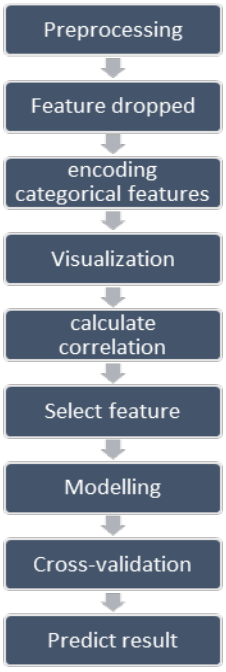
Report on Predicting Healthcare Employee Attrition: insights from Gradient Boosting

- ***Introduction-***

The COVID-19 pandemic has presented unique challenges to the healthcare sector, particularly in the form of increased **employee attrition**, predominantly among nurses. This contest draws inspiration from the **IBM Watson Wear Dataset.** The dataset is exhaustively compiled and highly relevant to big data, encompassing four key aspects:

1. Abundance: This is a crucial feature of big data.
2. Variety: The dataset includes up to 35 wide-ranging variables, illustrating the diversity of big data.
3. High speed: The velocity of big data generation and processing in this undertaking is exceedingly fast-paced.
4. Accuracy: The caliber and precision of data underscore the big data's potential to power decision-making and strategy formulation.



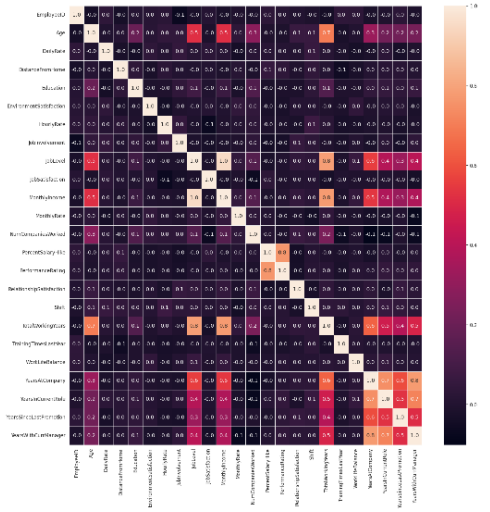
- ***Methodology-***

A. Data Preprocessing

1. Data **cleaning**: Obtain variable descriptions in the dataset and address data gaps, anomalies, and inaccuracies.
2. Data **reduction**: Use principal component analysis (PCA) techniques to eliminate irrelevant variables such as "Over18" from the dataset, thereby reducing data dimensionality.
3. **Discretization**: converting categorical (continuous) variables into discrete variables to make them countable.
4. Data **classification** and **visualization**: Classify and visualize numerical and descriptive information, and find the relationships with “Attrition”.
5. **Standardization**: Normalize the variables to a standard range of **0-1** for calculating the correlation coefficient between each variable and "Attrition".
6. Feature selection: Use correlation coefficients to determine the most significant and pertinent variables.

B. Classification Algorithm

Due to the high-dimensional data and complex, interactive relationships within the dataset, we chose to utilize the Gradient Boosting Classifier. This ensemble learning method divides the data into training and test sets and utilizes one-hot encoding to resolve the issue of classifiers struggling to handle attribute data effectively. Then it combines the tree algorithm, and in each iteration, cross-validation is employed to enhance accuracy (with **Overall Test Accuracy: 0.94841**)



- ***Results-***

These competitor models and algorithms include Naive Bayes, logistic regression, decision trees, and k-NN classification, etc.

1. **Naive Bayes** is proficient at processing large datasets with independent features, delivering fast and accurate results.
2. **Logistic regression**, on the other hand, is better suited for dealing with linear relationships.
3. **Decision Tree**: is more proper for intuitive classification, although not as effective as Gradient Boosting Classifier for complicated datasets.
4. **K-Nearest Neighbor (K-NN)** technique is slower for larger datasets, and the error rate diminishes as the value of "k" increases.

- ***Discussion:***

Advantages	Shortcomings
Predictive capabilities are often better than other algorithms.	High computational cost
Strong interpretability because of ensemble learning	Slow training especially for large datasets
Resist overfitting	Weak scalability
Good at handling complex features	Strong noise sensitivity

The convenience and adaptability, along with the capability to comprehend intricate and non-linear patterns, serve as the pros that I appreciate. However, apprehension arises due to the likelihood of overfitting. It may be a beneficial extension to adopt regularization methods, like **Weight decay** (also known as L2 regularization), **Pruning**, **Dropout**, etc., which will impose constraints and penalties.

- ***Conclusion-***

One conclusion drawn from the study is that employees who have low job satisfaction, low income, and high overtime have a greater chance of attrition. Another emerged from the study is that employees who received a promotion within the last few years or had a higher performance rate were more likely to stay. It contributes to the field of HR analytics by developing models that predict employee attrition, thereby helping organizations retain their valuable employees.