Prediction and Analysis of Data on Job Satisfaction of Employees with Machine Learning

Dion Chang

Abstract

The purpose of this project was to build a machine learning model to predict whether an employee is satisfied with their job as well as to find association rules using the Apriori algorithm in order to find noticeable behaviors of employees through the employee perception survey data set. Employee retention has been notably poor, and as a result it negatively impacts organizations. However, there are also many cases where although employees dislike their current roles they cannot easily leave. The issue with that is it does not directly affect statistics for turnover rate, but may result in a worse impact on the company. Therefore, job satisfaction would be the class variable to be predicted in this study instead of turnover rate. The dataset for this project was taken from Western Australia's public employee feedback surveys of the years 2015 and 2016. The results from this project was that the logistic regression model performed the best out of the other two algorithms (decision tree and Naïve Bayes) when predicting if the employee likes or dislikes their job using the following seven features: skill level for the job, job clarity, how much contributions are recognized, career developmental potential, how well-managed is the organization, work-life balance, and effectiveness of communication between departments. Employers providing meaningful and challenging tasks as well as giving reasonable workload and flexible deadlines would help employees to grow in their careers. Findings from this analysis can benefit employers and employees by improving job satisfaction as well as the work environment in organizations. As a result, this can decrease turnover rate and make positive impacts on the company.

Introduction

The two objectives of this project are to apply supervised learning techniques to predict if employees are satisfied or dissatisfied with their current job, and to use unsupervised learning to identify noticeable behaviors of employees from data. **It was determined that the logistic regression model with seven features was the most efficient to predict job satisfaction with an accuracy of 84%. From the Apriori algorithm, actionable insights were also discovered through learning the behavior of employees from the data. Improving career growth as well as keeping up with the level of skill for the job, ensuring job expectations are clear, and appreciation of employees (recognize contributions) can ultimately improve employee satisfaction levels in their jobs.**

One of the main notable problems in many businesses and organizations is poor employee retention. There are many factors that determine whether the employee stay or leave their current roles. In fact, that is why many firms tend to use great amount of resources in order to study the causes

of employee turnover (Flowers & Hughes, 1973). However, there are also cases where although the employee dislikes their current role they still continue to stay. This is due to many possible factors such as family responsibilities, source of income, lack of external opportunities, or being too afraid to leave as they are too comfortable in their job.

Employers and researchers mainly focus on studying or predicting the turnover rate of employees instead of whether or not they are currently satisfied with their job. As mentioned in an online article by Flowers and Hughes (1973), there are employees that are not satisfied with their job but continue to stay in their current role anyway. As a result, it does not affect the turnover statistics even though it may still have the same or a worse impact on the company. Therefore, this project focuses on job satisfaction rather than turnover rate. This would be beneficial for employers planning to improve the work environment in their department or the entire company.

In this project, the class variable chosen to be predicted by exploring several supervised learning algorithms was job satisfaction. The data type would be considered a Boolean (yes or no) as it has one of two possible values: the employee likes their job (satisfied with their job), or they do not like their job (dissatisfied with their job). Methods used to predict the job satisfaction of employees are logistic regression, decision tree, and Naïve Bayes. Explanatory variables used by the prediction models was the following: level of skill used, clarity of expectations, level of contribution recognition, career development opportunities, how well-managed is the organization, work-life balance, and communication level. The features chosen for the prediction models could be related to job satisfaction. Furthermore, the explanatory variables are ordinal data which means that they are categorical data in which their values are ordered.

The dataset used for this study is a survey designed for employees in a certain organization in order to capture their views regarding the factors affecting their workplace (Public Sector Commission (Western Australia), 2016).

The features for the model are considered to be categorical (ordinal) data because the format of the survey was for each question (or statement), employees would give a number rating which indicates how strongly they disagree or agree with that statement. As an example, for the statement: “My job allows me to utilize my skills, knowledge, and abilities.” The employee would then give a reverse scoring between 1 – 7 where 1 means “strongly agree” while 7 is “strongly disagree” (Public Sector Commission (Western Australia), 2016). The “Data” section of this paper will provide more detail on the data set that was used for machine learning.

For unsupervised learning, the goal was to discover patterns for employee behaviour. Thus, it would make sense to do association rule mining. In this project, the Apriori algorithm was used to find associations between values of all features used for the classification of job satisfaction. Since supervised learning would be used to predict the class variable, the class variable was excluded in this case.

Related Work

A study performed by Vezzoli (2010) involved ensemble learning algorithms such as Random Forest and TreeBoost as well as a new algorithm called CRAGGING (Cross-validation Aggregating) to identify important variables that have a major impact on job satisfaction. However, their main concern after obtaining the final model in their study was the lack of interpretability (Vezzoli, 2010). However, the paper has also pointed out that ensemble learning algorithms in general reduces the interpretability of the model due to its complexity and as a result can make it difficult to provide any business insights. The dataset the author used for this work was of Italian Social Cooperatives workers. Extrinsic and intrinsic features used in this work was divided to study how each of them influence job satisfaction. Extrinsic refers to pay or career advancement, while intrinsic was how the employee's work was useful to or if their work was recognized by others within the company. What was interesting in the analyses performed by Vezzoli (2010) was that intrinsic aspects appeared as the more important contribution to job satisfaction over extrinsic aspects. However, because the dataset used for this study was from Social Cooperatives workers it was really no surprise that the intrinsic features would play a bigger role in their job satisfaction compared to profit-oriented firms.

Job satisfaction of health care workers have been previously studied. Krogstad et al. (2006) predicted the job satisfaction of doctors, nurses, and auxiliaries working in Norwegian hospitals by setting up a multiple linear regression model. The multiple linear regression model consisted of five independent variables: top management, competence, work organization, professional development, and local leadership. The dependent variable was a measure of job satisfaction through a Likert scaling ranging from 1 (very dissatisfied) to 5 (very satisfied). For their data collection, 1814 doctors, nurses, and auxiliaries from 11 Norwegian hospitals participated in a survey to provide feedback on their work experiences. They have discovered that local leadership was the factor that was most significant in predicting high job satisfaction for all three groups of doctors, nurses, and auxiliaries. As per each group, professional development was most important for doctors. Support and feedback from their direct supervisors was the most important to nurses. Finally, professional development and local leadership were both equally important in predicting job satisfaction of auxiliaries. Another study conducted by Kuzey (2018) used Support Vector Machine (SVM) to determine how the performance of the company was impacted by the factors of job satisfaction. In other words, how the features that contribute to employee satisfaction: management's attitude, colleagues, pay/reward, and job security affect organizational performance. From SVM, it was determined that ranking of importance that impacted job satisfaction was the following: management's attitude, pay/reward, job security, and colleagues.

Many other studies have focused on employee turnover instead of job satisfaction. Girmanova and Gasparova (2018) analyzed data on employee turnover from a company's HR department using association rules and decision trees with the R programming language. The study has concluded that communication and regular conversations between superiors and employees can help detect problems early in order to reduce the amount of employee turnover. In looking for a more robust model to predict employee turnover, Punnoose & Ajit (2016) applied Extreme Gradient Boosting (XGBoost) as a novel method in their work. Their analysis was based on turnover that are voluntary. They found that age, tenure, job satisfaction, pay, and fairness weighed the most for voluntary turnover (Punnoose & Ajit, 2016). XGBoost resulted in a higher accuracy for predicting turnover in their work. In a study by Sexton, McMurty, Michalopoulos, and Smith (2004), it focused on using a neural network to predict employee turnover in an attempt to retain them for a small manufacturing company.

Studies such as Sikaroudi & Ghousi (2015) and Zhao, Hryniewicki, Cheng, Fu, and Zhu (2018) explores multiple methods in an attempt to determine the employee turnover prediction model that gave the highest accuracy. The article by Zhao et al. (2018) focused on supervised learning methods while Sikaroudi & Ghousi (2015) also used supervised learning as well as unsupervised learning by Apriori algorithm for prediction.

There have also been types of work related to employment, but not exactly just for predicting employee turnover or job satisfaction. Zhang & Tan (2019) discovered association rules between the academic performance and career choices of 228 recent graduate students from the School of Information of Zhejiang University of Finance and Economics in the year of 2017. The purpose of this work was to be a reference for recent graduate students when picking their career choices as well as a guide for planning their career. From the results, it was obvious that the areas students excel at would significantly affect their career path.

Overall, this project on job satisfaction was different compared to previous work was due to the following: the dataset, feature selection, and algorithms. Majority of the data collected from previous works was through healthcare, manufacturing, social cooperatives, schools. The data collected for this project was from the Public Sector Commission, an agency of the Government of Western Australia. As for the features, unlike Vezzoli (2010) where extrinsic and intrinsic factors were separated to predict job satisfaction the features for this project combined both aspects. Age, race, employment status, length in company, and gender was not included as features to predict job satisfaction in this project while they were factors in other studies. The explanatory variables in this project are the following: skill level used for the job, clarity of expectations, contribution recognition, career development, how well-managed is the organization, work-life balance, and communication effectiveness. Instead of using complex algorithms such as novel ensemble learning methods or neural network to make predictions, this project used (comparatively) simpler supervised learning techniques to predict job satisfaction: logistic regression, decision tree, and Naïve Bayes.

Zhang & Tan (2019) used the Apriori algorithm to discover association rules between career choices and grades from each course subject taken by fresh graduate students, while Sikaroudi & Ghousi (2015) used it to predict the class variable of whether employees left the company. How this project uniquely uses the Apriori algorithm was to discover association rules between the features while excluding the class variable, job satisfaction (like or dislike job). In other words, the Apriori algorithm was used to look for relationships between the explanatory variables that determined job satisfaction.

Data

The data set used in this study was taken from the Public Sector Commission (an agency from the Government of Western Australia) Employee Perception Surveys of both 2015 and 2016, with a combined total of 15288 responses received (11405 responses from the year 2015 and 3883 responses from the year 2016). Links provided here: https://data.gov.au/data/dataset/public-sector-commission-wa-employee-perception-survey-2015 (Public Sector Commission (Western Australia), 2015), and https://data.gov.au/data/dataset/b814c55a-d9d1-4145-af88-0fb78a354f8a (Public Sector Commission

(Western Australia), 2016). Out of the 109 questions or sections in the survey, 13 of them were initially chosen as features as shown in Table 1. This was later decreased to seven features to solve the problem with the curse of dimensionality (more details in the Results section of this paper). The final seven features used was the following: level_of_skill_used, expectations_clarity, contribution_recognition, career_development, well_managed, worklife_balance, and effective_communication.

Table 1. Explanatory Variables chosen to build a model to predict the Class Variable. Given Feature Name was the name of features used to represent the question or section in the dataset used to apply machine learning techniques with Python. The Possible Responses from Employee column was the format of the raw data from the survey. With the exception of education_level and salary_rate, 1 meant the best rating, 7 was the worst, and 8 indicated "Do not know or does not apply." This would be later changed through the data cleaning and preparation processes.

Explanatory Variables			
Survey Question/Section	**Given Feature Name**	**Possible Responses from Employee**	**Responses from Employee after Data Cleaning and Preprocessing**
"My job allows me to utilise my skills, knowledge and abilities"	level_of_skill_used	1 – Strongly agree 2 – Moderately agree 3 – Mildly agree 4 – Neither agree nor disagree 5 – Mildly disagree 6 – Moderately disagree 7 – Strongly disagree 8 – Do not know or does not apply	1 – Strongly disagree 2 – Moderately disagree 3 – Mildly disagree 4 – Neither agree nor disagree 5 – Mildly agree 6 – Moderately agree 7 – Strongly agree
"I am clear what my duties and responsibilities are"	expectations_clarity	1 – Strongly agree 2 – Moderately agree 3 – Mildly agree 4 – Neither agree nor disagree 5 – Mildly disagree 6 – Moderately disagree 7 – Strongly disagree 8 – Do not know or does not apply	1 – Strongly disagree 2 – Moderately disagree 3 – Mildly disagree 4 – Neither agree nor disagree 5 – Mildly agree 6 – Moderately agree 7 – Strongly agree
"I am recognised for the contribution I make"	contribution_recognition	1 – Strongly agree 2 – Moderately agree 3 – Mildly agree 4 – Neither agree nor disagree 5 – Mildly disagree 6 – Moderately disagree 7 – Strongly disagree 8 – Do not know or does not apply	1 – Strongly disagree 2 – Moderately disagree 3 – Mildly disagree 4 – Neither agree nor disagree 5 – Mildly agree 6 – Moderately agree 7 – Strongly agree
"I am satisfied with the opportunities available to me for career progression in my current agency"	career_development	1 – Strongly agree 2 – Moderately agree 3 – Mildly agree 4 – Neither agree nor disagree 5 – Mildly disagree 6 – Moderately disagree 7 – Strongly disagree 8 – Do not know or does not apply	1 – Strongly disagree 2 – Moderately disagree 3 – Mildly disagree 4 – Neither agree nor disagree 5 – Mildly agree 6 – Moderately agree 7 – Strongly agree
"I feel that my agency on the whole is well managed"	well_managed	1 – Strongly agree 2 – Moderately agree 3 – Mildly agree 4 – Neither agree nor disagree 5 – Mildly disagree 6 – Moderately disagree 7 – Strongly disagree	1 – Strongly disagree 2 – Moderately disagree 3 – Mildly disagree 4 – Neither agree nor disagree 5 – Mildly agree 6 – Moderately agree 7 – Strongly agree

		8 – Do not know or does not apply	
"You are able to access and use flexible work arrangements to assist in your work/life balance"	worklife_balance	1 – Strongly agree 2 – Moderately agree 3 – Mildly agree 4 – Neither agree nor disagree 5 – Mildly disagree 6 – Moderately disagree 7 – Strongly disagree 8 – Do not know or does not apply	1 – Strongly disagree 2 – Moderately disagree 3 – Mildly disagree 4 – Neither agree nor disagree 5 – Mildly agree 6 – Moderately agree 7 – Strongly agree
"In my work area, communication between senior managers and other employees is effective"	effective_communication	1 – Strongly agree 2 – Moderately agree 3 – Mildly agree 4 – Neither agree nor disagree 5 – Mildly disagree 6 – Moderately disagree 7 – Strongly disagree 8 – Do not know or does not apply	1 – Strongly disagree 2 – Moderately disagree 3 – Mildly disagree 4 – Neither agree nor disagree 5 – Mildly agree 6 – Moderately agree 7 – Strongly agree
"Your immediate supervisor treats employees from all diversity groups with equal respect"	respectful_supervisor	1 – Strongly agree 2 – Moderately agree 3 – Mildly agree 4 – Neither agree nor disagree 5 – Mildly disagree 6 – Moderately disagree 7 – Strongly disagree 8 – Do not know or does not apply	1 – Strongly disagree 2 – Moderately disagree 3 – Mildly disagree 4 – Neither agree nor disagree 5 – Mildly agree 6 – Moderately agree 7 – Strongly agree
"Your co-workers treat employees from all diversity groups with equal respect"	good_relationship_with_coworkers	1 – Strongly agree 2 – Moderately agree 3 – Mildly agree 4 – Neither agree nor disagree 5 – Mildly disagree 6 – Moderately disagree 7 – Strongly disagree 8 – Do not know or does not apply	1 – Strongly disagree 2 – Moderately disagree 3 – Mildly disagree 4 – Neither agree nor disagree 5 – Mildly agree 6 – Moderately agree 7 – Strongly agree
"My agency actively encourages ethical behaviour by all of its employees"	ethical_practices	1 – Strongly agree 2 – Moderately agree 3 – Mildly agree 4 – Neither agree nor disagree 5 – Mildly disagree 6 – Moderately disagree 7 – Strongly disagree 8 – Do not know or does not apply	1 – Strongly disagree 2 – Moderately disagree 3 – Mildly disagree 4 – Neither agree nor disagree 5 – Mildly agree 6 – Moderately agree 7 – Strongly agree
"Conflicts of interest are identified and managed effectively in my workplace"	conflicts_immediately_resolved	1 – Strongly agree 2 – Moderately agree 3 – Mildly agree 4 – Neither agree nor disagree 5 – Mildly disagree 6 – Moderately disagree 7 – Strongly disagree 8 – Do not know or does not apply	1 – Strongly disagree 2 – Moderately disagree 3 – Mildly disagree 4 – Neither agree nor disagree 5 – Mildly agree 6 – Moderately agree 7 – Strongly agree
"What is the highest level of formal education you have completed?"	education_level	1 – Primary/secondary school or technical/trade certificate or diploma 2 – University qualification	1 – Primary/secondary school or technical/trade certificate or diploma 2 – University qualification
"What is your current total annual gross salary (before tax)?"	salary_rate	1 – Less than $60,000 2 - $60,000 to $79,999 3 - $80,000 to $109,999 4 - $110,000 to $159,999 5 - $160,000 and over	1 – Less than $60,000 2 - $60,000 to $79,999 3 - $80,000 to $109,999 4 - $110,000 to $159,999 5 - $160,000 and over

Class Variable			
Survey Question/Section	**Given Feature Name**	**Possible Responses from Employee**	**Possible Responses from Employee after Data Cleaning and Preprocessing**
Please indicate your level of satisfaction with: My job overall	job_satisfaction_level	1 – Very satisfied 2 – Moderately satisfied 3 – Mildly satisfied 4 – Neither satisfied nor dissatisfied 5 – Mildly dissatisfied 6 – Moderately dissatisfied 7 – Very dissatisfied	**Likes_job = 0:** 1 – Very dissatisfied 2 – Moderately dissatisfied 3 – Mildly dissatisfied **Likes_job = 1:** 5 – Mildly satisfied 6 – Moderately satisfied 7 – Very satisfied

There was an imbalance of data where there was significantly more employees having high satisfaction ratings than low satisfaction. To alleviate this issue, all 2015 survey answers where employees expressed low satisfaction for their job was combined together with the 2016 survey (Badr, 2019). This process was done using Excel. An imbalance of data for this project would result in possibly having high accuracy, precision, and/or recall when predicting employees liking their job. However, if predicting job dissatisfaction it could present problems such as low accuracy, precision, and/or recall. The goal is for the machine learning model to predict both class variables (satisfied and dissatisfied with job) with high accuracy and precision.

For each section of the survey as shown in Table 2, employees gave a reverse scoring (Likert scale) from 1 (Very satisfied/Strongly agree) to 7 (Very dissatisfied/Strongly disagree). It is called "reversing scoring" because typically the higher the rating, the better or more satisfied with the section. A rating of 4 indicates a neutral answer and 8 indicates "do not know or does not apply." Rows with blank answers, "N/A's", and answer 8 were removed from the data set. Employees that gave a neutral answer (score of 4) under the "job_satisfaction_level" section of the survey was also removed from the data set.

As a result, the combined total of employee responses used in the data analysis was 4084 (3883 responses from 2016 plus 201 employees from 2015 with low satisfaction rating). There was a total of 2691 (65.9%) employees with positive job satisfaction and 1393 (34.1%) employees with negative job satisfaction. This CSV file or dataset was used for this machine learning project using Python. The name of this file was "public-sector-commission-eps-2016and2015.csv" as seen in the code on Jupyter Notebook (Appendix A and Appendix B). An example of the data set with the final seven features and job_satisfaction_level as the dependent variable was displayed in Table 2.

Table 2. A sample view of the dataset of employees' responses for each section. Each row represents an employee's answers for each question/section of the survey. Data table named "df" in Appendix A and B.

job_satisfaction_level	level_of_skill_used	expectations_clarity	contribution_recognition	career_development	well_managed	worklife_balance	effective_communication
2	1	1	1	3	6	3	2
1	1	1	1	1	1	1	1
6	3	3	4	4	7	2	7
7	7	7	7	7	7	7	7
7	5	7	5	7	6	3	7

In order to reverse the rating scale to 1 as the worst and 7 as the best, this process was done using Python. For example, an employee that gave a rating of 1 for work-life balance would be changed to 7. Therefore, the original scale was that 1 meaning very satisfied/strongly agree to 7 meaning very dissatisfied/strongly disagree was converted to 1 meaning very dissatisfied/strongly disagree to 7 meaning very satisfied/strongly agree. Table 3 was a sample view of what the dataset looks like in the Pandas table created in Python. The values from Table 2 was flipped to what was observed in Table 3.

Table 3. Sample view of the data table after changing the ratings from what was observed in Table 2 to the higher the number, the better the score for the given section. Named "df2" in Appendix A and B

job_satisfaction_level	level_of_skill_used	expectations_clarity	contribution_recognition	career_development	well_managed	worklife_balance	effective_communication
6	7	7	7	5	2	5	6
7	7	7	7	7	7	7	7
2	5	5	4	4	1	6	1
1	1	1	1	1	1	1	1
1	3	1	3	1	2	5	1

Since the desired class variables in this project are "likes job" and "dislikes job," a function was used in Python to convert the scale from the "job_satisfaction_level" section to a Boolean type (encoding). Positive satisfaction levels (employees who gave a rating of 5 or higher) was counted as 1 or "satisfied with job" while negative satisfaction levels (employees who gave a rating of 3 or less) was counted as 0 or "dissatisfied with job." In Table 4, the column "likes_job" highlighted in green would be the class variable attempted to be predicted with a model in this project. A value of 1 indicated the employee is satisfied with their current job while 0 meant that they are not satisfied with their current job.

Table 4. An example of the data set where "job_satisfaction_level" was converted to a Boolean data type under "likes_job." The class variable to be predicted was highlighted in green. A value of 1 indicated that the employee likes or was satisfied with their current job while a value of 0 indicated that the employee does not like or was dissatisfied with their job. Data frame "df2" in Appendix A and B.

job_satisfaction_level	level_of_skill_used	expectations_clarity	contribution_recognition	career_development	well_managed	worklife_balance	effective_communication	likes_job
6	7	7	7	5	2	5	6	1
7	7	7	7	7	7	7	7	1
2	5	5	4	4	1	6	1	0
1	1	1	1	1	1	1	1	0
1	3	1	3	1	2	5	1	0

In order to prepare to study of just the features used to predict the class variable, the data set was transformed or converted to what was observed in Table 5. Each row represents the recorded answers of an employee for the seven sections of the survey.

Table 5. Transformed data set of the features to prepare for the Apriori algorithm. The ratings were converted to the actual meaning of the employees' feedbacks to make it easier to understand when finding the association rules. Table 1 – Possible Responses from Employee column and the values in Table 4 (except for "likes_job" and "job_satisfaction_level") should match what is shown in this table. For example, in the first row (index = 0) "level_of_skill_used" was rated 7 and therefore the employee "strongly agrees" that their job allows them to fully utilize their skills. Data frame "df_a" in Appendix A and B.

	level_of_skill_used	expectations_clarity	contribution_recognition	career_development	well_managed	worklife_balance	effective_communication
0	level_of_skill_used = Strongly agree	expectations_clarity = Strongly agree	contribution_recognition = Strongly agree	career_development = Mildly agree	well_managed = Moderately disagree	worklife_balance = Mildly agree	effective_communication = Moderately agree
1	level_of_skill_used = Strongly agree	expectations_clarity = Strongly agree	contribution_recognition = Strongly agree	career_development = Strongly agree	well_managed = Strongly agree	worklife_balance = Strongly agree	effective_communication = Strongly agree
2	level_of_skill_used = Mildly agree	expectations_clarity = Mildly agree	contribution_recognition = Neither agree nor d...	career_development = Neither agree nor disagree	well_managed = Strongly disagree	worklife_balance = Moderately agree	effective_communication = Strongly disagree
3	level_of_skill_used = Strongly disagree	expectations_clarity = Strongly disagree	contribution_recognition = Strongly disagree	career_development = Strongly disagree	well_managed = Strongly disagree	worklife_balance = Strongly disagree	effective_communication = Strongly disagree
4	level_of_skill_used = Mildly disagree	expectations_clarity = Strongly disagree	contribution_recognition = Mildly disagree	career_development = Strongly disagree	well_managed = Moderately disagree	worklife_balance = Mildly agree	effective_communication = Strongly disagree
...	...	...	...	...	...	...	...
4079	level_of_skill_used = Neither agree nor disagree	expectations_clarity = Neither agree nor disagree	contribution_recognition = Strongly disagree	career_development = Strongly disagree	well_managed = Strongly disagree	worklife_balance = Strongly disagree	effective_communication = Strongly disagree
4080	level_of_skill_used = Moderately agree	expectations_clarity = Strongly agree	contribution_recognition = Moderately disagree	career_development = Mildly agree	well_managed = Moderately agree	worklife_balance = Mildly agree	effective_communication = Moderately agree
4081	level_of_skill_used = Neither agree nor disagree	expectations_clarity = Neither agree nor disagree	contribution_recognition = Strongly disagree	career_development = Strongly disagree	well_managed = Moderately disagree	worklife_balance = Mildly disagree	effective_communication = Neither agree nor di...
4082	level_of_skill_used = Mildly agree	expectations_clarity = Mildly agree	contribution_recognition = Mildly disagree	career_development = Mildly disagree	well_managed = Moderately agree	worklife_balance = Mildly disagree	effective_communication = Mildly disagree
4083	level_of_skill_used = Strongly agree	expectations_clarity = Strongly agree	contribution_recognition = Strongly disagree	career_development = Strongly disagree	well_managed = Moderately disagree	worklife_balance = Strongly disagree	effective_communication = Strongly disagree

A dataset with 13 explanatory variables (can be seen in Table 1) was also used for the study to compare how much of an impact those features made to the accuracy of the model and when the Apriori algorithm was used. The Results section of this paper goes more into detail.

Exploratory Data Analysis

Histograms of all 13 selected features was explored. In the count plots shown in Figure 1, a majority of the features show correlations with the class variable. From scores 1 to 4 for all features except "conflicts_immediately_resolved," "education_level" and "salary_rate," the number of employees who dislike their job was greater than the number of employees who like their job. While for ratings 5 and higher, the number of satisfied employees was greater than the number of dissatisfied employees. For conflict resolution (conflicts_immediately_resolved), there was an equal amount of satisfied and dissatisfied employees for ratings 3 and 4 but a greater amount of unhappy employees from ratings 1 to 2. There was an increase in numbers for both satisfied and dissatisfied employees for level of education (education_level). Education level had two categories: 1 – Primary/secondary school or technical/trade certificate or diploma and 2 – University qualification in which both of them has more satisfied employees than dissatisfied employees. Salary rate appeared to be normally distributed and in every category, there was more employees who like their job than employees who do not like their job.

Table 6 shows the summary of the average rating of each section given by employees satisfied with their job and dissatisfied with their job. The difference between the scores was obvious for the top eight on the list: job satisfaction level, skills utilized, clear expectations, recognized contribution, career

development, well-managed organization, work-life balance, and effective communication. While the scores given by two different type of employees for the factors respectful supervisor, relationship with coworkers, ethical practices, conflict resolution, education level, and salary was close.

For further interpretation of Table 6, by going through each factor the average satisfaction level of satisfied employees was “moderately satisfied” while it was “moderately dissatisfied” for dissatisfied employees. This made sense. For the other non-highlighted work factors, employees who on average mildly to moderately agree that their organization provide those features to them was satisfied with their job. On the other hand, employees who mildly disagree to neither agree nor disagree was not satisfied with their job. As seen in Table 6, there was not much difference in the highlighted work factors for both types of employees meaning they did not have much impact on job satisfaction. In fact, there was more dissatisfied employees with university qualifications than ones with technical certificates or high school diplomas. Another interesting find was the salaries of the employees. Consistent with Figure 1 where salary_rate showed a normal distribution, the salaries of satisfied employees compared to dissatisfied employees was similar. Because this data set was a survey completed by public sector employees, this suggested that monetary compensation has little influence on job satisfaction for employees working in government agencies. Studies have also suggested that employees would more likely to report higher job satisfaction levels the closer their job is related to their education (Lee & Sabharwal, 2016).

In order to check for multicollinearity where two or more explanatory variables are highly linearly related to eachother, a correlation matrix was created as shown in Table 7. Looking a the pairwise correlation values, the highest coefficient was between “contribution_recognition” and “career_development” which was 0.698771. Multicollinearity in a data set would be regarded as severe if the correlation coefficient between the features was at least 0.90 (Bowerman, et al., 2019). This was not true for this data set since the largest coefficient in Table 7 was less than 0.90.

Table 6. Comparison of the average scores (Appendix B) of work factors given by satisfied and dissatisfied employees. Factors highlighted in yellow showed the least difference between satisfied and dissatisfied employees especially for education level and salary.

Work factor	Satisfied with Job	Dissatisfied with Job	Δ Difference (Satisfied – Dissatisfied)
job_satisfaction_level	6.130435	2.228284	3.902151
level_of_skill_used	5.997770	3.634602	2.363168
expectations_clarity	6.136009	4.388370	1.747639
contribution_recognition	5.156076	2.663317	2.492759
career_development	4.478632	2.213209	2.265423
well_managed	4.861018	2.692032	2.168986
worklife_balance	5.611669	3.863604	1.748065
effective_communication	5.049052	2.789663	2.259389
respectful_supervisor	6.284653	5.235463	1.049190
good_relationship_with_coworkers	6.169082	5.443647	0.725435
ethical_practices	6.093274	5.016511	1.076763
conflicts_immediately_resolved	5.518395	3.800431	1.717964
education_level	1.584169	1.686289	-0.102120
salary_rate	2.690450	2.651831	0.038619

Table 7. Correlation matrix for the employee survey data. Pairwise correlations between the features used to predict job satisfaction are shown. All coefficients are less than 0.9 meaning multicollinearity was not severe between the independent variables. Code "x.corr()" in Appendix B.

	level_of_skill_used	expectations_clarity	contribution_recognition	career_development	well_managed	worklife_balance	effective_communication	respectful_supervisor	good_relationship_with_coworkers	ethical_practices	conflicts_immediately_resolved	education_level	salary_rate
level_of_skill_used	1	0.559235	0.602459	0.582643	0.512888	0.384257	0.508481	0.379405	0.302235	0.415736	0.453511	-0.048319	0.088491
expectations_clarity	0.559235	1	0.504209	0.451326	0.481366	0.338756	0.503434	0.344156	0.261561	0.410757	0.425057	-0.035048	-0.017058
contribution_recognition	0.602459	0.504209	1	0.698771	0.63194	0.516167	0.652432	0.440224	0.338992	0.45102	0.540943	-0.037957	0.056214
career_development	0.582643	0.451326	0.698771	1	0.609257	0.425575	0.572997	0.351345	0.300468	0.407488	0.497393	-0.060798	0.054183
well_managed	0.512888	0.481366	0.63194	0.609257	1	0.492068	0.649121	0.369086	0.307182	0.51989	0.596498	-0.059992	-0.008584
worklife_balance	0.384257	0.338756	0.516167	0.425575	0.492068	1	0.487506	0.389899	0.294164	0.350381	0.431643	-0.033523	0.036455
effective_communication	0.508481	0.503434	0.652432	0.572997	0.649121	0.487506	1	0.471082	0.358373	0.489228	0.58969	-0.041553	0.044992
respectful_supervisor	0.379405	0.344156	0.440224	0.351345	0.369086	0.389899	0.471082	1	0.597125	0.446082	0.454056	-0.005886	0.077747
good_relationship_with_coworkers	0.302235	0.261561	0.338992	0.300468	0.307182	0.294164	0.358373	0.597125	1	0.455408	0.408546	0.016328	0.093796
ethical_practices	0.415736	0.410757	0.45102	0.407488	0.51989	0.350381	0.489228	0.446082	0.455408	1	0.637306	-0.03781	0.032769
conflicts_immediately_resolved	0.453511	0.425057	0.540943	0.497393	0.596498	0.431643	0.58969	0.454056	0.408546	0.637306	1	-0.032349	0.069175
education_level	-0.048319	-0.035048	-0.037957	-0.060798	-0.059992	-0.033523	-0.041553	-0.005886	0.016328	-0.03781	-0.032349	1	0.366699
salary_rate	0.088491	-0.017058	0.056214	0.054183	-0.008584	0.036455	0.044992	0.077747	0.093796	0.032769	0.069175	0.366699	1

Figure 1. Histograms, or more specifically count plots of the employee survey data. Bars in blue represent employees dissatisfied with their job while orange represents employees satisfied with their job. Code in Appendix B under "Exploratory Data Analysis."

Results

Supervised Learning

Three machine learning models was explored to predict employee job satisfaction: logistic regression, decision tree, and Naïve Bayes. Different number of features was also used for comparison. The three algorithms were applied with all thirteen explanatory variables shown in Table 1 as well as seven of the best features which was level_of_skill_used, expectations_clarity, contribution_recognition, career_development, well_managed, worklife_balance, and effective_communication. Both feature sets were compared in order to see if removing certain features would affect the evaluation (accuracy, precision, and recall) of the models.

Supervised Learning – Logistic Regression

Logistic regression was the first algorithm used for this experiment. The model was initially created to fit the data to check the coefficients of the model in order to get a better idea of the best predictors for job satisfaction. Furthermore, the data was also split into training and testing sets to check for any improvements of the logistic regression model. It was noticed that after removing several features, the coefficients on the right table was barely affected (Table 8). The coefficients of respectful supervisor, relationship with coworkers, ethical practices, conflict resolution, education level, and salary rate shown in Table 8 showed that they did not contribute much in predicting the class variable (whether the employee likes their job) since their values are close to zero. Although education_level appeared to have a high negative value, it only ranges from 1 to 2 compared to the other features ranging from 1 to 7 and therefore was not selected as one of the seven features used to make predictions. The level of skill used for the job was determined to be the most important factor for job satisfaction. It is also interesting to note that salary and education level did not contribute much when predicting job satisfaction. For salary it would make sense since the data set was collected from a public sector where workers value other salary less, and education was a negative coefficient which could possibly mean that a majority of the employees think that their job does not meet the standards of their university qualifications. In other words, given their education background, employees may feel that they are overqualified for their current role.

Table 8. Comparison of the coefficients between 13 features (left) and 7 best features (right). These values were prior to splitting the data into the training and testing set for the logistic regression model. Named "coeff_d2" in Appendix A and B.

Features	Coefficients
level_of_skill_used	0.464673
expectations_clarity	0.209938
contribution_recognition	0.205228
career_development	0.100158
well_managed	0.123334
worklife_balance	0.207161
effective_communication	0.137075
respectful_supervisor	0.00127
good_relationship_with_coworkers	-0.009691
ethical_practices	-0.146294
conflicts_immediately_resolved	0.047392
education_level	-0.543386
salary_rate	0.018917

Features	Coefficients
level_of_skill_used	0.453104
expectations_clarity	0.191388
contribution_recognition	0.199236
career_development	0.105925
well_managed	0.114778
worklife_balance	0.203463
effective_communication	0.124227

For the data set with thirteen features, the accuracy score by just fitting the data into the logistic regression model was 0.83986, while the data set with seven features had an accuracy score of 0.83570. The data set with more features only has a slightly higher accuracy score compared to the data set with less features.

After cross validation was used to train and test the data set, classification reports for both feature sets were generated to evaluate the logistic regression model. A summary of the reports is shown in Table 9. The improvement in model accuracy of having thirteen features was determined to be small, so seven strongest features were chosen for the model. For both feature sets, the precision and recall (specificity) when predicting dissatisfaction was determined to be lower than the precision and recall (sensitivity) for predicting satisfaction.

Table 9. Accuracy, precision, and recall for both feature sets summarized from their classification report in Python. After the data was cross validated by splitting into a training and testing set for the logistic regression algorithm, it was determined that having more features only improved the accuracy by a tiny amount.

13 Features	**Precision**	**Recall**	**Accuracy**
Satisfied with job	0.86	0.91	0.85
Dissatisfied with job	0.81	0.73	

7 Features	**Precision**	**Recall**	**Accuracy**
Satisfied with job	0.86	0.91	0.84
Dissatisfied with job	0.80	0.71	

Confusion matrix tables (Figure 2) was created to look at the performance of the logistic regression model using the seven best features. It was unnecessary to display the confusion matrix for

the model with all thirteen features since it would look similar any way. The values represent the class variable “likes_job” in the Pandas data frame where 0 indicated negative (employee does not like job) and 1 was a positive result (employee likes job). By looking at the confusion matrix and the recall value (0.71) for predicting job dislike, the model was determined to be at risk of giving false positive results. Meaning it could possibly predict the employee being satisfied with their job even though they are really dissatisfied. The confusion matrix on the left shows the exact numbers in each box while the confusion matrix on the right gives a more visual display. The true negatives and positives was determined to have the darkest coloured squares as expected.

		Predicted		
		0	**1**	**All**
Actual	**0**	335	134	469
	1	82	797	879
	All	417	931	1348

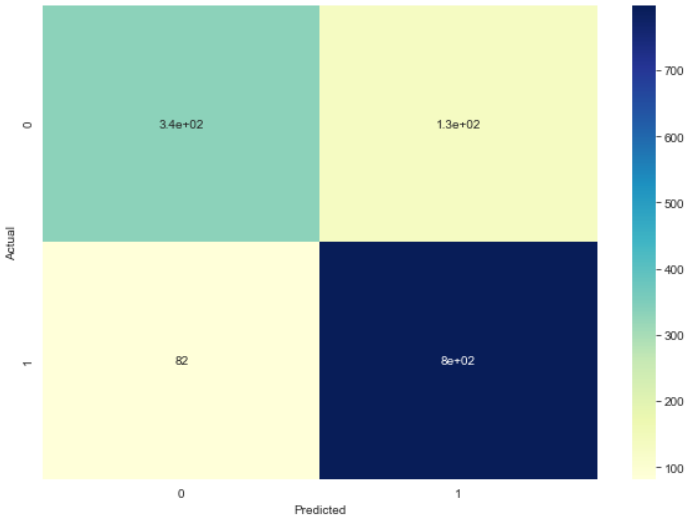

Figure 2. Confusion matrix to describe the performance on the logistic regression model on the employee job satisfaction data set with the seven best features. The value 0 indicates that the employee was not satisfied with their job while 1 indicates that the employee likes their job. The rows indicate the true (actual) result while columns represent predicted results from the model. The confusion matrix on the left shows the exact numbers in each box while the confusion matrix on the right gives a more visual display. The true negatives and positives was determined to have the darkest colour squares as expected.

Supervised Learning – Decision Tree

With the same two sets of features, a summary of the evaluations using the decision tree method was shown in Table 10. Accuracy was lower by about 0.01 or 1% compared to the logistic regression model (Table 9). There was no significant difference in accuracy having more features, so seven features were ultimately chosen for the model. Precision and Recall was respectively 0.76 and 0.74 which are close to each other. The specificity (recall) was also slightly higher than the recall from the logistic regression model which was 0.71. Therefore, if minimizing false positives were more important than false negatives then that would be then the decision tree model would have been chosen. However, overall the logistic regression model would still be the best choice for predicting job satisfaction or dissatisfaction.

The confusion matrix in Figure 3 also visually shows the improvement in the recall in the decision tree model compared to the logistic regression one in Figure 2. The top right square was a lighter colour meaning less false positives.

Table 10. Accuracy, precision, and recall for both feature sets summarized from their classification report in Python. After the data was cross validated by splitting into a training and testing set for the decision tree algorithm, it was determined that having more features only improved the accuracy by a tiny amount.

13 Features	**Precision**	**Recall**	**Accuracy**
Satisfied with job	0.86	0.89	0.84
Dissatisfied with job	0.79	0.73	

7 Features	**Precision**	**Recall**	**Accuracy**
Satisfied with job	0.86	0.87	0.83
Dissatisfied with job	0.76	0.74	

		Predicted		
		0	**1**	**All**
Actual	**0**	345	124	469
	1	110	769	879
	All	455	893	1348

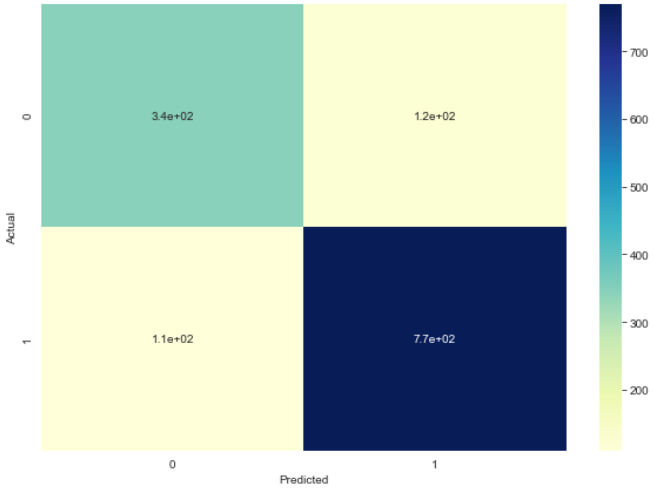

Figure 3. Confusion matrix tables to evaluate the performance of the decision tree model with the seven best features. The values represent the class variable "likes_job" where 1 means positive (true) and 0 means negative (false). Recall in this model was better than the one in the logistic regression model, but the score for precision was lower.

Supervised Learning – Naïve Bayes

Out of the three algorithms, Naïve Bayes overall performed the worst. However, what was interesting was that the data set with all thirteen independent variables was 0.11 or 11% more accurate than the one with less features. The reports in Table 11 showed that both feature sets would do a decent job at predicting satisfied employees, but would be a bad classifier for predicting dissatisfied employees especially with seven features. This was also visually confirmed in Figure 4. The two darkest shades should be the top left (true negative) and bottom right (true positive). For the data set with seven features the top right square was dark which meant that it the model would likely predict the employee likes their job even it should predict that they dislike it.

Table 11. Accuracy, precision, and recall for both feature sets summarized from their classification report in Python. After the data was cross validated by splitting into a training and testing set for the Naïve Bayes algorithm, it was determined that having more features only improved the accuracy by about 11%. The top table with thirteen features was determined to have an accuracy of 79% which was higher than the bottom table with seven features where the accuracy was 68%.

13 Features	**Precision**	**Recall**	**Accuracy**
Satisfied with job	0.82	0.87	0.79
Dissatisfied with job	0.73	0.65	

7 Features	**Precision**	**Recall**	**Accuracy**
Satisfied with job	0.69	0.92	0.68
Dissatisfied with job	0.61	0.23	

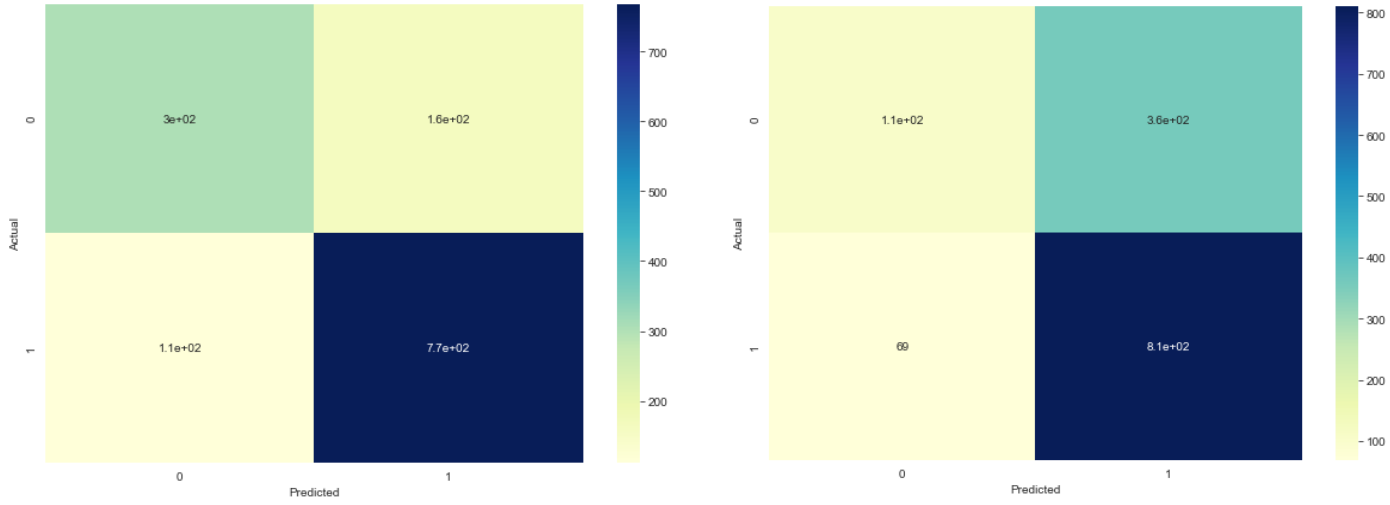

Figure 4. Side-by-side comparison of the two confusion matrix tables with the Naïve Bayes algorithm. The one on the left represents the data set with thirteen explanatory variables while the one on the right represents the data set with seven explanatory variables. The confusion matrix on the right is prone to false positives and does not accurately predict true negatives. The values represent the class variable "likes_job" where 1 means positive (true) and 0 means negative (false).

Supervised Learning – Model Selection and Further Explanations of the Results

Cross validation method of splitting training and testing the data set was used for the three algorithms to prevent overfitting, where the model has good accuracy on the training dataset but poor accuracy on the testing dataset. For both data sets with thirteen and seven features, ROC (Receiver Operating Characteristic) curves of each model was generated for comparison. It was used for the purpose of measuring the performance of the classifier by obtaining the AUC (Area Under the Curve) score. The higher the AUC, the better the model is at predicting employee job satisfaction. Figure 5 shows the ROC curves where it is a plot with the true positive rate against the false positive rate. For both feature sets, logistic regression was proven to be the best model for predicting whether the employee likes their job. The second best was decision tree, and the worst was the Naïve Bayes model.

Table 12 gave a summary of the accuracy and AUC scores of the models for both data sets. There was not much of a difference in accuracy and AUC between their logistic regression and decision tree models. However, the biggest difference in accuracy and AUC was with the Naïve Bayes algorithm. Several possible reasons for the data set with seven features having lower scores than the one with thirteen features. One of them could be the assumption in Naïve Bayes that the features are independent, and there may have been multicollinearity with some of the independent variables. In reality it is nearly impossible to have completely independent predictors. Another reason could be some features that might have been strong predictors with the Naïve Bayes model was removed, resulting in a significant decrease in accuracy. Overall this could have caused issues when classifying satisfied and dissatisfied employees based on probabilities.

Decision tree was used since it could be built for classification models such as this one. It performed better than the Naïve Bayes model but slightly worse than the logistic regression model. A risk of using decision trees was that it could have possibly overfitted the data.

Therefore, the results have shown that the logistic regression was the best model for predicting if employees are satisfied or not satisfied with their job. This was also proven as a suitable model since the purpose of this experiment was to predict a value between 0 (employee does not like job) and 1 (employee likes job), corresponding to the probability that the Boolean outcome variable was 1. In order to avoid the curse of dimensionality, the most efficient logistic regression model was with the data set with seven best features instead of thirteen features. Furthermore, as shown in Figure 5 and Table 12 having more features only improved the accuracy and AUC scores by an insignificant amount.

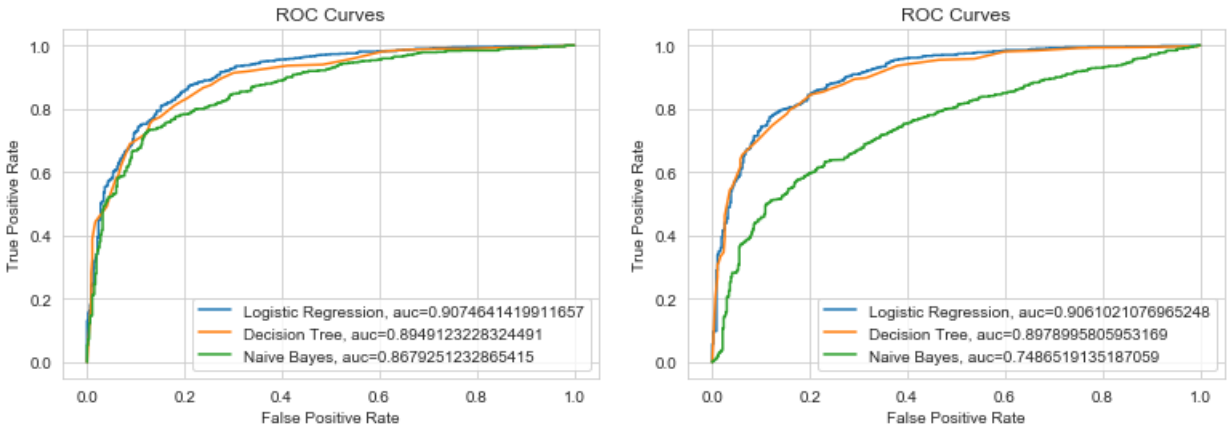

Figure 5. Left: ROC curves of the three algorithms used for the data set with thirteen explanatory variables. Right: ROC curves of the three algorithms used for the data set with seven explanatory variables. The AUC (Area Under the ROC Curve) score was also given in the legend. For both graphs, the ranking of algorithms from best to worst was the following: logistic regression, decision tree, and Naïve Bayes.

Table 12. Summary of the accuracy and AUC scores for each model. Table on the left are the scores with the data using thirteen features while the table on the right was for the data with seven features. Overall there was not much of a difference between the scores except for the Naïve Bayes model. The data set with seven features using the logistic regression model would be the most efficient for making predictions on employee job satisfaction.

13 Features		
Model	**Accuracy**	**AUC**
Logistic Regression	0.85	0.907
Decision Tree	0.84	0.895
Naïve Bayes	0.79	0.868

7 Features		
Model	**Accuracy**	**AUC**
Logistic Regression	0.84	0.906
Decision Tree	0.83	0.898
Naïve Bayes	0.68	0.749

Unsupervised Learning – Apriori Algorithm

The seven following features was used to attempt to find meaningful association rules using the Apriori algorithm: skill level used for the job (level_of_skill_used), expectation clearness (expectations_clarity), how much contributions are recognized (contribution_recognition), career development (career_development), how well-managed is the organization (well_managed), work-life balance (worklife_balance), and how effective is the communication within the organization (effective_communication). Due to the curse of dimensionality, the data set with all thirteen features was omitted from this experiment as it was too difficult to find meaningful association rules with it. It had resulted in too many rules, and therefore sticking to seven features was the better option.

At the beginning of the analysis of the data, several support and confidence parameters was experimented with. The parameters that provided the most insight was when support threshold was set to 0.02 (2%) and confidence threshold was 0.9 (90%). A total of 59 rules was given. Table 14 displayed the top ten rows where the rules were the most robust or has the highest lift value. The confidence

values were also strong. Lift was used as the measure of quality for the association rules, and having a lift higher than 1 indicates that the rule is high quality and reliable (Girmanova & Gasparova, 2018).

Four of the rules out of fifty-nine (as shown in Table 13) was found to be most insightful and interpretable with those set parameters in no particular order. In the first rule, it translates to if the employee strongly does not agree that there is a good work-life balance and their job does not require much skill then they would strongly disagree that there are any career development opportunities within the organization. The likeliness of the employee going by the first rule would be 90.7407% (confidence of 0.907407) which is high. The important of this association rule was determined to quite high as the lift value was 3.601411. The following shows a sample interpretation of support, confidence, and lift as formulas with the first rule. (1) represents support, equation (2) represents confidence, and (3) represents lift:

(1)

$$\text{Support}\{\text{worklife\_balance} = \text{Strongly disagree}, \text{level\_of\_skill\_used} = \text{Strongly disagree}, \text{career\_development} = \text{Strongly disagree}\} = 0.023996$$

(2)

$$\text{Confidence}\{\text{worklife\_balance} = \text{Strongly disagree}, \text{level\_of\_skill\_used} = \text{Strongly disagree} \rightarrow \text{career\_development} = \text{Strongly disagree}\} = \frac{0.023996}{\text{Support}\{\text{worklife\_balance=Strongly disagree}, \ \text{level\_of\_skill\_used=Strongly disagree}\}} = 0.907407$$

(3)

$$\text{Lift}\{\text{worklife\_balance} = \text{Strongly disagree}, \text{level\_of\_skill\_used} = \text{Strongly disagree} \rightarrow \text{career\_development} = \text{Strongly disagree}\} = \frac{0.023996}{\text{Support}\{\text{worklife\_balance} = \text{Strongly disagree}, \ \text{level\_of\_skill\_used} = \text{Strongly disagree}\} \text{ x Support}\{\text{career\_development} = \text{Strongly disagree}\}} = 3.601411$$

Rule number 2 in Table 13 showed that with 90% confidence if the employee thinks that the organization is very well managed, strongly agrees that there is career development, and a very high rating for work-life balance, then they would strongly agree that their job allows them to effectively utilize their skills. Rule number 3 has the highest confidence of 94.1558% out of the four rules where if the employee strongly agrees that the organization is well managed, has many opportunities for career development, and communication is effective, then they would strongly agree that the expectations of their responsibilities would be clear. Rule number 4 has the highest lift value of 5.639139 out of the rules in Table 13. If the employee strongly agrees that communication is effective, the organization is well-managed, the level of skill for the job is appropriate, and there are opportunities for career development, then they would strongly agree that they would be recognized for their contributions.

Table 13. Four rules that stood out from the Apriori algorithm between the seven features used to predict job satisfaction. Set minimum support at 0.02 and minimum confidence at 0.9.

Number	Rules	Support	Confidence	Lift
1	{worklife_balance = Strongly disagree, level_of_skill_used = Strongly disagree} → {career_development = Strongly disagree}	0.023996	0.907407	3.601411
2	{well_managed = Strongly agree, career_development = Strongly agree, worklife_balance = Strongly agree} → {level_of_skill_used = Strongly agree}	0.039912	0.900552	3.195357
3	{well_managed = Strongly agree, career_development = Strongly agree, effective_communication = Strongly agree} → {expectations_clarity = Strongly agree}	0.035504	0.941558	2.570404
4	{effective_communication = Strongly agree, well_managed = Strongly agree, level_of_skill_used = Strongly agree, career_development = Strongly agree} → {contribution_recognition = Strongly agree}	0.030607	0.905797	5.639139

Table 14. Sample view of the 10 most robust rules (highest lift) with strong confidence out of the 59 rules (last line of code in Appendix A). The support parameter was set at 0.02 or 2% minimum and confidence threshold was set at 0.9.

	antecedents	consequents	support	confidence	lift
42	(well_managed = Strongly agree, level_of_skill_used = Strongly agree, expectations_clarity = Strongly agree, career_development = Strongly agree, effective_communication = Strongly agree)	(contribution_recognition = Strongly agree)	0.029628	0.909774	5.663901
22	(effective_communication = Strongly agree, well_managed = Strongly agree, level_of_skill_used = Strongly agree, career_development = Strongly agree)	(contribution_recognition = Strongly agree)	0.030607	0.905797	5.639139
55	(well_managed = Strongly agree, level_of_skill_used = Strongly agree, expectations_clarity = Strongly agree, career_development = Strongly agree, effective_communication = Strongly agree, worklife_balance = Strongly agree)	(contribution_recognition = Strongly agree)	0.025465	0.904348	5.630117
58	(well_managed = Strongly agree, career_development = Strongly agree, contribution_recognition = Strongly agree, effective_communication = Strongly agree, worklife_balance = Strongly agree)	(level_of_skill_used = Strongly agree, expectations_clarity = Strongly agree)	0.025465	0.928571	4.419913
45	(contribution_recognition = Strongly agree, well_managed = Strongly agree, career_development = Strongly agree, effective_communication = Strongly agree)	(level_of_skill_used = Strongly agree, expectations_clarity = Strongly agree)	0.029628	0.916667	4.363248
14	(worklife_balance = Strongly disagree, level_of_skill_used = Strongly disagree, contribution_recognition = Strongly disagree)	(career_development = Strongly disagree)	0.021303	0.915789	3.634679
35	(well_managed = Strongly disagree, level_of_skill_used = Strongly disagree, contribution_recognition = Strongly disagree, effective_communication = Strongly disagree)	(career_development = Strongly disagree)	0.026445	0.915254	3.632554
11	(expectations_clarity = Strongly disagree, contribution_recognition = Strongly disagree, effective_communication = Strongly disagree)	(career_development = Strongly disagree)	0.022282	0.910000	3.611701
12	(well_managed = Strongly disagree, contribution_recognition = Strongly disagree, expectations_clarity = Strongly disagree)	(career_development = Strongly disagree)	0.022037	0.909091	3.608093
0	(worklife_balance = Strongly disagree, level_of_skill_used = Strongly disagree)	(career_development = Strongly disagree)	0.023996	0.907407	3.601411

Conclusions

In summary, it was determined that the data set with seven features: skill level used for the job (level_of_skill_used), clarity of expectations (expectations_clarity), how recognized are employee contributions (contribution_recognition), opportunities for career growth and promotions (career_development), how well-managed is the organization/agency (well_managed), how is the employee's work-life balance (worklife_balance), and how well is the communication within the organization (effective_communication) were the most efficient for both supervised and unsupervised learning in this project. As shown in Table 8, the level of skills used for the job was the biggest factor for predicting job satisfaction since it has the highest coefficient value. This confirmed the fact that employee satisfaction increases as their job enables them to put their knowledge and skills to the right use, which allows them to perform at a higher level (Lee & Sabharwal, 2016).

The logistic regression model was proven to be the most accurate model for predicting job satisfaction or dissatisfaction of an employee with an accuracy 84% and an AUC score of 0.906. Therefore, from the supervised learning portion of this project employers should learn that given the coefficient values of the features in Table 8 they should focus on the factors that weighed the most when predicting employee job satisfaction. For example, the four highest coefficients ranged from 0.191 to 0.453 and they were (from lowest to highest) expectations clarity, contribution recognition, work-life balance, and level of skill used. With those four best predictors in mind, to improve employee satisfaction, the employer could give more challenging tasks to their employees, let them know what they expect (deadline, goals, etc.), praise them when they are performing well, and give employees a reasonable workload.

For noticeable behaviors of employees from the survey data set, four rules from the Apriori algorithm that were found to be most interesting was presented in Table 11. Rule 1 translates to if the employee thinks their skillset are not being tapped and their work-life balance is poor, then they would believe that there are no opportunities for career development in the organization. The second rule was that the employee would feel that their skills are put to the right use if they believe that the organization is well managed, there is potential for career growth, and they have an excellent work-life balance. In the third rule, the employee would think that if the organization is well managed, has a lot of room for career growth, and communication is excellent then they would be clear on what their employer expectations are. Finally, the employee would feel appreciated or their contributions are recognized if they think that communication between departments are effective, the organization is managed well, their skills are properly utilized, and there are career development opportunities.

Therefore, to improve job satisfaction in the organization the employer should give their employees more meaningful or challenging tasks so they would be able to demonstrate their skills and knowledge. The employer should give reasonable amount of workload or be more flexible with deadlines to improve work-life balance. As a result, employees may feel that they are building valuable skills to move up in their careers.

Overall, many things were learned from this project. In machine learning, data does most of the heavy lifting. This was evident during the data cleaning and pre-processing phase of the project. Changing the algorithm and the number of features only affected the accuracy of the job satisfaction

prediction model slightly. However, when more data was gathered there was significant improvement in the accuracy, precision, and recall of the model. As mentioned earlier in this paper, the model initially performed well when predicting satisfied employees but it was poor when predicting dissatisfied employees. After gathering and adding more data of dissatisfied employees from another year to the data set to be used, it significantly improved the model's performance in predicting dissatisfied employees. What was also learned was that the source of data could also affect the results during the analysis. For example, the weight of features may differ if the data set was taken from a profit-driven firm instead of a public sector. Instead of intrinsic values such as contribution recognition and skill development being the strongest predictors, an extrinsic value such as salary rate could be the biggest factor in predicting job satisfaction. There are many fancy data mining techniques for classification that this project has not touched base on such as XGBoost and ensemble learning which may be more accurate in making predictions. As for association rules, different insights could also be discovered if new combinations of features or data sets were to be used. As an example, age, gender, position, how long the employee is in the same company, and performance rating of the employee could be the new set of features in the data set for building prediction models and finding association rules as comparison to previous features used.

Employees and employers can both benefit from this analysis since it provides actionable insights for employers to improve job satisfaction and work environment, which coincides with decreased turnover rate of employees.

References

Badr, W. (2019, Feb 22). *Having an Imbalanced Dataset? Here is How You Can Fix It.* Retrieved from Towards Data Science: https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb

Bowerman, B. L., Hummel, R. M., Drougas, A. M., Moninger, K. B., Duckworth, W. M., Schur, P. J., & Froelich, A. G. (2019). *Business Statistics and Analytics in Practice.* New York: McGraw-Hill Education.

Deb, S. (2019, Jun 20). *Apriori Algorithm - Know How to Find Frequent Itemsets*. Retrieved from edureka!: https://medium.com/edureka/apriori-algorithm-d7cc648d4f1e

Flowers, V. S., & Hughes, C. L. (1973, July). *Why Employees Stay*. Retrieved from Harvard Business Review: https://hbr.org/1973/07/why-employees-stay

Garg, A. (2018, Sep 3). *Complete guide to Association Rules*. Retrieved from Towards Data Science: https://towardsdatascience.com/association-rules-2-aa9a77241654

Girmanova, L., & Gasparova, Z. (2018). Analysis of Data on Staff Turnover Using Association Rules and Predictive Techniques. *Quality Innovation Prosperity*, 82-99. doi:10.12776/QIP.V22I2.1122

Krogstad, U., Hofoss, D., Veenstra, M., & Hjortdahl, P. (2006). Predictors of job satisfaction among doctors, nurses and auxiliaries in Norwegian hospitals: relevance for micro unit culture. *Human Resources for Health*, 1-8. doi:10.1186/1478-4491-4-3

Kuzey, C. (2018). Impact of Health Care Employees' Job Satisfaction on Organizational Performance Support Vector Machine Approach. *Journal of Economics and Financial Analysis*, 45-68. doi: http://dx.doi.org/10.1991/jefa.v2i1.a12

Lee, Y.-j., & Sabharwal, M. (2016). Education–Job Match, Salary, and Job Satisfaction Across the Public,, Non-Profit, and For-Profit Sectors: Survey of recent college graduates. *Public Management Review*, 40-64. doi:10.1080/14719037.2014.957342

Public Sector Commission (Western Australia). (2015). Public Sector Commission WA Employee Perception Survey (2015). [Data File]. Retrieved from https://data.gov.au/data/dataset/public-sector-commission-wa-employee-perception-survey-2015

Public Sector Commission (Western Australia). (2016). Public Sector Commission WA Employee Perception Survey (2016). [Data file]. Retrieved from https://data.gov.au/data/dataset/b814c55a-d9d1-4145-af88-0fb78a354f8a

Punnoose, R., & Ajit, P. (2016). Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. *International Journal of Advanced Research in Artificial Intelligence*, 22-26. Retrieved from https://pdfs.semanticscholar.org/fa49/19810eaee67e851ad13775b78c94217a7908.pdf

Rencberoglu, E. (2019, April 1). *Fundamental Techniques of Feature Engineering for Machine Learning*. Retrieved from Towards Data Science: https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114

Sexton, R. S., McMurtrey, S., Michalopoulos, J. O., & Smith, A. M. (2004). Employee turnover: a neural network solution. *Elsevier*, 2635-2651. Retrieved from https://doi.org/10.1016/j.cor.2004.06.022

Sikaroudi, A. M., Ghousi, R., & Sikaroudi, A. E. (2015). A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing). *Journal of Industrial and Systems Engineering*, 106-121.

Suadicani, P., Bonde, J. P., Olesen, K., & Gyntelberg, F. (2013). Job Satisfaction and Intention to Quit the Job. *Oxford Academic*, 96-102. Retrieved from https://doi.org/10.1093/occmed/kqs233

Vezzoli, M. (2010). Exploring the Facets of Overall Job Satisfaction Through a Novel Ensemble Learning. *Electronic Journal of Applied Statistical Analysis, 4*(1), 23-38. doi:10.1285/i20705948v4n1p23

Zhang, L., Tan, X., Zhang, S., & Zhang, W. (2019). Association Rule Mining for Career Choices Among Fresh Graduates. *Science Publishing Group*, 37-43. doi:10.11648/j.acm.20190802.13

Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018). Employee Turnover Prediction with Machine Learning: A Reliable Approach. In K. Arai, S. Kapoor, & R. Bhatia, *Intelligent Systems and Applications* (pp. 737-758). Cham: Springer. doi:10.1007/978-3-030-01057-7_56

Appendix A: Python Code – Dataset with Seven Features

```
# Importing numpy and pandas
import numpy as np
import pandas as pd

# Importing plot libraries
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline
sns.set_style('whitegrid')

# Import employee survey dataset
df = pd.read_csv('public-sector-commission-eps-2016and2015.csv')
df

# Ratings except education_level and salary_rate originally reversed: 1 -> 7
= Best -> worst
# Changing to 1 -> 7 as from worst to best

df2 = df[['job_satisfaction_level','level_of_skill_used', 'expectations_clari
ty','contribution_recognition','career_development', 'well_managed','worklife
_balance', 'effective_communication', 'respectful_supervisor','good_relations
hip_with_coworkers', 'ethical_practices','conflicts_immediately_resolved']].r
eplace({1:7,2:6,3:5,4:4,5:3,6:2,7:1})

df2.head()

# Adding columns education_level and salary_rate back in
df2['education_level'] = df['education_level']
df2['salary_rate'] = df['salary_rate']

df2.head()

# Checking if ratings are changed correctly
df.head()

# Creating function to classify that the employee likes the job or not like.
def job_satisfaction(x):
    if x >= 5:
        return 1
    else:
        return 0

# Adding 'likes_job' column where 0 if employee does not like job and 1 meani
ng they like their job.
df2['likes_job'] = df2['job_satisfaction_level'].apply(job_satisfaction)
```

```
# Check table
df2

# To solve problem with curse of dimensionality. Removed these features.
# Lowest coefficients. Did not contribute much in predicting job satisfaction
.

del df2['salary_rate']
del df2['education_level']
del df2['ethical_practices']
del df2['good_relationship_with_coworkers']
del df2['respectful_supervisor']
del df2['conflicts_immediately_resolved']

df2

# Features. Dropping target variables.

x = df2.drop(['likes_job','job_satisfaction_level'], axis=1)
x

# Columns
x.columns

# Multicollinearity quick check
x.corr()

# Class variable

y = df2['likes_job']

# Machine Learning Imports. Logistic Regression

from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

# For evaluation of accuracy later
from sklearn import metrics

# Creating Logistic Regression Model. Not training model yet. For accuracy co
mparison.

log = LogisticRegression()

# Fitting data
log.fit(x,y)

# Accuracy check
log.score(x,y)
```

```
# Coefficients from the model

coeff_df2 = pd.DataFrame(x.columns, columns = ['Features'])
coeff_df2['Coefficients'] = np.ravel(log.coef_)

coeff_df2
```

Logistic Regression - Training and Testing Data Set

```
# Splitting the data. Test size set to 33% and arbitrary random_state = 100 t
o ensure reproducible results and split.

X_train, X_test, y_train, y_test = train_test_split(x, y, test_size = 0.33, r
andom_state = 100)

# Creating Logistic Regression model
log2 = LogisticRegression()

# Fitting new model
log2.fit(X_train, y_train)

# Predict the classes of the testing data set. Employee likes job or not.
y_predict = log2.predict(X_test)

# Accuracy score
metrics.accuracy_score(y_test, y_predict)

# Classification report
from sklearn.metrics import classification_report

print(classification_report(y_test, y_predict))

# Confusion matrix

pd.crosstab(y_test,y_predict,rownames=['Actual'],colnames=['Predicted'],margi
ns=True)

# Visualized confusion matrix

from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_predict)
plt.figure(figsize = (10,7))
sns.heatmap(cm, annot = True, cmap="YlGnBu")
plt.xlabel('Predicted')
plt.ylabel('Actual')
```

Decision Tree

```
# Decision tree imports
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier, export_graphviz
from sklearn.model_selection import train_test_split
```

```
# Train test split
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.33, ran
dom_state=100)

# Decision tree model
c = DecisionTreeClassifier(min_samples_split=101)

# Fit model
c.fit(X_train,y_train)

# Predict classes
y_pred = c.predict(X_test)

# Accuracy score
metrics.accuracy_score(y_test, y_pred)

# Classification report
print(classification_report(y_test, y_pred))

# Confusion matrix

pd.crosstab(y_test,y_pred,rownames=['Actual'],colnames=['Predicted'],margins=
True)

# Visualized Confusion matrix

cm = confusion_matrix(y_test, y_pred)
plt.figure(figsize = (10,7))
sns.heatmap(cm, annot = True, cmap="YlGnBu")
plt.xlabel('Predicted')
plt.ylabel('Actual')
```

Naive Bayes

```
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size = 0.33, r
andom_state = 100)

# Naive Bayes model

from sklearn.naive_bayes import MultinomialNB

# Fitting the model. MultinomialNB() used since they are categorical values i
nstead of numerical.
naive_bayes = MultinomialNB()
naive_bayes.fit(X_train, y_train)

# Predict test data set.

predictions = naive_bayes.predict(X_test)

# Accuracy score
metrics.accuracy_score(y_test, predictions)
```

```
# Classification report
print(classification_report(y_test, predictions))

# Confusion matrix

pd.crosstab(y_test,predictions,rownames=['Actual'],colnames=['Predicted'],mar
gins=True)

# Visual Confusion matrix

cm = confusion_matrix(y_test, predictions)
plt.figure(figsize = (10,7))
sns.heatmap(cm, annot = True, cmap="YlGnBu")
plt.xlabel('Predicted')
plt.ylabel('Actual')

# ROC Curves and obtaining AUC score
# ROC Curve and obtaining AUC score for Logistic Regression
y_pred_proba = log2.predict_proba(X_test)[::,1]
fpr, tpr, _ = metrics.roc_curve(y_test, y_pred_proba)
auc = metrics.roc_auc_score(y_test, y_pred_proba)
plt.plot(fpr,tpr,label="Logistic Regression, auc="+str(auc))

# ROC Curve and obtaining AUC score for Decision Tree
y_pred_proba2 = c.predict_proba(X_test)[::,1]
fpr2, tpr2, _ = metrics.roc_curve(y_test, y_pred_proba2)
auc2 = metrics.roc_auc_score(y_test, y_pred_proba2)
plt.plot(fpr2,tpr2,label="Decision Tree, auc="+str(auc2))

# ROC Curve and obtaining AUC score for Naive Bayes
y_pred_proba3 = naive_bayes.predict_proba(X_test)[::,1]
fpr3, tpr3, _ = metrics.roc_curve(y_test, y_pred_proba3)
auc3 = metrics.roc_auc_score(y_test, y_pred_proba3)
plt.plot(fpr3,tpr3,label="Naive Bayes, auc="+str(auc3))

# Labels. Moving legend location.
plt.legend(loc=4)
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.title('ROC Curves')
plt.show()
```

Unsupervised Learning - Apriori Algorithm

```
# Apriori imports

import csv
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori, association_rules
```

```
# Removing class variables. Only comparing features.
df_a = df2.drop(['job_satisfaction_level','likes_job'],axis=1)

df_a

# Transforming rating (numerical variables) into categorical (strongly disagr
ee to strongly agree)

df_a = df_a[['level_of_skill_used', 'expectations_clarity','contribution_reco
gnition','career_development', 'well_managed','worklife_balance', 'effective_
communication']].replace({1:'Strongly disagree',2:'Moderately disagree',3:'Mi
ldly disagree',4:'Neither agree nor disagree',5:'Mildly agree',6:'Moderately
agree',7:'Strongly agree'})

df_a

# Converting data to make it appropriate for Apriori. Each cell format would
be "-feature- = -response-"
df_a['level_of_skill_used'] = df_a['level_of_skill_used'].map('level_of_skill
_used = {}'.format)
df_a['expectations_clarity'] = df_a['expectations_clarity'].map('expectations
_clarity = {}'.format)
df_a['contribution_recognition'] = df_a['contribution_recognition'].map('cont
ribution_recognition = {}'.format)
df_a['career_development'] = df_a['career_development'].map('career_developme
nt = {}'.format)
df_a['well_managed'] = df_a['well_managed'].map('well_managed = {}'.format)
df_a['worklife_balance'] = df_a['worklife_balance'].map('worklife_balance = {
}'.format)
df_a['effective_communication'] = df_a['effective_communication'].map('effect
ive_communication = {}'.format)

df_a



# Transform data for the Apriori algorithm. Data has to be as an array, hence
.to_numpy()
te = TransactionEncoder()
te_array = te.fit(df_a.to_numpy()).transform(df_a.to_numpy())

te_array

te.columns_

# See transformed data. Created pandas dataframe.
te_df = pd.DataFrame(te_array, columns = te.columns_)

te_df
```

```
# Expanding table for clear view of item sets and association rules.
pd.set_option('display.max_colwidth', 1)

# Set minimum support threshold and show itemsets

freq_items = apriori(te_df, min_support = 0.02, use_colnames = True)
freq_items

# If, then rules. Also setting minimum confidence thresholds

rules = association_rules(freq_items,metric='confidence',min_threshold=0.9)
rules[['antecedents','consequents','support','confidence','lift']]

# View rules from highest lift value to lower.

rules[['antecedents','consequents','support','confidence','lift']].sort_value
s(by=['lift'],ascending = False)
```

Appendix B: Python Code – Dataset with Thirteen Features

```
# Importing numpy and pandas
import numpy as np
import pandas as pd

# Importing plot libraries
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline
sns.set_style('whitegrid')

# Import employee survey dataset
df = pd.read_csv('public-sector-commission-eps-2016and2015.csv')
df

# Agreeable level that employee is satisfied with their job.
# 1 for Strongly agree and 7 for Strongly disagree.

df['job_satisfaction_level'].value_counts()

df.columns

# Ratings except education_level and salary_rate originally reversed: 1 -> 7
= Best -> worst
# Changing to 1 -> 7 as from worst to best

df2 = df[['job_satisfaction_level','level_of_skill_used', 'expectations_clari
ty','contribution_recognition','career_development', 'well_managed','worklife
_balance', 'effective_communication', 'respectful_supervisor','good_relations
hip_with_coworkers', 'ethical_practices','conflicts_immediately_resolved']].r
eplace({1:7,2:6,3:5,4:4,5:3,6:2,7:1})

df2.head()

# Adding columns education_level and salary_rate back in
df2['education_level'] = df['education_level']
df2['salary_rate'] = df['salary_rate']

df2.head()

# Checking if ratings are changed correctly
df.head()

df2.job_satisfaction_level.value_counts()

# Creating function to classify that the employee likes the job or not like.
def job_satisfaction(x):
    if x >= 5:
```

```
        return 1
    else:
        return 0

# Adding 'likes_job' column where 0 if employee does not like job and 1 meani
ng they like their job.
df2['likes_job'] = df2['job_satisfaction_level'].apply(job_satisfaction)

# Check table
df2

df2.columns

# Exploratory Data Analysis

# Histogram/count plot for each feature. 6 by 2 format.
fig2 = plt.figure(figsize=(15,25))

# Level of skills the job requires
ax1 = fig2.add_subplot(621)
ax1.set_xlabel('level_of_skill_used')
ax1.set_ylabel('count')
sns.countplot(data = df2, x = 'level_of_skill_used', hue = 'likes_job')

# Clear Expectations
ax2 = fig2.add_subplot(622)
ax2.set_xlabel('expectations_clarity')
ax2.set_ylabel('count')
sns.countplot(data = df2, x = 'expectations_clarity', hue = 'likes_job')

# How often recognized for contributions
ax3 = fig2.add_subplot(623)
ax3.set_xlabel('contribution_recognition')
ax3.set_ylabel('count')
sns.countplot(data = df2, x = 'contribution_recognition', hue = 'likes_job')

# Management level
ax4 = fig2.add_subplot(624)
ax4.set_xlabel('well_managed')
ax4.set_ylabel('count')
sns.countplot(data = df2, x = 'well_managed', hue = 'likes_job')

# Work/life balance
ax5 = fig2.add_subplot(625)
ax5.set_xlabel('worklife_balance')
ax5.set_ylabel('count')
sns.countplot(data = df2, x = 'worklife_balance', hue = 'likes_job')

# Communication between departments
ax6 = fig2.add_subplot(626)
```

```
ax6.set_xlabel('effective_communication')
ax6.set_ylabel('count')
sns.countplot(data = df2, x = 'effective_communication', hue = 'likes_job')

# Supervisor rating
ax7 = fig2.add_subplot(627)
ax7.set_xlabel('respectful_supervisor')
ax7.set_ylabel('count')
sns.countplot(data = df2, x = 'respectful_supervisor', hue = 'likes_job')

# Relationship with peers
ax8 = fig2.add_subplot(628)
ax8.set_xlabel('good_relationship_with_coworkers')
ax8.set_ylabel('count')
sns.countplot(data = df2, x = 'good_relationship_with_coworkers', hue = 'like
s_job')

# Company Ethics
ax9 = fig2.add_subplot(629)
ax9.set_xlabel('ethical_practices')
ax9.set_ylabel('count')
sns.countplot(data = df2, x = 'ethical_practices', hue = 'likes_job')

# How quickly are conflicts resolved
ax10 = fig2.add_subplot(6,2,10)
ax10.set_xlabel('conflicts_immediately_resolved')
ax10.set_ylabel('count')
sns.countplot(data = df2, x = 'conflicts_immediately_resolved', hue = 'likes_
job')

# Education Level
ax11 = fig2.add_subplot(6,2,11)
ax11.set_xlabel('education_level')
ax11.set_ylabel('count')
sns.countplot(data = df2, x = 'education_level', hue = 'likes_job')

# Salary
ax12 = fig2.add_subplot(6,2,12)
ax12.set_xlabel('salary_rate')
ax12.set_ylabel('count')
sns.countplot(data = df2, x = 'salary_rate', hue = 'likes_job')

df2['education_level'].value_counts()

df2['salary_rate'].value_counts()


# Average score for each feature by two groups: satisfied and dissatisfied em
ployees
```

```
df2.groupby('likes_job').mean()

# Features. Dropping target variables.

x = df2.drop(['likes_job','job_satisfaction_level'], axis=1)
x

# Multicollinearity quick check
x.corr()

# Class variable

y = df2['likes_job']

# Machine Learning Imports. Logistic Regression

from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

# For evaluation later
from sklearn import metrics

# Creating Logistic Regression Model. Not training model yet. For accuracy co
mparison.

log = LogisticRegression()

# Fitting data
log.fit(x,y)

# Accuracy check
log.score(x,y)

# Flattened array of feature coefficients
np.ravel(log.coef_)

# Coefficients from the model

coeff_df2 = pd.DataFrame(x.columns, columns = ['Features'])
coeff_df2['Coefficients'] = np.ravel(log.coef_)

coeff_df2
```

Logistic Regression - Training and Testing Data Set

```
# Splitting the data. Test size set to 33% and arbitrary random_state = 100 t
o ensure reproducible results.

X_train, X_test, y_train, y_test = train_test_split(x, y, test_size = 0.33, r
andom_state = 100)

# Creating Logistic Regression model
log2 = LogisticRegression()
```

```
# Fitting new model
log2.fit(X_train, y_train)

# Predict the classes of the testing data set. Employee likes job or not.
y_predict = log2.predict(X_test)

# Accuracy score
metrics.accuracy_score(y_test, y_predict)

# Classification report
from sklearn.metrics import classification_report

print(classification_report(y_test, y_predict))

# Confusion matrix

pd.crosstab(y_test,y_predict,rownames=['Actual'],colnames=['Predicted'],margi
ns=True)

# Visualized confusion matrix

from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_predict)
plt.figure(figsize = (10,7))
sns.heatmap(cm, annot = True, cmap="YlGnBu")
plt.xlabel('Predicted')
plt.ylabel('Actual')
```

Decision Tree

```
# Decision tree imports
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier, export_graphviz
from sklearn.model_selection import train_test_split

# Train test split
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.33, ran
dom_state=100)

# Decision tree model
c = DecisionTreeClassifier(min_samples_split=101)

# Fit model
c.fit(X_train,y_train)

# Predict classes
y_pred = c.predict(X_test)

# Accuracy score
metrics.accuracy_score(y_test, y_pred)
```

```
# Classification report
print(classification_report(y_test, y_pred))

# Confusion matrix

pd.crosstab(y_test,y_pred,rownames=['Actual'],colnames=['Predicted'],margins=
True)

# Visualized Confusion matrix

cm = confusion_matrix(y_test, y_pred)
plt.figure(figsize = (10,7))
sns.heatmap(cm, annot = True, cmap="YlGnBu")
plt.xlabel('Predicted')
plt.ylabel('Actual')
```

Naive Bayes

```
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size = 0.33, r
andom_state = 100)

# Naive Bayes model

from sklearn.naive_bayes import MultinomialNB

naive_bayes = MultinomialNB()
naive_bayes.fit(X_train, y_train)

# Predict testing set
predictions = naive_bayes.predict(X_test)

# Accuracy score
metrics.accuracy_score(y_test, predictions)

naive_bayes.predict([[5,4,6.5,7,3,4,2,1,4,5,6,5,4]])

# Classification report
print(classification_report(y_test, predictions))

# Confusion matrix

pd.crosstab(y_test,predictions,rownames=['Actual'],colnames=['Predicted'],mar
gins=True)

# Visualized Confusion matrix

cm = confusion_matrix(y_test, predictions)
plt.figure(figsize = (10,7))
sns.heatmap(cm, annot = True, cmap="YlGnBu")
plt.xlabel('Predicted')
plt.ylabel('Actual')
```

```
# ROC Curves of all three models. AUC scores beside them. For logistic regres
sion, decision tree, and Naive Bayes

y_pred_proba = log2.predict_proba(X_test)[::,1]
fpr, tpr, _ = metrics.roc_curve(y_test, y_pred_proba)
auc = metrics.roc_auc_score(y_test, y_pred_proba)
plt.plot(fpr,tpr,label="Logistic Regression, auc="+str(auc))

y_pred_proba2 = c.predict_proba(X_test)[::,1]
fpr2, tpr2, _ = metrics.roc_curve(y_test, y_pred_proba2)
auc2 = metrics.roc_auc_score(y_test, y_pred_proba2)
plt.plot(fpr2,tpr2,label="Decision Tree, auc="+str(auc2))

y_pred_proba3 = naive_bayes.predict_proba(X_test)[::,1]
fpr3, tpr3, _ = metrics.roc_curve(y_test, y_pred_proba3)
auc3 = metrics.roc_auc_score(y_test, y_pred_proba3)
plt.plot(fpr3,tpr3,label="Naive Bayes, auc="+str(auc3))

# Labels. Moving legend location.
plt.legend(loc=4)
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.title('ROC Curves')
plt.show()
```

Unsupervised Learning - Apriori algorithm

```
# Apriori imports

import csv
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori, association_rules

# Removing class variables. Only comparing features.
df_a = df2.drop(['job_satisfaction_level','likes_job'],axis=1)

df_a

# Transforming rating (numerical variables) into categorical (strongly disagr
ee to strongly agree)

df_a = df_a[['level_of_skill_used', 'expectations_clarity','contribution_reco
gnition','career_development', 'well_managed','worklife_balance', 'effective_
communication', 'respectful_supervisor','good_relationship_with_coworkers','e
thical_practices','conflicts_immediately_resolved']].replace({1:'Strongly dis
agree',2:'Moderately disagree',3:'Mildly disagree',4:'Neither agree nor disag
ree',5:'Mildly agree',6:'Moderately agree',7:'Strongly agree'})

df_a
```

```
# Converting data to make it appropriate for Apriori. Each cell format would
be “-feature- = -response-”
df_a['level_of_skill_used'] = df_a['level_of_skill_used'].map('level_of_skill
_used = {}'.format)
df_a['expectations_clarity'] = df_a['expectations_clarity'].map('expectations
_clarity = {}'.format)
df_a['contribution_recognition'] = df_a['contribution_recognition'].map('cont
ribution_recognition = {}'.format)
df_a['career_development'] = df_a['career_development'].map('career_developme
nt = {}'.format)
df_a['well_managed'] = df_a['well_managed'].map('well_managed = {}'.format)
df_a['worklife_balance'] = df_a['worklife_balance'].map('worklife_balance = {
}'.format)
df_a['effective_communication'] = df_a['effective_communication'].map('effect
ive_communication = {}'.format)
df_a['respectful_supervisor'] = df_a['respectful_supervisor'].map('respectful
_supervisor = {}'.format)
df_a['good_relationship_with_coworkers'] = df_a['good_relationship_with_cowor
kers'].map('good_relationship_with_coworkers = {}'.format)
df_a['ethical_practices'] = df_a['ethical_practices'].map('ethical_practices
= {}'.format)
df_a['conflicts_immediately_resolved'] = df_a['conflicts_immediately_resolved
'].map('conflicts_immediately_resolved = {}'.format)

df_a

# Adding education_level. It only ranges from 1-2 unlike other features. Repl
aced numbers to actual meaning for Apriori.

df_a['education_level'] = df2[['education_level']].replace({1:'Education = Pr
imary/secondary school or technical/trade certificate or diploma',2:'Educatio
n = University qualification'})

# Adding back salary_rate. Five available numbers. Replaced numbers to actual
salary ranges.

df_a['salary_rate'] = df2[['salary_rate']].replace([1,2,3,4,5],['Salary = Les
s than $60,000','Salary = $ 60,000-70,000','Salary = $ 80,000-109,999','Salar
y = $ 110,000-159,999','Salary = $ 160,000 and over'])

df_a

df_a.to_numpy()

# Transform data for the apriori algorithm
te = TransactionEncoder()
te_array = te.fit(df_a.to_numpy()).transform(df_a.to_numpy())

te_array

te.columns_
```

```
# See transformed data
te_df = pd.DataFrame(te_array, columns = te.columns_)

te_df

# Expanding table for clear view of item sets and association rules.

pd.set_option('display.max_colwidth', 1)

# Set minimum support threshold and show itemsets. Too many rules. Curse of d
imensionality.

freq_items = apriori(te_df, min_support = 0.03, use_colnames = True)
freq_items

# If, then rules. Also setting minimum confidence thresholds

rules = association_rules(freq_items,metric='confidence',min_threshold=0.9)
rules[['antecedents','consequents','support','confidence','lift']]

# Default value of display.max_rows is 100 i.e. at max 100 rows will be print
ed.
# Set it to 100 to display all rows in the dataframe
pd.set_option('display.max_rows', 100)

rules.head(100)
```