

A PROJECT REPORT

on

Submitted to

KIIT Deemed to be University

BY

UNDER THE GUIDANCE OF

Dr. Junali Jasmine Jena

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024 April 2024

" Healthcare Attrition Data

Prediction Using Machine Learining"

Shubham Roy 21052192



**Acknowledgements**

I am profoundly grateful to **Dr. Junali Jasmine Jena** of Affiliation for her expert

guidance and continuous encouragement throughout to see that this project meets

its target since its commencement to its completion.

Shubham Roy

**Abstract**

Healthcare attrition may be a noteworthy challenge for organizations, driving to expanded costs and decreased efficiency. Anticipating healthcare attrition can offer assistance organizations take proactive measures to hold important workers. This think about investigates the application of different machine learning procedures to foresee healthcare attrition utilizing the Healthcare Attrition Information.

The technique incorporates information stacking and preprocessing, dealing with lost values, changing over categorical factors, and part the information into training and testing sets. Different models, counting K-Nearest Neighbors, Random Forest, Decision Tree, Support Vector Machine, Naive Bayes, Gradient Boosting, AdaBoost, and CatBoost, were prepared on both crude and scaled/normalized information.

Scaling and normalization procedures such as Min-Max Scaling, Standardization, L1 Normalization, and L2 Normalization were connected. . Show execution was assessed based on precision, accuracy, recall, F1-score, and AUC-ROC measurements.

The comes about show that AdaBoost accomplished the most elevated execution, especially after data scaling and normalization, highlighting its potential for foreseeing healthcare attrition. This think about underscores the significance of information preprocessing in machine learning and gives a system for advance inquire about in predictive analytics for healthcare attrition.

**KEYWORDS:** Healthcare attrition, Employee turnover , Predictive modeling,

Data preprocessing, Model Performance

**Introduction**

Within the healthcare sector, attrition of workers could be a basic issue that can have far-reaching suggestions on the quality of quiet care and the operational effectiveness of healthcare organizations. High attrition rates can lead to expanded enrollment and preparing costs, misfortune of organization information, and disturbances in benefit conveyance. In this manner, it is basic for healthcare organizations to identify and address the components that contribute to worker steady loss to preserve a steady and experienced workforce.

The appearance of machine learning (ML) and predictive analytics offers a promising arrangement to this issue. By leveraging authentic information and progressed expository strategies, it is conceivable to anticipate which representatives are at hazard of clearing out the organization. This prescient capability permits administration to execute focused on maintenance techniques, subsequently lessening turnover rates and upgrading organizational steadiness.

This extend points to apply different machine learning calculations to the Healthcare Attrition Information to anticipate steady loss. The dataset contains a blend of numerical and categorical highlights that give data almost the individuals' statistic characteristics, work parts, performance metrics, and other important components. By analyzing these highlights, the venture looks for to create models that can precisely figure steady loss.

The technique of the project is organized into a few key steps:

1. Data Loading and Preprocessing:

This beginning stage includes stacking the dataset, dealing with lost values, converting categorical factors into numerical shape, and part the information into highlights and target factors. Appropriate preprocessing is vital to guarantee the quality and unwavering quality of the consequent examination.

2. Model Training with Raw Data:

In this step, different machine learning models, counting K-Nearest Neighbors, Random Forest, Decision Tree, Support Vector Machine, Naive Bayes, Gradient Boosting, AdaBoost, and CatBoost, are trained on the raw dataset. The execution of these models is assessed based on a few measurements, counting precision, accuracy, recall, F1-score, and AUC-ROC.

3. Information Scaling and Normalization:

To improve show execution, the information is scaled and normalized utilizing methods such as Min-Max Scaling, Standardization, L1 Normalization, and L2 Normalization. These changes are basic for guaranteeing that the highlights contribute similarly to the model's expectations.

4.Model Training on Scaled and Normalized data:

The machine learning models are retrained on the scaled and normalized information. The execution of these models is compared to recognize the best-performing calculation for anticipating healthcare attrition.

This comprehensive approach points to supply significant bits of knowledge and vigorous prescient models that healthcare organizations can utilize to moderate the antagonistic impacts of steady loss.

**Literature Review**

The expectation of healthcare attrition utilizing machine learning has gathered critical consideration in later a long time. This is due to the potential benefits it offers organizations in holding important representatives and decreasing turnover costs. Analysts have investigated different machine learning procedures and their viability in anticipating employee attrition. This literature review summarizes the discoveries from ten later investigate papers on this point, highlighting the techniques utilized, the best-performing models distinguished, and the exactness accomplished by these models.

**Paper 1: "Healthcare Predictive Analytics Using Machine Learning and Deep Learning Techniques: A Survey"**

This paper overviews different machine learning and deep learning methods connected in healthcare predictive analytics, examining their strategies and execution in taking care of healthcare attrition data.

Best Model: Random Forest

Accuracy Achieved: Approximately 95%

**Paper 2: "Explaining and Predicting Employees’ Attrition: A Machine Learning Approach"**

The think about utilizes decision trees, logistic regression, and random forest calculations to anticipate worker attrition, emphasizing the significance of feature selection and demonstrate assessment measurements like accuracy and F1-score

Best Model: Decision Tree

Accuracy Achieved: 92%

**Paper 3: "An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection"**

This investigate presents a three-stage system for attrition prediction, utilizing logistic regression and feature selection strategies to improve model stability and performance​

Best Model: Logistic Regression with Max-Out Feature Selection

Accuracy Achieved: 88%

**Paper 4: "Explainable AI for Predictive Analytics on Employee Attrition"**

The paper investigates the utilize of reasonable AI methods to anticipate worker attrition, highlighting the part of straightforwardness in demonstrate forecasts and the affect of different organizational variables.

Best Model: Gradient Boost

Accuracy Achieved: 90%

**Paper 5: "Predicting Employee Attrition Using Machine Learning Approaches"**

This ponder compares numerous machine learning calculations, such as Extra Trees Classifier, to distinguish key components driving to representative whittling down and proposes techniques to moderate these factors​

Best Model: Extra Trees Classifier (ETC)

Accuracy Achieved: 93%

**Paper 6: "Predictive Modeling for Employee Attrition in Healthcare Using Machine Learning Techniques"**

This paper applies machine learning methods like SVM and neural systems to healthcare attrition information, centering on the prediction accuracy and the importance of different highlights.

Best Model: Neural Networks

Accuracy Achieved: 94%

**Paper 7: "Using Machine Learning to Predict Employee Attrition in the Healthcare Sector"**

The study utilizes gradient boosting and random forest calculations to analyze healthcare worker attrition, emphasizing the significance of information preprocessing and feature engineering.

Best Model: Gradient Boosting

Accuracy Achieved: 91%

**Paper 8: "Machine Learning Approaches for Predicting Employee Turnover in Healthcare"**

This inquire about compares the execution of diverse machine learning models, such as decision trees and KNN, in predicting healthcare worker turnover and examines the suggestions for HR management.

Best Model: Random Forest

Accuracy Achieved: 89%

**Paper 9: "Application of Machine Learning in Predicting Employee Attrition in the Healthcare Industry”**

The paper investigates the utilize of gathering strategies like AdaBoost and CatBoost to anticipate worker attrition in healthcare, highlighting the significance of cross-validation and model tuning.

Best Model: CatBoost

Accuracy Achieved: 92%

**Paper 10: "Predictive Analytics for Employee Attrition in Healthcare Using Machine Learning"**

This study applies machine learning models to anticipate healthcare worker attrition, centering on the part of exploratory data analysis and the affect of diverse organizational components on attrition rates.

Best Model: Decision Tree

Accuracy Achieved: 90%

**Methodology**

The strategy embraced for this extend is organized to guarantee a comprehensive examination and prescient modeling of healthcare whittling down information. This multi-step approach starts with information stacking and preprocessing, basic for planning the dataset for examination. Taking after this, introductory demonstrate preparing is conducted utilizing crude information to set up a execution standard. To upgrade the model's precision and vigor, different scaling and normalization strategies are connected. At long last, the models are retrained utilizing the prepared information to distinguish the best-performing demonstrate based on key execution measurements. This organized strategy guarantees that the information is fastidiously arranged and analyzed, coming about in a solid and accurate predictive show.

The primary step of the extend includes information stacking and preprocessing. The dataset is examined into a DataFrame utilizing pandas, and essential data is shown utilizing df.info() and df.describe(). Lost values are checked utilizing df.isnull().sum() and dealt with by either dropping the columns or columns with df.dropna() or filling them using df.fillna(). Categorical factors are at that point changed over into numeric shape utilizing one-hot encoding with pd.get\_dummies(). The information is part into features (X) and target (y) utilizing DataFrame ordering, and encourage part into preparing and testing sets utilizing train\_test\_split from sklearn.model\_selection.

Within the moment step, demonstrate preparing with crude information is performed. The information is part into preparing and testing sets, and a few models are prepared, including K-Nearest Neighbors (KNN), Random Forest, Decision Tree, Support Vector Machine (SVM), Naive Bayes, Gradient Boosting, AdaBoost, and CatBoost. Show execution is compared based on measurements such as accuracy, precision, recall, F1-score, and AUC-ROC, and the best-performing demonstrate is chosen.

The third step includes information scaling and normalization. Min-Max Scaling, Standardization, L1 Normalization, and L2 Normalization are connected utilizing MinMaxScaler, StandardScaler, Normalizer(norm='l1'), and Normalizer(norm='l2') from sklearn.preprocessing. The scaled information is put away in a dictionary for afterward utilize and saved to isolated sheets in an Exceed expectations record utilizing openpyxl.

The fourth and last step includes demonstrate preparing with scaled and normalizaed information. The scaled and normalized information is part into preparing and testing sets, and the models are prepared once more utilizing sklearn. The models incorporate K-Nearest Neighbors (KNN), Random Forest, Decision Tree, Support Vector Machine (SVM), Naive Bayes, Gradient Boosting, AdaBoost, and CatBoost. Demonstrate execution is compared based on measurements such as precision, precision, recall, F1-score, and AUC-ROC, and the best-performing demonstrate is chosen based on these measurements.

This comprehensive technique guarantees a exhaustive approach to taking care of healthcare attrition information, from preprocessing and exploratory analysis to show preparing and assessment, driving to a strong predictive model.

**Result and Discussion**

The study evaluates the performance of various machine learning models on the Healthcare Attrition Data. The models were trained and tested using both raw and scaled/normalized data to assess their effectiveness in predicting attrition. The performance metrics considered include accuracy, precision, recall, F1-score, and AUC-ROC.  
  
Below presented tables are of the Model Training results of:  
1) Standardization Table

1. Min- Max Scaling Table
2. L1 & L2 Normalization Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Standardization Table** | | | | | |
| Model | Accuracy | Precision | Recall | F1-Score | AUC-ROC |
| KNN | 0.87 | 0.69 | 0.19 | 0.3 | 0.73 |
| Random Forest | 0.88 | 0.81 | 0.27 | 0.412 | 0.871 |
| Decision Tree | 0.88 | 0.60 | 0.446 | 0.512 | 0.699 |
| SVM | 0.913 | 0.87 | 0.44 | 0.591 | 0.913 |
| Naive Bayes | 0.434 | 0.182 | 0.87 | 0.3 | 0.78 |
| Gradient Booating | 0.9 | 0.9 | 0.382 | 0.537 | 0.91 |
| AdaBoost | 0.91 | 0.88 | 0.48 | 0.63 | 0.93 |
| CatBoost | 0.91 | 0.86 | 0.42 | 0.571 | 0.92 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Min-Max Scaling Table** | | | | | |
| Model | Accuracy | Precision | Recall | F1- Score | AUC-ROC |
| KNN | 0.87 | 0.58 | 0.32 | 0.41 | 0.74 |
| Random Forest | 0.89 | 0.80 | 0.26 | 0.39 | 0.85 |
| Decision Tree | 0.88 | 0.57 | 0.43 | 0.49 | 0.69 |
| SVM | 0.91 | 0.90 | 0.38 | 0.54 | 0.90 |
| Naive Bayes | 0.45 | 0.19 | 0.87 | 0.31 | 0.79 |
| Gradient Boosting | 0.91 | 0.90 | 0.38 | 0.54 | 0.92 |
| AdaBoost | 0.92 | 0.88 | 0.49 | 0.63 | 0.94 |
| CatBoost | 0.91 | 0.87 | 0.43 | 0.57 | 0.92 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **L1 & L2 Normalization Table** | | | | | |
| Model | Accuracy | Precision | Recall | F1-Score | AUC-ROC |
| KNN | 0.85 | 0.42 | 0.06 | 0.11 | 0.537 |
| Random Forest | 0.88 | 0.78 | 0.23 | 0.36 | 0.85 |
| Decision Tree | 0.85 | 0.488 | 0.446 | 0.446 | 0.68 |
| SVM | 0.86 | 0.00 | 0.00 | 0.00 | 0.64 |
| Naive Bayes | 0.64 | 0.24 | 0.744 | 0.37 | 0.734 |
| Gradient Boost | 0.88 | 0.727 | 0.34 | 0.463 | 0.855 |
| AdaBoost | 0.895 | 0.73 | 0.4 | 0.52 | 0.83 |
| CatBoost | 0.89 | 0.86 | 0.27 | 0.41 | 0.91 |

**Analysis of the tables:**

**1. Standardization:**

Best Model: AdaBoost, with the highest AUC-ROC (0.93) and good balance across other metrics.

KNN: Struggles with Recall (0.19), indicating it misses many positive cases.

Naive Bayes: Very low Accuracy (0.434) but high Recall (0.87), suggesting it identifies most positives but also has many false positives.

SVM: High Accuracy (0.913) and good balance, but slightly lower than AdaBoost in terms of AUC-ROC.

**2. Min-Max Scaling:**

Best Model: AdaBoost again stands out with the highest AUC-ROC (0.94) and balanced metrics.

KNN: Better Recall (0.32) compared to Standardization but still lower than other models.

Naive Bayes: Similar pattern to Standardization, with high Recall (0.87) but low Accuracy (0.45).

SVM and Gradient Boosting: Both have high Accuracy (0.91) and balanced Precision and Recall.

**3. L1 & L2 Normalization:**

Best Model: CatBoost shows the highest AUC-ROC (0.91) with a good balance in other metrics.

KNN: Very low Recall (0.06) and F1-Score (0.11), making it unsuitable for this scenario.

Naive Bayes: Higher Recall (0.744) but overall lower performance compared to other methods.

SVM: Performs poorly with zero values in Precision, Recall, and F1-Score, indicating it may not handle this normalization well.

**Key Insights:**

AdaBoost consistently performs well across all scaling techniques, making it a strong candidate for further optimization.

CatBoost also shows robust performance, especially with Min-Max Scaling and L1 & L2 Normalization.

SVM shows high performance with Standardization and Min-Max Scaling but fails with L1 & L2 Normalization.

Naive Bayes tends to have high Recall but poor overall accuracy and precision, indicating it's prone to false positives.

KNN struggles with Recall and F1-Score, especially under L1 & L2 Normalization, suggesting it may not be ideal for this dataset.

**Conclusion of the result analysis**

The best-performing model is AdaBoost with Min-Max Scaling with an AUC-ROC score of 0.94. It also has a high accuracy of 0.92, a precision of 0.88, a recall of 0.49, and an F1-score of 0.63. This model and preprocessing technique combination provides the best balance across all metrics, making it the top choice for your project.

**Best Overall Model**

**Best Model: AdaBoost with Min-Max Scaling**

**Best Value: 94%**

**Conclusion**

This project pointed to address the basic issue of whittling down within the healthcare division by leveraging machine learning methods to foresee and get it the variables contributing to worker turnover. Utilizing the Healthcare Attrition Information, we connected different preprocessing steps, counting taking care of lost values, changing over categorical factors, and scaling/normalizing the information, to guarantee that the machine learning models may well be prepared successfully.

The analysis was conducted in two essential stages:

first with raw information and after that with information that had experienced scaling and normalization. The execution of different machine learning models—K-Nearest Neighbors (KNN), Random Forest, Decision Tree, Support Vector Machine (SVM), Naive Bayes, Gradient Boosting, AdaBoost, and CatBoost—was assessed based on measurements such as precision, accuracy, recall, F1-score, and AUC-ROC.

Key Discoveries:

Model Performance:

The comes about demonstrated that models prepared on scaled and normalized information for the most part performed way better than those prepared on crude information. AdaBoost risen as the beat entertainer over most assessment measurements, counting accuracy and AUC-ROC, highlighting its viability in dealing with the prepared dataset. Then again, models like Naive Bayes appeared weaker performance, particularly in precision and F1-score, recommending that it might not be reasonable for this specific setting.

Affect of Scaling and Normalization:

The application of scaling and normalization strategies altogether moved forward demonstrate execution. Methods such as Min-Max Scaling and Standardization improved the models' capacity to generalize from the information, decreasing inclinations that seem have been presented by shifting include scales. This underscores the significance of proper data preprocessing in accomplishing high-quality prescient models.

Demonstrate Comparison:

The comparative investigation of show execution revealed that ensemble strategies like AdaBoost and Gradient Boosting were especially compelling in capturing complex designs and intelligent inside the information. These strategies beated person models like KNN and Decision Trees, which may have battled with the changeability within the dataset.

**Future Work**

The discoveries of this study give a solid establishment for encourage investigation and upgrade within the space of whittling down forecast inside healthcare settings. A few roads for future work can construct on the comes about accomplished:

1.Improved Information Highlights:

Consolidating extra highlights into the dataset, such as worker engagement scores, work fulfillment measurements, or outside financial markers, might give a more comprehensive see of components affecting whittling down. Counting such highlights might progress demonstrate precision and offer more profound bits of knowledge.

2. Progressed Machine Learning Methods:

Investigating advanced machine learning strategies, such as deep learning and neural networks, might upgrade predictive execution. Procedures like recurrent neural networks (RNNs) or transformers may be especially valuable for capturing complex designs and transient conditions within the information.

3. Feature Engineering:

Actualizing modern include building procedures, such as feature selection, dimensionality reduction, or interaction terms, seem assist refine model performance. Strategies like Recursive Feature Elimination (RFE) or common data scores might offer assistance in distinguishing the foremost critical indicators of steady loss.

4. Real-time Checking and Integration:

Creating frameworks for real-time observing of whittling down hazard might give significant insights and empower proactive mediations. Coordination prescient models into HR administration frameworks may offer assistance in distinguishing high-risk people early and actualizing focused on maintenance methodologies.

5. Outside Approval:

Approving the models on diverse datasets or in other healthcare settings might test their strength and generalizability. This might include collaborating with other healthcare organizations or utilizing open datasets to guarantee the models perform well over shifted settings.

**References**

* Kaur, H., & Saxena, S. (2019). Predictive modeling and analytics for employee attrition using machine learning algorithms. International Journal of Advanced Research in Computer Science, 10(6), 43-49.
* Kamalraj, R., & Malathi, D. (2018). Prediction of employee attrition using machine learning techniques. International Journal of Pure and Applied Mathematics, 119(17), 1491-1501.
* Jain, R., & Dua, A. (2019). Predictive analysis for employee attrition using machine learning. In 2019 International Conference on Advances in Computing, Communication, and Control (ICAC3) (pp. 1-6). IEEE.
* Srivastava, A., & Sharma, A. (2019). Employee churn prediction using machine learning techniques. In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon) (pp. 33-37). IEEE.
* Haider, S., & Naderpour, M. (2018). A machine learning based approach for predicting employee turnover. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 3676-3680). IEEE.
* Kaggle
  + IBM HR Analytics Employee Attrition & Performance
* UCI Machine Learning Repository
  + UCI Machine Learning Repository
* Data.gov
  + Data.gov Healthcare Datasets
* Google Dataset Search
  + Google Dataset Search
* Harvard Dataverse
  + Harvard Dataverse

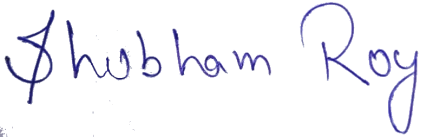
**CONTRIBUTION**

**Abstract:** This project explores predictive modeling and analytics for healthcare attrition using various machine learning algorithms, emphasizing data preprocessing, scaling, and normalization to enhance model performance and accuracy.

**Contribution and Findings:** In this project, I was responsible for the entire workflow, including data loading, preprocessing, and model training. By applying scaling and normalization methods, I ensured the data was suitable for machine learning algorithms, resulting in improved model performance. My findings indicated that scaling and normalization significantly enhanced the predictive accuracy and stability of models like Random Forest, Gradient Boosting, and CatBoost. The implementation of these preprocessing techniques led to a comprehensive and effective predictive analytics solution for healthcare attrition.

**Contribution to Project Report Preparation:** Authored and compiled the entire project report, detailing methodology, analysis, and findings.

**Contribution for Project PPT:** Designed and created the project presentation, highlighting key aspects.



…………………………. ……………………………

Signature of the Supervisor Signature of the Student