**Data Mining for Business (BUDT758T)**

Project Title: **Employee Attrition and Resignation Prediction for Healthcare**

(Be explicit about the explanatory or predictive flavor of your, using terms such as “explaining”, “predicting,” etc.).

Team Members: _ Kyungho Yu_____

__Jaehyun Lee_____

__Sangwon Seo_____

_ Yullie Yang_____

(SIGN THE FOLLOWING STATEMENT AND INCLUDE IT ON THE COVER PAGE OF YOUR PROPOSAL)

## *ORIGINAL WORK STATEMENT*

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

| | **Typed Name** | **Signature** |
|---|---|---|
| Contact Author | | |
| | | |
| | | |
| | | |
| | | |

## II. **Executive Summary**

The healthcare industry is rapidly expanding worldwide, and the role of nurses and related personnel is crucial for delivering high-quality patient care. However, a shortage of manpower can have a direct impact on service quality, making it essential to manage nurse turnover effectively. This report leverages data analysis and predictive modeling techniques to identify the main reasons for nurse attrition and provide actionable insights for hospitals and staffing companies to reduce turnover rates.

The study highlights several key factors contributing to nurse attrition rates, such as over time, age, and total working year. By addressing these factors, companies can improve their service quality management and provide better care for their patients. The recommendations provided in this report can serve as a valuable guide for healthcare industry professionals seeking to reduce nurse turnover rates and improve overall service quality. By leveraging data analysis and predictive modeling techniques, this report offers a practical and effective approach to address the challenges of nurse attrition in the healthcare industry.

## III. **Data Description**

Data source: The data was collected from Nursing Solutions Inc. and made available on Kaggle.
‘**Employee Attrition for Healthcare’**
https://www.kaggle.com/datasets/jpmiller/employee-attrition-for-healthcare?resource=download
This is a dataset with 1676 observations and 35 variables in ‘csv’ file. Here are the variables with a brief description:
Sample size ($n$) : 1676 , number of variables ($k$): 35

- EmployeeID: employee identification number
- Age: employee age
- Attrition: whether or not the employee has left the company (Yes/No)
- BusinessTravel: how often the employee travels for business (Travel_Rarely/Travel_Frequently/Non-Travel)
- DailyRate: daily rate of pay
- Department: the department the employee belongs to
- DistanceFromHome: distance in miles from home to work
- Education: employee's highest education level (1-5: below college degree to doctorate)
- EducationField: field of study for the employee's highest education
- EmployeeCount: always 1 (variable can be dropped)

- EnvironmentSatisfaction: employee's satisfaction level with their work environment (1-4: low to high)
- Gender: employee's gender
- HourlyRate: hourly rate of pay
- JobInvolvement: employee's involvement in their job (1-4: low to high)
- JobLevel: employee's job level (1-5: entry level to executive)
- JobRole: employee's role in the company
- JobSatisfaction: employee's satisfaction level with their job (1-4: low to high)
- MaritalStatus: employee's marital status
- MonthlyIncome: employee's monthly income
- MonthlyRate: monthly rate of pay
- NumCompaniesWorked: number of companies the employee has worked for prior to their current position
- Over18: whether the employee is over 18 (variable can be dropped)
- OverTime: whether the employee works overtime (Yes/No)
- PercentSalaryHike: the percentage of the employee's salary increase from the previous year
- PerformanceRating: employee's performance rating (1-4: low to high)
- RelationshipSatisfaction: employee's satisfaction level with their relationships at work (1-4: low to high)
- StandardHours: standard hours per day (variable can be dropped)
- Shift: the shift the employee works (1-4: day to night)
- TotalWorkingYears: employee's total number of years working
- TrainingTimesLastYear: number of times the employee received training last year
- WorkLifeBalance: employee's satisfaction level with their work-life balance (1-4: low to high)
- YearsAtCompany: number of years the employee has been working for the company
- YearsInCurrentRole: number of years the employee has been in their current role
- YearsSinceLastPromotion: number of years since the employee's last promotion
- YearsWithCurrManager: number of years the employee has been with their current manager

The following variables were dropped from the dataset:

- EmployeeID
- StandardHours
- EmployeeCount
- Over18
- DailyRate
- MonthlyIncome
- MonthlyRate

To remove three variables had the same values on each row, indicating that they did not provide any additional information. The variable "employeeId" was a unique identification factor that did not impact the model's predictions. To substitute the three variables related to income with "hourlyrate", which is a more meaningful representation of an employee's income.

The following variables were treated as categorical:

- Attrition
- BusinessTravel
- Department
- Education
- EducationField
- EnvironmentSatisfaction
- Gender
- JobInvolvement
- JobLevel
- JobRole
- JobSatisfaction
- MaritalStatus
- OverTime
- Shift
- PerformanceRating
- RelationshipSatisfaction
- TrainingTimesLastYear
- WorkLifeBalance

The following variables were treated as numeric:

- Age
- DistanceFromHome
- HourlyRate
- NumCompaniesWorked
- PercentSalaryHike
- TotalWorkingYears
- YearsAtCompany
- YearsInCurrentRole
- YearsSinceLastPromotion
- YearsWithCurrManager

**Sample Observations of data**

| Age | Attrition | BusinessTravel | DailyRate | Department | DistanceFromHome | Education |
|---|---|---|---|---|---|---|
| 41 | No | Travel_Rarely | 1102 | Cardiology | 1 | 2 |
| 49 | No | Travel_Frequently | 279 | Maternity | 8 | 1 |
| 37 | Yes | Travel_Rarely | 1373 | Maternity | 2 | 2 |
| 33 | No | Travel_Frequently | 1392 | Maternity | 3 | 4 |
| 27 | No | Travel_Rarely | 591 | Maternity | 2 | 1 |

Fig[1] Sample Observations of data

**Distinct characteristics of the data set**

This dataset is noteworthy because it doesn't have any missing values, meaning that analysis could be conducted using complete and accurate data. Additionally, the variables in this dataset are straightforward and easy to understand, making it an excellent resource for studying the connections between different factors in the healthcare industry. The interpretation of the analysis would be intuitive and easy to

understand as well. In addition, The dataset provides in-depth information on various aspects of healthcare, offering a unique opportunity to gain insights into the sector's complex workings. Healthcare researchers and practitioners will find the dataset compelling due to its wealth of data on critical healthcare variables presented in a clear format. Its extensive coverage and lack of missing data make it a valuable resource for understanding the healthcare industry.

**III. Research Questions**

In today's competitive job market, high employee turnover rates have become a major concern for hospitals and staffing companies. To address this issue, our research aims to predict employee attrition , identify the primary factors that affect employee attrition and provide practical solutions to reduce turnover rates.

To achieve our research objectives, we will conduct a comprehensive analysis of employee behavior, including job satisfaction, workload, compensation, and growth opportunities. Data will be gathered from various sources, such as surveys, interviews, and databases, and use advanced statistical techniques such as machine learning and regression analysis to identify the root causes of high turnover rates.

Our research findings will provide valuable insights to organizations seeking to improve their recruitment and retention strategies. Based on our analysis, we will offer practical recommendations and solutions to hospitals and staffing companies to address high employee turnover rates. These may include improving compensation packages, offering opportunities for career advancement and professional development, and creating a positive workplace culture that fosters employee satisfaction and engagement.

By adopting these solutions, organizations can not only lower their turnover rates but also attract and retain high-quality employees, leading to better organizational performance and success. Our research will provide a roadmap for organizations to build a happier, more productive workforce and improve their bottom line.

In conclusion, our research will offer valuable insights into the factors that affect employee attrition and provide practical solutions to reduce turnover rates. By improving recruitment and retention strategies, organizations can create a positive workplace culture that fosters employee satisfaction and engagement, leading to better organizational performance and success.

## IV. Methodology

The main focus of this study was to predict whether the nurse would go through attrition using different data mining models. To do so, we implemented four models to build classification model: K-Nearest Neighbors (KNN), Decision Tree, Naive Bayes, and Random Forest. Those four models are the main models used for predicting classification models thanks to their efficiency in terms of accuracy and handling large sample sizes. For the performance metrics, accuracy and specificity was explored.

To start, K-Nearest Neighbors (KNN) is a classification model that is associated with nearest neighbors assigning the label considering the majority class. After comparing similarities, the model assigns class to the new data point.

Next one is the decision tree, which can be applied not only for the classification model but also for the regression model. Similar to a tree structure, the model has each node for attribute and each branch for the outcome. To figure out the class label, we can look at the leaf node.

The third model is the Naive Babes model, which implements the Babes theorem for prediction. The main implication while using this model is to treat each feature as an independent variable.

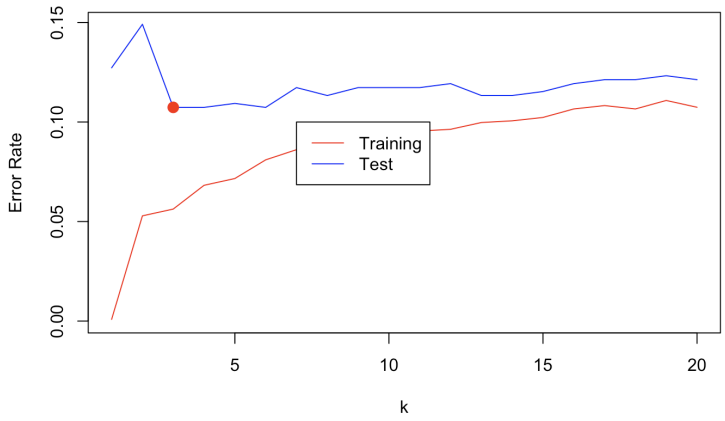
Finally, as a form of ensemble learning method, Random Forest was used for this project's classification. Through constructing a large number of decision trees, the Random Forest model comes out with statistical information of each tree model. Random Forest model is highly recommended while improving accuracy.

With these four data mining techniques, we were able to make classification predicting the nurse attrition. For each model, we began by loading the necessary R packages. Also, setting a seed helped us for fair comparison of the model as well as for reproducibility. The data set was splitted into training data set and test dataset and the size was 70%  and 30% of the dataset, respectively.

## V. Results and Finding

### K-Nearest Neighbors (KNN)

Using the 'knn' function available in R packages, we built the KNN model in R. After observing results  k from 1 to 20, we chose the best value of k that produced the lowest test error rate. Then we used a confusion matrix for comparing the training and test dataset. In the plot, we can see the error rate for both of the training and test data set and the best k was k =3. From k=-3, the confusion matrix, accuracy score and error rate were built. For the KNN model, the accuracy score was 0.8946 and 0.7143 for specificity.

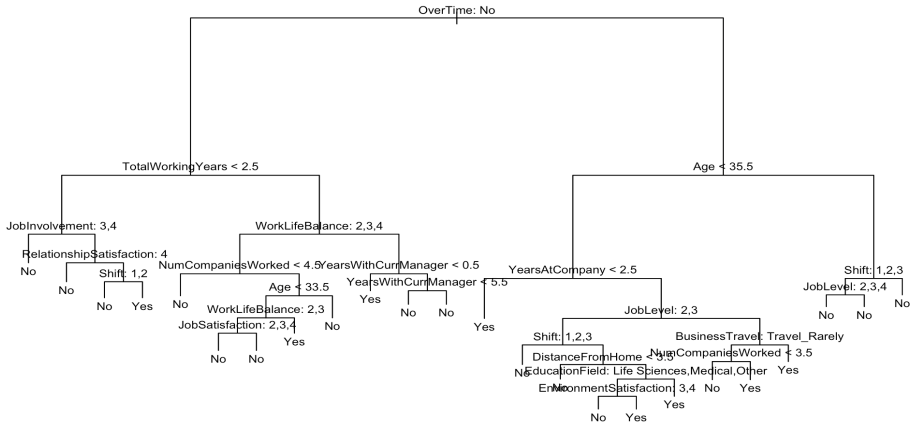



Fig[2] K-Nearest Neighbors (KNN)  Accuracy Plot

## Decision Tree

Like the KNN model above, building the decision tree started with the necessary package in R. For the decision tree model, ‘tree’ function was mainly used, as well as ‘plot’ and ‘text’ functions for visualizing the tree. We computed the accuracy and error rates for each training data set and the test data set.

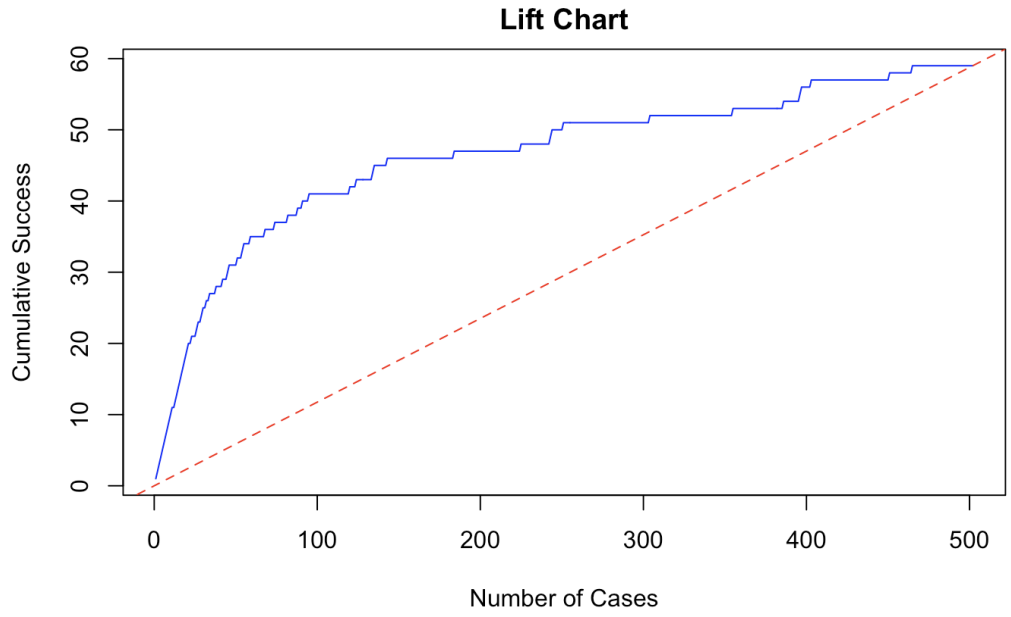
For the training data set, the accuracy score was 0.9505 and the accuracy score for the test data set was 0.8850. For specificity, the test data set was 0.3692. In contrast to the accuracy score of 0.8850 in the test data set, the specificity rate is quite low, implying that we have to consider the possibility of false positives.




Fig[3] Decision Tree Model

**Naive Bayes**

As well, we started by loading the necessary R packages for Naive Bayes model. From “caret” package, "createDataPartition" function was used to partition the "df5" dataset into training and testing samples. Then, from the "e1071" package we built the NB model using the "naiveBayes" function .
After that, we trained the model using the training dataset and used that to predict the testing data set. Again, with the confusion matrix, we were able to compute accuracy score and specificity, which was 0.8187 and 0.6949 respectively. To visualize the model performance, the Lift Chart was mainly used. 0.8187 is a high accuracy rate implying the usefulness of Naive Bayes model for our classification project. However, the specificity is relatively low, requiring additional performance measures.




Fig[4] Naive Bayes Lift Chart

**Random Forest**

In the random forest model, various groups of hyperparameters, such as ntree and mtry are used to train the model. From 50 to 200, the ntree was nested and for mtry between the square root of the number of predictors minus 3 and the square root of the number of predictors plus 3 were used. In the testing dataset, the model with the highest accuracy score was chosen and that model had 200 trees and mtry= 8. As a result, our random forest model had an output of 0.902 of accuracy and 0.778 of specificity. It is possible to assume that our random forest model had performed well in the classification tasks and the relatively high specificity also implies that the model is also effective at catching the true negatives.

**Results Comparison**[1]

| | Test Accuracy | Specificity |
|---|---|---|
| KNN | 0.8946 | 0.7143 |
| Decision Tree | 0.8850 | 0.3692 |
| Naive Bayes | 0.8187 | 0.6949 |
| Random Forest | 0.9006 | 0.7778 |

Fig[5]Results Comparison Accuracy score and Specificity for each model

**Model Selection**

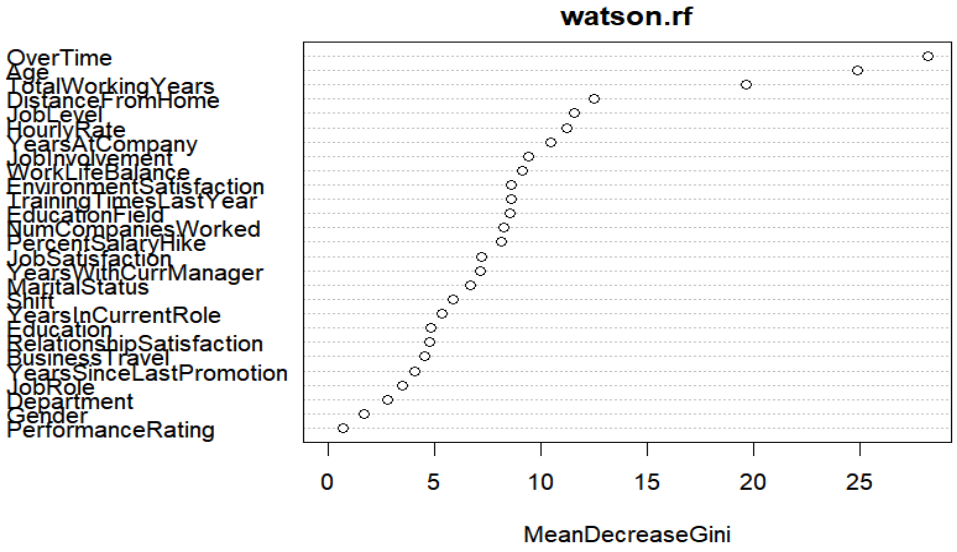
From the results, the Random Forest had the highest accuracy of 0.9006. Next was KNN with an accuracy of 0,8946. Decision Tree had the third best model with the accuracy of 0.8850 and Naive Bayes model had the lowest accuracy of 0.8187.

In terms of specificity, Random Forest has the highest value of 0.7778. KNN had the second best specificity of 0.7143. Naive Bayes with a value of 0.6949, and Decision Tree had the lowest value of 0.3692.

Therefore, we concluded that the Random Forest model was the best model for our classification project. Not only the highest accuracy score, but also the model's high specificity suggest that we would be able to detect true negatives with this model. As well, it is recommended to consider other factors like model complexity to provide unbiased results of the model.

[1] Numbers are rounded to four decimal places

**Random Forest Feature Importance**




Fig[6] Random Forest Feature Importance

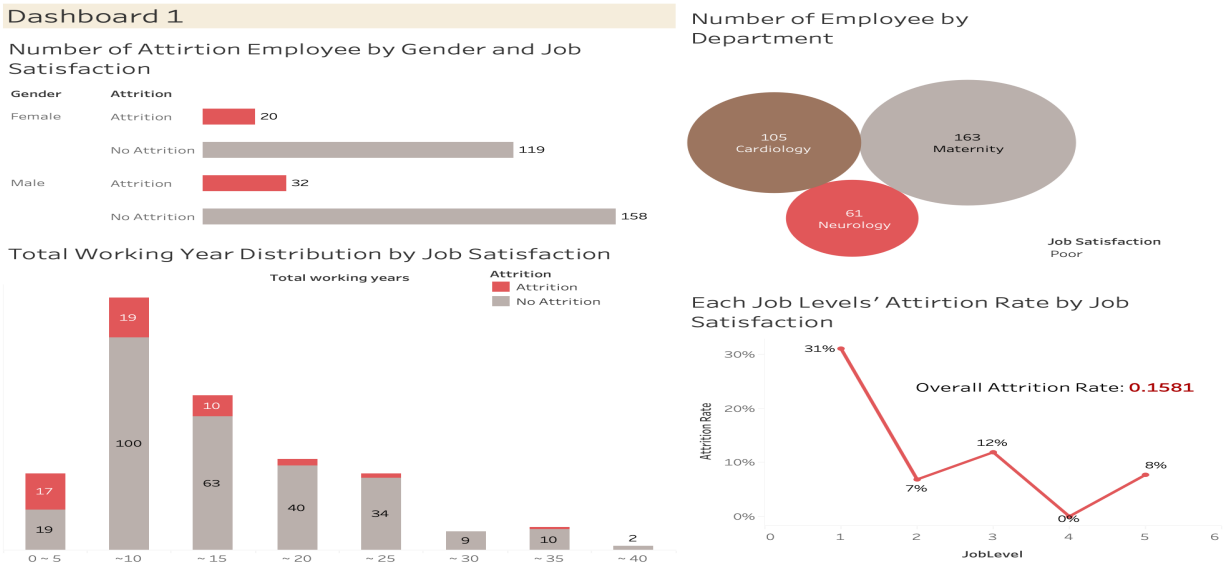
According to the Random Forest model, the factors that play a significant role in employee attrition are age, the total number of years worked, and overtime. These findings highlight the importance of understanding the root causes of nurse attrition and taking actions to address them.

## VI. Conclusion

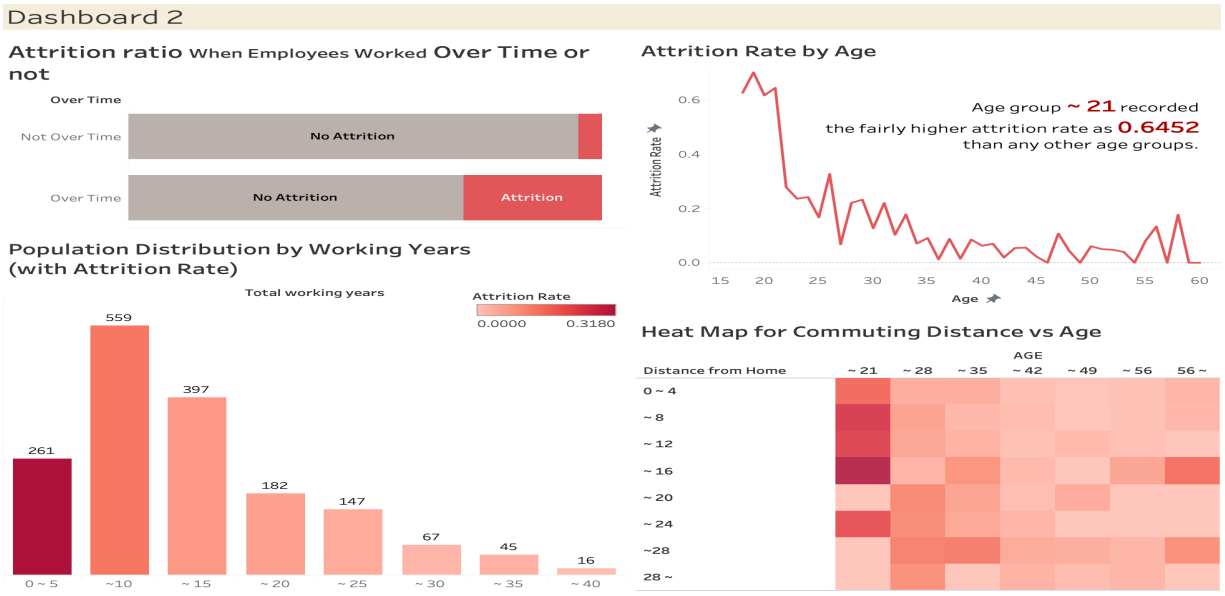
As part of our research on predicting employee attrition in the healthcare industry, we have conducted a comprehensive analysis of the available data using various machine learning techniques for classification. Through rigorous evaluation, we have determined that the RandomForest model yields the highest performance in terms of accuracy and specificity. Our results have highlighted that overtime work and age are critical factors in predicting employee attrition. Specifically, our analysis indicates that overtime work increases the likelihood of attrition, while younger employees are more susceptible to leaving their jobs. Based on our findings, we propose two actionable solutions for healthcare companies seeking to reduce attrition rates. Firstly, companies should prioritize minimizing overtime work by creating a working environment that prioritizes employee work-life balance. Secondly, healthcare companies should endeavor to understand and cater to the unique needs and expectations of younger employees to provide a better working environment. By implementing these solutions, healthcare companies can enhance their employee retention rates and create a more stable workforce. Ultimately, we expect our models and recommendations to contribute to the provision of better medical services.

## VII. Appendix (Any additional information to be submitted):

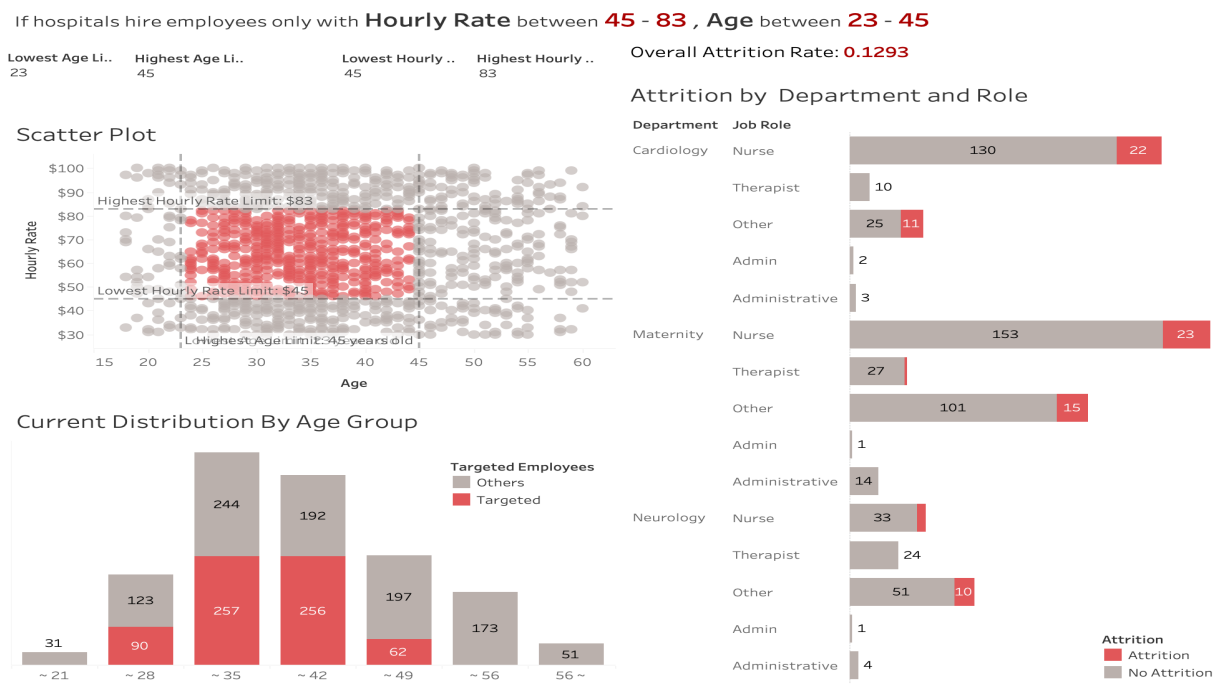
.




Fig[7] Job satisfaction versus employee attrition in Tableau Dashboard




Fig[8] Most crucial factors of employee attrition ("total working year," "age," and "over time") in Tableau Dashboard




Fig[8] “Hourly Rate” and “Age” effect on employee retention in Tableau Dashboard

# References

"2023 NSI National Health Care Retention & RN Staffing Report." *NSI_National_Health_Care_Retention_Report.*, https://www.nsinursingsolutions.com/Documents/Library/NSI_National_Health_Care_Retention_Report.pdf.

*Alma Mater Studiorum Universit a Di Bologna - Amslaurea.unibo.it*. https://amslaurea.unibo.it/21599/1/Tesi_Conte_Vittoria.pdf.

Ayers, Margaret, et al. "Adopting AI in Drug Discovery." *BCG Global*, BCG Global, 8 Feb. 2023, https://www.bcg.com/publications/2022/adopting-ai-in-pharmaceutical-discovery.

Charanjeet Singh Ahluwalia. "Employee Attrition: Human Resource Concern." *RPubs*, https://rpubs.com/CJ_09/Emp_Attrition_Final.

*Collaboration Systems and Technologies | Hawaii International ...* https://aisel.aisnet.org/hicss-52/cl/.

JohnM. "Employee Attrition for Healthcare." *Kaggle*, 15 Feb. 2023, https://www.kaggle.com/datasets/jpmiller/employee-attrition-for-healthcare?resource=download.

Li Z;Yu Q;Zhu Q;Yang X;Li Z;Fu J; "Applications of Machine Learning in Tumor-Associated Macrophages." *Frontiers in Immunology*, U.S. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/36211379/.

Liu, Yung-Chun, et al. "The Steelmaking Process Parameter Optimization with a Surrogate Model Based on Convolutional Neural Networks and the Firefly Algorithm." *MDPI*, Multidisciplinary Digital Publishing Institute, 25 May 2021, https://www.mdpi.com/2076-3417/11/11/4857/html.

Piotr. "Random Forest Feature Importance Computed in 3 Ways with Python." *MLJAR*, Piotr Płoński, 29 June 2020, https://mljar.com/blog/feature-importance-in-random-forest/.

Prasad, Kislaya. " Naive Bayes, Performance Measures ,Tree-Based Models in R, ClassificationFirst Look, K Nearest Neighbor (Revised) ."

Reachiteasily. "ReachIt Easily on Tumblr." *Tumblr*, 1 Jan. 2022, https://www.tumblr.com/reachiteasily.