Fanyun Xu

2033366

Introduction - Employees are a valuable asset for any organization, playing a pivotal role in its success and stability. This project is centered around exploring the healthcare industry, aiming to predict the likelihood of employees resigning from their positions. The study holds significance in ensuring the stability of the entire industry and maintaining the quality of its workforce.

Utilizing machine learning for predicting employee attrition allows companies to enhance their proactive intervention capabilities. By understanding the factors contributing to employee turnover, organizations can strategize on retention initiatives and ensure the continuity of a skilled workforce. The project encompasses various technological aspects, including data visualization, data preprocessing, data transformation, and predictive modeling.

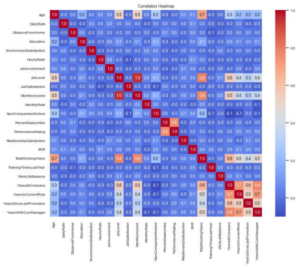
Methodology – **Data Preprocessing Stage：** In the data preprocessing stage, the basic information of the dataset was first outputted, including the data types of each feature and the presence of missing values. Subsequently, the two types of features in the dataset were categorized, and their relevant information was displayed. One type of feature is continuous, with multiple possible values. We presented information such as cardinality, minimum and maximum values, quartiles, median, standard deviation, and other related statistics for these features. As for categorical features, which describe finite categories within the data, we showcased their cardinality, the most frequently occurring label, and its frequency.

Through these two tables, we gained further insights into the characteristics of the dataset and removed features with a cardinality of 1 as they did not contribute to predictions. Additionally, we visually represented the dataset by plotting histograms. Given the large number of features, for dimensionality reduction, we calculated numerical values such as correlation attribute, gain ratio attribute, and symmetrical uncertainty attribute. Based on these values, we selected 15 features for subsequent predictive analysis. This series of data preprocessing steps aims to prepare the data for optimal use in training and predicting with machine learning models.

Classification Algorithm： During the data prediction phase, we employed two different machine learning models: Support Vector Machine (SVM) and Decision Tree. Prior to model training, the entire dataset was divided into training and validation sets with an 80:20 ratio. This division aims to utilize the training set for model training while using the validation set to assess the model's performance, especially in terms of accuracy on unseen data.

For each model, training was conducted using the training set, and the model's performance on unknown data was evaluated using the validation set. This evaluation process helps determine whether the model is overfitting (performing well on the training set but poorly on the validation set) or underfitting (performing poorly on both the training and validation sets).

After completing the model training, we utilized the trained models to make predictions on the test set. The prediction results were then saved in a CSV file for subsequent analysis and use. This process leverages the machine learning training-validation-testing workflow, ensuring that the models exhibit good generalization performance across different datasets.



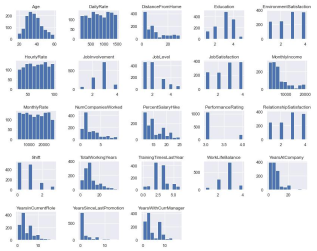


Fig.1 Correlation Matrix

Fig.2 Histograms of Data Set

Results - In using SVM, we observed its optimal performance on the dataset, achieving an accuracy of 96.8%. The superiority of this result may stem from SVM's effective modeling capabilities for non-linear relationships in high-dimensional spaces, as well as its robustness to outliers. The robust classification performance of Support Vector Machines makes them a powerful tool for tackling complex problems. In contrast, the decision tree performs slightly less well, with an accuracy of 92.8%. This could be due to the tendency of decision trees to overfit the training set when handling complex data relationships, resulting in poorer generalization on the test set.

Discussion: **Good Aspects:** The project's proactive use of SVM for predicting employee attrition in healthcare demonstrates high accuracy (96.8%), showcasing the model's efficacy in handling non-linear relationships. A comprehensive technological approach from data preprocessing to modeling adds depth to the study.

Bad Aspects: While SVM excels, the decision tree slightly lags (92.8%) due to potential overfitting issues. Balancing accuracy and interpretability becomes crucial in choosing models.

Contributor Thoughts: Proactive attrition prediction aligns with industry needs. Balancing SVM's accuracy and decision tree's interpretability is key for practical implementation.

Conclusion: The project's strategic approach aids healthcare retention efforts. SVM's robust performance and decision tree's considerations provide insights for future predictive analytics in healthcare.