ATTRITION DATASET ANALYSIS
TEAM: "THE DATA DIGGERS"
Team Members: Malika Thakur, Jayant Kumar, Swati Shiriyannavar, Neil Sharma

Almost all companies are more concerned about human resources. Selecting proper employees depending upon their performance is one of the issues of managers. Hence, "**employee attrition"** is one of the pressing issues. In most cases, employee turnover is very costly hence management builds an evaluation system in an attempt to keep the best performing employees.

Data Mining is an important field of knowledge discovery. Our team has developed a classification model to predict the performance of employees using data mining models.
Data Mining has several steps associated with it which we used.
The following are the steps we underwent to get the model.

1. **Data Acquisition**: It is a process of gathering data from different sources.
2. Data Cleaning: This is the most time-consuming stage. The data had so many inconsistencies and noise. So to build the model we cleaned and pre-processed the data by removing unnecessary data and null values.
3. Train Dataset: We created data features from data and split data randomly into training and test data.
4. Train Machine Learning Model: We built the Machine Learning model using all kinds of algorithms on the basis of the training dataset. We used models like KNN, Random Forest, SVM, etc.
5. Test Model: We evaluated and validated the training and test dataset and used a series of Machine Learning algorithms.
6. Deploy Model: We created the usable model and deployed it. The later model was constantly improved.

LIST OF ALGORITHMS USED:

- K- Nearest Neighbors
- Gaussian Naive Bayes
- Random Forest Classifier
- Decision Tree Classifier
- Artificial Neural Network
- Support Vector Machine

DATA PRE-PROCESSING

In any Machine Learning process, Data Preprocessing helps us to dig out all the missing values and remove unwanted and null values. It helps to transform data into that state which helps the machine to understand easily. Hence features of the data can be easily interpreted by algorithms. There are 27 attributes in the dataset, the most relevant classes are combined to make the outliers. For importing the dataset Pandas data frame and Scipy libraries are used. After importing data is being scaled using standard scalers. Dataset analysis is done by bivariate plots for Status: Active or Terminated for a better understanding of the data along with the multifarious representation. The dataset is divided into a training set and a testing set.

VARIABLE DESCRIPTION

This table describes all the variables in the dataset and classification of ones we used for predicting the performance rate.

DATA ANALYSIS

Employee Personal Demographics:
We can observe from the slides that the attrition rate is higher for:

- **Younger employees** with the 25-35 age group.

- **Female** employees.
- **Disabled** employees.
- Employees with **education level between 1-3.**
- **White** Employees.
- **Not divorced** (Married/Single)
- Not a disabled Veteran.

Employee Work Demographics:
We can observe from the slides that the attrition rate is higher for:

- Employees with **1st Job.**
- The number of **teams changed is < 2.**
- Travel required is minimal.
- Employees **Job group: Physical flow.**
- Employees with **less** performance **rating.**

CORRELATION MATRIX

A correlation matrix is a table showing the correlation relationship between sets of variables.

CONFUSION MATRIX

A **confusion matrix** is a table that describes the performance of a classification model. Each row of the matrix represents predicted values while each column represents actual values (or vice versa).

K- NEAREST NEIGHBORS

It is a nonparametric approach. It is one of the simplest machine learning techniques that can be used for both classification and regression. Usually, KNN tries to capture the neighborhood in the sense of distances, many other algorithms try to capture a different essence. It computes the majority votes from K nearest neighbors. The hyperparameter K is very important in the algorithm and the value of K should be odd, for ease of calculating the majority vote. I have used the different values of K to find the optimal solution to the problem and got the best accuracy for K = 5, because if the value of K is too small, there are not enough samples participating in the voting, on the other hand, if the value of K is too large then it may overfit. It gives an average accuracy of 56.38%.

GAUSSIAN NAIVE BAYES

In this algorithm, the probability is computed for each class label and the class with the largest probability is selected. However, if certain features are only available on the test sample and are not observed in the training set, then this model will fail. Therefore, to ensure that the training data has enough variation, we use the K-fold cross-validation approach. This is based on Bayes' theorem. This model gave the lowest accuracy of 55.7%.
The likelihood is estimated as follows -

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

RANDOM FOREST CLASSIFIER

Random Forest is a large number of relatively uncorrelated models (trees) operating as a committee that will outperform any of the individual constituent models. The low correlation between models is the key.

Decisions trees are very sensitive to the data they are trained on, small changes to the training set can result in significantly different tree structures. The random forest takes advantage of this by allowing each individual tree to randomly sample from the dataset with replacement, resulting in different trees. This process is known as bagging. This gave the highest stable accuracy of 66% with the highest precision.

DECISION TREE CLASSIFIER

Decision Tree Classifier uses a tree-like graph of decisions. It can cover both areas of regression and classification. This tree is drawn from top to bottom with its root node at the top. It's a tree-like structure where each node represents the test on an attribute, the branch node denotes the outcome of the test, and the child node shows a class label. It makes uses of the CART algorithm. CART stands for Classification And Regression Tree algorithm. Decisions trees are very sensitive to the data they are trained on, small changes to the training set can result in significantly different tree structures. This gave the second-highest accuracy of 60.22%.

ARTIFICIAL NEURAL NETWORK

Artificial Neural Networks primarily consist of three layers: Input Layer, Hidden Layers, and Output Layer. A weight optimizing algorithm together with back-propagation is employed to tune the random weights. To calculate errors in our prediction, back-propagation compares the output against the random weights to the desired labels for the specific data row. It gave a second-lowest accuracy of 56.1%.
We can refer the following formula to calculate the number of neurons in the hidden layers –

Nh = Ns / (α*(Ni+No))

Ni = Number of input neurons

No= Number of output neurons

Ns = Number of samples in the training data set.

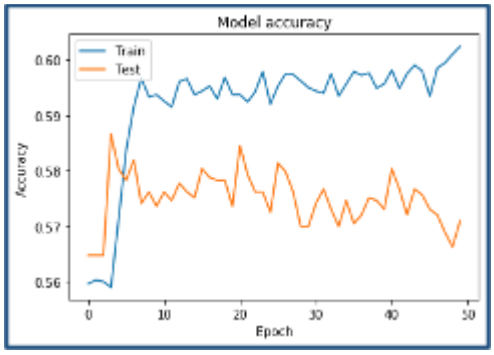

FIGURE: ACCURACY VS EPOCH PLOT FOR ANN

SUPPORT VECTOR MACHINES

SVM is primarily used for binary classification, specifically patterns that are linearly separable. We consider points known as the support vectors; the optimal hyperplane is selected in such a manner that these points are at a minimal distance from each other. Classifiers that make use of this attribute are hence called support vector machines. SVMs work behind the elementary concept of generating an optimal hyperplane to maximize the separation between the classes, outlier vs inlier in my case. SVM assumes that all the samples in the dataset follows Gaussian distribution and are contributing individually to the features. This accounted for an accuracy of 57.38%.
SVM assumes that all the samples in the dataset follow Gaussian distribution and are contribute individually to the features.

$$g(x)=\left(\sum_{x_i \in S} \alpha_i z_i x_i\right)^t x + w_0$$

RESULTS AND CONCLUSION

For companies, this model can be used in predicting the performance of the new incoming employees. Data of the current employees and even previous ones can be used to have the correct performance rate.

From our research, we estimated that Random Forest Classifier helped us to achieve the highest accuracy of 66% as seen from the table. Random Forest is a good model to get high performance with less need for interpretation. It can handle all kinds of features like binary features, categorical features, and numerical features. It works well with the less pre-processing tasks. Random forests work well with high dimensional data. Prediction speed is faster than training speed hence the Random Forest model was more efficient.