REPORT OF CASE STUDY RESULTS

EMPLOYEE CHURN

GSB ACADEMY x DIBIMBING.ID 2022

By: Muhammad Nabiil

BUSINESS UNDERSTANDING

Employee churn analysis is a study of the 'churn' level of employees in a company. Employees are said to 'churn' when they stop working and move to another company. This analysis is carried out as an initial prediction to anticipate employees not to 'churn'. This is done because the costs incurred by a company in recruiting and training new employees are quite expensive. Employees who leave will be replaced with new employees, or it can also be called employee turnover. Employee Churn is often detrimental to the company, and many researchers have examined this problem. Therefore, the company wants to identify what factors influence whether or not an employee 'churns'. There are several things that cause an employee to leave the company, such as work environment, work location, gender equality, salary equality, and others.

DATA UNDERSTANDING

The data used is employee data of a company that has a target variable 'churn' and consists of 9 attributes.

A. Company Employee Data

Employee data which contains a summary of the personal data of employees who have moved or who have changed jobs with a total of 1033 records and 9 attributes. The following is a table containing information on the data used.

Table 1 Employee data information

Attributes	Record (Non-Null)	Description
office_distance_from_house	1033	Distance (in meters) from the office to the employee's house
bonus_salary_percentage	1033	Percentage of salary bonus received by employees in the last 6 months
job_satisfaction	1033	Level of employee satisfaction
education_level	1033	Employee education level
overtime_hour	1033	Average length of overtime (in hours)
company_latitude	1033	The latitude coordinates of the company's headquarters
company_longitude	1033	The longitude coordinates of the company's headquarters
gender	1033	if 0, means female, if 1, means male
churn	1033	if 1, means Churn (employee moved), if 0, means not Churn (employee did not move)

Based on the table above, there are no attributes that have empty data (null values). This can be caused because the data entered is complete. However, there may be empty data or undetected error data due to inappropriate data types. Therefore, we will check the data type and data format so that they match what they should.

DATA PREPARATION

The initial data preparation process is carried out by changing the 'gender', 'churn', and 'job_satisfaction' attributes into object data types. After changing the attribute data type, the

numeric data type for employee data consists of 5 attributes and categorical data consists of 4 variables. In addition, checking for null values and data duplication is also performed on this data. The results show that the attributes in the employee data do not have empty values and also duplicate data. After the data is ready to be used, then descriptive statistical data will be seen on the data to be processed. Descriptive statistical data is carried out separately, namely for categorical data and numerical data. The following is a descriptive statistical result of employee data. The results show that the majority of employees have a job_satisfaction value of 3, education_level diploma. In addition, it can be seen that the data we have is unbalanced in the 'churn' category with a ratio of 70:30.

Tabel 2 Deskriptif statistik

	job_satisfaction	education_level	gender	churn
Count	1033	1033	1033	1033
Unique	6	4	2	2
Top	3	Bachelor/diploma	Male	No
freq	408	551	517	724

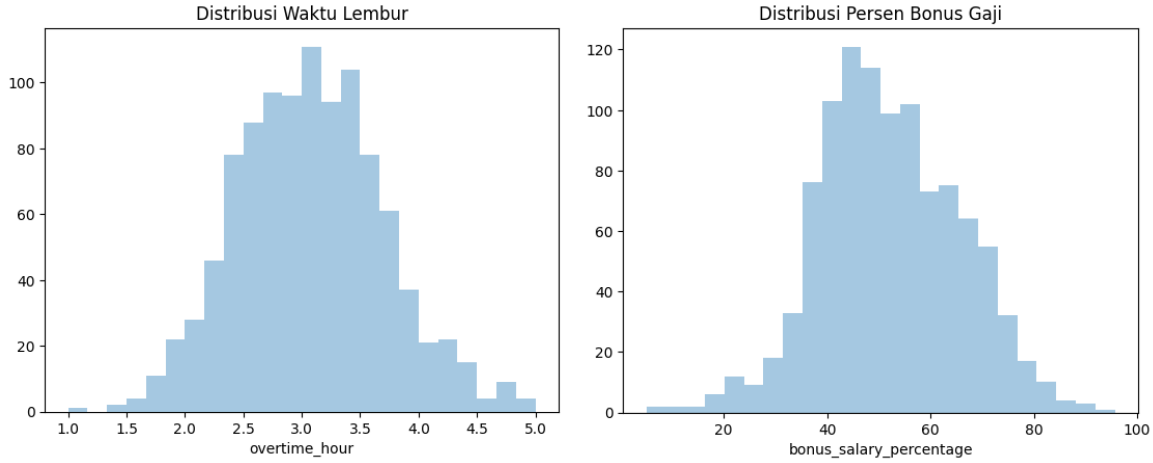
	office_distance_from_house	bonus_salary_percentage	overtime_hour
Count	1033	1033	1033
Mean	10771.87	52.125	3.08
Std	3810.28	13.62	0.63
Min	583	5	1
25%	8202	42.60	2.63
50%	10530	50.98	3.06
75%	13185	61.73	3.48
Max	24786	95.70	5

PREDICTION MODEL AND EVALUATION

A. Exploratory Data Analysis (EDA)

Exploratory Data Analysis is a method of analyzing data for a substantive understanding of data that answers questions about what is happening in relation to the data being analyzed (Syaripul, 2016). This EDA has an emphasis on graphical representation of the data with the aim of providing a statistical summary of the data to be processed. EDA can be done by carrying out univariate, bivariate, or univariate analysis.

The histogram represents the distribution of numerical data that relates the value of a variable to its frequency. The histogram of employee data in a company aims to determine the distribution of each characteristic or data attribute of each employee. In figure 1, the graph shows the normal distribution of the three numerical variables. For the overtime_hour variable, the majority of employees work overtime for 2.5-3.5 hours. The percentage of salary bonus received by employees is 40-60% and the majority of employees' home to office distance is 7500-12500 m.



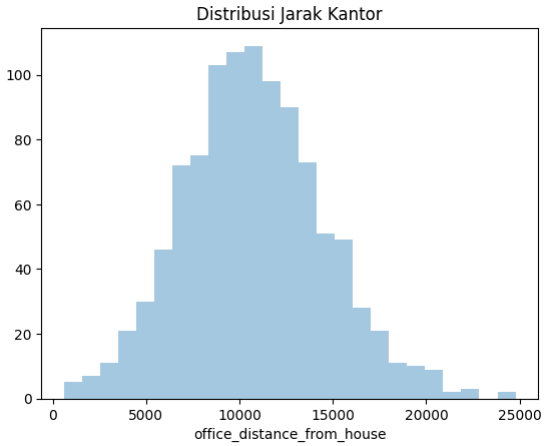


Figure 1 Numerical attribute histogram distribution graph

Bar charts represent information about a categorical variable. There are three variables represented in the bar chart, namely job_satisfication, education_level, and gender. Based on the bar chart in Figure 2, the majority of workers have job satisfaction levels of 2 and 3, as well as educational levels at the bachelor, diploma and master degree levels. Male and female employees have an equal percentage in both gender.

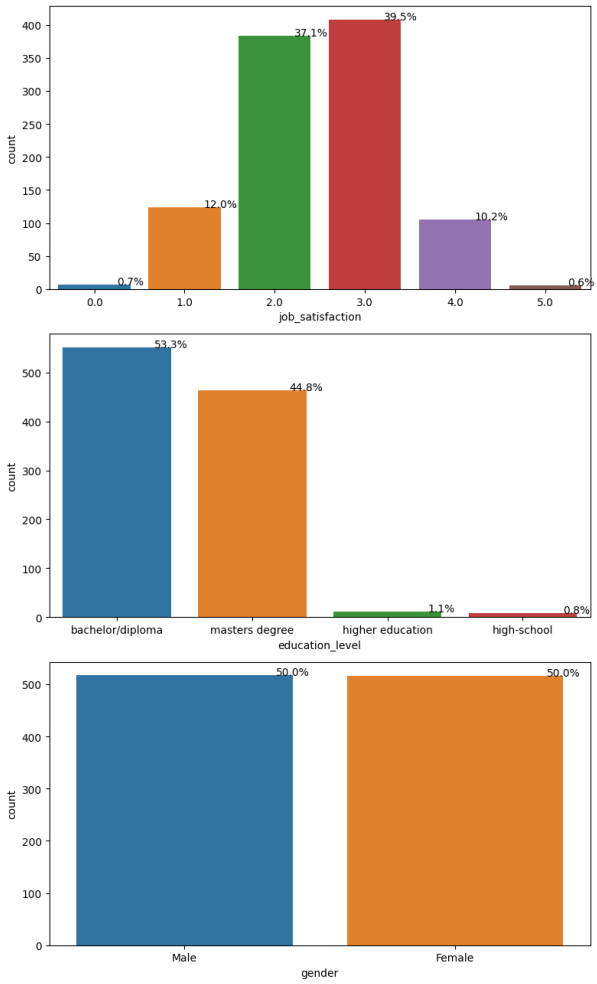


Figure 2 Bar graph of categorical data (univariate analysis)

Bivariate analysis is an analysis conducted to determine the relationship between 2 variables. In this analysis, two measurements were made for each observation. Bivariate analysis can be done by representing both variables with data visualization. In bivariate analysis between categorical data, we will see how each variable relates to the target variable, namely, 'churn'. Based on figure 3, employees with a satisfaction level of more than equal to 3 have a higher proportion of not moving companies. Based on the level of education, the majority of workers have completed their education at the bachelor, diploma and master degree levels.

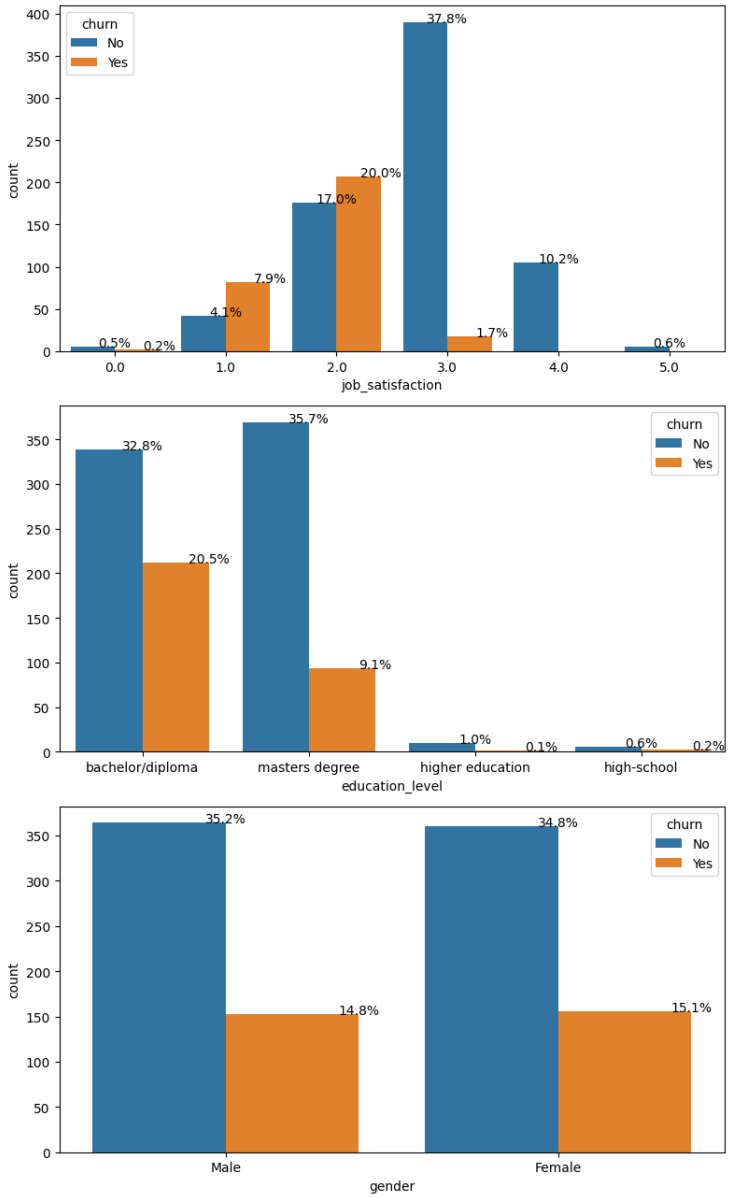


Figure 3 Bar graph of categorical data (bivariate analysis)

Next is bivariate analysis between numeric data and target variables that have categorical data types. The relationship between these categorical variables can be visualized using a PDF (Probability Dense Function) as well as a line box graph. Probability Density Function (PDF) is used to define the probability of a random variable falling in a different range of values. The function describes the probability density function of the normal distribution and how the mean and standard deviation exist. Based on Figure 4, we can conclude that employees with salary bonuses in the 30-50% range have a higher percentage of moving to companies, while employees

with salary bonuses in the 40-75% range have a higher percentage of staying with companies. Based on the distance from home to the office, employees with a distance of 7500-17500 m have a higher chance of moving companies, while employees with a closer working distance have a higher chance of staying at the company. For overtime, employees who move or stay have the same distribution.

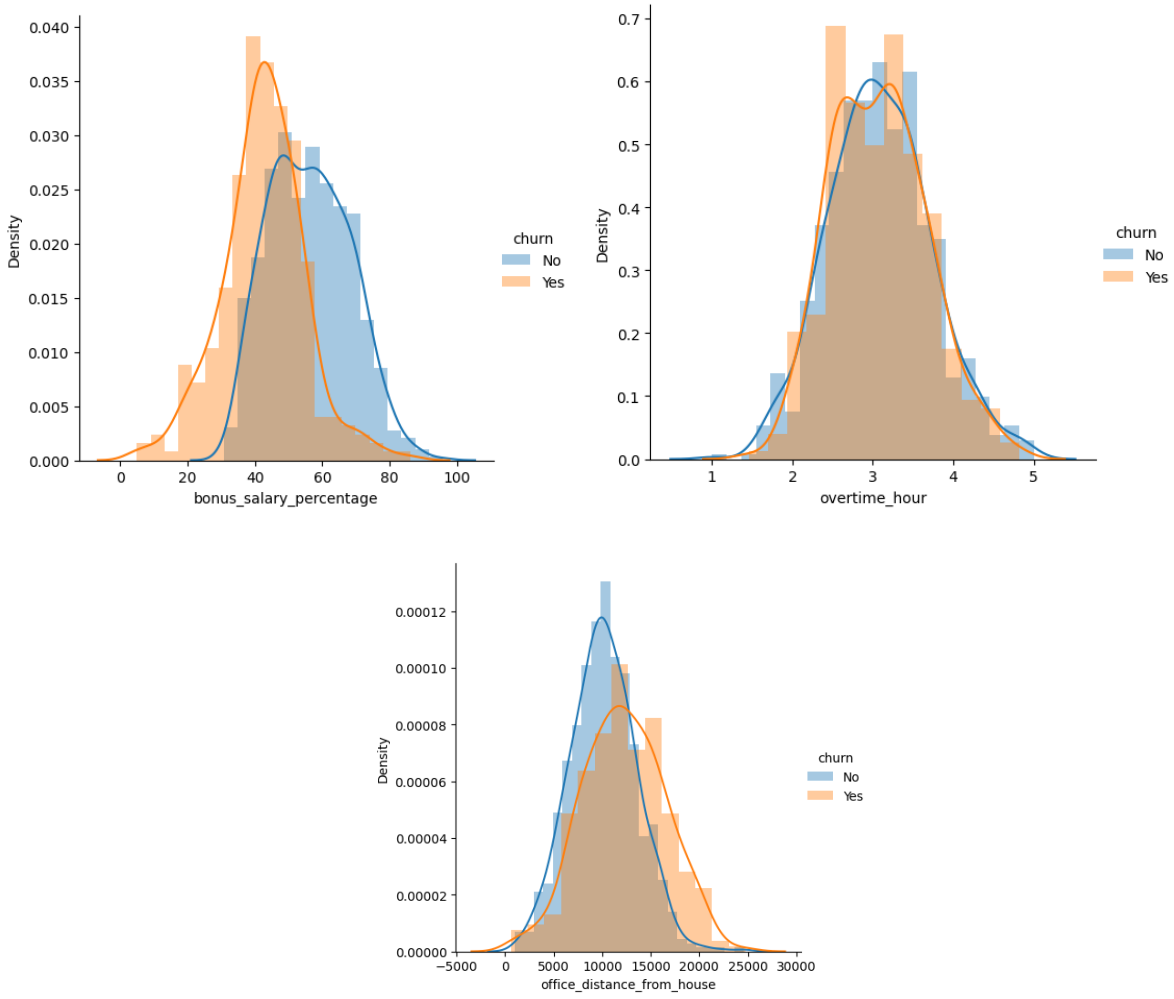


Figure 4 PDF graph of numeric data

The results of the analysis of the distribution of numerical variables and their relationship to churn show that shifting employees have farther values and distribution of distance from home to office and a lower percentage of salary bonuses. In addition, churned employees have a level of

job satisfaction below 4 with long office distances and low bonuses. Long distance from home to office and low bonuses for churned employees also applies based on employee education level.

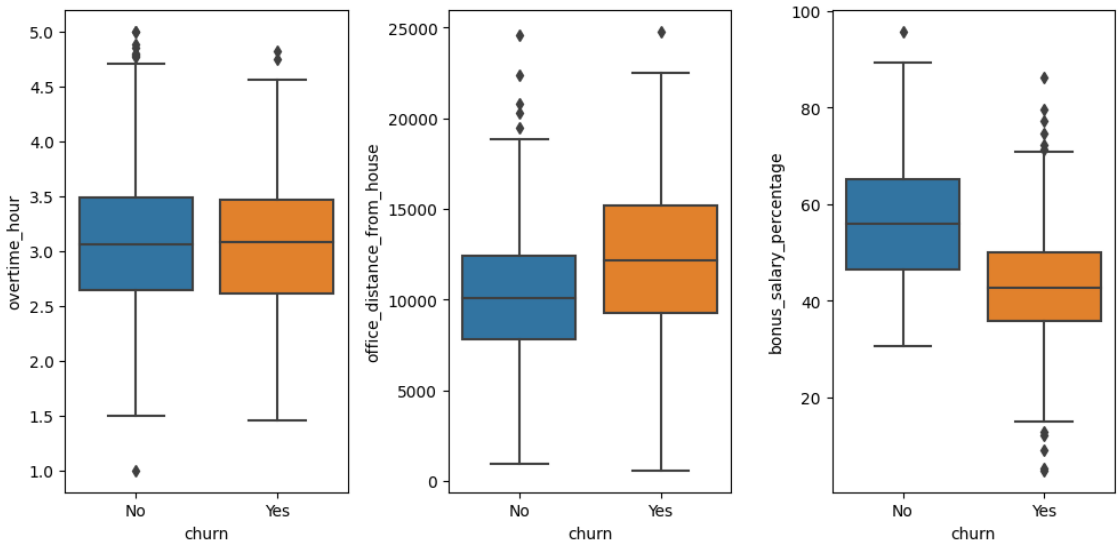


Figure 5 Graph of numerical data distribution based on churn

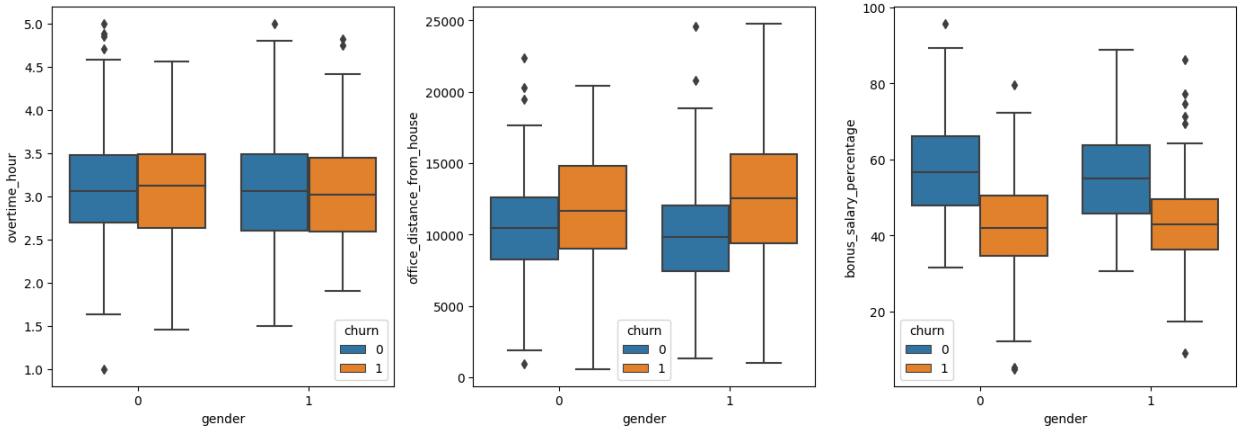


Figure 6 Graph of the distribution of the relationship between numerical data and gender based on churn

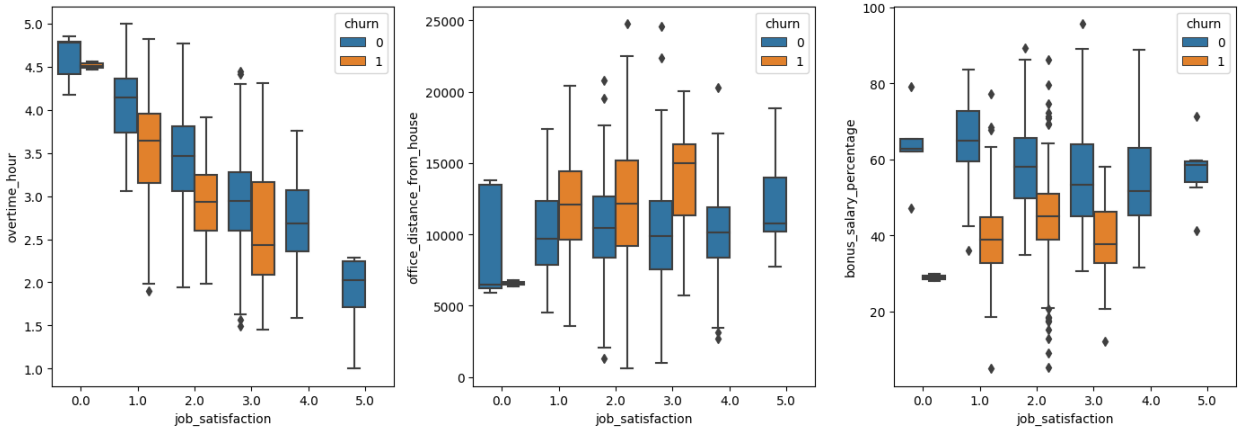


Figure 7 Graph of distribution of numerical data relationship with job satisfaction based on churn

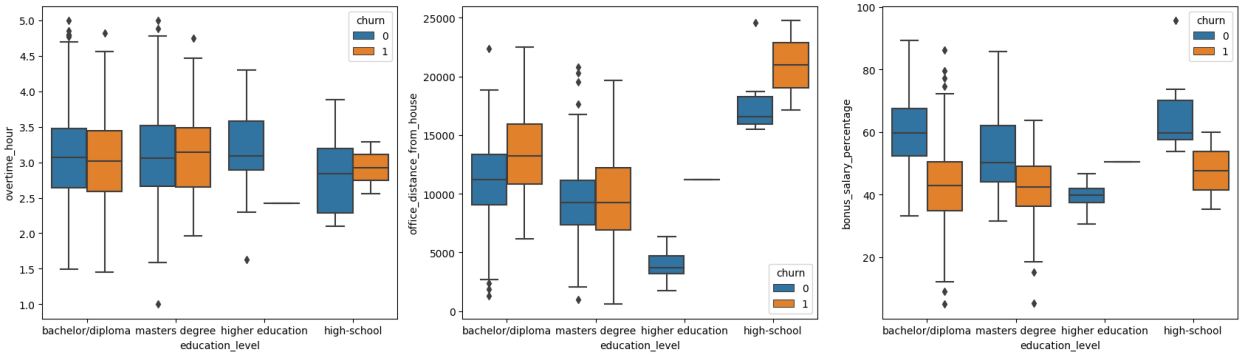


Figure 8 Graph of distribution of numerical data relationship with education level based on churn

Multivariate analysis is an analysis consisting of more than two variables where these variables are interrelated. Based on figure 9, the majority of workers who move companies are employees who get a 40-50% bonus and the workplace is 10000-15000 m from their home. For the relationship between overtime work and salary bonuses, the majority of workers who change companies are employees who receive salary bonuses in the range of 40-50% and overtime hours of around 3-4 hours. Lastly is the relationship between office distance and overtime for employees who change companies. The majority of workers who change companies are workers who get 3-4 hours of overtime and the office is 10000-15000 m from home.

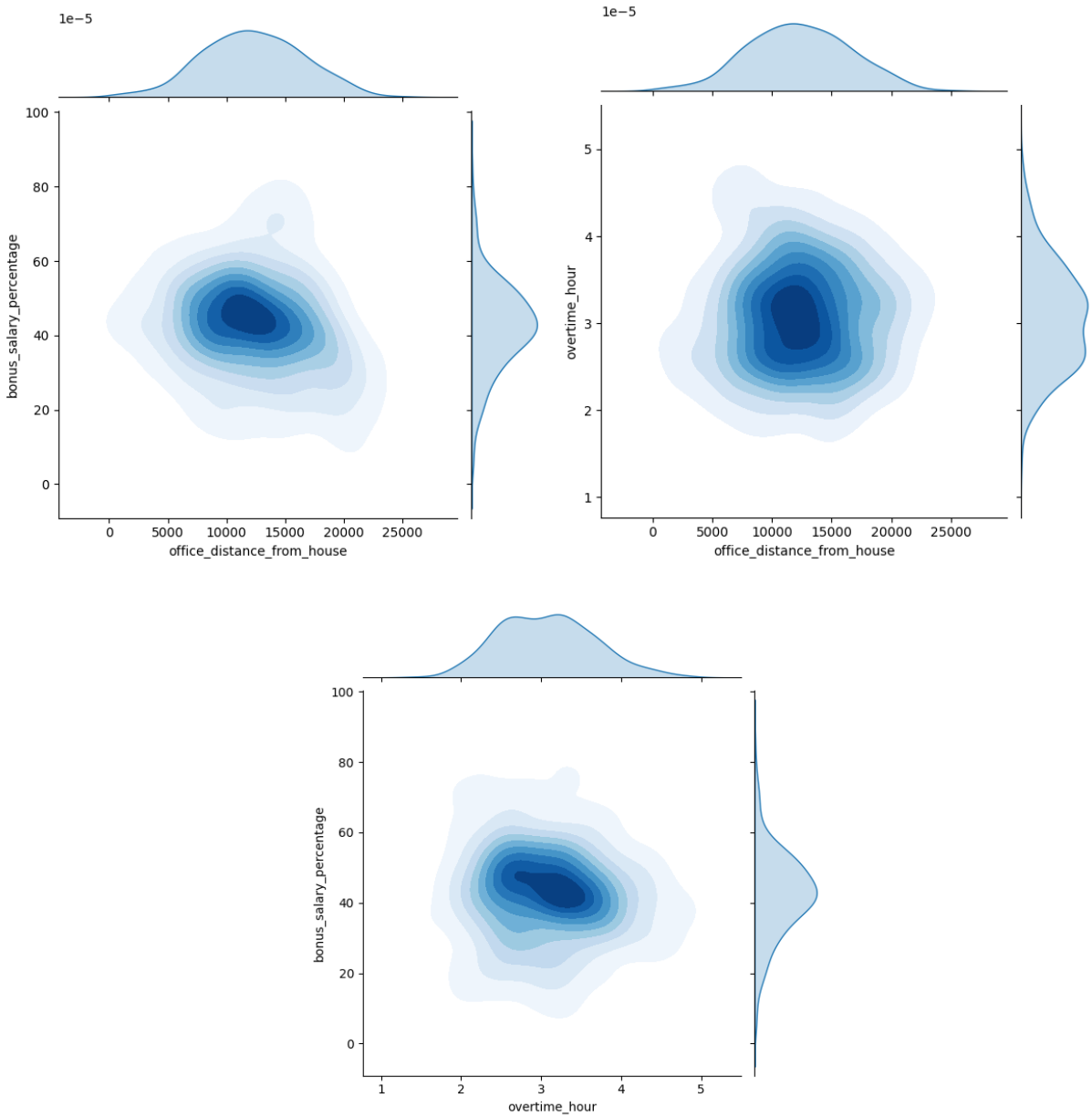


Figure 9 Graph of multivariate analysis of numerical data

The correlation matrix is a matrix that shows the simple correlation r between all possible pairs of variables involved in the analysis. Value or all possible pairs of variables involved in the analysis. The correlation matrix helps to predict the evolution of the relationship between the variables. The correlation matrix allows you to have a global view of the more or less strong relationship between several variables. The values or numbers on the main diagonal are all the same, namely 1. Figure 10 shows that the numerical variables in the employee data have a relatively weak correlation between the variables.

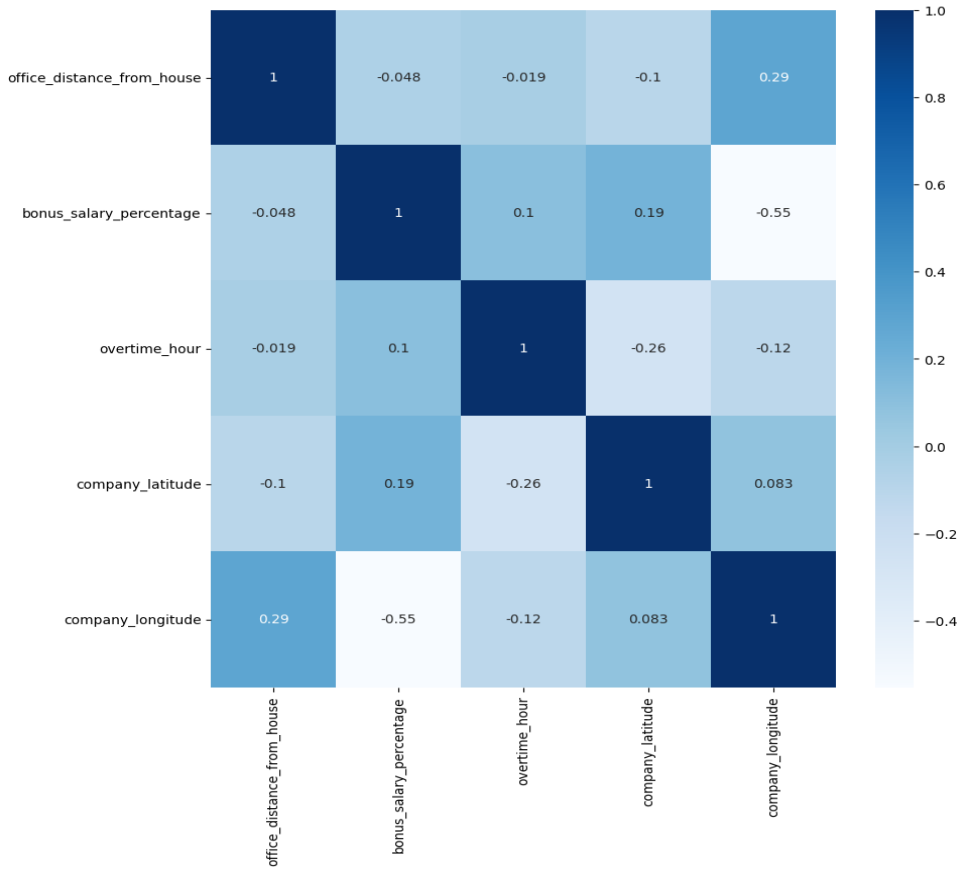


Figure 10 Correlation matrix

B. Modeling & Evaluation

The first is baseline modeling to find out which approach produces a model with high accuracy. The basic models tested are Logistic Regression, Decision Tree, Random Forest, KNN, and Support Vector Classifier, and XGBoost. The determination of training data and data testing in this modeling process is to divide the dataset into two, namely 70% for training data and 30% for testing data. Modeling is done by 5 fold cross validation of the basic model of the algorithm used. By validating, we can see how the model improves with the number of validations that are carried out. Just because a model has high accuracy in cross validation does not mean it is the best model when tested on data testing.

Table 3 Evaluation metrics on the baseline model

Model	F1-Score (%)	Accuracy (%)	Cross Validation (%)

K-Nearest Neighbors	93.12	95.81	94.33
Support Vector Machine Classification	92.63	95.48	94.33
Random Forest	91.98	95.16	95.02
XGBoost	90.32	94.19	95.58

Furthermore, the baseline models that will be used for tuning are K-Nearest Neighbors, Random Forest, and Support Vector Classifier, XGBoost. The tuning results show that the best algorithm to predict employee churn in this case is K-Nearest Neighbors. KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression). the best parameter obtained from the grid search process are *{'algorithm': 'auto', 'n_neighbors': 7, 'p': 1, 'weights': 'uniform'}*. K-Nearest Neighbors model produces a very good accuracy score of 95% and F1 score of 93%. The model to be used is the K-Nearest Neighbors model because it has the highest level of accuracy and the small difference between the accuracy of the training and the test data.

Table 4 Evaluation metrics on the tuned model

Model	F1-Score (%)	Accuracy (%)	Best Parameters
K-Nearest Neighbors	93.26	95.80	{'algorithm': 'auto', 'n_neighbors': 7, 'p': 1, 'weights': 'uniform'}
Support Vector Machine Classification	92.47	95.48	{'n_estimators': 1000, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': 100, 'bootstrap': False}
Random Forest	91.98	95.16	{'kernel': 'poly', 'degree': 3, 'C': 100}

XG Boost	88.40	93.22	{'colsample_bytree': 0.8, 'gamma': 1, 'learning_rate': 0.5, 'min_child_weight': 0.01, 'n_estimators': 500, 'reg_alpha': 1, 'reg_lambda': 5, 'sampling_method': 'uniform', 'subsample': 0.55}

The ROC curve is a graph showing the performance of a classification model across all classification thresholds. This curve displays two parameters, namely True Positive Rate and False Positive Rate. Figure 11 shows that the random forest algorithm with an AUC value of 0.96. This algorithm has a 96% chance of differentiate between shifting and non-shifting employees.

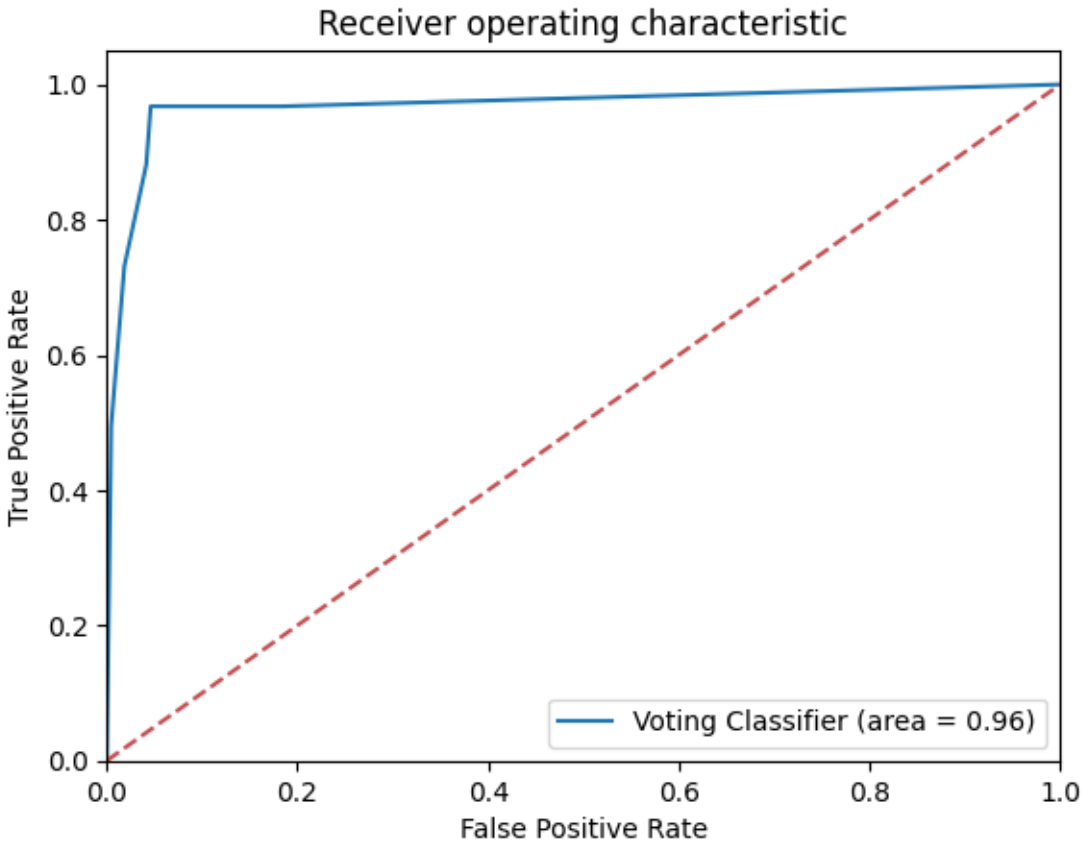


Figure 11 ROC graph of the K-Nearest Neighbors model

Furthermore, the prediction results from the classification model are diagnosed by using a confusion matrix to see how well the model has been made. In the confusion matrix, there are false positives and false negatives. The False Positive of this case is employees who change companies, but it is predicted that they will not beautify themselves. The False Negative of this case is an employee who does not change companies, but is predicted to move. In this case, the more important type of 'error' to pay attention to is False Negatives (Type 2 Errors) because this type of error can be detrimental to the company when someone who lives in the company is predicted to move. Figure 12 shows quite accurate results where the predictions that enter into false positives and false negatives are very few and the accuracy of the test data has a small difference with the train data.

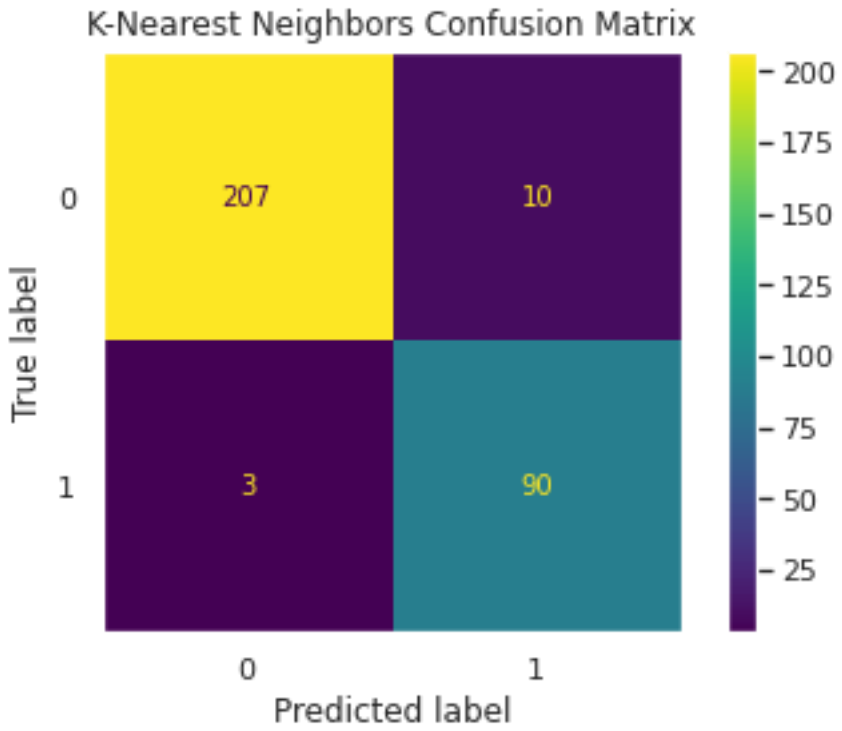


Figure 12 Confusion matrix model K-Nearest Neighbors (0 = churn, 1 = not churn)

For further evaluation regarding the model that has been tuned, the accuracy, recall, precision and F1-score values generated from the confusion matrix will be sought in this model. The results show quite high accuracy and good precision on both labels. We can see that the F1-score is quite high, 95%, so it can be said that the model created can predict whether or not an employee churns

well enough. The following is the result of the classification report for the K-Nearest Neighbors model used.

Table 4 Classification report model K-Nearest Neighbors

	Precision	Recall	F1-score	Support
0	0.99	0.95	0.97	217
1	0.90	0.97	0.93	93
accuracy			0.96	310
macro avg	0.94	0.96	0.95	310
weighted avg	0.96	0.96	0.96	310

Each attribute used in the model has a role in classification. However, there are attributes that have a very important role to a lesser role in determining this classification. To see the role of each attribute, each attribute's importance is measured in predictions using the feature impotance method. This method uses the Random Forest algorithm which has been parameter tuned to produce the best Random Forest model. Figure 13 shows that the 5 attributes that play a very important role in the classification process are company_longitude, bonus_salary_percentage, job_satisfaction_3.0, job_satisfaction_2.0, overtime_hour, and office_distance_from_house.

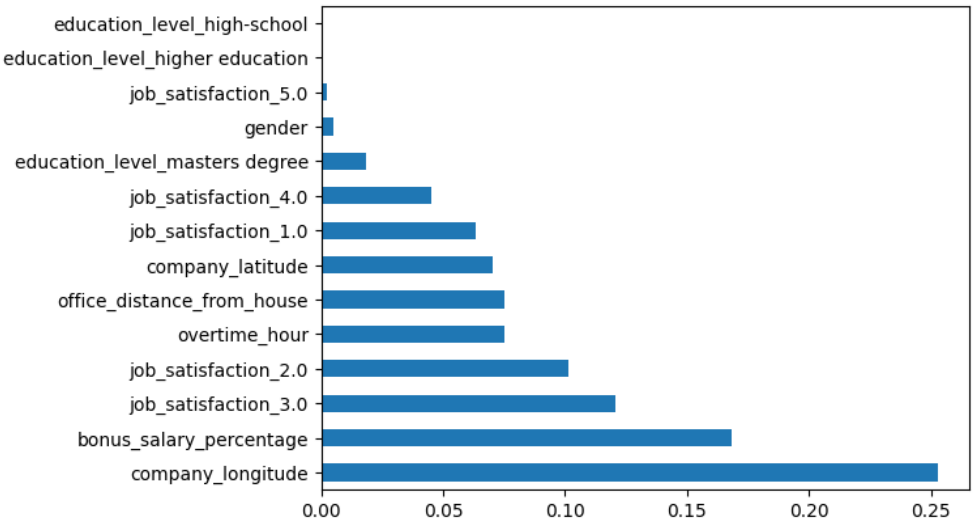


Figure 13 Feature importances graph

REFERENCES

Syaripul, N. A., Bachtiar, A. M. (2016). Visualisasi data interaktif data terbuka pemerintah Provinsi DKI Jakarta: topik ekonomi dan keuangan daerah. *Jurnal Sistem Informasi (Journal of Information System),* 12(2), 83-89.

Tri, M. F., Nataliani, Y. (2021). Analisis pengarruh penilaian asesor terhadap kinerja guru mata pelajaran dengan k-Means Culstering. *Indonesian Journal of Computing and Modeling*, 4(1), 15:22.