**VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY**

**UNIVERSITY OF INFORMATION TECHNOLOGY**

Logo

Description automatically generated

**FINAL PROJECT REPORT**Subject : Data Mining

**TOPIC:**

**PREDICT EMPLOYMENT TERMINATION**

Student:

1. Nguyen Cao Khoa *(Leader)* ID: 19521694
2. Le Dinh Quoc Huy ID: 19521614
3. Bui Ngoc Thanh ID 19522220
4. Tran Minh Tot ID: 19522369

Lecturer: Mrs. Cao Thi Nhan  
 Mr. Nguyen Ho Duy Tri

Class: IS252.N11.TMCL

**Ho Chi Minh city, December 2022**

Contents

[CHAPTER 1 : INTRODUCTION 6](#_Toc121976186)

[**1.1** **Problems** 6](#_Toc121976187)

[**1.2** **Project goal** 7](#_Toc121976188)

[**1.3** **Developer tools & Technology** 7](#_Toc121976189)

[CHAPTER 2: DATA PREPROCESSING 8](#_Toc121976190)

[**2.1.**  **Description of original data** 8](#_Toc121976191)

[**2.1.1** **Statistics of property values:** 8](#_Toc121976192)

[**2.1.2** **Property Statistics Table** 17](#_Toc121976193)

[**2.1.3** **Classification number** 21](#_Toc121976194)

[**2.2**  **Data processing** 21](#_Toc121976195)

[**2.2.1** **Import Library** 21](#_Toc121976196)

[**2.2.2** **Read data** 21](#_Toc121976197)

[**2.2.3** **Check data types of columns** 22](#_Toc121976198)

[**2.2.4** **Check for null value** 22](#_Toc121976199)

[**2.2.5** **Information description of numeric data** 23](#_Toc121976200)

[**2.2.6** **Information description of string data** 23](#_Toc121976201)

[**2.2.7** **Check OUTLIERS of attribute columns** 23](#_Toc121976202)

[**2.2.8** **Data classification** 24](#_Toc121976203)

[**2.2.9** **Use Label Encoder to convert columns from OBJECT to NUMERIC** 24](#_Toc121976204)

[**2.2.10** **Overview of the values of Discrete Data** 25](#_Toc121976205)

[**2.2.11** **Overview of values of Continuous Data** 26](#_Toc121976206)

[CHAPTER 3: VISUALIZATION 27](#_Toc121976207)

[**3.1** **Divide attributes into groups according to common characteristics and separate Attrition decision attributes separately** 27](#_Toc121976208)

[**3.2** **Group 1 - Relationship between employee background information and employee departure or stay** 27](#_Toc121976209)

[**3.2.1** **Age** 27](#_Toc121976210)

[**3.2.2** **Distance to work** 29](#_Toc121976211)

[**3.2.3** **Years of service** 31](#_Toc121976212)

[**3.2.4** **Gender and marital status** 31](#_Toc121976213)

[**3.3** **Group 2 - The relationship between the employee's job information and the employee's departure or stay** 32](#_Toc121976214)

[**3.3.1** **Qualification** 32](#_Toc121976215)

[**3.3.2** **Overtime** 33](#_Toc121976216)

[**3.4** **Group 3 - Employee-company information relationship with employee's departure or stay** 34](#_Toc121976217)

[**3.4.1** **Number of years working since promotion** 34](#_Toc121976218)

[**3.4.2** **Work and life balance** 35](#_Toc121976219)

[**3.5** **Group 4 - The relationship between company information and employees' departure or stay** 36](#_Toc121976220)

[**3.5.1** **Evaluation of satisfaction about the working environment** 36](#_Toc121976221)

[**3.5.2** **Satisfaction assessment of relationships** 37](#_Toc121976222)

[**3.6** **Group 5 - The relationship between finance and employee departure or stay** 38](#_Toc121976223)

[**3.6.1** **Monthly income** 38](#_Toc121976224)

[**3.6.2** **Hourly wage** 39](#_Toc121976225)

[**3.6.3** **Daily wage** 40](#_Toc121976226)

[**3.6.4** **Salary** 41](#_Toc121976227)

[**3.7** **SUMMARY** 42](#_Toc121976228)

[CHAPTER 4: Test for data balance and data balance with SMOTE 44](#_Toc121976229)

[**4.1** **Test for data balance** 44](#_Toc121976230)

[**4.2** **Data balance with SMOTE** 45](#_Toc121976231)

[CHAPTER 5 : ALGORITHMS IMPLEMENTATION 47](#_Toc121976232)

[**5.1** **Algorithm used** 47](#_Toc121976233)

[**5.1.1** **K-nearest Neighbors** 47](#_Toc121976234)

[**5.1.2** **Random Forest** 48](#_Toc121976235)

[**5.1.3** **Decision Tree** 48](#_Toc121976236)

[**5.1.4** **Naive Bayes** 51](#_Toc121976237)

[**5.2** **ALGORITHMS IMPLEMENTATION** 54](#_Toc121976238)

[**5.2.1** **Separate train and test data (train data=70%, test=30%) Separate train and test data (train data=70%, test=30%)** 54](#_Toc121976239)

[**5.2.2** **K-Nearest Neighbor (KNN)** 55](#_Toc121976240)

[**5.2.3** **Random Forest** 57](#_Toc121976241)

[**5.2.4** **Decision Trees (ID3)** 59](#_Toc121976242)

[**5.2.5** **Decision Trees (CART)** 61](#_Toc121976243)

[**5.2.6** **Naive Bayes** 64](#_Toc121976244)

[**5.2.7** **Logistic Regression** 66](#_Toc121976245)

[CHAPTER 6 : COMPARE ALGORITHMS 68](#_Toc121976246)

[**6.1 Algorithms run time** 68](#_Toc121976247)

[**6.2** **Algorithms Accuracy** 69](#_Toc121976248)

[CHAPTER 7 : BUILD FORECAST SOFTWARE 71](#_Toc121976249)

[**7.1** **Find the attribute with high confidence to choose as the predictor attribute** 71](#_Toc121976250)

[**7.2** **Software code** 73](#_Toc121976251)

[**7.2.1** **Code display** 73](#_Toc121976252)

[**7.2.2** **Code processing part** 76](#_Toc121976253)

[**7.2.3** **Software Interface** 77](#_Toc121976254)

[CHAPTER 8 : CONCLUSION 78](#_Toc121976255)

[**8.1** **Advantages and limitations of each algorithm** 78](#_Toc121976256)

[**8.1.1** **Decision Tree** 78](#_Toc121976257)

[**8.1.2** **Random Forest** 79](#_Toc121976258)

[**8.1.3**  **Logistic Regression** 80](#_Toc121976259)

[**8.1.4** **K-nearest Neighbors** 80](#_Toc121976260)

[**8.1.5**  **Naive Bayes** 80](#_Toc121976261)

[**8.2** **Direction of Development** 81](#_Toc121976262)

[CHAPTER 9 : REFERENCES 82](#_Toc121976263)

**GRATITUDE**

In fact, there is no success that is not tied to the support and help, whether more or less, directly or indirectly from others. With the deepest gratitude, we would love to say thank to lecturers in University of Information Technology, especially lecturers of Faculty of Information Systems for giving us much basis knowledge, practical skills and facilitates so that we can complete our project.

         In particular, our team would like to express our sincere thanks to Mrs. Cao Thi Nhan - theoretical lecturer and Mr. Nguyen Ho Duy Tri - practical lecturer of Data Mining who wholeheartedly helped, directly instructed and guided the group throughout the process a project.

           Through a semester doing this project, group of authors have manipulated every basis knowledge and combined with learning, researching new knowledges. Thence, group of authors have manipulated as much as possible everything which were collected to complete the project in a best way. However, during the time doing this project, group of authors hardly avoid getting mistakes. So, we would love to be gained every feedback from lectures to help completing our knowledge then we could use it like a preparation to continue doing any difference project in the future.

Many thanks!

**LECTURER’S COMMENT**

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

# **CHAPTER 1 : INTRODUCTION**

* 1. **Problems**

In an organization, company or enterprise, in the process of operation, they cannot maintain their labor resources and employees forever. Currently, our country is in the stage of industrialization, modernization, renewal of economic model and structure, improvement of labor productivity and competitiveness of enterprises and organizations; making great contributions to improving the country's economic potential. In which, workers play an important role and are the core to bring the innovation process into depth and promote industry. So, having a solid source of labor is extremely necessary because the termination of the contract is inevitable and the consequences of the termination of the labor contract are related to the problem of labor law. economy and society because each person, regardless of position, and different units are important subjects in creating and promoting the advantages of the organization. If the organization grows, it will help the economy better. Therefore, when someone terminates the contract at a certain position, it is necessary to have another person to replace that position, especially in important positions in the organization, finding the right person in a short time is essential and difficult.

Faced with this problem, our team has come up with a solution to apply information technology and specifically in the field of Data Mining, which is taken from the Employment Termination data source, from which it is possible to create machines. to make predictions about the employees who may terminate the contract, thereby helping the heads to have more insight into issues in the company such as working environment, position of each employee, about the departure of employees, thereby making decisions, future orientations as well as earlier preparation of their human resources. Supplying demand for loans for economic development.

* 1. **Project goal**

Modeling the business costs of maintaining data sources, thereby generating predictions about whether the company is understaffed or not. From there, it helps leaders have more perspectives on issues in the company such as the working environment, salary and bonus or whether they are forcing employees to work too hard, thereby making decisions and orientations in the future. as well as an earlier change in the way the company operates to ensure its full and best human resources.

* 1. **Developer tools & Technology**

In the process of implementation, the group used a number of software for researching and developing the topic:

* Information collection and analysis using the python library and programming language
* Data sources: [Modeling the Business Cost of Retention](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 this study.

**CHAPTER 2: DATA PREPROCESSING**

## **2.1.** **Description of original data**

The dataset is about employee attrition based on attributes like: age, hours worked per day, department worked, education, gender, etc.

Includes 1470 lines, 35 properties .

1. **Statistics of property values:**

* Symbols: **#** (Number), **✓** (Logical Value), **A** (Letters)
* Source: [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. **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. **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.

## **2.2 Data processing**

* **Purpose:**
* Data transformation
* Data capture
* Data review
  + 1. **Import Library**

Text

Description automatically generated

* + 1. **Read data**

A screenshot of a computer

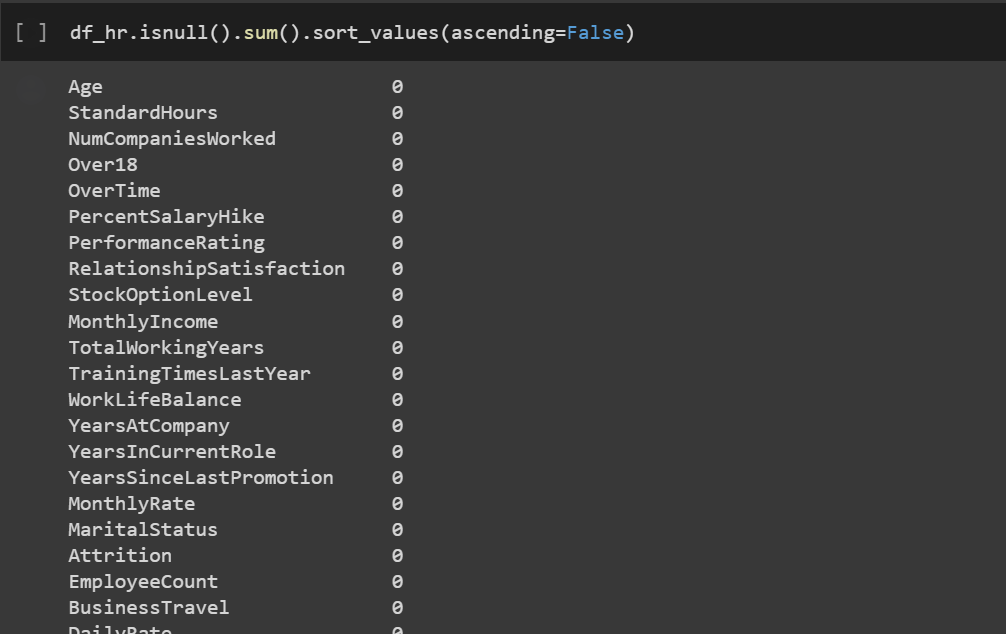
Description automatically generated with medium confidence

* + 1. **Check data types of columns**

Text

Description automatically generated

* + 1. **Check for null value**



* + 1. **Information description of numeric data**

A picture containing text

Description automatically generated

* + 1. **Information description of string data**

Text

Description automatically generated

* + 1. **Check OUTLIERS of attribute columns**

Graphical user interface, application

Description automatically generated

* The columns all have OUTLIERS, but there are no illogical OUTLIERS

=> No need to handle the OUTLIERS

* + 1. **Data classification**

Graphical user interface

Description automatically generated

* + 1. **Use Label Encoder to convert columns from OBJECT to NUMERIC**

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

* + 1. **Overview of the values of Discrete Data**

A picture containing diagram

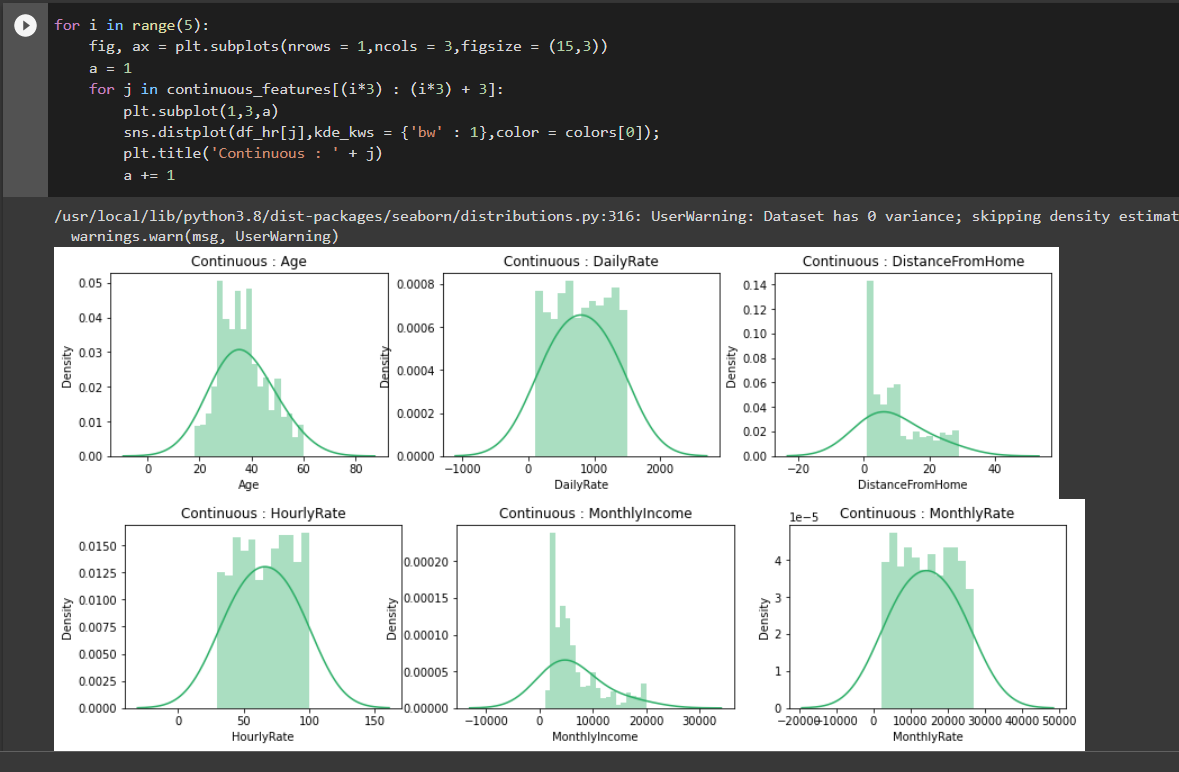
Description automatically generated

* Columns Over18 and EmployeeCount have only one value
  + - Delete column
    - The EmployeeNumber column is just the ordinal numbers of the employees so they are not necessary
    - Delete column

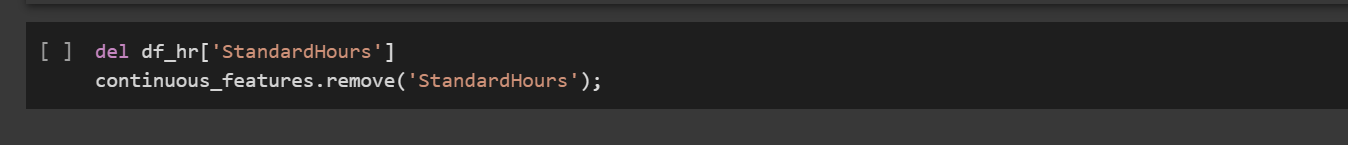
Text

Description automatically generated

* + 1. **Overview of values of Continuous Data**



* The StandardHours column has only one value
  + Delete column



**CHAPTER 3: VISUALIZATION**

* 1. **Divide attributes into groups according to common characteristics and separate Attrition decision attributes separately**

**Text

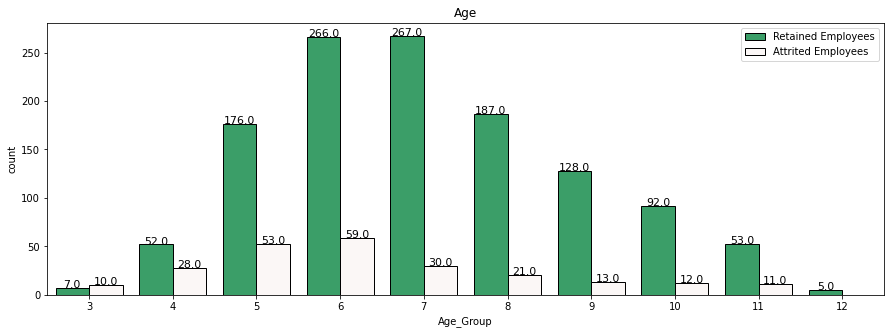
Description automatically generated**

**3.2 Group 1 - Relationship between employee background information and employee departure or stay**

**3.2.1 Age**

Text

Description automatically generated

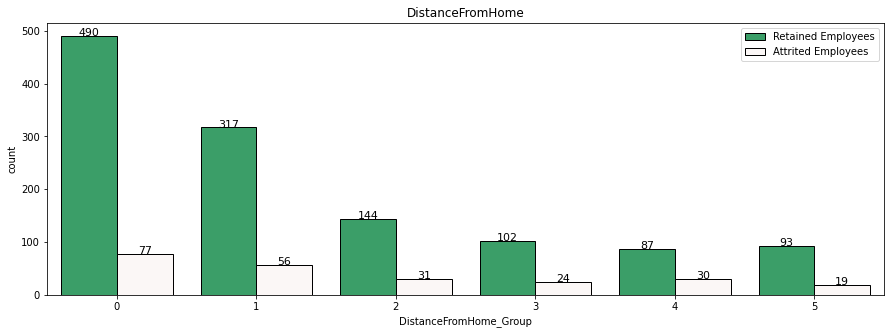


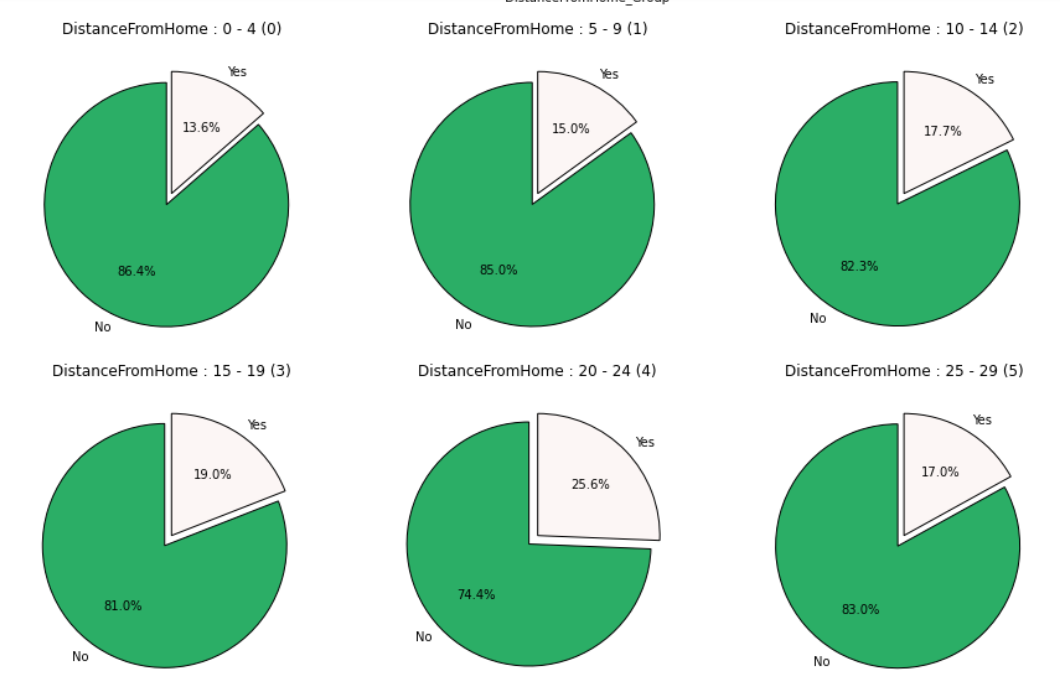
COMMENT

* Attrition occurs in most age groups.
* For the Age values 30 - 34, the highest number of employees leaving was 59.
* Employees aged 25 - 29 ranked second with 53 employees leaving the company.
* Age values 20 - 24 & 35 - 40 are roughly equal in number to 28 & 30.
* Employees over the age of 40 may be those who have been dismissed.
  + 1. **Distance to work**

Text

Description automatically generated



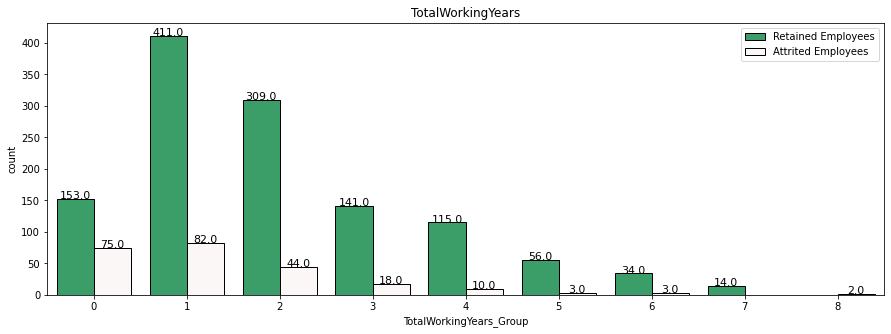


COMMENT:

* Looking at the bar chart, it can be seen that the employees in the group closest to the company (group 0) have the most number of employees leaving.
* However, when checking the attrition percentage, we can see that employees living in a distance from 0 to 4 have less attrition.
  + As the value of DistanceFromHome increases, so does the employee attrition rate!
    1. **Years of service**

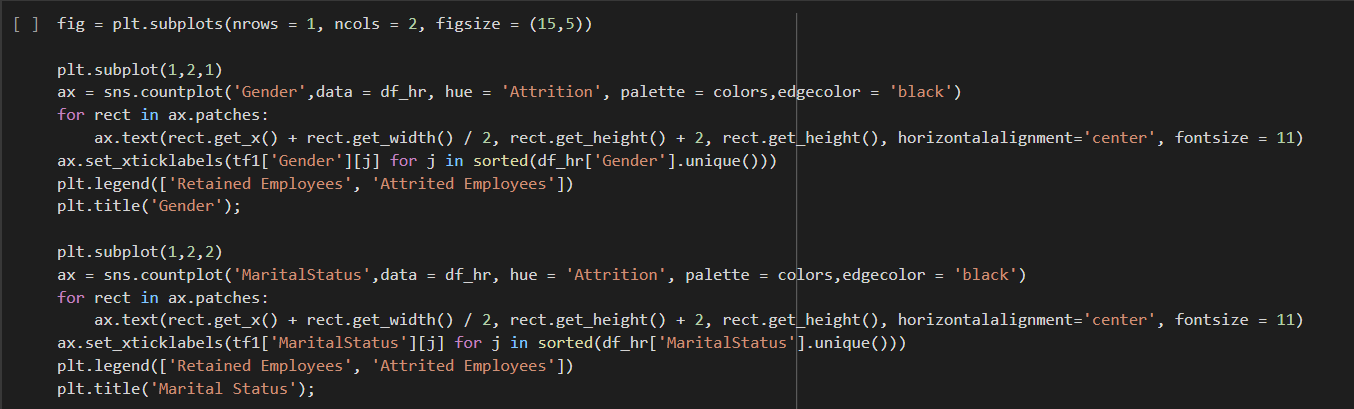
Text

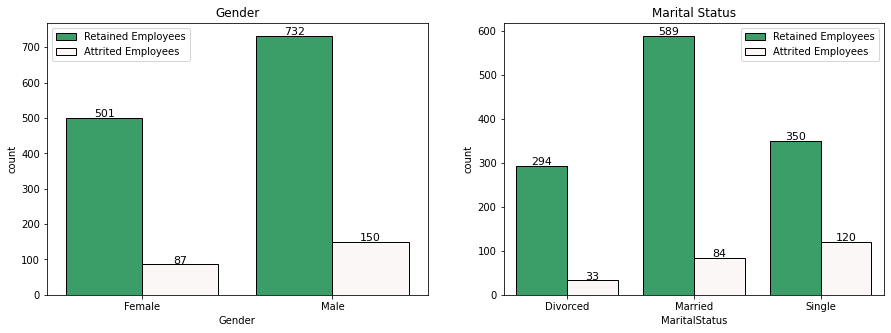
Description automatically generated



COMMENT:

* We can say that employees within the first 10 years of work experience are very likely to be fired!
* As work experience increases, the chance of attrition of manpower decreases!
  + 1. **Gender and marital status**



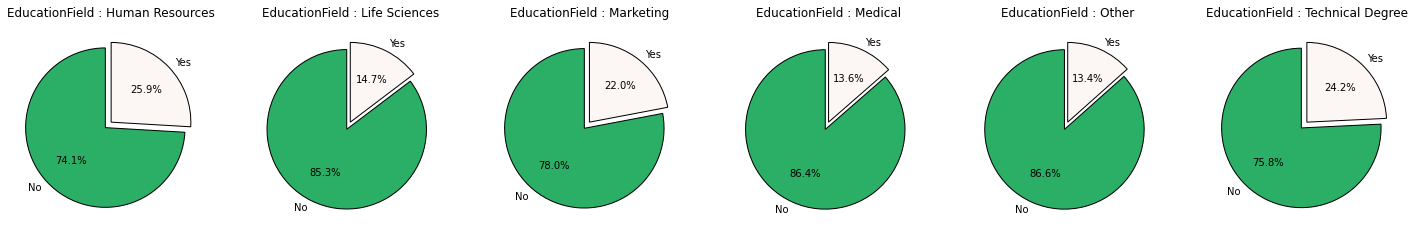


COMMENT:

* The number of male employees being laid off is higher than the number of female employees.
* Single employees have the most attrition.
* Married employees take 2nd place.
* Divorced in last place.
  1. **Group 2 - The relationship between the employee's job information and the employee's departure or stay**
     1. **Qualification**

A screenshot of a computer

Description automatically generated with medium confidence

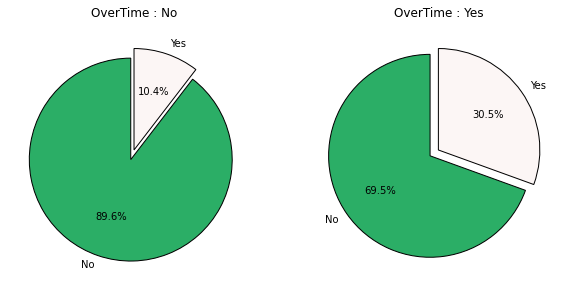


COMMENT:

* We can see that employees whose Education Field is Human Resources, Technical Degree & Marketing have a higher chance of being eliminated.
  + 1. **Overtime**

A picture containing graphical user interface

Description automatically generated

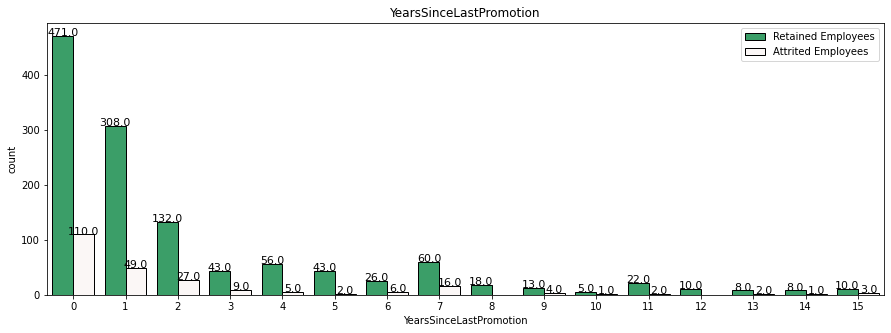


COMMENT

* We can see that people who work overtime are more likely to leave the company due to stress!
* The turnover rate is 30%, which is 3 times higher than employees who do not work overtime.
  1. **Group 3 - Employee-company information relationship with employee's departure or stay**
     1. **Number of years working since promotion**

A screenshot of a computer screen

Description automatically generated

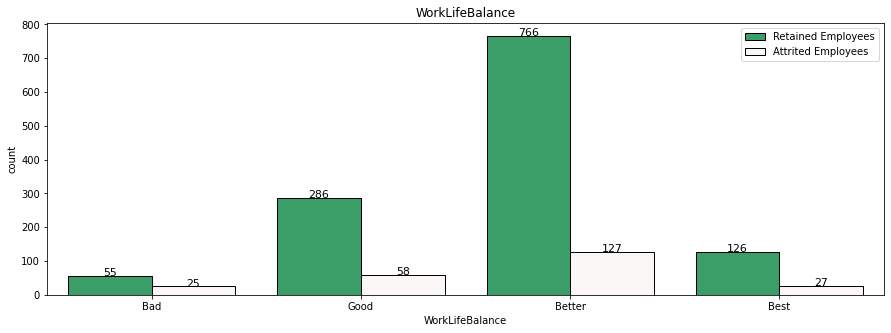


COMMENT:

* We can see that there are a large number of attrition cases for the year value 0.
* That could be the majority value that represents newcomers to the company.
* 1 & 2 years since the last promotion also recorded a significant number of cases of employees leaving.
* In the 7 years since the last promotion, there have been quite a few staff departures.
  + 1. **Work and life balance**

Text

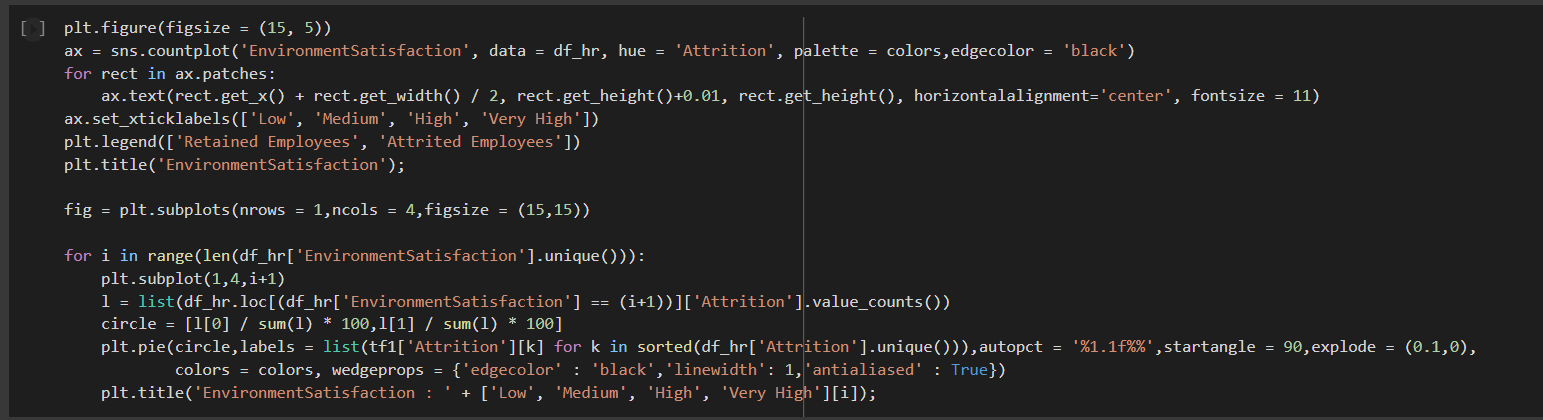
Description automatically generated

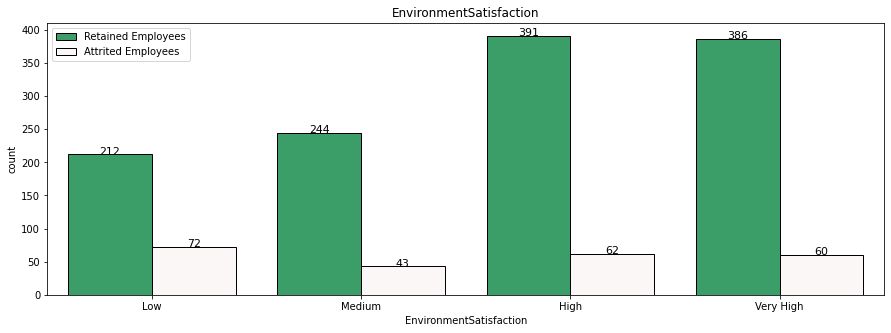
 A picture containing graphical user interface

Description automatically generated

COMMENT:

* As expected, the bad WorkLifeBalance resulted in a massive attrition rate of 31.2%.
* Surprisingly Best WorkLifeBalance has the second highest attrition percentage value.
  1. **Group 4 - The relationship between company information and employees' departure or stay**
     1. **Evaluation of satisfaction about the working environment**



 A picture containing application

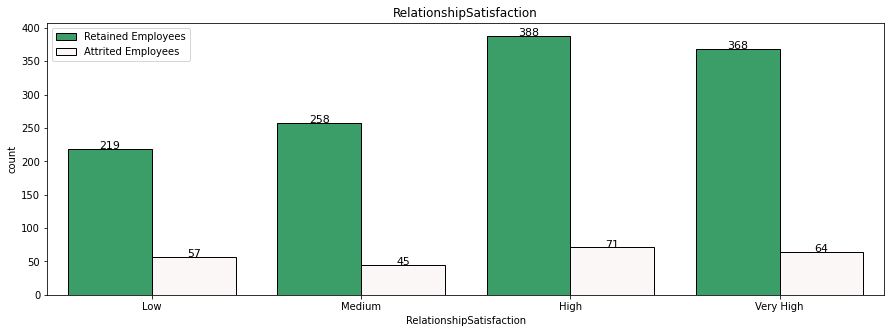
Description automatically generated

COMMENT:

* The High & Very High Environmental Satisfaction values were recorded the most times.
* As expected, they have low attrition rates compared to Low and Medium Satisfaction Levels.
* Attrition rate improves as Environmental Satisfaction improves!
  + 1. **Satisfaction assessment of relationships**

Text

Description automatically generated

 A picture containing application

Description automatically generated

COMMENT

* The chart above for Relationship Satisfaction is very similar to Workplace Satisfaction.
* As Relationship Satisfaction values improve, the attrition rate decreases.
  1. **Group 5 - The relationship between finance and employee departure or stay**
     1. **Monthly income**

A picture containing graphical user interface

Description automatically generated

Chart, line chart

Description automatically generated

Chart

Description automatically generated with low confidence

COMMENT:

* The graph highlights the overall decline in the number of values.
* Monthly Earnings values between 1000 - 2000 are present in large numbers.
* Values in the range 3000 - 4000 are second with more than 200 values included in this range.
  + 1. **Hourly wage**

A picture containing timeline

Description automatically generated

Chart, line chart

Description automatically generated

Chart, bar chart

Description automatically generated

COMMENT:

* For HourlyRate, values from 30 - 100 appear for a total of more than 175+ values each.
* The attrition rates of these values are also low and very close to each other.
* For HourlyRate greater than 100, there is very little value and therefore the consumption is also high. For HourlyRate, values from 30 - 100 appear for a total of more than 175+ values each.
* The attrition rates of these values are also low and very close to each other.
* For HourlyRate greater than 100, there is very little value and therefore the consumption is also high.
  + 1. **Daily wage**

Text

Description automatically generated

Chart, line chart

Description automatically generated

Chart, bar chart

Description automatically generated

COMMENT:

* The number of employees consumed is close to each other. There is a certain decrease in the number of values.
* Values from 600 - 700 have the lowest count.
  + 1. **Salary**

A picture containing graphical user interface

Description automatically generated

Chart, line chart

Description automatically generated

Chart, bar chart

Description automatically generated

COMMENT:

* The values of human attrition are very close. Values between 21000 - 22000 have the highest count.
  1. **SUMMARY**
* Group 1 - Basic information of employees:
* Age : 20 - 44
* Gender: Male > Female
* MaritalStatus : Single > Married > Divorced
* DistanceFromHome : 20 - 24 > 15 - 19 > 25 - 29 > 10 - 14 > 5 - 9
* TotalWorkingYears : Very high opportunity in the first 10 years of working
* Group 2 - Job information:
* EducationField : Human Resources > Technical Degree > Marketing > Life Sciences > Medical > Other
* OverTime : Yes > No
* Group 3 - Employee information - job:
* YearsSinceLastPromotion : 0 > 1 > 2.
* WorkLifeBalance : Bad > Best > Good > Better
* Group 4 - Company information:
* EnvironmentSatisfaction : Low > Medium > High > Very High
* RelationshipSatisfaction : Low > High > Medium > Very High
* Group 5 - Financial information:
* MonthlyIncome : 2000 - 3000
* HourlyRate : 50 - 60. The values are very close together.
* DailyRate : 300 - 400. The values are very close together.
* MonthlyRate : Very close and small vertices are present.

**CHAPTER 4: Test for data balance and data balance with SMOTE**

* 1. **Test for data balance**

Text

Description automatically generated

Shape

Description automatically generated

The dataset is not balanced with a 5.2:1 bias towards Employee Remaining: Employee Leaving.

* + balance dataset using SMOTE Analysis!
  1. **Data balance with SMOTE**

Graphical user interface, text, application

Description automatically generated

* Balance the data by increasing the minority to the majority.
* In this case is to increase the minority data points because the data is very few.
* For unbalanced datasets, data will be duplicated to deal with potential bias in the predictions.

Dataframe has decision attribute (Attrition)

A screenshot of a computer

Description automatically generated with medium confidence

The decision attribute has been separated

A picture containing background pattern

Description automatically generated

# **CHAPTER 5 : ALGORITHMS IMPLEMENTATION**

* 1. **Algorithm used**

**5.1.1 K-nearest Neighbors**

* K-nearest Neighbors is one of the simplest supervised learning algorithms widely used in data mining. The idea of ​​this algorithm is that it doesn't learn anything from the learning dataset (so KNN is classified as lazy learning), all computation is done when it needs to predict the label of new data.
* The class (label) of a new data object can be predicted from the classes (labels) of its k nearest neighbors.
* Given a training dataset D with classes, classify X into classes based on the k elements that are most similar to X (voting rule: majority vote).
* Implementation steps:
* We have D as a set of labeled data points and A as unclassified data.
* Measure the distance (Euclidian, Manhattan, Minkowski, Minkowski or Weight) from new data A to all other data classified in D.
* Choose K (K is the parameter that you define) the smallest distance.
* Check the list of classes with the shortest distance and count the number of each class appearing.
* Get the correct class (the class that appears most times).
* The class of the new data is the layer received in step 5.
* Dependent:
* Distance measure to determine similarity.
* Value k, number of neighbors => k<=|D|1/2
* Euclidean measure:

A picture containing text, watch, gauge

Description automatically generated

* + 1. **Random Forest**

The 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 decision trees during training, then combining the returned results of each tree to make the final prediction decision.

* + 1. **Decision Tree**
* A decision tree is a tree structure such that:
* Each node in the network corresponds to a test on an attribute
* Each branch represents the test result
* Leaf nodes represent classes or class distributions
* The highest node in the tree is the root node. Decision tree shape:
* Basic strategy:
* Start from single node showing all samples
* If the samples belong to the same class, the node becomes a leaf node and is labeled with that class
* In contrast, using the attribute measure to select the attribute will best separate the samples into classes
* A branch is created for each value of the selected attribute and the samples are partitioned by use the same process recursively to create a decision tree
* The process ends only if any of the following conditions are true
* All templates for a given node belong to the same class.
* There are no more attributes that the sample can rely on for further partitioning.
* No samples left at node
* ID3 is an algorithm used in decision trees. This algorithm uses information gain to build a decision tree. The largest Information Gain attribute will be selected as the root node.
* A picture containing text, watch, clock

  Description automatically generatedInformation Gain:
* Amount of information needed to classify an element in S based on attribute A: InfoA(S)

Icon

Description automatically generated with medium confidence

* Information gain is the difference between the original Info(S) information value (before partitioning) and the new InfoA(S) information value (after partitioning with A).
* Entropy:
* Entropy measures the amount of information in an attribute of a collection of sample set values.
* Entropy is used to determine which node is split next in the algorithm.
* The higher the entropy, the better the classification improvement.
* Formula:

Text

Description automatically generated

𝑛 is the number of different values of the attribute A under consideration

𝐴i is the number of samples corresponding to each value of the attribute

𝑓S(𝐴i) is the ratio of the number of samples with attribute to S

CART: Unlike ID3 which uses Information Cain formula, Cart algorithmuses Gini formula. The attribute with the smallest Gini value will be the root node.

* Gini index of the set S:A picture containing text, clock, clipart

  Description automatically generated

P(j|S) is the frequency of j in S

* Gini of attribute:

A picture containing text

Description automatically generated

In case: ni is the number of samples in note I, n is the number of samples in note A

* + 1. **Naive Bayes**

1. **Bayes theorem**

* Bayes' Theorem (Bayes' Theorem) is a mathematical theorem that calculates the probability of a random event A, given that the related event B has occurred.
* This theorem is named after the 18th century English mathematician Thomas Bayes.
* This is one of the extremely useful tools, a close friend of Data Scientists who work in data science.
* Bayes theorem allows to calculate the probability of a random event A given that related event B has occurred. This probability is denoted P(A|B), and read as “the probability of A if there is B”. This quantity is called conditional probability or posterior probability because it is derived from a given value of B or depends on that value.
* According to Bayes' theorem, the probability that A occurs when B is known will depend on 3 factors:
* The probability that A occurs on its own, regardless of B. It is denoted by P(A) and read as the probability of A. This is called the marginal probability or a priori probability, it is “a priori” ” in the sense that it is not interested in any information about B.
* Probability of occurring B on its own, regardless of A. It is denoted by P(B) and read as “probability of B”. This quantity is also called a normalizing constant, because it is always the same, regardless of the event A is trying to know.
* Probability of B happening when A is known. It is denoted by P(B|A) and read as “probability of B if there is A”. This quantity is called the likelihood that B will occur, given that A has occurred. Pay attention not to confuse the probability that B will occur when A is known and the probability that A will occur when B is known.
* We can restate it with the following formula:Nếu A và B là 2 sự kiện độc lập, The probability that A and B occur at the same time is:

P(A,B) = P(A) P(B)

* In case:
* P(A)P(A) is the probability of a distinct A occurring.
* P(B)P(B) is the probability that B occurs separately.
* Text

  Description automatically generatedIf A and B are two related events, and the probability that event B occurs is greater than 0, we can define the probability that A will occur, given that B occurs as follows:
* A picture containing text

  Description automatically generatedBayes' theorem is based on the definition of conditional probability above, expressed in the form of a formula as follows:

The symbol ¬A is not A (or A's complement). We have P(A)+P(¬A) = 1

From there: P(B) =P(B,A) + P(B,¬A) = P(B∣A)P(A) + P(B∣¬A)P(¬A)

Bayes' theorem is written in variant form as follows:

A picture containing shape

Description automatically generated

1. **Naive Bayes classification algorithm**

Naive Bayes Classification (NBC) is a classification algorithm based on probability calculation applying Bayes theorem. This algorithm belongs to the group of supervised learning algorithms.

Each data sample is represented by X=(x1, x2,..., xn) with attributes A1, A2,..., An

Grades C1, C2, ..., Cm. Given an unknown sample X

Subclassing Naive Bayes will determine that X belongs to class Ci if andonly if:

Logo, company name

Description automatically generated with medium confidence

A picture containing text

Description automatically generated

* According to Bayes' theorem:

Since P(X) is constant for all classes, only the maximum P(X|Ci) x P(Ci) is needed. If P(Ci) is not known, we need to assume P(C1)=P(C2)=...= P(Cm) and we will maximize P(X|Ci). Otherwise, we maximize P(X|Ci) x P(Ci)

However, the problem of calculating P(X|Ci) is impossible!

Diagram

Description automatically generatedAdmit Naive: assume attribute independence

It is possible to approximate P(x1|Ci), ..., P(xn|Ci) from the training samples.

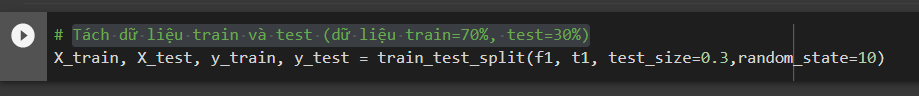
If Ak is a qualitative attribute, then P(xk|Ci) = sik/si where sik is the number of training samples of Ci with the value xk for Ak and si is the number of samples belonging to class Ci

If Ak is continuous, then it is assumed to have a Gaussian distribution:

Text

Description automatically generated

* 1. **ALGORITHMS IMPLEMENTATION**
     1. **Separate train and test data (train data=70%, test=30%) Separate train and test data (train data=70%, test=30%)**



* + 1. **K-Nearest Neighbor (KNN)**

Text

Description automatically generated

Calendar

Description automatically generated

* Running time of KNN algorithm: 0.7233s
* Accuracy: 80.146%
* Continue to implement the KNN algorithm
* Draw the confusion matrix

Text

Description automatically generated

Chart, treemap chart

Description automatically generated

* Through the confusion matrix of the KNN algorithm Picture model, we know
* The sensitivity (Precision) of the algorithm Picture model: 61.3138%
* Specificity (Recall) of algorithm Picture model: 82.0846%
* F1-Score: 70.1949%
  + 1. **Random Forest**

Implement the Random Forest algorithm

Text

Description automatically generated

* Random Forest algorithm picture tissue accuracy : 90.9489%
* Random Forest algorithm runtime: 0.3751s

Draw the confusion matrix for the Random Forest algorithm model

Text

Description automatically generated

Chart, treemap chart

Description automatically generated

* Through the confused matrix of random forest algorithm picture tissue, we know
* Precision of the algorithmic picture tissue : 92.683%
* Recall of algorithmic Picture tissue : 86.645%
* F1-Score: 89.5623%
  + 1. **Decision Trees (ID3)**

Implement Decision Trees Algorithm (ID3)

Text

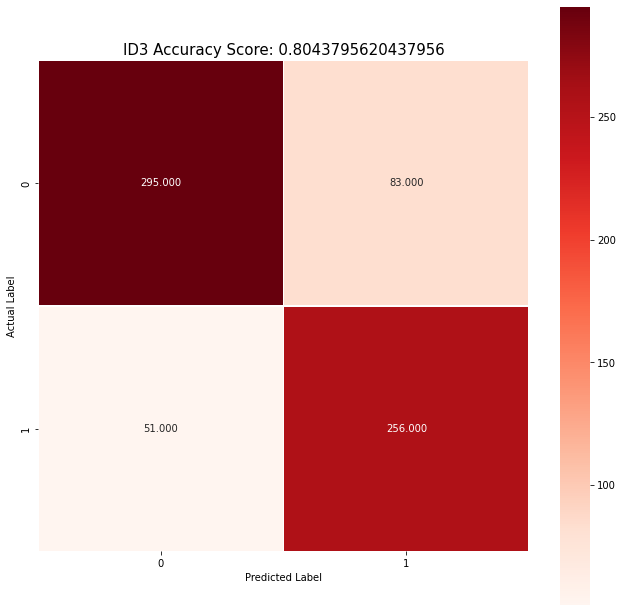
Description automatically generated

* Algorithm accuracy: 80.4379%
* Running time of the algorithm: 0.0285s

Draw the confusion matrix for the Decision Tree algorithm (ID3)

Text

Description automatically generated



* Through the confused matrix of the Picture model algorithm Decision Tree (ID3), we know
* Precision of the algorithmic Picture tissue: 75.5162%
* Recall of algorithmic Picture tissue : 83.3876%
* F1-Score: 79.2569%

Draw Decision Tree (ID3)

Text

Description automatically generated

Graphical user interface, chart, treemap chart

Description automatically generated

* + 1. **Decision Trees (CART)**

Implement Decision Trees algorithm (CART)

Text

Description automatically generated

* Algorithm accuracy: 80.4379%
* Algorithm runtime: 0.0271s

Draw the confusion matrix for the Decision Tree algorithm (CART)

A screenshot of a computer

Description automatically generated

Graphical user interface, application

Description automatically generated

* Through the confused matrix of the Decision Tree algorithm picture tissue (CART), we know
* Precision of the algorithmic Picture tissue: 76.7802%
* Recall of algorithmic Picture tissue : 80.7817%
* F1-Score: 78.7301%

Draw Decision Tree (CART)

Graphical user interface, text

Description automatically generated

Chart, treemap chart

Description automatically generated

* + 1. **Naive Bayes**

Implement Naive Bayes algorithm

Text

Description automatically generated with medium confidence

* Algorithm accuracy: 73,2847%
* Running time of algorithm: 0.0084s

Draw the confusion matrix for the Naive Bayes algorithm  
Text

Description automatically generated

Chart, treemap chart

Description automatically generated

* Through the confused matrix of the Naive Bayes algorithmic Picture tissue, we learn
* Precision of the algorithmic Picture tissue: 66.0622%
* Recall of algorithmic Picture tissue : 83.0618%
* F1-Score: 73.5931%
  + 1. **Logistic Regression**

Import library

Text

Description automatically generated

Check algorithm execution time

Text

Description automatically generated

Implement Logistic Regression algorithm

**Text

Description automatically generated**

* Algorithm accuracy: 77.9562%
* Running time of algorithm: 2.3619s

Draw the confusion matrix for the Logistic Regression algorithm

A screenshot of a computer

Description automatically generated

Graphical user interface, application

Description automatically generated with medium confidence

* Through the confused matrix of the Logistic Regression algorithmic Picture tissue, we learn
* Precision of the algorithmic Picture tissue: 74.3902%
* Recall of algorithmic Picture tissue : 79.4788%
* F1-Score: 76.8503%

# **CHAPTER 6 : COMPARE ALGORITHMS**

**6.1 Algorithms run time**

Text

Description automatically generated with low confidence

Chart, bar chart

Description automatically generated

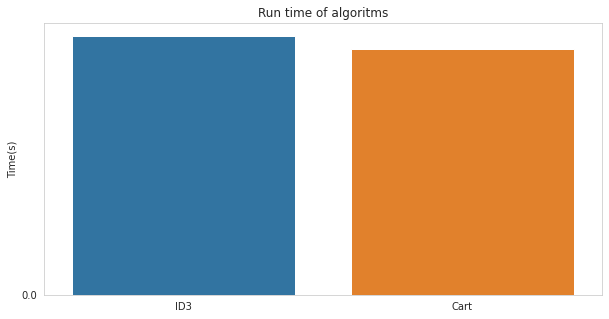
Comment:

* Logistic Regression algorithm has the highest running time
* Naive Bayes algorithm has the lowest running time
* ID3 and Cart have approximately the same running time

Closer examination of the runtime between ID3 and Cart

Text

Description automatically generated



Comment:

* ID3 has slightly better running time than Cart

## **6.2 Algorithms Accuracy**

Text, timeline

Description automatically generated

Chart, bar chart

Description automatically generated

Comment:

* Random Forest has the highest accuracy
* Then KNN and ID3
* Cart and Naive Bayes have the lowest accuracy and are approximately the same
  + Choose Random Forest algorithm to make prediction

# **CHAPTER 7 : BUILD FORECAST SOFTWARE**

## **7.1 Find the attribute with high confidence to choose as the predictor attribute**

Text

Description automatically generated

A picture containing text

Description automatically generated

Text

Description automatically generated

* + Select column
* StockOptionLevel
* MonthlyIncome
* JobInvolvement
* MonthlyRate
* DailyRate

A screenshot of a computer

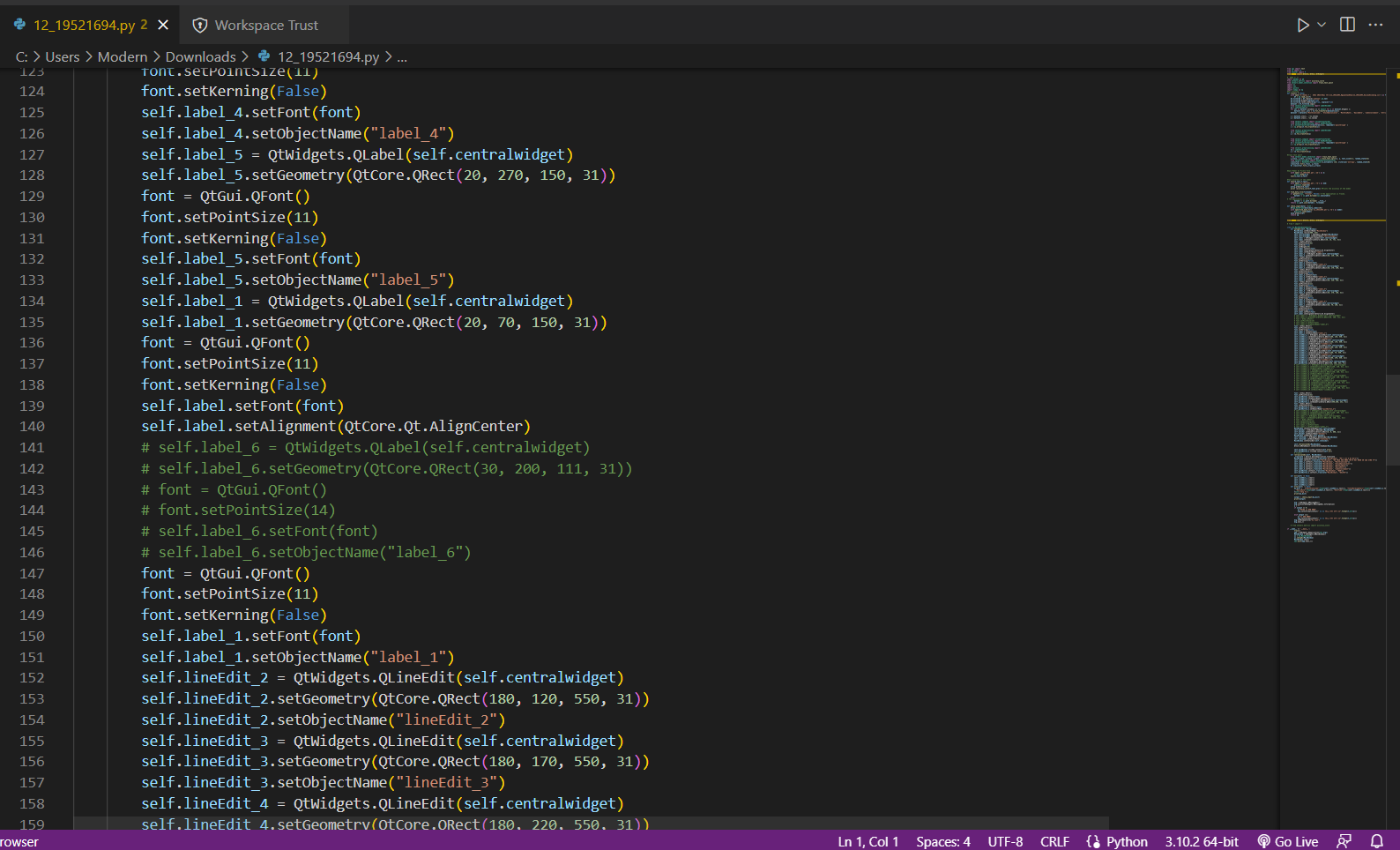
Description automatically generated with medium confidence

## **Software code**

### **7.2.1 Code display**

**Text

Description automatically generated**

****

**Text

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

* + 1. **Code processing part**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

* + 1. **Software Interface**

Graphical user interface, text, application, email

Description automatically generated

# **CHAPTER 8 : CONCLUSION**

**8.1 Advantages and limitations of each algorithm**

### **8.1.1 Decision Tree**

Advantages:

* 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

**8.1.2 Random Forest**

Advantages:

* 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.

**8.1.3 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.

**8.1.4 K-nearest Neighbors**

Advantages:

* The algorithm is simple, easy to deploy.
* Small computational complexity.
* Handle well with noise data sets

Limitations

* With small K is prone to interference leading to incorrect results
* It takes a long time to do so due to having to calculate the distance with all the objects in the data set.
* It is necessary to convert the data type into qualitative elements.

**8.1.5 Naive Bayes**

Advantages:

* Independent assumptions: works well for multiple problems/data domains and applications
* Simple but good enough to solve many problems such as text layering, spam filtering ...
* Easy to use and fast when it comes to guessing the label of test data. It's pretty good in multi-class prediction (test later).
* When assuming that the features of the data are independent of each other, Naive Bayes runs better than other algorithms such as logistic regression and also needs less data.
* Allows the succession of prior knowledge and obeserved data.
* It is good that there is a numerical difference between the classification classes.
* Model training (parameter estimation) is easy and fast.

Limitations:

* The accuracy of Naive Bayes compared to other algorithms is not high.
* In the real world, it is almost impossible when the features of test data are independent of each other.
* Problem zero (stated how to solve it above).
* The model is not trained by a strong and rigorous optimization method.
* The parameters of the model are estimates of the probability of single conditions. Do not take into account the interaction between these estimates.

**8.2 Direction of Development**

* Research and learn more about datasets, re-adjust properties and preprocessors for greater accuracy.
* Continue to apply other layering algorithms such as SVM, Artificial Neuron Network, ... To find the optimal algorithm.

# **CHAPTER 9 : REFERENCES**

**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/)

**Learn the algorithm:** [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)

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

**Machine learning types:** [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.)