Prediction of Employee attrition using Gradient Boosting Algorithm

ABSTRACT
In today's organizational landscape, predicting employee attrition has emerged as a critical concern. The departure of trained, technical, and pivotal staff members poses significant challenges, including financial setbacks incurred in their replacement. To address this, organizations harness current and historical employee data to discern prevalent attrition triggers. Employing established classification methodologies such as Decision Tree, Logistic Regression, Random Forest, Support vector machine and Gradient boosting Algorithms are constructed using human resource data. Leveraging feature selection techniques, these models facilitate proactive measures to mitigate attrition risks. By accurately forecasting attrition, companies not only fortify their workforce stability but also enhance economic resilience through diminished human resource expenditures. This proactive approach not only aids in retaining valuable talent but also fosters sustainable growth by fostering an environment conducive to employee retention and organizational stability.

Keywords: Employing attrition, Machine learning, Gradient Boosting, Prediction, Feature selection

1. INTRODUCTION

Employee retention is a multifaceted issue influenced by various factors such as workplace environment, gender and pay equity, and personal circumstances like family relocation or health concerns [1]. Attrition poses significant challenges for organizations, especially when skilled employees depart for better opportunities. This leads to financial losses and disrupts workflow. Analysis present and historical employee data helps identify patterns and underlying causes of attrition, enabling proactive measures for retention. Understanding the complexities surrounding employee departures is crucial for implementing effective strategies to foster a supportive work environment and mitigate turnover, ultimately ensuring organizational stability and success [2].

Identifying employee attrition aids in predicting and addressing retention issues. Utilizing this data allows organizations to mitigate attrition rates and implement strategies for employee retention.

In our analysis, we employ various data classification methodologies to effectively address employee attrition. Decision Tree, characterized by its tree structure comprising branches, root node, and leaf nodes, facilitates attribute testing to determine class labels. Naive Bayes, leveraging Bayes Theorem, assumes independence among features, allowing for classification based on individual feature probabilities [3]. For instance, in classifying fruits, features like colour, shape, and size contribute independently to the probability of a fruit being identified as an apple. Logistic Regression, a statistical technique, evaluates datasets with independent variables to predict outcomes. By utilizing these methodologies, we can accurately identify patterns and factors contributing to employee attrition, enabling organizations to implement targeted retention strategies[4]. This comprehensive approach empowers decision-makers to proactively mitigate attrition rates and foster a supportive work environment conducive to employee retention and organizational success.

System architecture, the backbone of any software solution, undergoes meticulous design phases crucial for translating requirements into tangible systems. The design phase initiates the transformation from problem domain to solution domain, ensuring alignment with system requisites[5]. Its significance lies in its profound impact on software quality, influencing subsequent stages like testing and maintenance. Central to the design phase is the creation of a blueprint document outlining the system's structure, akin to an architectural plan guiding implementation, testing, and maintenance endeavors. This phase typically bifurcates into both detailed and conceptual system design[6].

Detailed Design, is known as high-level design, delineates the system's modules, their specifications, and their interrelations to achieve desired outcomes. Conceptual System Design delves deeper into the internal logic of each module outlined in System Design. While System Design focuses on module identification, Detailed Design delves into logic planning for each module [7].

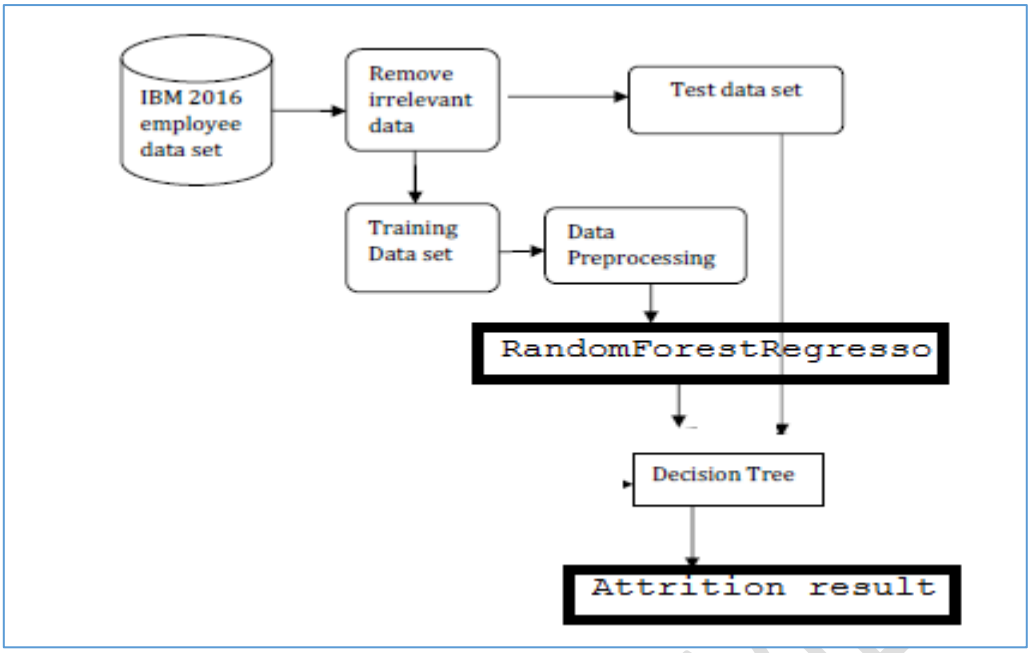

Figure 1. System Architecture

The following paper provide the literature review in section 2. Section 3 present the methodology. Section 4 stated that results and analysis. Final section conclude the paper.

2. LITERATURE REVIEW

Various software development tactics focus on the entire life cycle of systems development (the development life cycle), which includes the full process in creating or redesigning systems[8]. It involves systematic stages from initiation to maintenance, guiding engineers through planning, design, implementation, testing, and deployment. SDLC ensures structured, efficient development and optimization of systems[9].

Predictive analysis in current mechanisms often employs dictated data mining techniques. Knowing that it is prohibitively costly and challenging to hire back qualified workers, employee attrition exerts a significant financial, operational, and resource burden on firms, underscoring the necessity for innovative retention strategies[10].

Standard attrition identification advances by employing unstructured text data, which boosts scientific performance. using determining the presence of potential customers with greater dropout risk, this study aids those with marketing authority make

intelligent decisions that will increase the sustainability of client tactics and profitability [7].

3. METHODOLOGY

3.1 Language/Technology used for Implementation

3.1.1 Python

Python to be one of the premier languages available and why it serves as an excellent choice for beginners venturing into programming.
Here are some key features of Python:

- It is an open-source general-purpose language.
- Python supports multiple programming paradigms including Object Oriented, Procedural, and Functional.
- It offers ease of interfacing with languages such as C, Objective-C, Java, and FORTRAN.
- While interfacing with C++ can be a bit challenging, it's made easier via tools like SWIG.
- Python provides a superb interactive environment for programming.

Python itself is characterized by being a high-level, interpreted, interactive, and object-oriented scripting language. Notably, it is designed to prioritize readability, employing English keywords frequently in contrast to other languages which heavily rely on punctuation. Additionally, Python boasts a more streamlined syntax compared to many other programming languages[11].

• **Python is Interpreted** - Python code is executed at runtime by the interpreter, eliminating the need for compiling before execution. This characteristic is shared with languages like PERL and PHP.

• **Python is Interactive** - Users can engage directly with the Python interpreter, writing and executing code interactively from a Python prompt.

• **Python is Object-Oriented** - Python embraces the Object-Oriented programming paradigm, allowing for code encapsulation within objects.

• **Python is Beginner-Friendly** - Python serves as an excellent language for novice programmers, facilitating the development of diverse applications ranging from basic text processing to web browsers and games.

3.2 Algorithms

3.2.1 Logistic Regression

Logistic Regression is a statistic method used which analysis the relationship between dependent variables and one or more independent variables.

3.2.2 Support Vector Machine

A Support Vector Machine is a supervised machine learning algorithm that finds the hyper plane that separates the data into different classes. The hyperplane is choose to maximize the margarine, which is distance between the hyperplane and the nearest data points.

3.2.3 Decision Tree

A Decision Tree is a Supervised Machine Learning Algorithm that predicts a target variable based on series of decision rules. It is constructed by splitting the dataset's into subsets based on the value of a feature. So that each subset is as homogeneous as possible with respect to the target variable. The Resulting Tree can be used to make predictions for new data points [8].

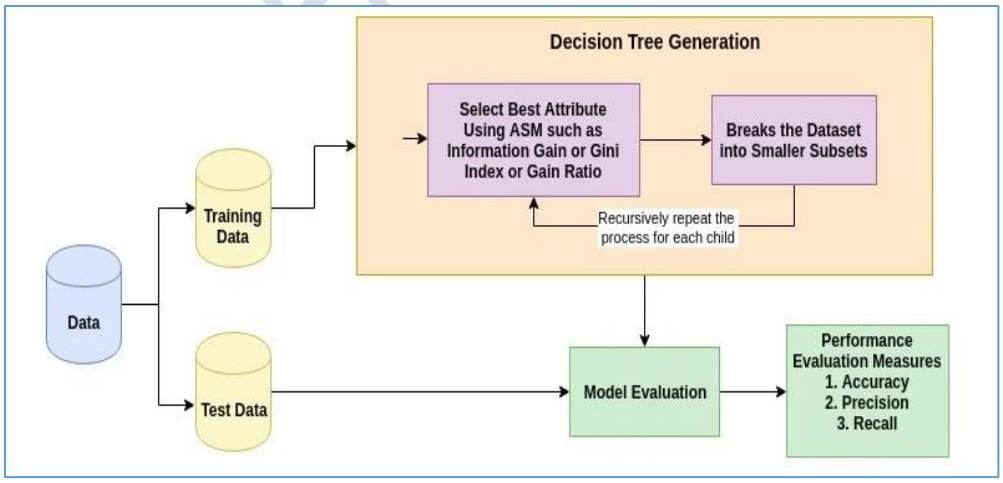

Figure 2. Decision tree Architecture

3.2.4 Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. In the domain of machine learning, it can be applied to problems with both regression and classification. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model [9].

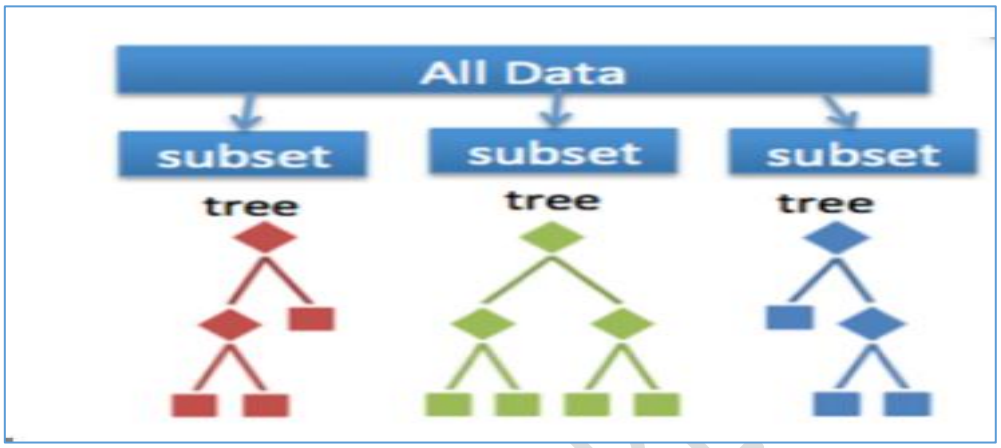

Figure 3. Random Forest tree Architecture

3.2.5 Gradient Boosting Algorithm

Gradient Boosting is a formidable boosting handle that turns consecutive weak learners into strong learners. Each new model is trained using gradient descent to minimize the loss function, such as average squared error or entropy cross of the prior model. According to the current ensemble's predictions, the algorithm calculates the gradient of the loss function in every loop. It then trains a new weak model to minimize this gradient. After adding the new model's predictions to the ensemble, the procedure goes through until a stopping specifications is established [10].

4. RESULTS AND ANALYSIS

Employing established classification methodologies such as Decision Tree, Logistic Regression, Random Forest, Support vector machine and Gradient boosting Algorithms are constructed using human resource data. Leveraging feature selection techniques, these models facilitate proactive measures to mitigate attrition risks. By accurately forecasting attrition, companies not only fortify their workforce stability but also enhance economic resilience through diminished human resource expenditures. This proactive approach not only aids in retaining valuable talent but

also fosters sustainable growth by fostering an environment conducive to employee retention and organizational stability.

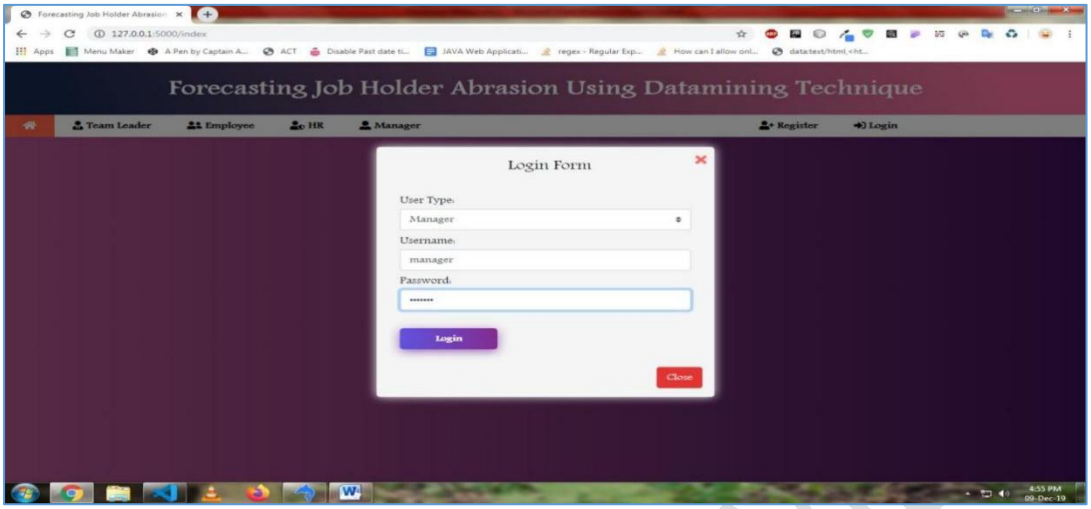

Figure 4. Login page

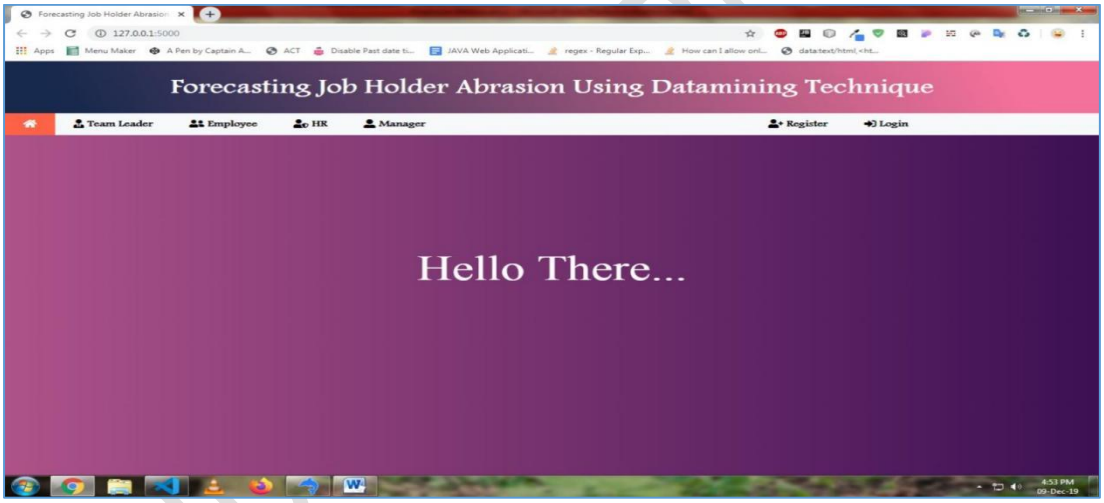

Figure 5. Message display page

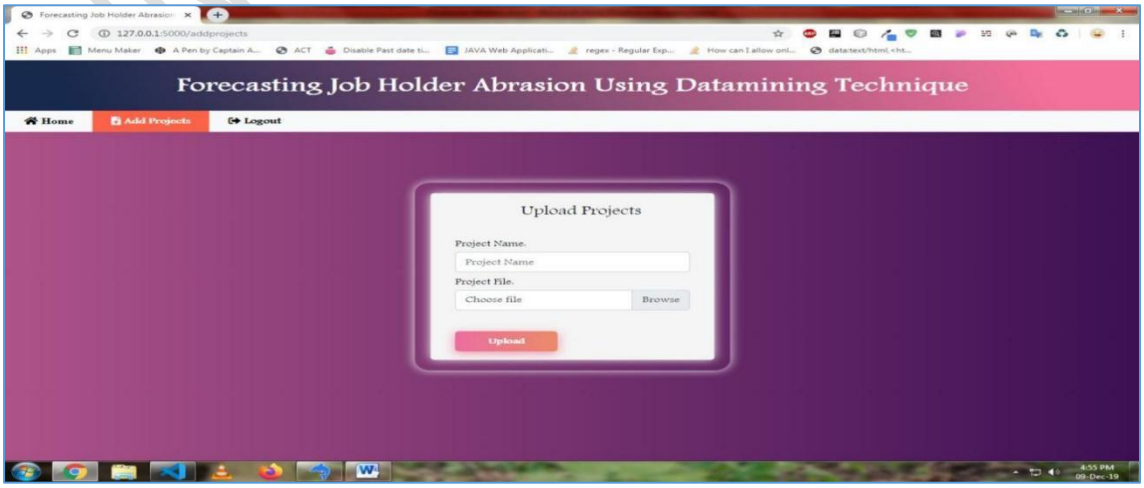

Figure 6. Given credentials

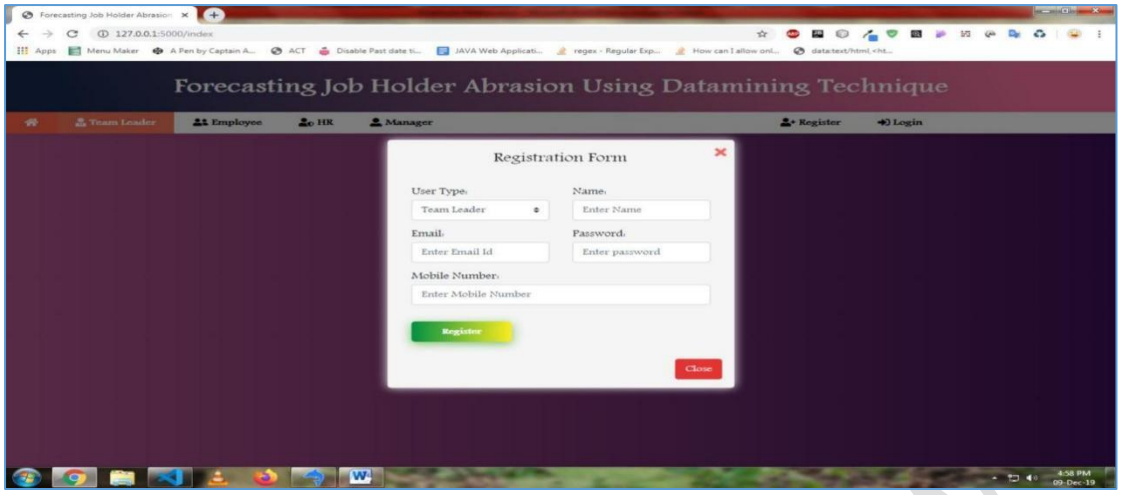

Figure 7. Fill Registration form

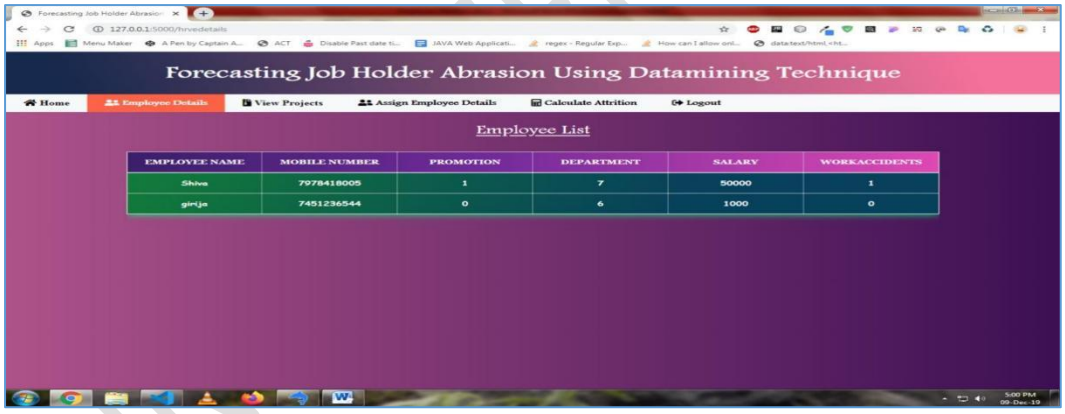

Figure 8. Employee list

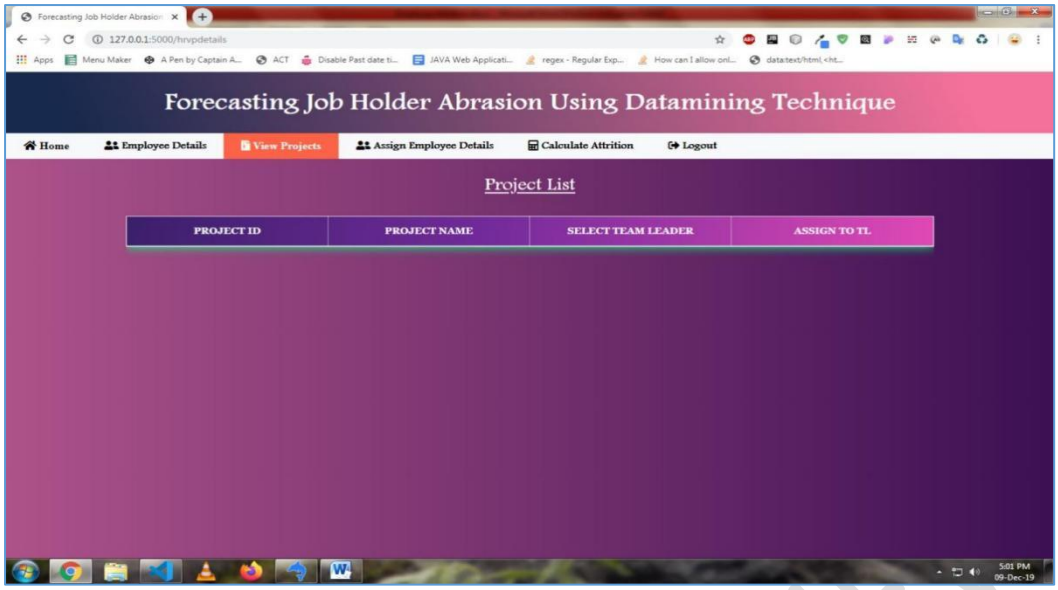

Figure 9. Project List

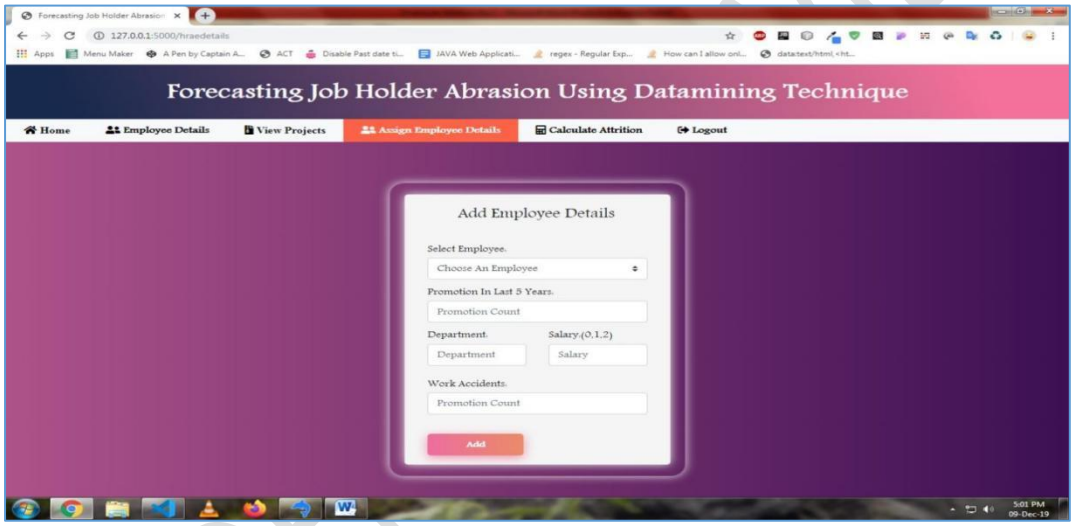

Figure 10. Adding Employee details to form

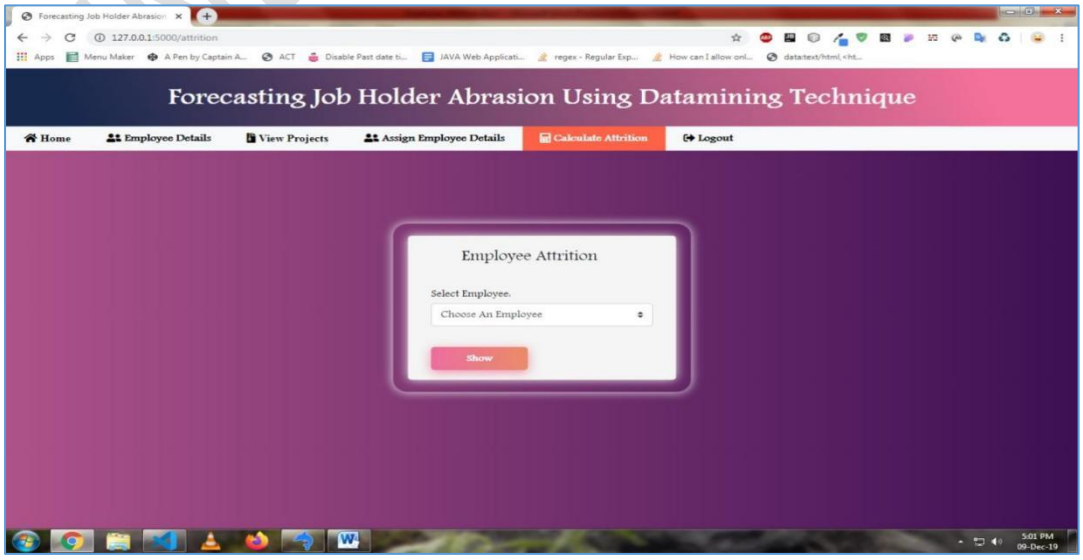

Figure 11. Employee attrition

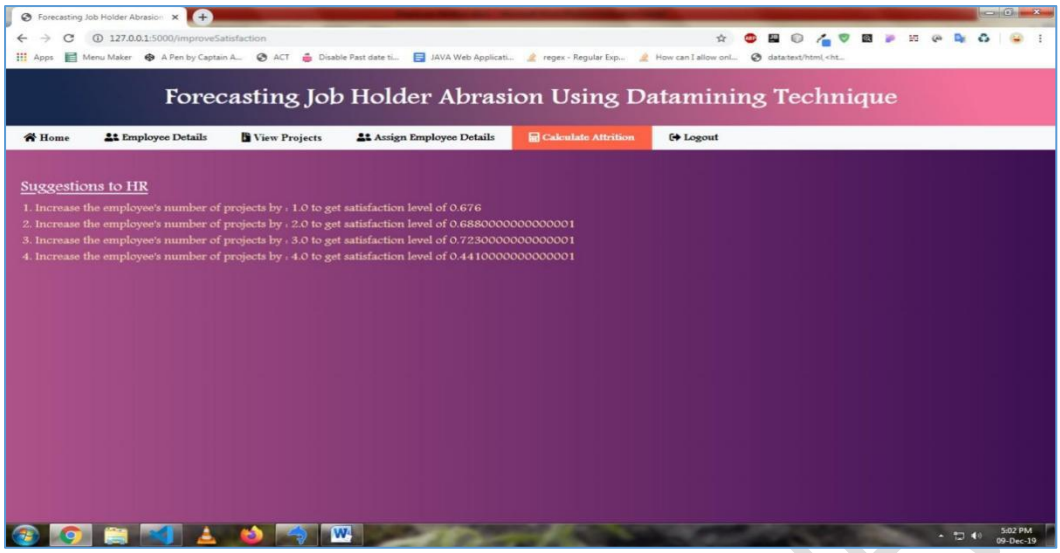

Figure 12. Suggestion to HR

5. CONCLUSION

Employee attrition poses significant challenges for organizations, leading to financial, temporal, and effort-related losses. The departure of trained and experienced personnel creates gaps that are hard to fill swiftly and economically. Recognizing its impact, organizations increasingly rely on analyzing past and current employee data to predict future attrition and understand turnover drivers. Through data extraction algorithms, accurate predictive models for attrition emerge, aiding in proactive retention strategies. This process extends beyond merely distinguishing potential leavers from stayers. Utilizing advanced data analysis techniques enables organizations to assign attrition probabilities to individual employees, facilitating targeted retention efforts and fostering a more stable workforce.

Future enhancements for the system entail continual adaptation to evolving user needs and technological advancements. Upgrades will ensure adaptability to changing environments and emerging technologies. Enhanced security measures, such as implementing single sign-on solutions, will address future security concerns effectively. Regular updates and advancements will maintain the system's relevance and functionality in an ever-changing landscape, ensuring optimal performance and user satisfaction.

REFERENCES

[1] S. Kaur and R. Vijay, "Job Satisfaction – A Major Factor Behind Attrition or Retention in Retail Industry", Imperial Journal of Interdisciplinary Research, vol. 2, no. 8, 2016.

[2] Subhani Shaik,"Analysis of Photo Plethysmography Signals with Artificial Neural Networks Using Curvelet Transform", Jour of Adv Research in Dynamical & Control Systems, Vol. 11, No. 1, 2019.

[3] Shiva Keertan J and Subhani Shaik," Machine Learning Algorithms for Oil Price Prediction", International Journal of Innovative Technology and Exploring Engineering, Volume-8 Issue-8, 2019.

[4] KP Surya Teja, Vigneswar Reddy and Subhani Shaik," Flight Delay Prediction Using Machine Learning Algorithm XGBoost", Jour of Adv Research in Dynamical & Control Systems, Vol. 11, No. 5, 2019.

[5] Dr. Sunil Butada and Subhani Shaik, "IPL Match Prediction using Machine Learning", IJAST, Vol.29, Issue 5, April-2020.

[6] Dr. R. Vijaya Kumar Reddy, Dr. Shaik Subhani, Dr. G. Rajesh Chandra, Dr. B. Srinivasa Rao," Breast Cancer Prediction using Classification Techniques", International Journal of Emerging Trends in Engineering Research, Vol. 8, No.9,2020.

[7] Mr. Sujan Reddy, Ms. Renu Sri and Subhani Shaik," Sentimental Analysis using Logistic Regression", International Journal of Engineering Research and Applications (IJERA), Vol.11, Series-2, July-2021.

[8] Ms. Mamatha, Srinivasa Datta and Subhani Shaik,"Fake Profile Identification using Machine Learning Algorithms", International Journal of Engineering Research and Applications (IJERA), Vol.11, Series-2, July-2021.

[9] R. Vijaya Kumar Reddy, Subhani Shaik, B. Srinivasa Rao," Machine learning based outlier detection for medical data" Indonesian Journal of Electrical Engineering and Computer Science, Vol. 24, No. 1, October 2021.

[10] Dr. Vijayalakshmi K and Dr. Subhani Shaik, "predicting employee attrition methodologies of K-fold technique" I. J. Mathematical Sciences and Computing, March2023, 1, 23-36.

[11] Subhani Shaik, P. Santhosh Kumar S. Vikram Reddy K. Sai Srinivas Reddy and Sunil Bhutada," Machine Learning based Employee Attrition Predicting", Asian Journal of Research in Computer Science, Volume 15, Issue 3, Page 34-39, March, 2023.