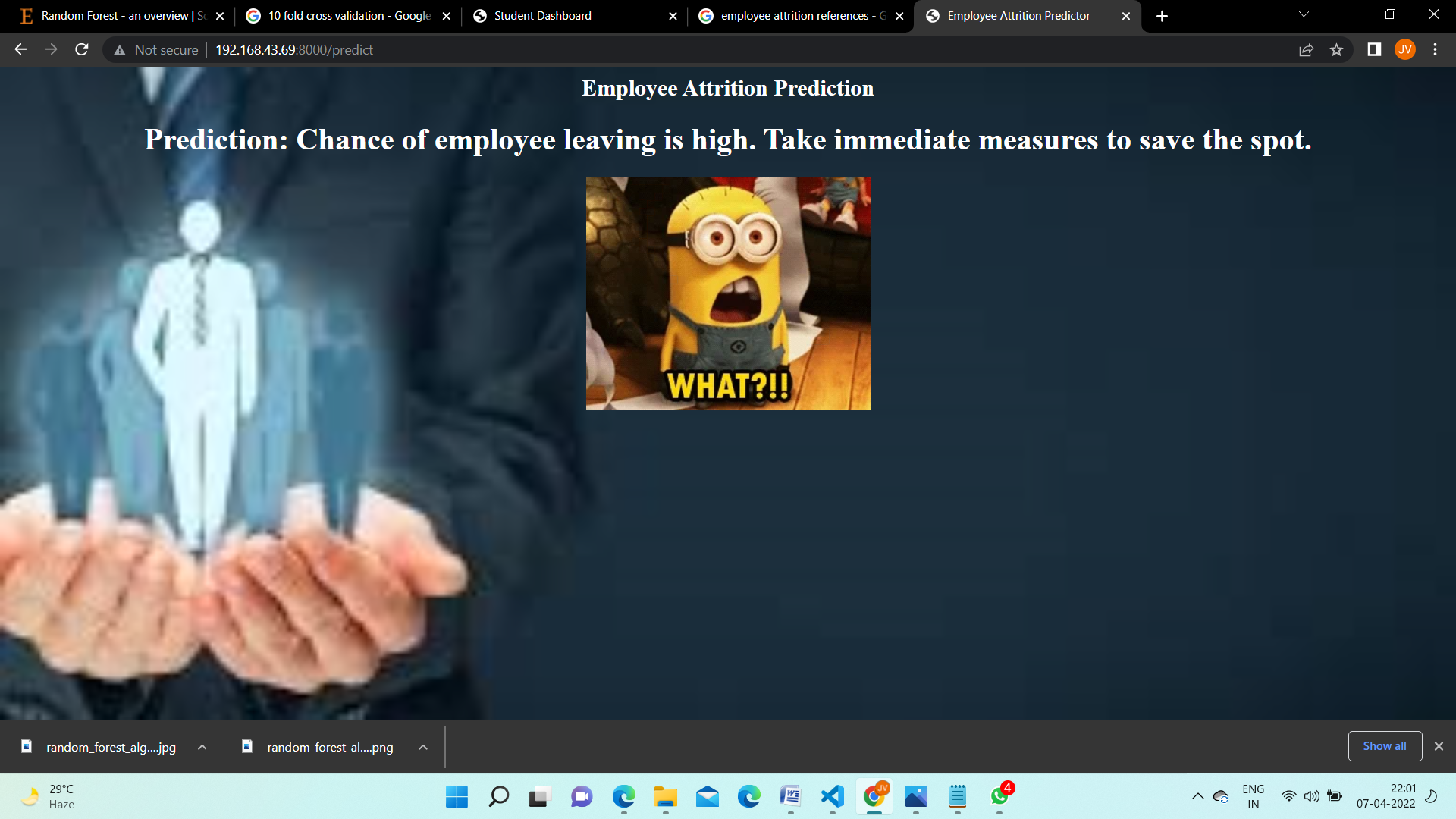
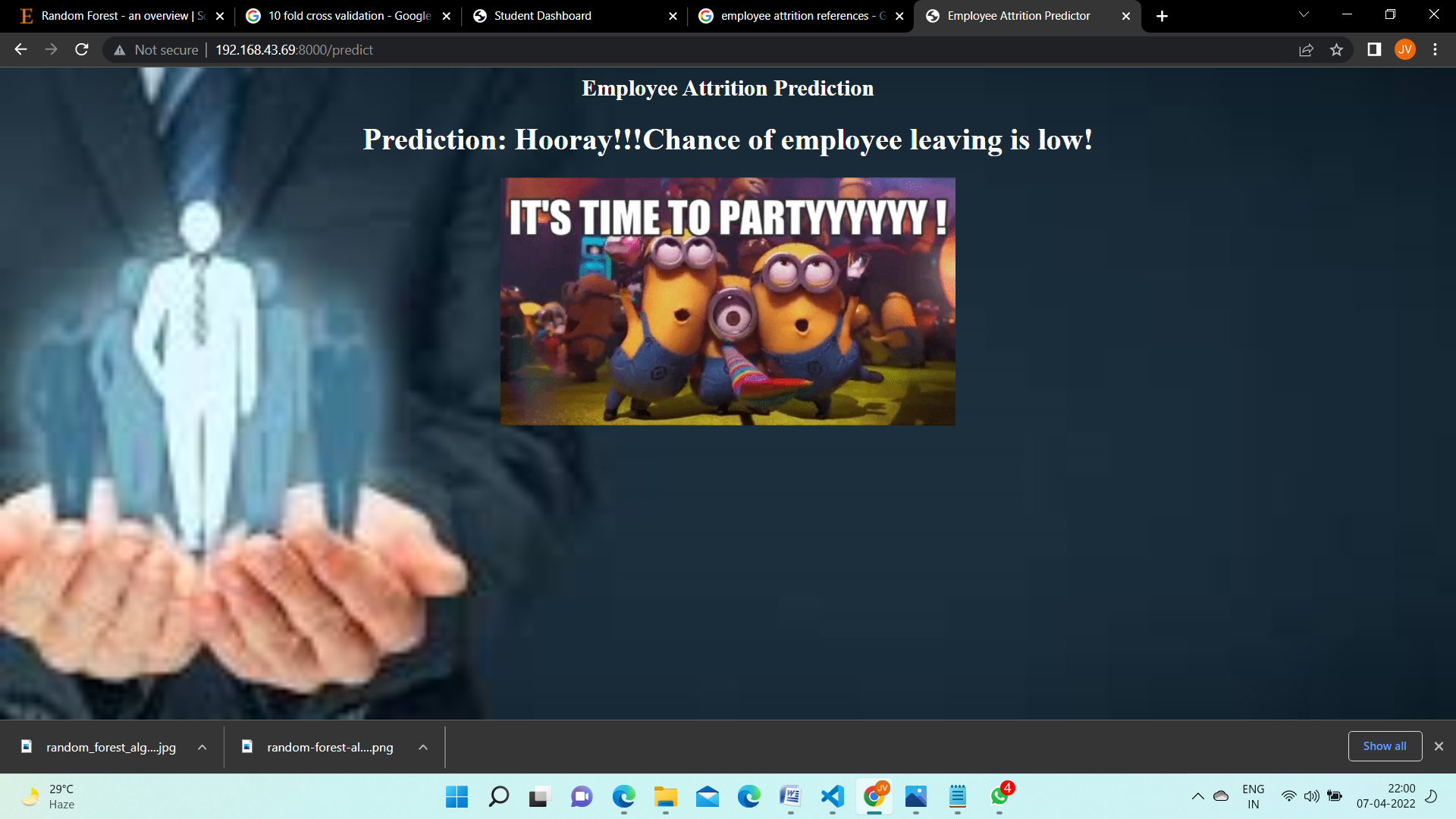
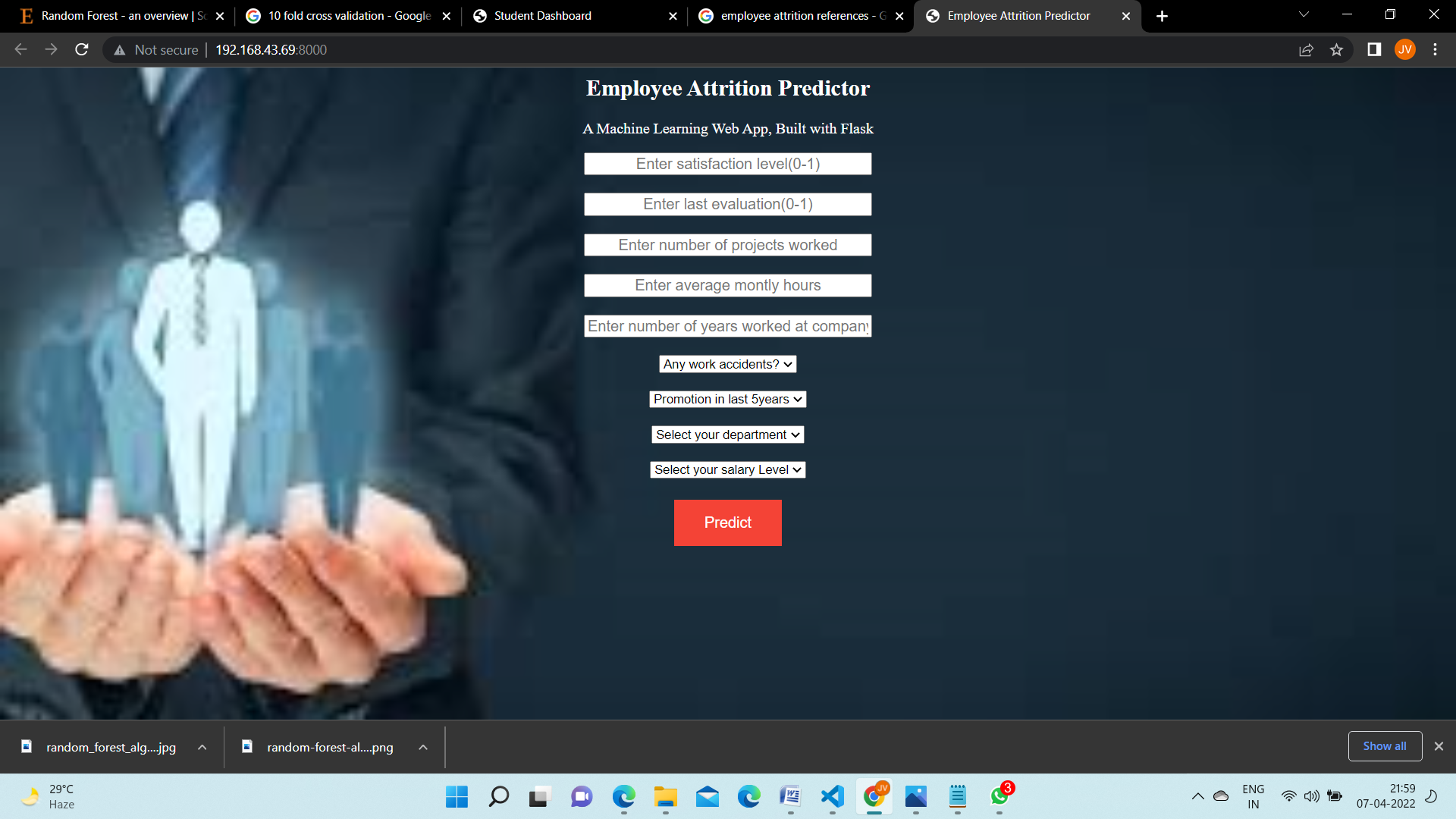
According to the findings of the study, the Random Forest approach is used to predict Employee Attrition more accurately. According to the study, their precision was 78 percent, and their accuracy was 90 percent. Compared to prior research, the accuracy percent of the models used in this investigation is considerably higher, indicating that the models used in this study are more reliable than those used in previous studies. When cross validation measurements are used in the prediction of Employee Attrition, the RF method out performs the other processes. Future research may build on this work by developing a web application that incorporates these algorithms and using a bigger dataset than the one utilized in this study. This will aid in the achievement of improved outcomes as well as the accuracy and efficiency . This will enhance the dependability of the framework as well as the framework’s presentation. The hope is that it would encourage people to seek early Alert for to saty in a company and to make improvements in their lives.

**Data Availability**

The data utilized to support these research findings is accessible online at <https://www.kaggle.com/giripujar/hr-analytics>

**Appendix**

**Main Code & Output Screenshots**

****