# 

**Analysis and predict the absenteeism and improve decision making for Employees in working place**

Md. Kamrul Hasan

A Thesis in the Partial Fulfillment of the Requirements

for the Award of Bachelor of Computer Science and Engineering (BCSE)



Department of Computer Science and Engineering

College of Engineering and Technology

IUBAT – International University of Business Agriculture and Technology

Summer2021

**Analysis and predict the absenteeism and improve decision making for Employees in working plac**e

Md. Kamrul Hasan

A Thesis in the Partial Fulfillment of the Requirements for the Award of Bachelor of Computer Science and Engineering (BCSE)

The thesis has been examined and approved,

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Dr. Utpal Kanti Das

Chairman and Professor

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. Muhammad Hasibur Rashid Chayon

Co-supervisor, Coordinator and Associate Professor

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Dr. Abhijit Saha

Supervisor and Professor

Department of Computer Science and Engineering

College of Engineering and Technology

IUBAT – International University of Business Agriculture and Technology

Summer 2021

## **Letter of Transmittal**

16 March 2022

The Chair

Thesis Defense Committee

Department of Computer Science and Engineering

IUBAT–International University of Business Agriculture and Technology

4 Embankment Drive Road, Sector 10, Uttara Model Town

Dhaka 1230, Bangladesh

**Subject:** Letter of Transmittal.

Dear Sir,

With due respect, this is pleasure to present my thesis report entitled “Analysis and predict the absenteeism and improve decision making for Employees in working place”. I have prepared this report as partial fulfillment of the requirement for the degree of Bachelor of Computer Science and Engineering. Now, I am looking forward to your kind estimation regarding this thesis report. I will remain enormously obliged to you if you kindly go through this report and evaluate my performance.

I hope that you would find the report comprehensive and competent augmented.

Yours sincerely,

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Md. Kamrul Hasan

18103012

## **Student’s Declaration**

I, Md. Kamrul Hasan declare that the work presented in this thesis paper titled, “Analysis and predict the absenteeism and improve decision making for Employees in working place” and the research carried out by me under the supervision of Dr. Abhijit Saha, Professor, and co-supervision of Co-supervisor, Coordinator and Associate Professor, Department of Computer Science and Engineering, IUBAT. I declare, no parts of this report have been submitted anywhere for any degree, diploma or certificate.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Md. Kamrul Hasan

18103012

## **Supervisor’s Certification**

This is to certify that the thesis report on “Analysis and predict the absenteeism and improve decision making for Employees in working place” has been carried out by Md. Kamrul Hasan bearing ID#18103012 student of Department of Computer Science and Engineering of IUBAT-International University of Business Agriculture and Technology, as a partial fulfillment of the requirement for the degree in Bachelor of Computer Science and Engineering. The report has been prepared under my guidance and is a record of work carried out successfully. To the best of my knowledge and as per his declaration, no parts of this report has been submitted anywhere for any degree, diploma or certificate.

Now he is permitted to submit the report. I wish, his success in all his future endeavors.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. Abhijit Saha

Professor

Department of Computer Science and Engineering

IUBAT–International University of Business Agriculture and Technology

## **Abstract**

Absenteeism is defined as the expected absence of work and represents a decline in the productivity and quality of work of the company. High levels of competitiveness in the market, professional development combined with organizational development, and pressure to reach even more daring goals lead to increased employee overload and ultimately work. To see the employee’s working behavior, manner, regularity or irregularity and having such information in advance can improve our decision making how by reorganizing the work manner. In a way that will not allow us to a lack of productivity and increase the quality of work generated in the company or firm, we want to know for how many working hours an employee could be away from work based on information such as how far long they live from their workplace how many children and pets they have. Do they have higher education or not etc.? Their sickness or health condition is good or not. By analyzing this kind of information, we can improve the work method. These are the main reason why the topic was chosen for the research. In this work we also inspect the use of predictive analytics as a decision making for growing employee well-being in the work-area by recognizing groups of employees at risk of sickness. In this research paper we actually analysis the different variables and find out the reasons from the employee dataset and we inspect the use of predictive analytics as a decision making for growing employee well-being in the work-area by recognizing groups of employees at risk of sickness or other issues. And also find out the age vs absence probability percentage to sort out which age employees are absence from workplace. The mail goal of this research is to come up with a useful and reliable model for the organization to observe their employee’s behavior in the case of absenteeism. To estimate the performance of the proposed model a prototype has been implemented and also tested to evaluate its accuracy. Proposed model focuses that the model is reliable and fulfill the needs of the organizations.

## **Acknowledgments**

First and foremost, I would like to thank Almighty Allah for giving me the strength and courage to finish this work.

During my work on this thesis, many people supported me. At this point, I would like to express my gratitude to them. The satisfaction that accompanies the successful completion of this thesis would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success. I am grateful to my honorable thesis supervisor Dr. Abhijit Saha, Professor, Department of Computer Science and Engineering, IUBAT, for the guidance, inspiration and constructive suggestions which were helpful in the preparation of this thesis. I also convey special thanks and gratitude to Dr. Utpal Kanti Das, Professor and Chairman and to Dr. Muhammad Hasibur Rashid Chayon, Co-supervisor, Coordinator and Associate Professor, Department of Computer Science and Engineering, IUBAT.

**Table of Contents**

**Letter of Transmittal iii**

**Student’s Declaration iv**

**Supervisor’s Certification v**

**Abstract vi**

**Acknowledgments vii**

**List of Figures xi**

**List of Tables xii**

**Chapter I. Introduction** 1

1.1 Backgorund and Context 5

1.2 Problem Statement 6

1.3 Reasearch Questions 7

1.4 Research Objectives 7

**Chapter II. Literature Review** **9**

2.1Job Satisfaction and Absenteeism interface in Corporate Sector –

A study 9

2.2 Factors contributing to work-related absenteeism during the

COVID-19 pandemic 12

2.4 Key Debates and Controversies 15

**Chapter III. Research Methodology** **17**

3.1 Model Design Approach 17

3.1.1 Describing Regression Analysis 18

3.2 Proposed Model Requirements 20

3.2.1 Hardware Requirements. 20

3.2.2 Software Requirements. 21

3.3 Flow Chart of Proposed model 23

3.4 Testing 23

**Chapter IV. Result and Discussion** **28**

4.1 Excessive Absenteeism Result 28

4.2 Discuss analyzing the predicted output 29

**Chapter V. Conclusion** **35**

**References** **37**

## **List of Figures**

Figure 1.1 Data Science process flow. 1

Figure 2.1 Block Diagram of Proposed Model. 19

Figure: 3.2 Computer hardware component. 20

Figure: 3.3 Operating System. 21

Figure: 3.4 Jupyter Notebook. 21

Figure: 3.5 Google Colab. 22

Figure: 3.6 tableau public 22

Figure: 3.7 Flow chart of proposed model. 23

Figure: 3.8 Testing train accuracy. 25

Figure: 3.9 Testing circle. 26

Figure: 3.10 Test accuracy. 26

Figure 4.1 data load in Tabeau Public. 29

Figure 4.2 age data initialization. 30

Figure 4.3 seclected the dimension. 30

Figure 4.4 add prediction field with the age. 31

Figure 4.5 shows age vs prediction value. 32

Figure 4.6 shows probability values vs age 33

Figure 4.7 age vs avg. probability 34

## **List of Tables**

Table 2.1: Summarized Result of the Seven Factors of Motivation/Job Satisfaction 11

Table 4.1 age vs avg. absence probability result 28

## **Chapter I. Introduction**

Data Science is a set of methodologies for taking in thousands of forms of data that are available to us today, and using them to draw meaningful conclusions. Data is being collected all around us. Data is collected from organization’s people who are working, every like, click, email, credit card swipe, or tweet is a new piece of data that can be used to better describe the present or better predict the future.

Figure 1.1 Data Science process flow

Data science is a systematic process used by data scientists to analyze, visualize, and model large amounts of data. Data science processes help data scientists use tools to find invisible patterns, extract data, and turn information into meaningful and actionable insights for the business. This helps businesses and businesses make decisions that help them retain their customers and make a profit. In addition, data science processes help discover hidden patterns of structured and unstructured raw data. This process helps turn a problem into a solution by treating the business problem as a project. Now let's learn what the data science process is and what steps are involved in the data science process.

There are many opportunities for ‘Data Science’ application development. Some of the main applications of ‘Data Science’ are:

**E-commerce:** E-commerce and retail benefit greatly from data science. Here are some of the ways data science has transformed the e-commerce industry: Data science is often used to identify potential customer bases. Predict products and services using predictive analytics. Data science is also used to identify popular product styles and predict their trends. With data science, businesses optimize their pricing structures for consumers. Fraud detection, a central role in machine learning in the industry, is tuned to find fraudulent merchants and scammers in transactions. Data science is also heavily used in collaborative filtering, which forms the backbone of advanced recommender systems. Using this technique, e-commerce platforms can provide insights to customers based on purchases made by people in the same style as the customer's past purchases.

**Healthcare:** In the fitness-care industry, statistics technology is making brilliant leaps. The diverse industries in fitness-care using statistics technology are clinical Image Analysis, genetics and genomics, drug discovery, predictive modeling for diagnosis, fitness bots or digital assistants.

**Transport:** Another important application of data science is transportation. In the transportation sector, data science is actively working to create a safer driving environment for drivers. It also plays an important role in optimizing vehicle performance and increasing driver autonomy. In addition, data science in the transportation sector is actively increasing its diversity with the advent of self-driving cars. Data science has established a strong position in the transportation industry through extensive analysis of fuel consumption patterns, driver behavior and active vehicle monitoring. Self-driving cars are the hottest topic in the world today.

**Manufacturing:** In the 21st century, data scientists are new factory workers. As a result, data scientists occupy an important position in the manufacturing industry. Data science is widely used in the manufacturing industry to optimize production, reduce costs and increase profits. In addition, with the addition of technologies such as the Internet of Things (IoT), data science has enabled organizations to anticipate potential problems, monitor systems, and analyze consecutive data streams. In addition, the industry can use data science to monitor the energy costs of the and optimize the production of the time.

**Finance**: Data science has played an important role in automating various financial tasks. The financial industry works with data science, just as banks do automate risk analysis. The financial industry needs to automate risk analysis in order to make strategic business decisions. Use machine learning to identify, monitor, and prioritize risks. These machine learning algorithms improve cost efficiency and model sustainability through training a large amount of available customer data.

**Banking:** Banking is one of the biggest uses of data science. Big data and data science have helped banks keep up with the competition. Data science allows the banks to take care of their assets efficiently. In addition, banks can make smarter decisions through fraud detection, customer data management, risk modeling, real-time predictive analytics, customer segmentation and more. Banks also value the lifetime value of their customers. This allows banks to monitor the number of customers. It provides them with some predictions that commercial banks will derive through their customers.

Data Science has a lot of applications. In spite of many applications and advantages Data Science has challenges. Some common challenges are given below:

We know the people miss work for a many of reasons. There are different types of challenges in this area based on these reasons. According to the Bureau of Labor Statistics (BLS), the absence rate from work for full-time employees is 2.9%. The impact of absenteeism is certainly an enormous enough problem that organizations should understand how it affects individual, team and organizational performance.

The absenteeism can affect individual productivity. Simply put, if someone works less, they're likely to be less productive. Employers should consider root causes, which include burnout, disengagement, also as people who may require accommodations, like child care or illness or varies diseases.

Consistent with the Society for Human Resource Management (SHRM)**,** overtime is employed to hide 47 percent of employee absences and associates are seemed to be 29.5 percent less productive when covering for absent employees. Even supervisor productivity is impacted. consistent with SHRM, supervisors spend quite four hours per week handling absences and preparing for adjusting workflow to stay things moving.

Absenteeism can reduce profit margins in two ways. First, increased costs reduce profit margins unless revenues increase. for instance, if organizations are spending extra money on overtime pay and contract workers, direct costs go up and profit margins are likely to shrink.

Second, absenteeism can decrease revenue if employees with specific roles aren't present. Employees who sell services or build and deliver a product like workers in manufacturing, software engineering, consulting or sales - simply have less time to hit their goals when absent, potentially decreasing revenue.

**1.1 Background and context**

Absence from work during normal working hours, resulting in temporary to execute regular working activity is said to be absenteeism. In this thesis paper we will look forward to exploring the relationship between the general information about the employees which information are stored in the organization database and the reasons for the absence and to predict their absenteeism rate with maximum level of accuracy. By finding this relationship, the organization gains a highly competitive advantage tool that could be used to address the consequences of the employee’s absence and help employee management to improve the process of recruitment and crisis management.

There are many different paths of exploring organization’s data but in this study, we used classification to analyze and explore our data and also to help in our decision-making prediction. This classification is an aspect of supervised machine learning in which we train different models with classified data and after the models have been trained, we test them with a few percentages of our collected data by using the models to predict new classes and determine the accuracy of the models.

These models predict the absenteeism with high accuracy. Logically here come some additional questions based on what information should we predict whether an employee is expected to be absent or not? How would we measure absenteeism? Now what we can say, should we rather think about trying to predict excessive absenteeism? We will answer these and other questions as we proceed with our analysis for the moment. Just a remember that as a whole we will be to explore whether a person presenting certain characteristics is expected to be away from work at some point in time or not. If it happens then how to improve this situation and making perfect decision using machine learning technique.

**1.2 Problem Statement**

The problem is that the business environment of today is more competitive than it used to be. This Leads to increased pressure in the workplace. Therefore, it is reasonable to expect that unachievable business goals and major risk of unemployment. It raised stress, anxious levels at high. Often the continuous presence of such factors becomes harm to a person’s health. Sometimes this may result in minor illness which of course in not desired. It can happen that the employee can suffer a long-term condition, like depression and other health issue. we will look at predicting absenteeism from work. More precisely we would like to know whether or not an employee can be expected to be missing for a specific number of hours in a given work day and improve the decision making for recognize the work manner.

**1.3 Research Question**

* should we predict whether an employee is expected to be absent or not?
* How would we measure absenteeism?
* Can it measure accurately?
* should we rather think about trying to predict excessive absenteeism?
* Is this functionality being very complex?

**1.4 Research Objectives**

* To propose an analysis that can be study of the causes of employee absenteeism in any workplace and improve decision making for reorganize the work manner.
* To recognize the main reasons principal to absenteeism of each employee.
* To invention out the sickness or stress level of the employees and make groups in the organization at the different stage of ages.
* To analyze the employee data using machine learning technique with statistical method like as regression analysis.
* To study of the analysis to take help of quantitative analysis and qualitative analysis with different types of variables and different types of tools.
* To study of the analysis to predict the absenteeism that can help to make improving the decision for any organization.
* To come out with age vs probability for measuring absenteeism in percentages, which age employees are more absent from their workplace.

This thesis is divided in to five chapters. Chapter I describes about the background of Data Science, absenteeism data analysis using supervised machine learning and current situation. Chapter II describes the previous work done in the field of absenteeism analysis in workplace. Chapter III directs the research methodology that describes the proposed model for absenteeism analysis. Chapter IV illustrates the result and discussion of the proposed research. Chapter V concludes the solved problem and result.

## **Chapter II. Literature Review**

There are different authors already have published paper on absenteeism analysis and prediction. Among them three research papers have been described here briefly and some of them also considered as references here. These three papers are “Job Satisfaction and Absenteeism interface in Corporate Sector – A study” (Prof. V.P. Thirulogasundaram, Dr.P.C.Sahu, 2014), “Factors contributing to work-related absenteeism during the COVID-19 pandemic”(Grigore, O. M. 2020).

* 1. Job Satisfaction and Absenteeism interface in Corporate Sector – A study

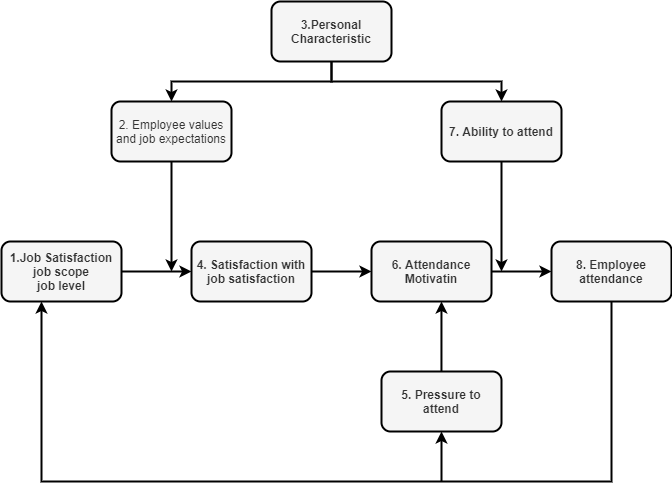
In this paper the author proposed the main objective of this study is to find out the reality of absenteeism of employees on their duty. Impact of motivation and job satisfaction on absenteeism.

Figure 2.1 Block Diagram of Proposed Model

Figure 2.1 shows that Work satisfaction and absenteeism are two human resources requirements that are directly related to and indicate each other's status. Statistically, increasing employee job satisfaction reduces absenteeism. Similarly, if your organization has a high absentee rate, your employees may be less satisfied. Identifying the factors involved and the relationship between job satisfaction and absenteeism has been a challenge for recruiters as long as the job and employer are present.

Researchers define job satisfaction as employee satisfaction in their current job. Factors that influence satisfaction include, among other things, opportunities for professional development, safe working conditions, supportive culture, rewards and benefits. Most employees do not demand integrity for all influential factors, but rather a reasonable level of acceptable compromise. If the general consensus of employees is below average, morale is reduced and common problems with job satisfaction and absenteeism become apparent. Few employees are motivated to pursue a job that makes them feel underestimated, intimidated, stuck, or underpaid. We also often look for legitimate or semi-valid reasons for not going to work.

**The analysis and interpretation of data**

Seven major extrinsic factors of job motivation relative to absenteeism are considered and based on the data obtained from the field survey; each of the sources of job motivation plays a significant impact on absenteeism. This is presented in Table 1

### Table 1: Summarized Result of the Seven Factors of Motivation/Job Satisfaction

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Motivational factors relative to absenteeism** | **Pay (%)** | **Promotion (%**  **)** | **Work Interest (%)** | **Supervision (%**  **)** | **Co- workers (%)** | **Working condition (%** | **Fairness (%)** |
| **Strongly disagree** | 2 | 4 | 0 | 6 | 10 | 2 | 3 |
| **Disagree** | 6 | 1 | 2 | 4 | 14 | 6 | 1 |
| **Undecided** | 10 | 2 | 1 | 1 | 20 | 2 | 1 |
| **Agree** | 24 | 2 | 1 | 2 | 20 | 5 | 2 |
| **Strongly agree** | 58 | 4 | 7 | 5 | 36 | 4 | 5 |
| **Mean** | 4. | 3 | 4. | 4. | 3. | 4 | 4. |

**Source:** Compilation of Primary data

Working conditions refer to such aspects as temperature, lighting, noise and ventilation. 90% (mean 4.2) of the respondents demonstrated that they prefer physical surroundings that are safe, clean, comfortable and with a minimum degree of distractions. With all these in place, job satisfaction is guaranteed and absenteeism placed on a barest minimum threshold.

One factor related to job satisfaction is the extent to which employees perceive that they are being treated fairly. Employees seek for policies and systems that they perceive to be fair as this will likely result in an increase in job satisfaction. Sequel to this, 72% (mean 4.1) of the respondents agree that fairness is a major key that drives absenteeism away in an organization.

This study further explores a link between job dissatisfaction and absenteeism. Result shows that 30% of the respondents strongly agree that absenteeism mean dissatisfaction while 22% agree to this effect. This indicates that over half of the respondents agree that absenteeism means job dissatisfaction. Job dissatisfaction on the other hand results when the basic extrinsic sources of job satisfaction are missing in a particular organization.

**Limitations:** Job satisfaction is not only the reason for absenteeism. In this design Not applied the quantitative and qualitative analysis to get accurate result. In this design does not use maximum numbers of variable.

2.2 Factors contributing to work-related absenteeism during the COVID-19 pandemic

In this paper author proposed perspective the profile of the absent employee, impact on companies, influence of the personal, attitudinal, and organizational factors. Feature of presence policies implemented in the context of the coronavirus pandemic

**Causes of work-related absenteeism**

A wide range of reasons can lead to an employee being absent from work. These causes are often complex and difficult to understand by the organization's managers. A study by Bolton and Hughes (2001) identifies the following reasons as the most common causes of absenteeism from the workplace. Their order is determined as a result of the most common causes mentioned by employees:

**Medical conditions and routine controls**: Minor conditions such as colds, headaches, or back pain can make it difficult to concentrate, so the employee seeks medical assistance. In most cases, employees could come to work after handling the problem with a simple remedy. This can be related to the Theory of attachment (Harvey & Nicholson, 1999) in which it is mentioned that the employees' perception of their health and the severity of the illness can cause the employee to come to work or to absent. Acute medical conditions may require an employee to perform medical examinations or treatments whose side effect prevents the individual from working.

**Accidents**: Regardless of where they happen, during leisure or work, employees will often not be able to do their job properly or even reach the organization's premises.

**Family responsibilities**: Employees may be required to stay at home and take care of a child or an elderly when special situations arise (for example, the person caring for the old is sick or school closures due to weather conditions) or if a child/an elder is ill.

**Alcohol or drug problems**: Drug-dependent workers can experience a temporary inability to concentrate, which could prevent them from going to work. They can also be out of work to avoid a manager detecting their problem, taking refuge in taking frequent sick leave.

**Interview for another job**: Employees can say they do not feel well or leave without leave to attend an interview for a new job, to participate in a job fair, or to work on their CVs.

**Transport conditions and problems:** Extreme weather conditions, such as flooding or heavy snow, can make roads impracticable or isolate employees at home, making it impossible to get to work. Failure of public transport or some strikes can cause delays or absenteeism from work.

**Stress:** Green glass and Burke (2003) defines occupational stress as the difference perceived by the individual between the tasks distributed to him and his ability to carry them through. The stress caused by workplace conditions strongly influences the satisfaction level of the company's employees. The feeling of stress can be worsened by the demands of managers, colleagues, third parties with which the organization works, and the behavior of family members. Over time, the effect of stress and negative emotions leads the employee to absenteeism as a way to recover and restore his physical and mental energy.

**Burnout:** Burnout can result from continuous overloading due to the nature of work or prolonged working hours without rest. Organizations with a reduced number of employees or with employees performing work tasks in more than one area may have a higher incidence of burnout and staff fatigue.

**Participation in training, conferences, or congresses (Nel et al., 2008):** Employees participating in workshops and training courses contributing to their professional development will be excused from work to attend these events. This type of absenteeism is planned and is considered to be absenteeism that generates more benefits than losses.

**Organizational culture:** If the organization encourages a tolerant work environment concerning absenteeism, its employees can see it as a hidden benefit and an opportunity they do not dare to miss.

**Remote work/Virtual office (Mukhopadhyay, 2020):** When working in a virtual office several factors can contribute to absenteeism: outdated equipment and programs, poor internet connection, informal setting, incomprehensible tasks, and low morale from the lack of social interaction

This study focuses on the phenomenon of absenteeism in terms of factors that may cause employees to be absent from work: organizational, attitudinal, and personal (Muchinsky, 1977). It also takes into account absenteeism and presence in the context of the pandemic, as well as how absenteeism management policies have changed to support companies' efforts to ensure the safest possible working environment. The study aims to create a brief profile of absenteeism in companies operating in Iasi in the context of the coronavirus pandemic.

**Limitations:** Describing only age, education and year of experience variables, not used any machine learning technique or any statistical method like regression.

**2.4 Key Debates and Controversies**

A growing economic literature is devoting attention to absenteeism analyzing the effects on worker behavior of a large number of variables (see, among others, Dionne and Dostie, 2007; Barmby, Ercolani, and Treble, 2002). Some of these variables are related to individual characteristics (gender, age, education, health status, etc.), while others are related to contractual and institutional aspects (such as the generosity of sickness benefits, the degree of employment protection, firm size, type of job, labor market conditions, etc.). Since the worker’s effective state of health is typically costly to observe for the employer or for public authorities (and even for qualified physicians), sickness insurance creates a classical moral hazard problem for workers, who, given the prospect of gaining a wage without providing any effort are induced to take days off work. This opportunistic behavior tends to be encouraged by employment protection measures. A number of works have focused their attention on the relationship between firing costs and absence behavior showing that workers on fixed term contracts or on probation, for which contractual arrangements are characterized by less severe firing restrictions, present lower absence rates (Arai and Thoursie, 2005; Ichino and Riphahn, 2005; Scoppa, 2009). Other works have highlighted a positive relationship between firm size and absence rates, which can be explained in relation to the higher monitoring costs faced by larger firms (Winkelman, 1999). Moreover, absence behavior has also been shown to be negatively related to unemployment, since the threat of termination to prevent shirking tends to be related to labor market conditions (Leigh, 1985, Hesselius, 2007). Curington (1994) and Meyer et al. (1995) examine the effects of several legislative changes in benefit levels on absence using US data. They show that increases in these benefits produce an increase of employees’ opportunistic behaviour. Cucchiella and Gastaldi (2006) and Cucchiella et al. (2010) analyses the relation between workforce and risks in supply chain management while Campisi and Gastaldi (1996) show the role of workforce in I-O analysis. An international comparison of the effects of sickness benefits on individual absenteeism is provided by Frick and Malo (2008).

## **Chapter III. Research Methodology**

For this analysis we will use the machine learning technique and statistical methods to get the final result of our research. In this analysis we will use python programming language to train the data set. The research design in this study is analysis and descriptive research. The researcher has used both the primary as well as secondary data. We will use our analysis to take only secondary data as well.

**3.1 Model Design Approach**

In this paper, Analysis and predict the absenteeism and improve decision making for Employees in working place has been proposed. This proposed model is implemented based on a few steps of data science approach such as collecting dataset, data preprocessing, machine learning process, loading the absenteeism module and analyzing the predict outputs in tableau public. Those are describing below:

1. **Collecting Dataset:** This data set collected as secondary data from Kaggle.
2. **Data Preprocessing:** First, we will preprocess the data. We dedicate a momentous amount of time to this step as it is a crucial part of every analytical task. We will start working on the ‘Absenteeism data set’ file and take it to a usable state in a machine learning algorithm.
3. **Machine Learning Process:** This process will incorporate the task we did in the preprocessing section into the code need for making the further step. Here, to develop a model that will predict the probability of an individual being excessively absent from work and machine can provide the decision what action will be taken in future for improving the working environment. In this case study, this will be a logistic regression model. There are many machine learning tools and techniques will co-operate us at this stage of model. At the end, we will store our work as a Python module that we will call ‘absenteeism module’ and will thus preserve it in a form suitable for further analysis.
4. **Loading the absenteeism module:** After that in this section we will load the ‘absenteeism module’ and use its methods to obtain predictions.
5. **Analyzing the predicted outputs in Tableau:** Finally, we will use Tableau to analyses three separate dependencies between the inputs of our model. The visualizations we will obtain with this software will help us a great deal while looking for insights.

**3.1.1 Describing Regression Analysis**

This logistic regression model, as it will be the tool we will implement in our analysis while solving our business task model. Now what is logistic regression analysis? A popular tool in data analytics, machine learning, advanced statistics, and econometrics, is regression analysis. This is an equation which on one side has a variable, called a dependent variable, because its value will depend on the values of all variables we see on the other side. The variables on the right side are all independent, or explanatory. Their role is to explain the value of the dependent variable. There are more terms that can be used for these variables in the same context. The dependent variable can also be called a target, while the independent variables can be called predictors. The data scientists, will call the instructive variables ‘features. Recall that if you prefer, you could call them ‘attributes’ or ‘inputs’ as well. It said that, it is easy to explain what actually logistic regression is about. It is a kind of a regression model whose dependent variable is binary (0,1). That can be the final assume one of two values – 0 or 1, True or False, yes value or No value. Considering the values of all our features, we want to be capable to predict whether the dependent variable will take the value of 0 (zero) or 1 (one).

Here apart from logistic regression, there are many other types of equations that allow us to calculate the dependent variable in a different way. Logistic regression is just one of them – and it is massively used nowadays in data science or machine learning technique.

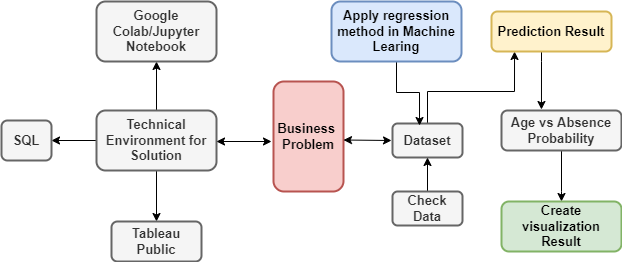


Figure 3.1 Block diagram of proposed model

Figure 3.1 shows the proposed model which will be implemented through based on machine learning and regression analysis. In the diagram we can see there is business problem in an organization. And this problem is employee absenteeism of an organization. This diagram shows the technical environment for solution of the business problem. Then collecting the data sets to apply the machine learning and regression approach to predict the result of the proposed model. There are many different paths of exploring organization’s data but in this study, we used classification to analyze and explore our data and also to help in our decision-making prediction. This classification is an aspect of supervised machine learning in which we train different models with classified data and after the models have been trained, we test them with a few percentages of our collected data by using the models to predict new classes and determine the accuracy of the models. And also predict the age vs probability absent percentages.

**3.2 Proposed Model Requirements**

The diagram of the proposed system demonstrated in Figure 3.1. The Jupyter or Google Colab is used for implemented data model through programming languages. It is a data science and machine learning platform for implementation. MySQL Workbench for data modeling and database operation. Tableu public for visualize the prediction result.

**3.2.1 Hardware Requirements**

Core i5 processor, 120 GB SSD and 8GB RAM of Computer for better performance.



SSD

RAM

Processor

Figure: 3.2 Computer hardware component

**3.2.2 Software Requirements**

**Operating System**: The operating system is available for Windows and Mac. Linux. Operating system "OS" an interface between a computer user and computer hardware. The operating system is software that controls all basic tasks such as file management, memory management, process management, I / O, peripherals and more.



Figure: 3.3 Operating System

**Jupyter Notebook:** The Jupyter Notebook is an open-source web application that allows data scientists to create and share documents that integrate live code, equations, computational output, visualizations, and other multimedia resources, along with explanatory text in a single document.



Figure: 3.4 Jupyter Notebook

**Google Colab:** Colab is a free Jupyter notebook environment that runs entirely in the cloud. Most importantly, it does not require a setup and the notebooks that you create can be simultaneously edited by your team members - just the way you edit documents in Google Docs. Colab supports many popular machine learning libraries which can be easily loaded in your notebook.



Figure: 3.5 Google Colab

**Tableau Public**: Tableau is a powerful and fastest growing data visualization tool used in the Business Intelligence Industry. It helps in simplifying raw data in a very easily understandable format. Tableau helps create the data that can be understood by professionals at any level in an organization. It also allows non-technical users to create customized dashboards. Data analysis is very fast with Tableau tool and the visualizations created are in the form of dashboards and worksheets.



Figure: 3.6 tableau public

**3.3 Flow Chart of Proposed model**

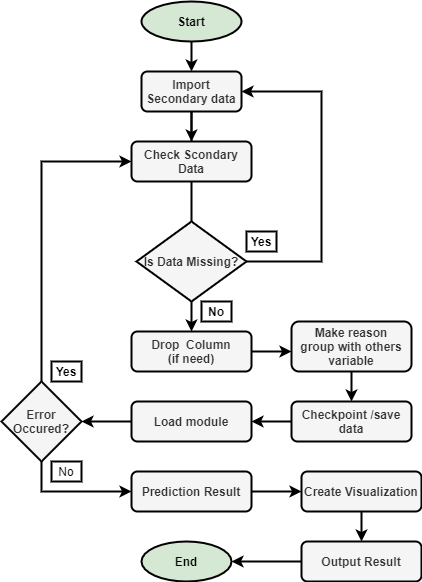


Figure: 3.7 Flow chart of proposed model

Figure 3.7 shows the flow chart of the proposed model. It includes the developed workflow for the applying machine learning technique. When import the data set in google colab or Jupyter notebook it checks the data through python language that whether data is missing or not. If data is missing, then it will again import the data set. If it yes, then it will go next step. Then drop a column from the data set according to out need. Make a reason group with others variables then check point/save the data. After that load the module. If the save module is not working well then it will through error and go to check data section again and if the module is working will then it will be ready to provide prediction result. For this prediction we will use the tableau public software for result visualization age vs probability to find out which age persons are absent from their workplace.

**3.4 Testing**

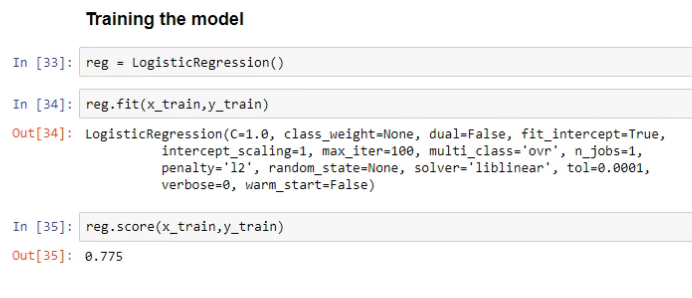
****

Figure: 3.8 train accuracy

Figure 3.8 shows when refering to the model accuracy we meant the train accuracy at this stage overtrain accuracy is around 77% and it is well. In our algorithm has seen this train data many times in fact thousands of times during the training process. So it has learned to model that quite well. Now will use the test data thorugh machine learning technique.

**GET BETTER ACCURACY**

**TWEEK**

Figure: 3.9 Testing circle

Figure 3.9 shows the testing circle of data test. It shows the test the model should be more correct for repeating testing.

Now , Final test on tested data:

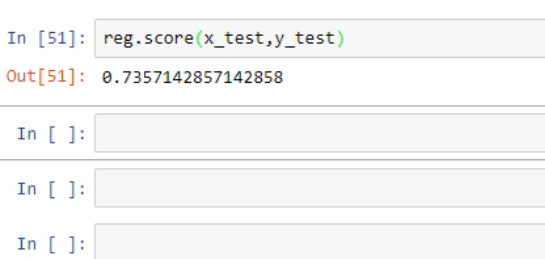


Figure: 3.10 test accuracy

Figure 3.10 in final test out we got different number is **around 74%** rahter than **77%**. Based on data that the model has never seen before. We can say that 74% of the cases the model will predict of a person is going to be excessively absent the test accuracy is always less than the train accuracy. By definition if we get a higher number then we either got lucky or made a mistake. Often it is dramatically it is lower than the train accuracy. This would mean that our model overfit it. It learned the train data very well but is prone to fail in real life with our small percentage diiference between the train and test accuracy.

Now apart from the accuracy we can get the outputs themselves using the predict method.It is also much more usefull for analysis instead of 0 and 1 we can get the probability of an output being 0 or 1. This will give us the probabilities of excessive absenteeism and this result is much simply 0 and 1 in reality.

Logistic regression models calculate these probabilities in the background.

* If probability is below 0.5, it places a 0.
* If probability is greater than 0.5, it places a 1.

## **Chapter IV. Result and Discussion**

**4.1 Excessive Absenteeism Result**

Table 4.1 age vs avg. absence probability result

|  |  |
| --- | --- |
| **Age** | **Average Absence Probability** |
| 28 | 59% |
| 30 | 27% |
| 31 | 35% |
| 33 | 45% |
| 34 | 50% |
| 36 | 45% |
| 37 | 36% |
| 39 | 30% |
| 40 | 20% |
| 43 | 51% |
| 46 | 57% |
| 48 | 10% |
| 50 | 64% |
| 53 | 8% |
| 58 | 63 |

Tabble 4.1, shows the age vs average absence probabillty result. The table has two columns, one is age and another is aberage absence probability. It show the persons age and excessive abesence result in percentage who are absence from their work place. Example the person whose age is 28 , his average absence probaility is 59% and rest of the data will be described in this way. It will be discussed in details in discussion section.

**Analyzing the predicted output**

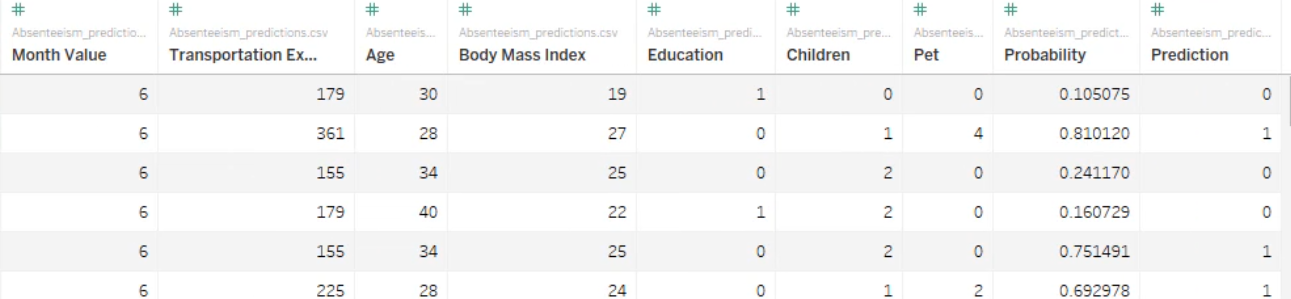
We will do in this section is analyze the abasenteeism model with the help of the Tableau software. Before we discussed about absenteeism model and absenteeism data set but both are not same. We will not analyze only the data but also the predictions of our logistic regression model based on 40 new observations. Thus we can explore the inputs that seem to be most important according to our model(data+predictions). Obtainin such information is key for the management of a company. This allows us to compare how powerfull this model

Figure : 4.1 data load in Tabeau Public

is in general but also with respect to other machine learning models such as neural networks random forests or SVM all right after this short clarification. We have to connect the data source with the Tableau. We would like this to be the csv file conataining the dataset with the predictive outputs. Now we will see that our data will be visualized immediately.

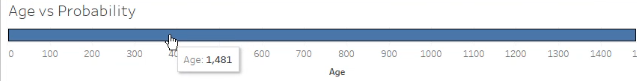
**Result Contd..****Age vs probability:** Now we can proceed by going to open the sheet and it will plot age vs probability.

Figure: 4.2 age data initialization

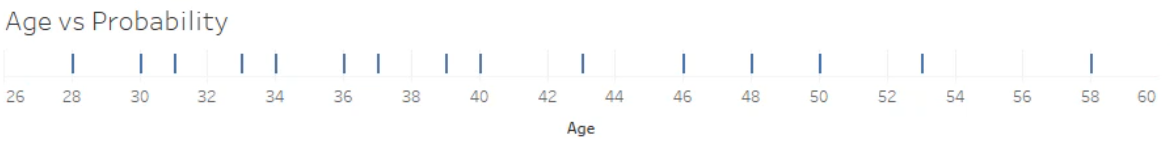
Figure 4.3, we can see a long thick line showing an age of 1481 as we hover above it. The observation age shows a larger number. The reason is that by default Tableau aggregates values that have been the role of measures. These are numeric quantitative values that can be measured and hence aggregating values means that one count them obtain their sum extract their average and more. That’s why Tableau gave us the sum of the ages of all individuals from our table automatically to correct this errorneous representation.

Figure 4.3 selected the dimension

Figure 4.4, we can selected the dimention and we can see the different ages for which we have data. The reason for this is that dimensions and Tableau contain qualitative values that can be categorized or segmented as a result. In this example we see a line mark for each distinct age recorded in our observations.

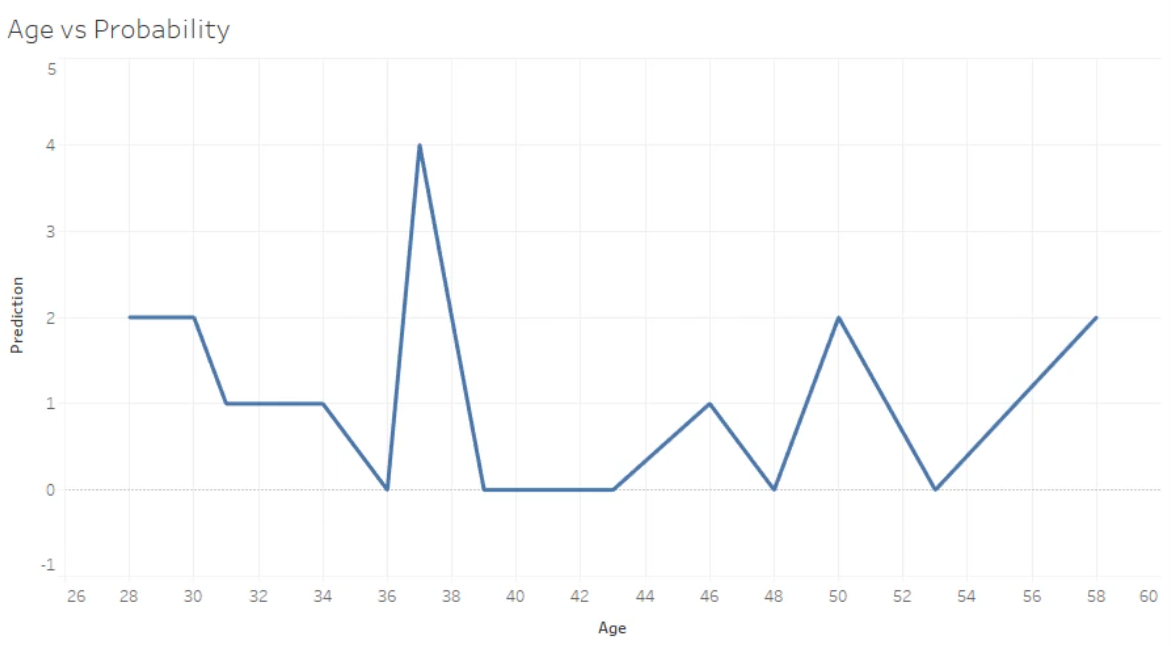


Figure 4.4 add prediction field with the age

Figure 4.5, now we will add the prediction field with the age section, its values will be summed automatically. So as we did with age we must set prediction to be treated as a dimention.

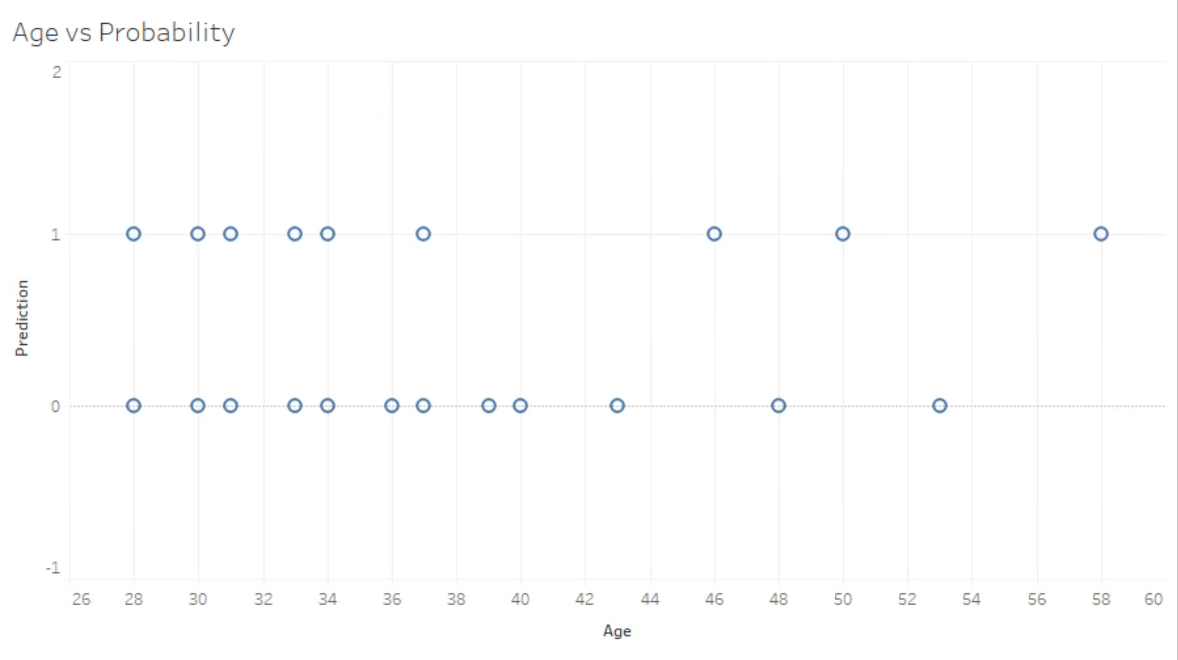


Figure 4.5 shows age vs prediction value

Figure 4.6, now we can see whether our model has predict that and individual will be excessively absent from work prediction value 1 or not prediction vlaue zero. By looking at this graph we can not conclude whether a person above or below a certain age is expected to be excessively absent from their workplace in general. We can not see how many observations are behind every dot in this visualization. A dot could represent 1, 5, 10 or more individuals of the same age.

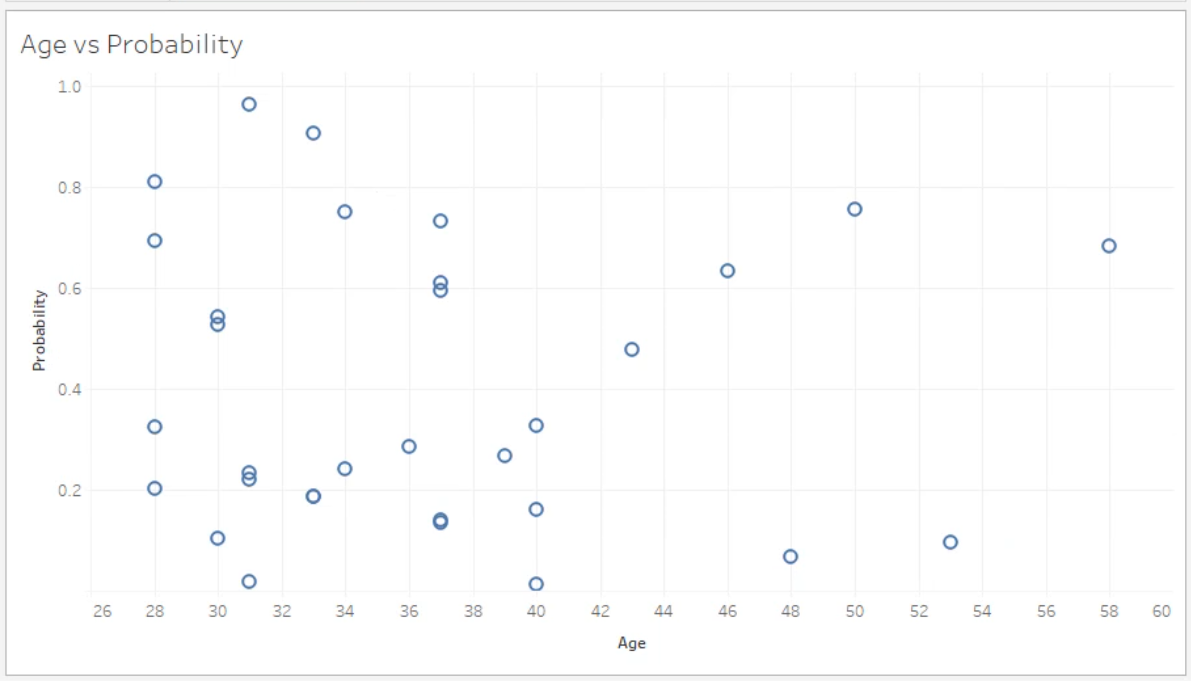


Figure 4.6 shows age vs absence probability

Figure 4.7, we can add probability vlaues instead of prediction values. The probability values are float and their values are between 0 and 1 inclusive and turn into it dimension. Compared to the previous graph this one shows that most of the individuals in our dataset were 40 years old or younger. This became clear only because our visualization now allows us to see multiple observation for every age. The type of graph Tableau for this particular data was shape. It is the one that corresponds to the so-called scatterplot which is a visualization of a set of points plotted on the horizontal and vertical axes.

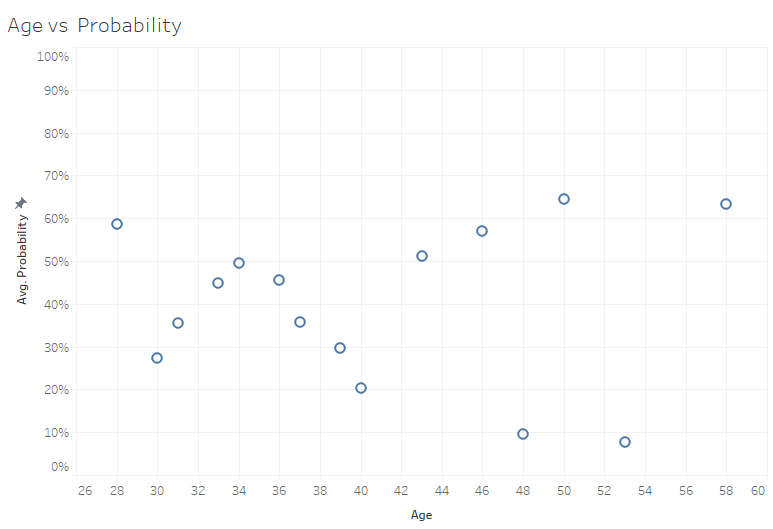


Figure 4.7 age vs avg. absence probability

Figure 4.8, now our task is to plot the average probability value against the feature age. We can turn the probability value back into a measure and designate we want see the average value displayed.If we can hover over every dot and see the age and the average probability value it corresponds to Table 4.1.

**4.2 Discussion**

The model will cover only the small part of the absenteeism analysis under the Data Science. It is implemented for reasons for absence and age vs absence probability through only logistic regression in machine learning technique which is a limitation on our reserch. During the developing model there have some practical considerations that we have faced. In this model we have applied logistic regression in machine learning through python programming language, some times it creates a lack because of heavy library package we have used in this model. We have used numpy library, pandas library, mathplot library, google colab which is connect through internet, also local computer where some issues can be occurred, so some complexity is there as well. Developing this model there was another complexity that we have faced during testing the data model to find out the train accuracy. The model accuracy we meant the train accuracy at this stage overtrain accuracy is around 77% and it is well. In our algorithm has seen this train data many times in fact thousands of times during the training process. So it has learned to model that quite well. Now will use the test data thorugh machine learning technique. in final test out we got different number is around 74% rahter than 77%. Based on data that the model has never seen before. We can say that 74% of the cases the model will predict of a person is going to be excessively absent the test accuracy is always less than the train accuracy.

By definition if we get a higher number then we either got lucky or made a mistake.Now we found it is lower than the train accuracy.This is means that our model is perfect for fit the find out result.

So, we can say our data model is learned the train data very well but is prone to fail in real life with our small percentage difference between the train and test accuracy.After that when we load the data model in tableau public for visualize the result that time we identify age vs prediction and age vs average absence probability percentage for the employee which is excessively absence from workplace.

Here we analysis and predict the absenteeism also implemented the absence reason and tried to implemented age vs absence probability. Our data model has future direction that wil have more analysis and predict the result accordingly reason vs probability and also transportation expense vs probability that the challenges we need to be overcome in the future.

## **Chapter V. Conclusion**

Already there have been done a lot of researches done on Absenteeism analysis the field of Data Science. Data science is a field of study that combines subjectable knowledge, programming skills, and knowledge of mathematics and statistics to derive meaningful insights from data. Data science practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to build artificial intelligence (AI) systems that perform tasks that normally require human intelligence. These systems provide insights that analysts and business users can turn into tangible business value.

In this work we tried to implement a data model that gives the result reasons of absence and tried to measure excessive absenteeism rate in percentage with the logistic regression in machine learning process. Where we find out the age vs average probability to measure this absenteeism.

Absenteeism is defined as expected absenteeism and represents a decline in a company's productivity and quality of work. High market competitiveness, professional development combined with organizational development, and the pressure to reach even more ambitious goals are on employees and ultimately on the job.

This proposed model will work in google colab. And also possible with the jupyter notebook, tableau public software. Here google colab used for data preprocessing and training the data model that we want to analysis in the purpose of our research and also same thing can be done with jupyter notebook. After all it will be loaded on tableau public software for data result visualization in graphs or charts etc. In our thesis our main concern to correctly measure the age vs average probability to be absent from workplace. According to the trainable data model we found the result in percentage which analysis done with respect to prediction values and also probability values.

The developed model gives the model train accuracy with 77% which is in tolerable state. And also give the result of excessive absenteeism in percentage which is in table 4.1. We can use the different machine learning algorithm to improve the accuracy of the model. We can apply this model in organizations to solve the business problems and also can make it better trainable large project in future.

Our proposed work will help strengthen the theory of model, challenge current assumptions, or create a basis for future research. Whenever the researchers will study our paper, it will permit the reader to evaluate them critically. Our proposed system connects the researchers to existing knowledge’s as well. It is guided by this relevant theory. this theory helps the researchers to identify the limits to those generalizations. A theoretical framework specifies the key variables which influence the phenomenon of interest. It highlights the need of examine how the key variable might differ under what the circumstances are happening. the model accuracy we meant the train accuracy at this stage over train accuracy is around 77% rather than 74 % test accuracy and it is well. In our algorithm has seen this train data many times in fact thousands of times during the training process. Finally we visualize the result which is shown the Table 4.1 identifies the age vs absence probability.

## **References**

Cucchiella, F., Gastaldi, M., & Ranieri, L. (2014). Managing absenteeism in the workplace: the case of an Italian multiutility company. *Procedia-Social and Behavioral Sciences*, *150*, 1157-1166.

Grigore, O. M. (2020). Factors contributing to work-related absenteeism during the COVID-19 pandemic. *Management Dynamics in the Knowledge Economy*, *8*(4), 401-418.

Gouws, W. J. S. (2015). *The relationship between absenteeism and physical workplace conditions at Tshepong mine* (Doctoral dissertation, University of the Free State).

Omari, G.C., Manyele, S.V. and Mwaluko, G., 2019. Analysis of Measured Employees’ Absenteeism in the Forensic Science Laboratory. Engineering, 11(3), pp.137-166.

Thirulogasundaram, V. P., & Sahu, P. C. (2014). Job satisfaction and absenteeism interface in corporate sector–A study. *IOSR J Humanities Social Sci*, *19*(3), 64-68.

VanWormer, J. J., Linde, J. A., Harnack, L. J., Stovitz, S. D., & Jeffery, R. W. (2012). Weight change and workplace absenteeism in the HealthWorks study. *Obesity facts*, *5*(5), 745-752.