**Ảnh có chứa quảng trường

Mô tả được tạo tự độngVIETNAM NATIONAL UNIVERSITY – HO CHI MINH CITY**

UNIVERSITY OF INFORMATION TECHNOLOGY

🙞🕮🙝



**FINAL PROJECT REPORT**

Subject: BUSINESS DATA ANALYSIS

TOPIC: ANALYSIS OF THE EMPLOYEE ATTRITION

AT THE COMPANY

Instructor:  **PhD. Nguyen Dinh Thuan**

**Teacher Nguyen Minh Nhut**

Implementation group:  **Group 4**

Members:

Ho Trong Khang 19521661

Nguyen Cao Khoa 19521694

Le Tuan Khanh 19521681

**Ho Chi Minh City, May 2022**

**VIETNAM NATIONAL UNIVERSITY – HO CHI MINH CITY**

UNIVERSITY OF INFORMATION TECHNOLOGY

🙞🕮🙝



**FINAL PROJECT REPORT**

Subject: BUSINESS DATA ANALYSIS

TOPIC: ANALYSIS OF THE EMPLOYEE ATTRITION

AT THE COMPANY

Instructor:  **PhD. Nguyen Dinh Thuan**

**Teacher Nguyen Minh Nhut**

Implementation group:  **Group 4**

Members:

Ho Trong Khang 19521661

Nguyen Cao Khoa 19521694

Le Tuan Khanh 19521681

**Ho Chi Minh City May 2022**

[THANK YOU 5.](#_Toc105377437)

[TEACHER'S COMMENTS 6](#_Toc105377438)

[CHAPTER I: TOPIC OVERVIEW 7](#_Toc105377439)

[**1.Ask problem 7**](#_Toc105377440)

[**2.Goal 8**](#_Toc105377441)

[**3.Tools used 8**](#_Toc105377442)

[CHAPTER II: DATA PREPROCESSING 8](#_Toc105377443)

[**1.Data description 8**](#_Toc105377444)

[**1.1 Data source 8**](#_Toc105377445)

[**1.2 Data field number 8**](#_Toc105377446)

[**1.3 Property number and property value 8**](#_Toc105377447)

[**1.4 Statistics of property values 9**](#_Toc105377448)

[**1.5 Spreadsheet Properties 18**](#_Toc105377449)

[**1.6 Class number 20**](#_Toc105377450)

[**2.Data processing money 20**](#_Toc105377451)

[2.1 Import library 21](#_Toc105377452)

[2.2 Enter data 21](#_Toc105377453)

[2.3 Data Description 21](#_Toc105377454)

[2.4 Data type test 22](#_Toc105377455)

[2.5 Overview of Data 22](#_Toc105377456)

[2.6 Delete the Over18 23](#_Toc105377457)  column

[2.7 Description of information of string type data 23](#_Toc105377458)

[2.8 Description of information of type 23](#_Toc105377459)  data

[2.9 Test Outliers 23](#_Toc105377460)

[2.10 Change the value of the Attrition column from string to boolean 24](#_Toc105377461)

[**2.11 Use Label Encoder to convert data columns from object to numeric type 25**](#_Toc105377462)

[**2.12 Conduct data classification 25**](#_Toc105377463)

[2.12.1 Genus square test (with properties as Categorical Features) 26](#_Toc105377464)

[2.12.2 Anova Test (with continuous features) 27](#_Toc105377465)

[CHAPTER III: ALGORITHMS AND EXPERIMENTS 28](#_Toc105377466)

[**1.The algorithm uses 28**](#_Toc105377467)

[**1.1 Decision Tree 28**](#_Toc105377468)

[1.1.1 ID3 30](#_Toc105377469)

[1.1.2 CART 31](#_Toc105377470)

[**1.2 Random Forest 31**](#_Toc105377471)

[**1.3 Logistic Regression 32**](#_Toc105377472)

[**1.4 K-nearest Neighbors 32**](#_Toc105377473)

[**2.Experiment on Jupyter Notebook 34**](#_Toc105377474)

[**2.1 Unnecessary Drop Column and split the decision property column into a separate column 34**](#_Toc105377475)

[**2.2 Separation of train and test data (train data accounted for 70%, test accounted for 30%) 34**](#_Toc105377476)

[**2.3 K-Nearest Neighbors Algorithm (KNN) 35**](#_Toc105377477)

[**2.4 Random Forest Algorithm 36**](#_Toc105377478)

[**2.5 Decision Trees Algorithm 39**](#_Toc105377479)

[2.5.1 ID3 Tree 39](#_Toc105377480)

[2.5.2 Cart Tree 42](#_Toc105377481)

[**2.6 Logistic Regression**](#_Toc105377482)  [**Algorithm 44**](#_Toc105377482)

[**2.7 Comparison, rating 46**](#_Toc105377483)

[CHAPTER IV: PREDICTIVE SOFTWARE 49](#_Toc105377484)

[**1.Software overview 49**](#_Toc105377485)

[**1.1 Algorithm used 49**](#_Toc105377486)

[**1.2 Attributes used to make predictions 49**](#_Toc105377487)

[**1.3 Interface and Testing 50**](#_Toc105377488)

[**2.Software Code 52**](#_Toc105377489)

[**2.1 Code part interface 52**](#_Toc105377490)

[**2.2 Processing part code 56**](#_Toc105377491)

[CHAPTER 5: CONCLUSION 58](#_Toc105377492)

[**1. Advantages and limitations 58**](#_Toc105377493)

[**5.1.1 Decision Tree 58**](#_Toc105377494)

[**5.1.2 Random Forest 59**](#_Toc105377495)

[**5.1.3 Logistic Regression 59**](#_Toc105377496)

[**5.1.4 K-nearest Neighbors 60**](#_Toc105377497)

[**2. Development Direction 60**](#_Toc105377498)

[CHAPTER 6: REFERENCE 60](#_Toc105377499)

[**Reference link 60**](#_Toc105377500)

# **ACKNOWLEDGEMENTS**

In fact, there is no success that is not tied to the support, help more or less, whether directly or indirectly of others. With the deepest gratitude, first of all, our group would like to express our sincere thanks to the teachers of the University of Information Technology - National University of Ho Chi Minh City. Ho Chi Minh City and the Teachers of Information Systems have helped the group have the basic knowledge as the foundation to implement this topic.

In particular, our group would like to express our sincere thanks to Mr. Nguyen Dinh Thuan – lecturer in theory and Mr. Nguyen Minh Nhut - a lecturer practicing business data analysis decided to have dedicatedly helped, directly instructed and guided the group throughout the process of making the subject project. As a result, we have acquired a lot of useful knowledge in the application as well as the skills of making projects. Without the instructions and teachings of the two teachers, our group thought that this project of the group was very difficult to complete. Again, I would like to sincerely thank you both. In addition, in order for the project to be completed, it is impossible to thank the people who made it, thank you to the team members who have worked hard and worked hard to complete the task on schedule.

During the one semester of the project, the authors tried to apply the foundational knowledge accumulated while combining with learning and researching new knowledge to apply to the topic **"Analysis of the employee attritionat the company".** However, due to limited knowledge and experience, the team of authors will find it difficult to avoid shortcomings. Therefore, the group is looking forward to receiving comments from the two teachers to improve the knowledge that the authors have learned and is a baggage for the authors to continue other topics in the future.

Thank you very much to both of you!

Student group implementation.

# **TEACHER'S COMMENTS**

………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………… ………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………… …………………………… ……………………………………………………………………

# **CHAPTER I: OVERVIEW OF TOPICS**

1. **Questions**

* In an organization, company or business, in the process of operation, they cannot sustain their labor resources, their employees. Currently, our country is in the stage of industrialization, modernization, renovation of models, economic structure, improving labor productivity, competitiveness of enterprises and organizations;

Great contribution to improving the country's economic potential. In which workers play an important role and the core to bring innovation into depth, promote industry. Therefore, it is essential to have a solid source of labor because the loss of personnel is inevitable and as a result of the loss of personnel is related to economic and social problems because each person, regardless of position, All different units are important subjects in creating and promoting the advantages of the organization. If the organization grows, it will make the economy better. So when someone leaves in a certain position, someone else is needed to replace it, especially important positions in the organization, it is very difficult to find the right person in a short time.

* Faced with this problem, our team has come up with a solution to apply information technology and specifically in the field of **Business Data Analysis** derived from **Modeling the business cost of retention** data source, thereby creating machines to make predictions about whether the company is understaffed or not, thereby helping the leaders to have more perspective on the problems in the company such as the working environment, bonuses or whether employees are forcing employees to work too much or not, from which there are decisions, orientations in the future as well as an earlier change in the way the company operates to ensure its human resources. full and best.

1. **Goal**

* Building data systems about natural language, using machine learning to train machines can make highly reliable information and predictions for human service.
* Predict the possibility of staff loss at the company, help the company recognize the problem and fix it early to grow for the future.

1. **Tools to use**

* In the process of implementation, the team used a number of software for understanding and building the topic:
  + The information collection and analysis section uses the library and python programming language.
  + Data: [Modeling the Business Cost of Retention | Kaggle](https://www.kaggle.com/code/jamestollefson/modeling-the-business-cost-of-retention/data)
* All of the above software is installed and used by the team on Microsoft Windows 10 Operating System. The compatibility of the above software with other operating systems is not within the scope of research of this topic.

# **CHAPTER II: PREPROCESSING DATA**

## Data description

### Data sources

* By James Tollefson

### Number of data fields

* Total data lines: 1700

### 1.3 Number of properties and property values

* Total properties: 35
* Data set characteristics: avariable
* Property number characteristics: characters, real numbers, integers
* Lost value: none

### 1.4 Statistics of property values

* Symbols: **#** (Number), **✓** (Logical Value), **A** (Letters)
* Nguồn: [Modeling the Business Cost of Retention](https://www.kaggle.com/code/jamestollefson/modeling-the-business-cost-of-retention/data)

Chart, histogram

Description automatically generated

Graphical user interface, chart

Description automatically generated

Graphical user interface, text

Description automatically generated

Chart

Description automatically generated with medium confidence

Graphical user interface, chart

Description automatically generated

Chart

Description automatically generated

Chart, histogram

Description automatically generated

A picture containing chart

Description automatically generated

Chart

Description automatically generated

Chart, histogram

Description automatically generated

Graphical user interface

Description automatically generated

Graphical user interface

Description automatically generated

A picture containing waterfall chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, waterfall chart

Description automatically generated

Chart, waterfall chart

Description automatically generated

Graphical user interface, text

Description automatically generated with medium confidence

### 1.5 Property Statistics Table

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| BTI | Attribute | Property meaning | Property type | Value of property | Average value | Median value | Fashion |
| 1 | Age | Employee's age | Ordinal | From 18-60 | 36.9 | 36 |  |
| 2 | Attrition | Attrition | Nominal | Yes,No |  |  |  |
| 3 | Business Travel | Business travel density | Nominal | Travel\_Rarely,  Travel Frequently, Non Travel |  |  | Travel  \_Rarely |
| 4 | Daily\_Rate | Wages per day | Ordinal | From $102-1499 | 802 | 802 |  |
| 5 | Department | Department | Nominal | Research  &Development,  Sales,  Human Resources |  |  | Research&Development |
| 6 | Distance From Home | Distance to the house | Ordinal | 1-29km | 9.19km | 7 |  |
| 7 | Education | Education level | Nominal | 1-5 | 2.91 | 3 |  |
| 8 | Education Field | Professional knowledge | Nomial | Life Science, Medical, Marketing, Techical Degree, Other |  |  | Life Science |
| 9 | Employee Count | Number of employees | Ordinal | 1 |  |  |  |
| 10 | Employee Number | Employee order number | Ordinal | 1-2068 | 1022 | 1022 |  |
| 11 | Environment Satisfaction | Working environment | Ordinal | 1-5 |  |  |  |
| 12 | Gender | Gender | Nominal | Male-Female |  |  | Male |
| 13 | Hourly-Rate | Wages per hour | Ordinal | 30-100$ | 65.9 | 66 |  |
| 14 | Job Involment | Productivity at work | Ordinal | 1-4 |  |  |  |
| 15 | Job Level | Work level | Ordinal | 1-5 |  |  |  |
| 16 | Job Role | Job roles | Nominal | Sale Executive, Research Scientist, Laboratory Technician,  Manufacturing Director, Healthcare Respentative |  |  | Sale Executive |
| 17 | Job Satisfaction | Job satisfaction | Ordinal | 1-4 |  |  |  |
| 18 | Marital Status | Marital status | Nominal | Married,  Single, Divorced |  |  | Married |
| 19 | Monthly Income | Income per month | Ordinal | 1009 – 20000$ | 6500 | 4930 |  |
| 20 | Monthly Rate | Monthly salary | Ordinal | 2094-27000$ | 143000 | 142000 |  |
| 21 | NumCompaniesWorked | The number of companies that have done it. | Ordinal | 0-9 | 2.69 | 2 |  |
| 22 | Over18 | Over 18 years old | Nominal | Y-N |  |  | And |
| 23 | OverTime | Work more | Nominal | Yes-No |  |  | Yes |
| 24 | Percent Salary Hike | Increase salary as a percentage | Ordinal | 11-25 | 15.2 | 14 |  |
| 25 | Performance Rating | Assess the performance | Ordinal | 3-4 |  |  |  |
| 26 | Relationship Satisfaction | Satisfaction with relationships. | Ordinal | 1-4 |  |  |  |
| 27 | Standard Hours | Standard working hours | Ordinal | 80 |  |  |  |
| 28 | Stock Option Level | Stock options level | Ordinal | 0-3 |  |  |  |
| 29 | Total Working Years | Total number of years worked | Ordinal | 0-40 |  |  |  |
| 30 | Training Times Last Year | The number of trainings last year. | Ordinal | 0-6 |  |  |  |
| 31 | Work Life Balance | The balance between work and life. | Ordinal | 1-4 |  |  |  |
| 32 | Years At Company | The number of years I've been at the company. | Ordinal | 0-40 |  |  |  |
| 33 | Years In Current Role | The number of years I have worked in my current position. | Ordinal | 0-18 |  |  |  |
| 34 | Years Since last promotion | Number of years since the last promotion | Ordinal | 0-15 |  |  |  |
| 35 | Years With Current Manager | Number of years working with current manager | Ordinal | 0-17 |  |  |  |

### 1.6 Classification number

* Classification properties: Attrition, Business Travel, Department, Education, Education Field, Environment Satisfaction, Gender, Job Involvement, Job Level, Job Role, Job Satisfaction, Marital Status, Over18, OverTime, Performance Rating, Relationship Satisfaction, Stock Option Level, Work Life Balance.

## Data processing

* **Purpose:**
* Data transformation
* Data capture
* Data review

#### **2.1 Import Library**

Graphical user interface, text, application, email

Description automatically generated

*Picture 2.2.1. Libraries are needed.*

#### **2.2 Enter data**

A picture containing text

Description automatically generated

*Picture 2.2.2. Use pandas library to enter CSV data*

#### **2.3 Data description**

Graphical user interface

Description automatically generated with low confidence

*Picture 2.2.3. Data description for the entire property*

#### **2.4 Check data type**

Table

Description automatically generated

*Picture 2.2.4. Check the data type for the entire property*

#### **2.5 Overview of data**

* Data is a combination of chain values and integers.
* The data does not contain null values.
* The Over18 column contains only 1 value, which can be deleted.
* The Attrition column is a data column for prediction so must be converted from string to number.

#### **2.6 Thực hiện xóa cột Over18**

Graphical user interface

Description automatically generated with medium confidence

*Picture 2.2.6. Delete the Over18 column*

#### **2.7 Description of information about string data**

Graphical user interface, text, application, chat or text message, email

Description automatically generated

*Figure 2.* *2.7. String type data information*

#### **2.8 Description of numeric data**

Table

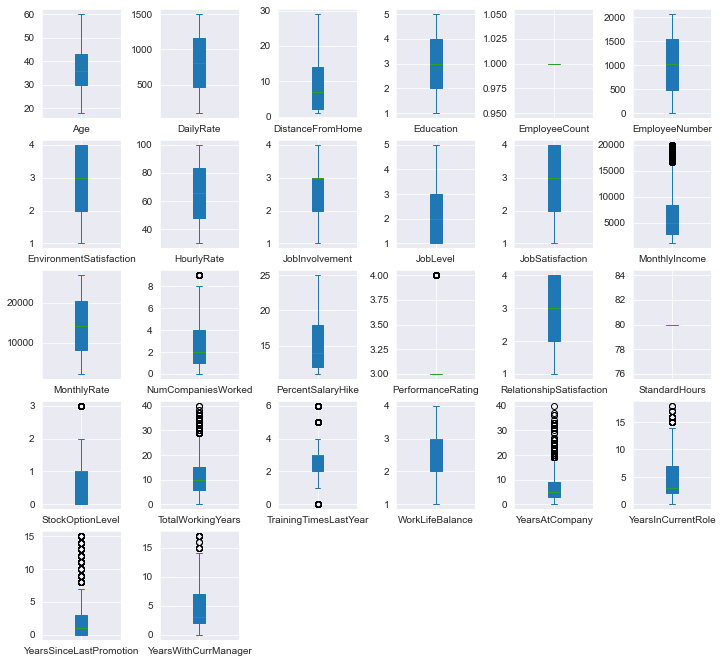
Description automatically generated

*Figure 2.* *2.8. String type data information*

#### **2.9 Test Outliers**

A picture containing chart

Description automatically generated



*Hình 2.2.*9*. Outliers của các cột thuộc tính*

* **Outlier review:**

Most columns have Outliers, but there are no illogical outliers.

**➜** No need to handle Outliers.

#### **2.10 Change the value of the Attrition column from string to boolean**



*Figure 2.* *2.10. Change the data type to a number for the Attrition column*

### 2.11 Use Label Encoder to convert data columns from object to numeric type

Text

Description automatically generated

*Picture 2.2.11. Change the data type for object-style columns*

Table

Description automatically generated

* 1. **Perform data classification**

Text

Description automatically generated

*Picture 2.2.12. Data classification*

#### **2.12.1 Chi square test (with Categorical Features)**

**Set the hypothesis:**

* Null hypothesis (H0) : output feature does not depend on input feature
* Alternate hypothesis (H1): output feature is dependent on input feature
* If we accept the H0 hypothesis, the label does not depend on the input feature being considered and vice versa.
* Alpha rating = 0.05
* The goal is to get out the features that have an impact on output.

Application

Description automatically generated with low confidence

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

#### **2.12.2 Anova Test (with Continuous Features)**

**Set the hypothesis:**

* Null hypothesis (H0): 2 groups have the same variance
* Alternate hypothesis (H1): 2 groups have different variances
* If we accept the H0 hypothesis, the current input feature does not affect the label and vice versa.
* Alpha rating = 0.05
* The goal is to get out the features that have an impact on output.

Text

Description automatically generated with medium confidence

Graphical user interface, text, application, email

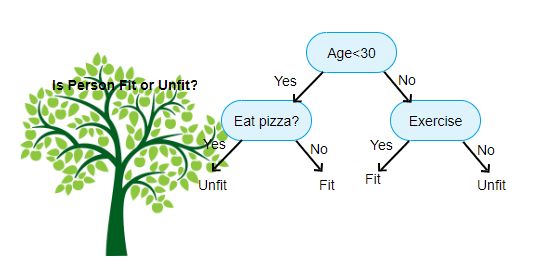
Description automatically generated

# **CHAPTER III: ALGORITHMS AND EXPERIMENTS**

## Algorithms Used

### Decision Tree

* **The Decision Tree is the structure of the tree so that:**
  + Each node in response to a test on a property
  + Each branch represents test results
  + Leaf buttons represent layers or layer distributions
  + The highest button in the tree is the root button.
* **The shape of the tree determines:**



*Figure 3.1.1*

* **Basic strategy:**
  + Start from the single button that represents all templates
  + If the patterns belong to the same class, the button becomes a leaf button and is labeled by that layer.
  + Conversely, using property measurements to select properties will best separate patterns into layers
  + A branch is created for each value of the selected property and the patterns are segmented according to
  + Use the same recursion process to create a decision tree
  + The process ends only if any of the following conditions are correct.
  + All templates for a given button belong to the same class
  + There are no more attributes that the sample can rely on for further planning
  + There are no more templates at the button.

**1.1.1 ID3**

* **Concept: ID3** is an algorithm used in the decision tree. This algorithm uses information gain to build the decision tree**. The largest Information Gain property will be selected as the root button.**
* Information gain:

A picture containing text, watch, clock

Description automatically generated

* The amount of information needed to class an element in S is based on property A:

**InfoA(S)**

Icon

Description automatically generated with medium confidence

* **Information Gain** is the difference between the original Info(S) information value.

(trước phân hoạch) và giá trị thông tin mới InfoA(S) (sau phân hoạch với A):



* Uncertainty (Entropy):
  + Entropy measures the amount of information in a property of a set of values collected by the sample set
  + Entropy is used to determine which nodes are separated next in the algorithm
  + The higher the entropy, the higher the likelihood of layering improvement
  + **Formula:**  
    Text

    Description automatically generated
  + n is the number of different values of property A under consideration
  + Ai is the sample number corresponding to each i value of property A
  + fS(Ai) is the ratio of the number of samples with Ai to S properties

**1.1.2 CART**

* **Concept:** Unlike id3 using the Information Cain formula, the Cart algorithm uses the Gini formula. **The property with the smallest Gini value will be made as a cork.**
* **Gini Stats of Episode S**

A picture containing text, clock, clipart

Description automatically generated

* P(j| S) is the frequency of j in S
* **Gini's attributes**

A picture containing text

Description automatically generated

* ni is the sample number in note I, n is the sample number in note A

**1.2 Random Forest**

* Random Forest algorithm is a machine learning algorithm that can be used to solve both classification and regression problems. It works by building a set of deciding trees during training, then combining the return results of each tree to make a decision that predicts the end of the g-bluntend.

### Logistic Regression

* **Concept:** The logistic regression method is a regression model that predicts the discrete target variable value corresponding to an x input vector. This is equivalent to classifying x inputs into the corresponding y groups.
* Logistic regression analysis is a statistical technique for looking at the connection between independent variables (variables or taxonomic variables) and dependent variables that are binary variables. In single linear regression, the independent variable x and y dependency is the continuous variable associated through the equation:
  + **y = a + bx + e**
  + Where: a is alpha, b is beta, e is epxilon
* In logistic regression, the y-dependent variable has only 2 states 1 and 0. In order to change the variable continuously, one calculates the probability of these two states. If p is called the probability that an event occurs, then 1-p is the probability that the event does not occur.
* The logistics regression equation states:

A picture containing text, watch, clock

Description automatically generated

* We calculate the probability of:

Chart

Description automatically generated

### XGBoost

**Concept:** XGBoost (Extreme Gradient Boosting) is an algorithm based on gradient boosting, but accompanied by tremendous improvements in terms of algorithm optimization, in terms of perfect combination of software and hardware power, helping to achieve superior results both in terms of training time as well as memory usage.

Open source with ~350 contributors and ~3,600 commits on Gihub, XGBoost shows its incredible application capabilities such as:

* XGBoost can be used to solve all problems from regression, classification, ranking, and solving problems defined by users.
* XGBoost supports on Windows, Linux and OS X.
* Supports all major programming languages including C++, Python, R, Java, Scala and Julia.
* Supports AWS, Azure, and Yarn clusters and works well with Flink, Spark, and other ecosystems.
* **Question :**

yIt is a random "output" or "response" variable.

\mathbf{x} = \{x_1, ..., x_n\}Is the random variable "input" or "explanatory".

\{y_i, \mathbf{x}_i\}It's a "training" data template.

F^*(\mathbf{x})is the target function that maps to.\mathbf{x}y

L(y, F(\mathbf{x})) là loss function:

* Squared-error: .(y - F)^2
* Absolute error:  (regression).|y - F|, y \in R^1
* Negative binomial log-likelihood:  (classification).log(1 + e^{-2yF}), y \in \{-1, 1\}

Our goal finds the target function so that the expectations of the error function are minimized.F^*

## Experiments on the Jupyter Notebook

* **Purpose:**
  + Chart, count, and view label ratios to get an overview of an employee's ability to leave the company.
  + Build the decision property, with the decision property being **STATUS.**

### Drop Column is not needed and separates the decision attribute column into a private column



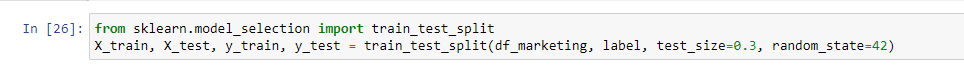
*Picture 3.2.1 Delete the column "StandarHours"*

Graphical user interface, text

Description automatically generated with medium confidence

*Picture 3.2.2 Split the decision property column*

### Separating train and test data (train data accounts for 70%, test accounts for 30%)



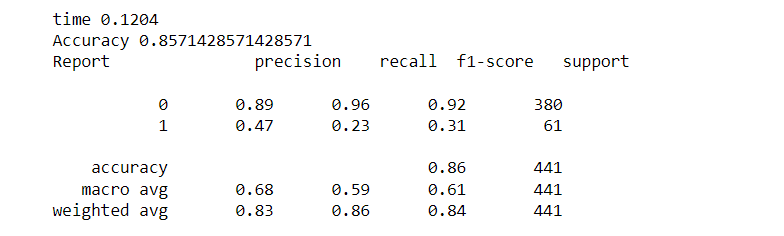
*Picture 3.2.3 Separates data into train and test*

### XGBoost Algorithm

Graphical user interface, text, application

Description automatically generated

*Picture 3.2.4* *XGBoost Algorithm*



*Picture 3.2.5 Results of XGBoost Algorithm model*

* Running time of XGBoost Algorithm: **0.** **1204s**
* Accuracy: **85.** **71%**

Text

Description automatically generated

*Picture 3.2.6* *XGBoost Algorithm*

Graphical user interface, application

Description automatically generated

*Picture 3.2.7 The confused matrix of the XGBoost Algorithm model*

* Through the confused matrix of the XGBoost algorithm model, we know:
  + Precision of the algorithm model: **88.564%**
  + Recall of the algorithm model: **95.78%**
  + F1-Score: **94.913%**

### Random Forest Algorithm

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

*Picture 3.2.8 Random Forest Algorithm and results*

* Random Forest algorithm model accuracy: **86.62%**
* Random Forest algorithm runtime: **0.3709**

Text

Description automatically generated with medium confidence

*Picture 3.2.9 Random Forest Algorithm*

Chart

Description automatically generated

*Picture 3.2.10 Confused Matrix of Random Forest Algorithm model*

* Through the confused matrix of the Random Forest algorithm model, we know:
  + Precision of the algorithm model: **87.037%**
  + Recall of the algorithm model: **98.94%**
  + F1-Score: **92.** **607%**

### Decision Trees Algorithm

#### **2.5.1 ID3 Tree**

A picture containing text

Description automatically generated

*Picture 3.2.10 Decision Tree Algorithm (ID3)*

* Algorithm accuracy: **79.81%**
* Running time of the algorithm: **0.021** **s**

Text

Description automatically generated

*Picture 3.2.11 Decision Tree Algorithm (ID3)*

Graphical user interface, application, Teams

Description automatically generated

*Picture 3.2.12 The Confused Matrix of the Decision Tree Algorithm model (ID3)*

* Through the confused matrix of the Decision Tree (ID3) algorithm model, we know:
  + Precision of the algorithm model: **89.43%**
  + Recall of algorithmic model: **86.84%**
  + F1-Score: **88.11%**

Text

Description automatically generated

Chart, timeline

Description automatically generated

*Picture 3.2.13 Decision Tree (ID3)*

#### **2.5.2 Cart Tree**

A picture containing text

Description automatically generated

*Picture 3.2.14 Decision Tree Algorithm (CART)*

* Algorithm accuracy: **78.46%**
* Running time of algorithm: **0.0289s**

Graphical user interface, text, application

Description automatically generated

Graphical user interface, chart

Description automatically generated

*Picture 3.2.15 The Confused Matrix of the Decision Tree Algorithm Model (CART)*

* Through the confused matrix of the Decision Tree (CART) algorithm model, we know:
  + Precision of the algorithm model: **89.** **256%**
  + Recall of the algorithm model: **85.** **263%**
  + F1-Score: **87.213%**

Text

Description automatically generated

Chart, timeline, treemap chart

Description automatically generated

*Picture 3.2.16 Decision Tree (CART)*

### 2.6 Logistic Regression Algorithm

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated with low confidence

*Picture 3.2.17 Logistic Regression Algorithm*

* Algorithm accuracy: **86.39%**
* Running time of algorithm: **2.0111s**

Text

Description automatically generated

Chart, treemap chart

Description automatically generated

*Picture 3.2.18 The Confused Matrix of Logistic Regression Algorithm Model*

* Through the confused matrix of the Logistic Regression algorithm model, we know:
  + Precision of the algorithm model: **88.647%**
  + Recall of algorithmic model: **96.578%**
  + F1-Score: **92.** **44%**

### 2.7 Comparison, Review

* Use the BarPlot graph to get an overview of runtime and accuracy between algorithms.
* Draw a chart comparing the running time of algorithms

Graphical user interface, text

Description automatically generated

Chart, bar chart

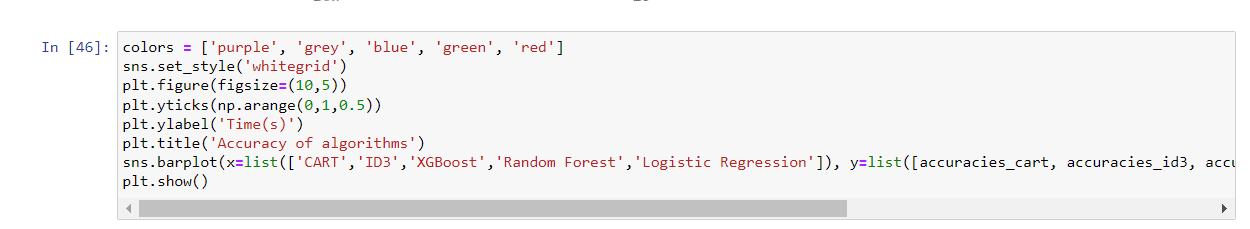
Description automatically generated

Chart

Description automatically generated

*Picture 3.2.19 Chart of running time of Algorithms*

* **Conclusion on the runtime chart:**
  + Decision Tree ID3 is the algorithm with the fastest running time for data sets. With only **0.021s**
  + Logistic Regression algorithm is slowest with **2.0111s**
* Draw a chart comparing the accuracy of algorithms



Chart, bar chart

Description automatically generated

*Picture 3.2.20 Chart of accuracy of Algorithms*

* **Conclusion on the accuracy chart:**
  + Based on the accuracy, we can see that the Random Forest algorithm has the highest accuracy with **86.61%.**
  + The Decision Tree CART algorithm has the lowest accuracy with **78.45%**

# **CHAPTER IV: PREDICTION SOFTWARE**

* 1. **Software overview**

### 1.1 Algorithm used

Based on the results obtained in the previous section, the team decided to use **Random Forest** algorithm for this software. According to the comparison results, this algorithm, although the speed is not too good, but it is strange i for the highest accuracy.

### 1.2 Attributes used to make predictions

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

* After calculating the reliability of the properties we obtained 5 properties: **DailyRate, MonthlyIncome, TotalWorkingYears, MonthlyRate** and **Overtime** to put into the software.

### 1.3 Display and Testing

* **Display:**

Graphical user interface, text, application, email

Description automatically generated

* **Testing:**
  + **Test dataset:**

Application

Description automatically generated with low confidence

* + **Result:**

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

* **We see:** The software results come out in line with the original data.

## Software code

### 2.1 Code display

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

### 2.2 Code processing part

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

# **CHAPTER 5: CONCLUSION**

**1. Advantages and limitations**

**5.1.1 Decision Tree**

* **Advantage:**
* The algorithm is simple, intuitive, not too complicated to understand the first time.
* The training dataset doesn't have to be too large to build an analytical model.
* Some decision tree algorithms are capable of processing missing data and faulty data without applying methods such as "imputing missing values" or removing. Less affected by the exception data.
* There is no need to make initial assumptions about the laws of distribution as in statistics, and as a result the results of the analysis obtained are the most objective, "natural".
* It can help us classify data objects according to multi-layered, multi-class classifications, especially if the target variable is a complex quantitative distortion.
* Can be applied flexibly to target variables, target variables.
* Delivers highly accurate forecast results, easy to implement, fast in training, no need to switch variables.
* Easy to interpret or explain to listeners, viewers who want to understand the results of analysis but have no knowledge of data science.
* Articulate the connection between variables, data attributes in the most intuitive way.
* In addition to economics, finance, decision tree algorithms can be applied in the fields of health, agriculture, biology.
* **Limitations:**
* The decision tree algorithm works effectively on a simple dataset that has few data variables that relate to each other, and vice versa if applied to complex datasets.
* When applied with complex datasets, many different variables and attributes can lead to overfitting patterns, which are too consistent with training data leading to the problem of not giving accurate classification results when applied to test data, and new data.
* The variance value is high, when there is a small change in the dataset can affect the structure of the model.
* The tree algorithm decides to apply only to classification trees if misclassification can lead to serious mistakes.
* The tree algorithm decides whether it is likely to be "biased" or biased if the dataset is not balanced.
* Training and testing datasets must be perfectly prepared, good quality must be balanced in layers, groups in target variables.
* There is no technical "support" or "reverse query" capability.

**5.1.2 Random Forest**

* **Advantage:**
* Improve with the decision tree algorithm, which solves the noise when the dataset changes.
* **Limitations**:
* The main disadvantage of Random Forest is the large volume of calculations, but with the increasing computing capacity of the computer (according to the exponential level), random forest's limitations are not a big problem.

**5.1.3 Logistic Regression**

* **Advantages:**
* Good layering when data is linearly separable.
* Easy to deploy and train.
* **Limitations**:
* Easily affected by noise.
* It is not possible to solve non-linear problems.
* Sensitive to overfitting.

**5.1.4 XGBoost**

* **Advantage:**
* Can be used to solve many problems such as regression, classification.
* Support on a variety of source codes.
* Supports a variety of programming languages.
* Flexibility.
* **Limitations**:
* The computing time of the algorithm.

**2. Development direction**

* Research and learn more about datasets, re-adjust properties and preprocessors for greater accuracy.
* Continue to apply other algorithms of Business Data Analyst such as Linear Regression or Arima model.

# **CHAPTER 6: REFERENCES**

## Reference link

**Reference links in the Business Data Analysis course module**

**Confused Matrix:** [What is Precision, Recall and F1-score? - The Conscious's notes (wordpress.com)](https://caihuuthuc.wordpress.com/2020/02/23/precision-recall-va-f1-score-la-gi/)

**Tìm hiểu thuật toán:** [CART: Classification and Regression Trees for Clean but Powerful Models | by Saul Dobilas | Towards Data Science](https://towardsdatascience.com/cart-classification-and-regression-trees-for-clean-but-powerful-models-cc89e60b7a85)

<https://ongxuanhong.wordpress.com/2017/12/21/xgboost-thuat-toan-gianh-chien-thang-tai-nhieu-cuoc-thi-kaggle>

<https://machinelearningcoban.com/tabml_book/ch_model/random_forest.html>

<https://viblo.asia/p/logistic-regression-bai-toan-co-ban-trong-machine-learning-924lJ4rzKPM>

**Data types:** [What is Ordinal Data? [Definition, Analysis & Examples] (careerfoundry.com)](https://careerfoundry.com/en/blog/data-analytics/what-is-ordinal-data/)

**Machine Learning :** [4 Machine Learning Approaches that Every Data Scientist Should Know | by Orhan G. Yalçın | Towards Data Science](https://towardsdatascience.com/4-machine-learning-approaches-that-every-data-scientist-should-know-e3a9350ec0b9#:~:text=Unsupervised%20learning%20is%20a%20type%20of%20machine%20learning,datasets%2C%20which%20do%20not%20contain%20labels.%20Figure%204.)