**HR Analytics**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER** | **TITLE** | **PAGE NO** |
|  | **ABSTRACT** |  |
|  | **INTRODUCTION** |  |
|  | * 1. General Introduction |  |
|  | * 1. Project Objectives |  |
|  | **SYSTEM** **PROPOSAL** |  |
|  | * 1. Existing System |  |
|  | * + 1. Disadvantages |  |
|  | * 1. Proposed System |  |
|  | 2.2.1 Advantages |  |
|  | * 1. Literature Survey |  |
|  | **SYSTEM DIAGRAMS** |  |
|  | * 1. Architecture Diagram |  |
|  | * 1. Flow Diagram |  |
|  | * 1. UML Diagrams |  |
|  | **IMPLEMENTATION** |  |
|  | * 1. Modules |  |
|  | * 1. Modules Description |  |
|  | **SYSTEM** **REQUIREMENTS** |  |
|  | * 1. Hardware Requirements |  |
|  | * 1. Software Requirements |  |
|  | * 1. Software Description |  |
|  | * 1. Testing of Products |  |
|  | **CONCLUSION** |  |
| **7.** | **FUTURE** **ENHANCEMENT** |  |
| **8.** | **SAMPLE** **CODING** |  |
| **9.** | **SAMPLE** **SCREENSHOT** |  |
| **10.** | **REFERENCES** |  |

**ABSTRACT**

The HR world is abuzz with talk of big data and the transformative potential of HR analytics. This article takes issue with optimistic accounts which hail HR analytics as a ‘must have’ capability that will ensure HR’s future as a strategic management function while transforming organisational performance for the better. It argues that unless the HR profession wises up to both the potential and drawbacks of this emerging field, and engages operationally and strategically to develop better methods and approaches, it is unlikely that existing practices of HR analytics will deliver transformational change. uman resource management bundles can favorably affect the performance of business firms (Boselie, Dietz, & Boon, 2005; Ferris, Hall, Royle, & Martocchio, 2004; MacDuffie, 1995). This is because the individual practices that make up these bundles can support each other in enhancing specific workforce characteristics, thereby creating combined synergistic effects that are substantially greater than those of individual best practices (Delery, 1998; Becker & Gerhart, 1996). Several researchers have attempted to classify HRM practices. The input HR Analysis dataset was collected from dataset repository. Then, the system is developed the different machine learning algorithms such as logistic regression and gradient boosting for predicting the candidate is selected or not.

**CHAPTER 1**

**INTRODUCTION**

* 1. **General Introduction:**

Analytics is the discipline which has developed at the intersection of engineering, computer science, decision making and quantitative methods to organise, analyse and make sense of the increasing amounts of data being generated by contemporary societies (Mortensen et al., 2015).

Analytics has been described as a ‘must have’ capability for the HR profession; a tool for creating value from people and a pathway to broadening the strategic influence of the HR function (CIPD, 2013). The central argument of this article is that the development of HR analytics is being hampered by a lack of understanding of analytical thinking by the HR profession. This problem is compounded by the HR analytics industry, which is largely based around products and services which too often fail to provide the tools for HR to create and capture the strategic value of HR data. Unless the HR profession wises up to both the potential and pitfalls of analytics, we contend that HR analytics is likely to have a number of negative consequences for the HR profession itself, for workers and for organisations.

Specifically, there is a risk that analytics will further embed finance and engineering perspectives on people management at boardroom level in ways that will restrict the strategic influence of the HR profession. It may also damage the quality of working life and employee wellbeing, without delivering sustainable competitive advantage to the organisations that adopt it.

This argument is a deliberately provocative one. It is based on a careful reading of the literature combined with what we have learnt from engagement with HR and analytics professionals rather than on a carefully constructed programme of academic research. When we discuss analytics with HR professionals with an interest in the subject we hear many of the themes and concerns that this article raises being echoed back at us. We hope that by being provocative, we can stimulate research that will point to a better way forward.

The rest of the article is organised as follows. Recent interest in HR analytics reflects growing interest in ‘big data’. We therefore begin by defining what is meant by data analytics and big data as they relate to HR. Second, we offer an overview of academic thinking on HR analytics and sketch its potential contribution. Third we argue that these ideas are not being adopted because of failings on the part of the HR profession combined with limitations in human resources information systems (HRIS) and significant problems with the analytics industry as it is currently constituted. Taken together these problems and failings are likely to prevent the promise of HR analytics being realised and will lead to a number of negative consequences. Finally we set out alternatives, and argue that industry/university collaborations offer a productive way forward.

* 1. **Objectives:**

The main objective of our project is,

* To classify or to predict or to detect the candidate is selected or not effectively.
* To implement the different classification algorithms for better performances.
* To enhance the overall performance for classification algorithms.

**CHAPTER 2**

**SYSTEM PROPOSAL**

* 1. **EXISTING SYSTEM:**

In existing system, Human resource management bundles consisting of multiple complementary practices are typically considered superior to individual best practices in infl uencing fi rm performance. This study investigates the relationship between three such bundles (empowerment, motivation, and skill-enhancing) and business outcomes (retention, operating performance, fi nancial performance, and overall performance ratings). A meta-analysis of 239 effect sizes derived from 65 studies reveals that HRM bundles have signifi cantly larger magnitudes of effects than their constituent individual practices, are positively related to business outcomes, and display effect sizes that are comparable to or larger than those of high-performance work systems. These fi ndings reaffi rm the case for fi rm-level investments in synergistic HRM combinations and highlight the importance of investing in complementary practices.

**2.1.1 DISADVANTAGES:**

* It doesn’t efficient for large volume of data’s
* Theoretical limits.
* Training time is high.
  1. **PROPOSED SYSTEM:**

In this system, the HR Analytics dataset was taken as input. The input data was taken from the dataset repository. Then, we have to implement the data pre-processing step. In this step, we have to handle the missing values for avoid wrong prediction, and to encode the label for input data. Then, we have to split the dataset into test and train. The data is splitting is based on ratio. Then, we have to implement the classification algorithm (i.e.) machine learning. The machine learning algorithms such as Logistic Regression and gradient boosting. Finally, the experimental results shows that the performance metrics such as accuracy and comparison results. Finally, the system can predict the candidate is selected or not.

**2.2.1 ADVANTAGES:**

* It is efficient for large number of datasets.
* Time consumption is low.
* The process is implemented with removing unwanted data.
* Prediction is accurate.

**2.3 LITERATURE SURVEY:**

**2.3.1 HR analytics and ethics, 2019**

**Author***:* K. Simbeck

**Methodology:**

The systematic application of analytical methods on human resources (HR)-related (big) data is referred to as HR analytics or people analytics. Typical problems in HR analytics include the estimation of churn rates, the identification of knowledge and skill in an organization, and the prediction of success on a job. HR analytics, as opposed to the simple use of key performance indicators, is a growing field of interest because of the rapid growth of volume, velocity, and variety of HR data, driven by the digitalization of work processes. Personnel files used to be in steel lockers in the past. They are now stored in company systems, along with data from hiring processes, employee satisfaction surveys, e-mails, and process data. With the growing prevalence of HR analytics, a discussion around its ethics needs to occur. The objective of this paper is to discuss the ethical implications of the application of sophisticated analytical methods to questions in HR management. This paper builds on previous literature in algorithmic fairness that focuses on technical options to identify, measure, and reduce discrimination in data analysis. This paper applies to HR analytics the ethical frameworks discussed in other fields including medicine, robotics, learning analytics, and coaching.

**Advantage:**

* Training time is low.

**Disadvantage:**

* It is not efficient for large number of data’s.

**2.3.2 HR Analytics to Support Managerial Decisions: A Case Study 2020**

**Author**: Liyuan Liu, Sanjoosh Akkineni, Paul Story, Clay Davis

**Methodology:**

Human Resource (HR) Analytics enables HRs to make strategic contributions and support managerial decisions. However, in most of the industry, HRs should have been on board with data analysis. There are several challenges: the HR data is messy and imbalanced, it is hard to harness both structured and unstructured data, some HR managers lack data mining skills and the lack of related empirical research that gives a detailed analytics guideline. The contribution of this study is that we develop a framework to support an industrial aluminum company to make the decisions and to improve strategy execution. The framework includes descriptive analysis, predictive analysis, and entity sentiment analysis. We analyzed an industrial aluminum company's data and found some actionable issues. Then we employed machine learning algorithms to predict employees' turnover and found risk factors. Moreover, we applied the entity sentiment analysis on the unstructured data collected from employees' engagement survey.

**Advantage*:***

* Less Efficiency

**Disadvantage*:***

* Training time is high.

**2.3.3 Predictive analytics of HR - A machine learning approach, 2021**

**Author:** - V. Kakulapati, Kalluri Krishna Chaitanya, Kolli Vamsi Guru Chaitanya & Ponugoti Akshay

**Methodology:**

One of the branches of analytics is HR analytics, which is developing the system HR units in organizations function, principal to sophisticated proficiency, and improved outcomes overall. The usage of analytics by human resources for many years. Though the assortment, processing, and data analysis have been generally manual and specified the nature of HR dynamics, the approach has been constraining HR. The prospect to effort predictive analytics in categorizing the employees furthermost likely to grow promoted. Here we apply machine learning techniques to analyze the employee information for improving his/her position in the organization. Compensation and job performance information from revenue rates and personnel characteristics to payroll and service history, never before have HR executives had such liberated right to use to individual details. In this work, we are applying random forest classification, which facilitates employee classification based on their monthly income and informal way to execute analytics on data. Further, we use clustering techniques based on the performance metrics similarity to analyze employee performance.

**Advantage*:***

* Training time is low.
* Better performance.

**Disadvantage**:

* The process is implemented without removing unwanted data.

**2.3.4 Human resources analytics: A systematization of research topics and directions for future research, 2020**

**Author:** Alessandro Margherita

**Methodology:**

The management of human resources is today significantly impacted by the emergence of the global workforce and the increasing relevance of business analytics as a strategic organizational capability. Whereas human resources analytics has been largely discussed in literature in the last decade, a systematic identification and classification of key topics is yet to be introduced. In particular, there is room for conceptual contributions aiming to provide a comprehensive definition of concepts and investigation areas related to HR analytics. Using a systematic literature review process, we deconstruct the concept of human resources analytics as presented in a vast although fragmented literature, and we identify 106 key research topics associated to three major areas, i.e. enablers of HR analytics (technological and organizational), applications (descriptive and diagnostic/prescriptive), and value (employee value and organizational value). We also speculate on an “exponential” view of HR analytics enabled by the affirmation of artificial intelligence and cognitive technologies.

**Advantage:**

* We can note that the Sensitivity for all ML models with the unbalanced data is lower than the Sensitivity for balanced data created by different resampling methods

**Disadvantage**:

* It creates a new instance by appropriately combining existing instances, thus making it possible to avoid the disadvantage of over fitting to a certain degree.

**CHAPTER 3**

**SYSTEM DIAGRAMS**

**3.1 SYSTEM ARCHITECTURE:**

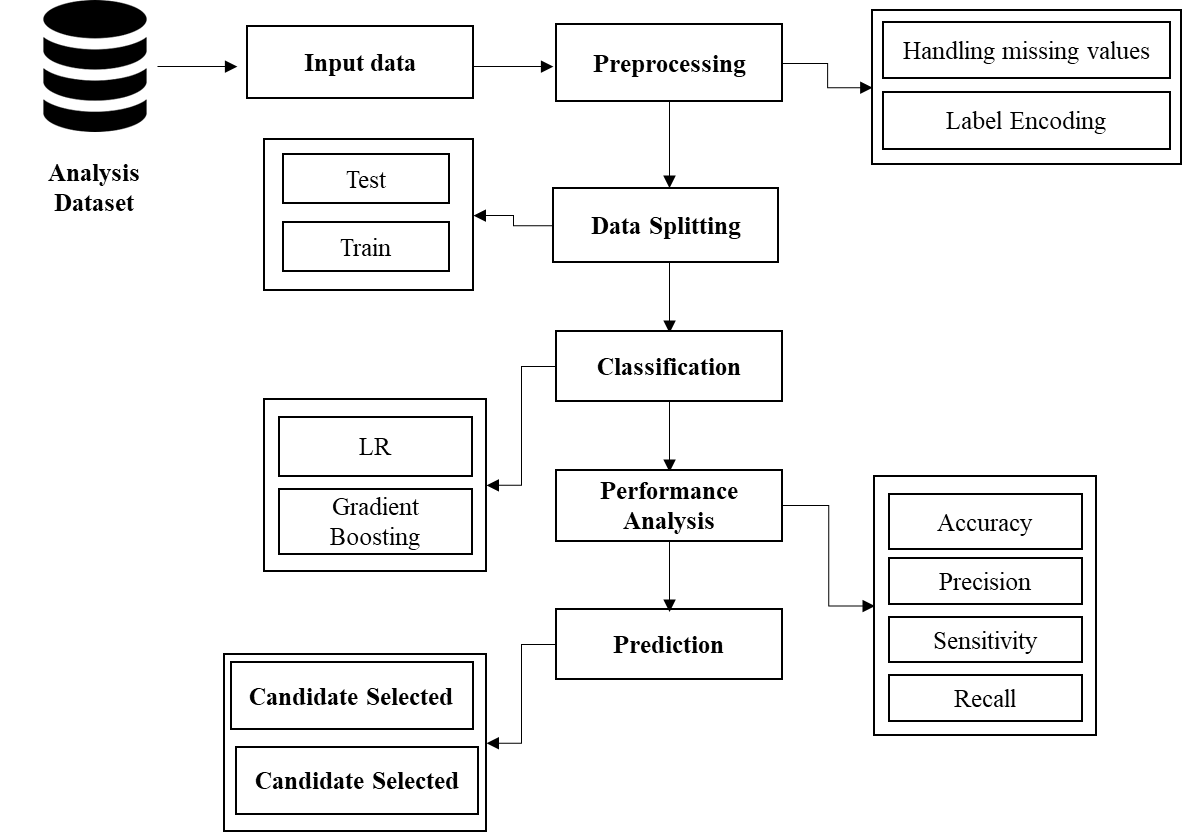
****

FIGURE 3.1: SYSTEM ARCHITECTURE

**3.2 FLOW DIAGRAM**

Input Data

Preprocessing

Data Splitting

Classification

Performance analysis

FIGURE 3.2: FLOW DIAGRAM

**3.3 UML DIAGRAMS:**

**3.3.1 USE CASE DIAGRAM:**

System

User

FIGURE 3.3.1: USE CASE DIAGRAM

**3.3.2 ACTIVITY DIAGRAM:**

Input Data

Preprocessing

Data splitting

Performance metrics

Classification

FIGURE 3.3.2: ACTIVITY DIAGRAM

**3.3.3 SEQUENCE DIAGRAM:**

**Input Data**

**Preprocessing**

**Data splitting**

**Classification**

Select data

Missing value

Test and Train

Load data

Label Encoding

LR and GB

FIGURE 3.3.3: SEQUENCE DIAGRAM

**3.3.4 ER DIAGRAM:**

**Data selection**

**Preprocessing**

**Classification**

**Prediction**

FIGURE 3.3.4: ER DIAGRAM

**3.3.5 CLASS DIAGRAM:**

Select data ()

Load data ()

View data ()

**Input Data**

Test ()

**Data Splitting**

Prediction ()

Accuracy ()

Precision ()

Recall()

**Performance**

**Preprocessing**

Missing values ()

Label encode ()

LR ()

Gradient Boosting ()

**Classification**

Train ()

Correlation ()

FIGURE 3.3.5: CLASS DIAGRAM

**CHAPTER 4**

**IMPLEMENTATION**

**4.1 MODULES:**

* Data selection
* Preprocessing
* Data splitting
* Classification
* Result generation

**4.2 MODULES DESCRIPTION:**

**4.2.1: DATA SELECTION:**

* The input data was collected from dataset repository.
* In our process, HR Analysis dataset is used.
* The data selection is the process of predicting the candidate is selected or not.
* The input dataset was taken from dataset repository such as UCI repository.
* In python, with the help of panda’s package, we can read or load our input dataset.
* Our dataset is in the format is ‘.csv’

**4.2.2: DATA PREPROCESSING:**

* Data pre-processing is the process of removing the unwanted data from the dataset.
* Pre-processing data transformation operations are used to transform the dataset into a structure suitable for machine learning.
* This step also includes cleaning the dataset by removing irrelevant or corrupted data that can affect the accuracy of the dataset, which makes it more efficient.
* Missing data removal
* Encoding Categorical data
* Missing data removal: In this process, the null values such as missing values and Nan values are replaced by 0.
* Missing and duplicate values were removed and data was cleaned of any abnormalities.
* Encoding Categorical data: That categorical data is defined as variables with a finite set of label values.
* That most machine learning algorithms require numerical input and output variables.

**4.2.3: DATA SPLITTING:**

* During the machine learning process, data are needed so that learning can take place.
* In addition to the data required for training, test data are needed to evaluate the performance of the algorithm in order to see how well it works.
* In our process, we considered 80% of the input dataset to be the training data and the remaining 20% to be the testing data.
* Data splitting is the act of partitioning available data into two portions, usually for cross-validator purposes.
* One Portion of the data is used to develop a predictive model and the other to evaluate the model's performance.
* Separating data into training and testing sets is an important part of evaluating data mining models.
* Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing.

**4.2.5: CLASSIFICATION:**

* In our process, we have to implement the different classification algorithm such as gradient boosting and LR.
* **Gradient boosting** is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.
* Logistic regression is an example of supervised learning. It is used to calculate or predict the probability of a binary (yes/no) event occurring. An example of logistic regression could be applying machine learning to determine if a person is likely to be infected with COVID-19 or not.

**4.2.6: RESULT GENERATION:**

The Final Result will get generated based on the overall classification and prediction. The performance of this proposed approach is evaluated using some measures like,

* **Accuracy:**

Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

AC= (TP+TN)/ (TP+TN+FP+FN)

* **Precision**

Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives.

Precision=TP/ (TP+FP)

* **Recall**

Recall is the number of correct results divided by the number of results that should have been returned. In binary classification, recall is called sensitivity. It can be viewed as the probability that a relevant document is retrieved by the query.

Recall=TP/ (TP+FN)

**CHAPTER 5**

**SYSTEM REQUIREMENTS**

**5.1 HARDWARE REQUIREMENTS:**

* System : Pentium IV 2.4 GHz
* Hard Disk : 200 GB
* Mouse : Logitech.
* Keyboard : 110 keys enhanced
* Ram : 4GB

**5.2 SOFTWARE REQUIREMENTS:**

* O/S : Windows 7.
* Language : Python
* Front End : Anaconda Navigator – Spyder

**5.3 SOFTWARE DESCRIPTION:**

**5.3.1 Python**

Python is one of those rare languages which can claim to be both *simple* and powerful. You will find yourself pleasantly surprised to see how easy it is to concentrate on the solution to the problem rather than the syntax and structure of the language you are programming in. The official introduction to Python is Python is an easy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python's elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms. I will discuss most of these features in more detail in the next section.

## **5.3.2 Features of Python**

### **Simple**

Python is a simple and minimalistic language. Reading a good Python program feels almost like reading English, although very strict English! This pseudo-code nature of Python is one of its greatest strengths. It allows you to concentrate on the solution to the problem rather than the language itself.

### **Easy to Learn**

As you will see, Python is extremely easy to get started with. Python has an extraordinarily simple syntax, as already mentioned.

### **Free and Open Source**

Python is an example of a FLOSS (Free/Libré and Open Source Software). In simple terms, you can freely distribute copies of this software, read its source code, make changes to it, and use pieces of it in new free programs. FLOSS is based on the concept of a community which shares knowledge. This is one of the reasons why Python is so good - it has been created and is constantly improved by a community who just want to see a better Python.

### **High-level Language**

When you write programs in Python, you never need to bother about the low-level details such as managing the memory used by your program, etc.

### **Portable**

Due to its open-source nature, Python has been ported to (i.e. changed to make it work on) many platforms. All your Python programs can work on any of these platforms without requiring any changes at all if you are careful enough to avoid any system-dependent features.

You can use Python on GNU/Linux, Windows, FreeBSD, Macintosh, Solaris, OS/2, Amiga, AROS, AS/400, BeOS, OS/390, z/OS, Palm OS, QNX, VMS, Psion, Acorn RISC OS, VxWorks, PlayStation, Sharp Zaurus, Windows CE and PocketPC!

You can even use a platform like [Kivy](http://kivy.org) to create games for your computer and for iPhone, iPad, and Android.

### **Interpreted**

This requires a bit of explanation.

A program written in a compiled language like C or C++ is converted from the source language i.e. C or C++ into a language that is spoken by your computer (binary code i.e. 0s and 1s) using a compiler with various flags and options. When you run the program, the linker/loader software copies the program from hard disk to memory and starts running it.

Python, on the other hand, does not need compilation to binary. You just run the program directly from the source code. Internally, Python converts the source code into an intermediate form called bytecodes and then translates this into the native language of your computer and then runs it. All this, actually, makes using Python much easier since you don't have to worry about compiling the program, making sure that the proper libraries are linked and loaded, etc. This also makes your Python programs much more portable, since you can just copy your Python program onto another computer and it just works!

### **Object Oriented**

Python supports procedure-oriented programming as well as object-oriented programming. In procedure-oriented languages, the program is built around procedures or functions which are nothing but reusable pieces of programs. In object-oriented languages, the program is built around objects which combine data and functionality. Python has a very powerful but simplistic way of doing OOP, especially when compared to big languages like C++ or Java.

### **Extensible**

If you need a critical piece of code to run very fast or want to have some piece of algorithm not to be open, you can code that part of your program in C or C++ and then use it from your Python program.

### **Embeddable**

You can embed Python within your C/C++ programs to give scripting capabilities for your program's users.

### **Extensive Libraries**

The Python Standard Library is huge indeed. It can help you do various things involving regular expressions, documentation generation, unit testing, threading, databases, web browsers, CGI, FTP, email, XML, XML-RPC, HTML, WAV files, cryptography, GUI (graphical user interfaces), and other system-dependent stuff. Remember, all this is always available wherever Python is installed. This is called the Batteries Included philosophy of Python.

Besides the standard library, there are various other high-quality libraries which you can find at the Python Package Index.

**5.4 TESTING PRODUCTS:**

System testing is the stage of implementation, which aimed at ensuring that system works accurately and efficiently before the live operation commence. Testing is the process of executing a program with the intent of finding an error. A good test case is one that has a high probability of finding an error. A successful test is one that answers a yet undiscovered error.

Testing is vital to the success of the system. System testing makes a logical assumption that if all parts of the system are correct, the goal will be successfully achieved. . A series of tests are performed before the system is ready for the user acceptance testing. Any engineered product can be tested in one of the following ways. Knowing the specified function that a product has been designed to from, test can be conducted to demonstrate each function is fully operational. Knowing the internal working of a product, tests can be conducted to ensure that “al gears mesh”, that is the internal operation of the product performs according to the specification and all internal components have been adequately exercised.

**5.4.1 UNIT TESTING:**

Unit testing is the testing of each module and the integration of the overall system is done. Unit testing becomes verification efforts on the smallest unit of software design in the module. This is also known as ‘module testing’.

The modules of the system are tested separately. This testing is carried out during the programming itself. In this testing step, each model is found to be working satisfactorily as regard to the expected output from the module. There are some validation checks for the fields. For example, the validation check is done for verifying the data given by the user where both format and validity of the data entered is included. It is very easy to find error and debug the system.

**5.4.2 INTEGRATION TESTING:**

Data can be lost across an interface, one module can have an adverse effect on the other sub function, when combined, may not produce the desired major function. Integrated testing is systematic testing that can be done with sample data. The need for the integrated test is to find the overall system performance. There are two types of integration testing. They are:

i) Top-down integration testing. ii) Bottom-up integration testing.

**5.4.3 TESTING TECHNIQUES/STRATEGIES:**

* **WHITE BOX TESTING:**

White Box testing is a test case design method that uses the control structure of the procedural design to drive cases. Using the white box testing methods, we

Derived test cases that guarantee that all independent paths within a module have been exercised at least once.

* **BLACK BOX TESTING:**

1. Black box testing is done to find incorrect or missing function
2. Interface error
3. Errors in external database access
4. Performance errors.
5. Initialization and termination errors

In ‘functional testing’, is performed to validate an application conforms to its specifications of correctly performs all its required functions. So this testing is also called ‘black box testing’. It tests the external behaviour of the system. Here the engineered product can be tested knowing the specified function that a product has been designed to perform, tests can be conducted to demonstrate that each function is fully operational.

**5.4.4 SOFTWARE TESTING STRATEGIES**

**VALIDATION TESTING:**

After the culmination of black box testing, software is completed assembly as a package, interfacing errors have been uncovered and corrected and final series of software validation tests begin validation testing can be defined as many,

But a single definition is that validation succeeds when the software functions in a manner that can be reasonably expected by the customer

**USER ACCEPTANCE TESTING:**

User acceptance of the system is the key factor for the success of the system. The system under consideration is tested for user acceptance by constantly keeping in touch with prospective system at the time of developing changes whenever required.

**OUTPUT TESTING**:

After performing the validation testing, the next step is output asking the user about the format required testing of the proposed system, since no system could be useful if it does not produce the required output in the specific format. The output displayed or generated by the system under consideration. Here the output format is considered in two ways. One is screen and the other is printed format. The output format on the screen is found to be correct as the format was designed in the system phase according to the user needs. For the hard copy also output comes out as the specified requirements by the user. Hence the output testing does not result in any connection in the system.

**CHAPTER 6**

**CONCLUSION**

We conclude that, the auto insurance claim dataset was taken as input. The input dataset was mentioned in our research paper. We are implemented the classification algorithms (i.e) machine learning algorithms. Then, machine learning algorithms such as LR and gradient boosting classification. Finally, the result shows that the accuracy for above mentioned algorithm and estimated the performances metrics such as accuracy for both algorithms and comparison graph.

**CHAPTER 7**

**FUTURE ENHANCEMENT**

Future work may be done in the next directions: Using hybrid classifiers to improve comparison and performance. Furthermore, feature selection approaches may be used to enhance model results and gain a deeper understanding of the important features. It will also be worthwhile to conduct this research for another insurance branch, whether to predict claim occurrences or to predict fraud because these kinds of data always are very heavily unbalanced.

**CHAPTER 8**

**SAMPLE CODING**

#======================= IMPORT PACKAGES =============================

import pandas as pd

from sklearn.model\_selection import train\_test\_split

import warnings

warnings.filterwarnings('ignore')

import matplotlib.pyplot as plt

from sklearn import preprocessing

#===================== DATA SELECTION ==============================

#=== READ A DATASET ====

data\_frame=pd.read\_csv("HRAnalytics.csv")

print("----------------------------------")

print(" 1.Data Selection ")

print("----------------------------------")

print()

print(data\_frame.head(20))

#===================== DATA PREPROCESSING ==============================

#=== CHECK MISSING VALUES ===

print("=====================================================")

print(" 2.Preprocessing ")

print("=====================================================")

print()

print("--------------------------------------------")

print(" Checking missing values ")

print("--------------------------------------------")

print()

print(data\_frame.isnull().sum())

print()

#=== LABEL ENCODING ===

label\_encoder = preprocessing.LabelEncoder()

print("---------------------------------")

print(" Before label encoding ")

print("---------------------------------")

print()

print(data\_frame['salary'].head(10))

data\_frame['salary']=label\_encoder.fit\_transform(data\_frame['salary'])

data\_frame['sales']=label\_encoder.fit\_transform(data\_frame['sales'])

print("-------------------------------------------")

print(" After label Encoding ")

print("------------------------------------------")

print()

print(data\_frame['salary'].head(20))

#============================= DATA SPILLTING =========================

X = data\_frame.drop("left",axis=1)

Y = data\_frame["left"]

print("----------------------------------------")

print("DATA SPLITTING")

print("------------------------------------")

print()

x\_train,x\_test,y\_train,y\_test=train\_test\_split(X,Y,test\_size=0.3,random\_state=1)

print()

print("Total Number Of data = ", len(X))

print()

print("Total Number Of Test data = ", len(x\_test))

print()

print("Total Number Of Train data = ", len(x\_train))

print()

#============================= CLASSIFICATION =========================

# === LOGISTIC REGRESSION =====

from sklearn import linear\_model

lr = linear\_model.LogisticRegression()

# fit the regressor with x and y data

lr.fit(x\_train, y\_train)

Y\_pred = lr.predict(x\_train)

from sklearn import metrics

Accuracy\_lr=metrics.accuracy\_score(y\_train,Y\_pred)\*100

print("----------------------------------------")

print("LOGISTIC REGRESSION --> LR")

print("------------------------------------")

print()

print("1. Accuracy =",Accuracy\_lr )

print()

print(metrics.classification\_report(y\_train,Y\_pred))

# ==== GRADIENT BOOSTING ====

from sklearn.ensemble import GradientBoostingClassifier

gbt = GradientBoostingClassifier()

gbt.fit(x\_train, y\_train)

y\_pred\_gbt = gbt.predict(x\_train)

from sklearn import metrics

acc\_gbt=metrics.accuracy\_score(y\_pred\_gbt,y\_train)\*100

print("----------------------------------------")

print(" GBDT ")

print("----------------------------------------")

print()

print("1. Accuracy = ", acc\_gbt)

print()

print(metrics.classification\_report(y\_pred\_gbt,y\_train))

print()

# ============== PREDICTION =====================

print("----------------------------------------")

print("PREDICTION ")

print("------------------------------------")

print()

for i in range(0,5):

if Y\_pred[i]==0:

print("-----------------------------")

print([i],"The Candidate is selected")

print("-----------------------------")

else:

print("----------------------------------")

print([i],"The Candidate is Non- selected")

print("----------------------------------")

# ===== COMPARISON =====

vals=[acc\_gbt,Accuracy\_lr]

inds=range(len(vals))

labels=["GB","LR"]

fig,ax = plt.subplots()

rects = ax.bar(inds, vals)

ax.set\_xticks([ind for ind in inds])

ax.set\_xticklabels(labels)

plt.show()

import seaborn as sns

plt.figure(figsize=(5, 5))

plt.title("Classification")

sns.countplot(x='left',data=data\_frame)

plt.show()

import seaborn as sns

plt.figure(figsize=(5, 5))

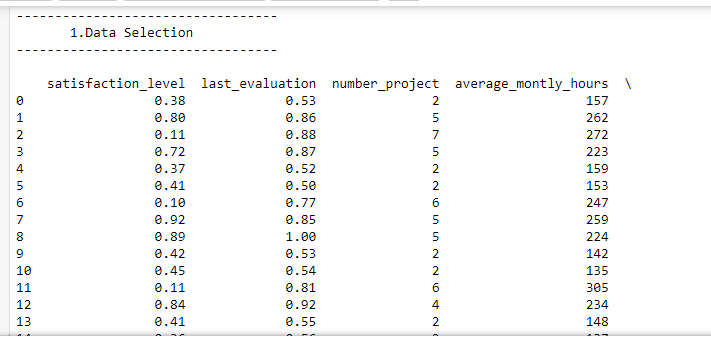
plt.title("Classification")

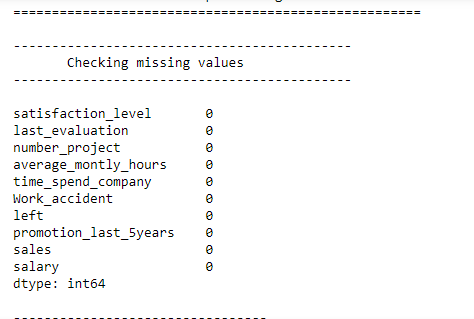
sns.countplot(x='salary',data=data\_frame)

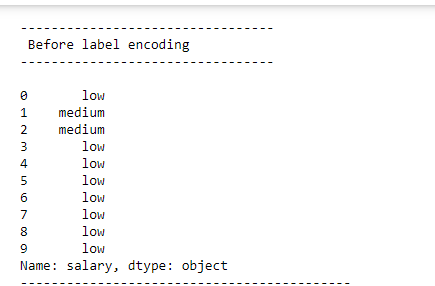
plt.show()

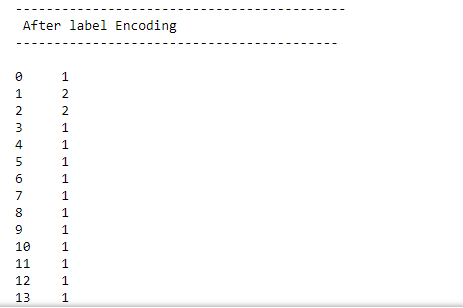
**CHAPTER 9**

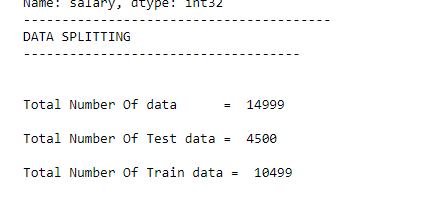
**SAMPLE SCREENSHOTS**

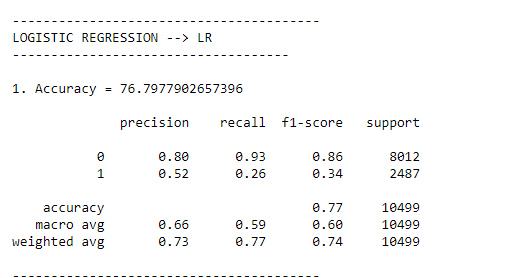


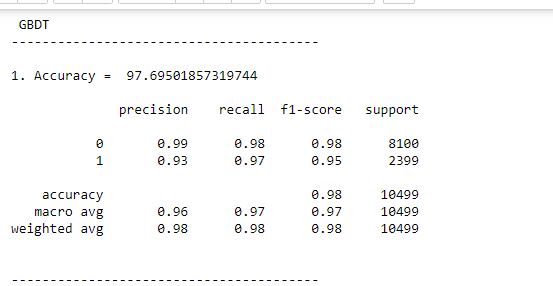


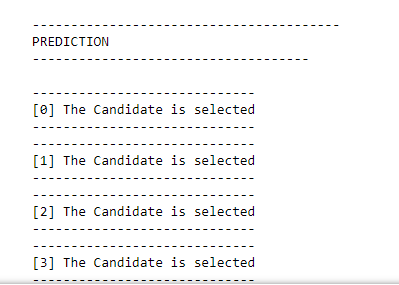












**CHAPTER 10**

**REFERENCES**

* Apgar, D. (2013). ‘The false promise of big data: can data mining replace hypothesis-driven learning in the identification of predictive performance metrics?’ Systems Research and Behavioural Science. Online first publication, doi: 10.1002/sres.2219.
* Aral, S., Bryjolfsson, E. and Wu, L (2012). ‘Three-way complementarities: Performance Pay, Human Resource Analytics, and Information Technology’. Management Science, 58(5): 913 – 931.
* Bersin (2014). Talent management for the global workforce: The market for talent management systems 2014. Oakland CA.
* Bersin J (2015a). ‘The geeks arrive in HR: people analytics is here.’ Forbes Magazine downloaded from http://www.forbes.com/sites/joshbersin/2015/02/01/geeks-arrive-in-hrpeople-analytics-is-here/ on 13/2/2015.
* Bersin J (2015b). ‘Why people management is replacing talent management.’ Joshbersin.com downloaded from http://joshbersin.com/2015/01/why-people-management-isreplacing-talent-management/#disqus\_thread on 13/2/2015.
* Boudreau, J. (2010). Retooling HR: Using proven business tools to make better management decisions. Boston: HBR Press.
* Boudreau, J. and Jesuthasan, R. (2011). Transformative HR: How great companies use evidence based change for sustainable competitive advantage. San Francisco: Jossey Bass.
* Boudreau, J. and Lawler, E. (2015a). Making Talent Analytics and Reporting into a Decision Science. Working paper, Centre for Effective Organisations. University of Southern California.
* Boudreau, J. and Lawler, E. (2015b). Talent Analytics Measurement and Reporting: Building a Decision Science or Merely Tracking Activity? Working paper, Centre for Effective Organisations. University of Southern California.