Table of Contents

[DATA PREPARATION 1](#_Toc155879160)

[Data Characterisation 1](#_Toc155879161)

[Missing Values 2](#_Toc155879162)

[Data Preparation and Evaluation 2](#_Toc155879163)

[1.2.1. Data Cleaning 2](#_Toc155879164)

[1.2.2. Exploratory Data Analysis 2](#_Toc155879165)

[1.2.3. Encoding 4](#_Toc155879166)

[1.2.4. Scaling: 4](#_Toc155879167)

[1.2.5. Dimensional reduction: 5](#_Toc155879168)

[PCA: 5](#_Toc155879169)

[LDA 6](#_Toc155879170)

[STATISTICAL TECHNIQUES: 7](#_Toc155879171)

[Descriptive Analysis 7](#_Toc155879172)

[Inferential Statistics 9](#_Toc155879173)

[MACHINE LEARNING 9](#_Toc155879174)

[Prediction 10](#_Toc155879175)

[Comparison 10](#_Toc155879176)

[REFERENCES 11](#_Toc155879177)

## DATA PREPARATION

Data preparation is defined as the process of cleaning data by reformatting, correcting errors, and combining data sets. (Abedjan 2015) this can be done to ensure the data in question is of good quality (data formats are in a standard form, outliers are taken care of etc)

### Data Characterisation

Data characterisation is crucial in data analysis because: it assists in deep understanding of the data (structure, content and format) and also provide insight into the nature of the attributes or variables in the dataset.

The dataset for this analysis is a dataset containing information about employees in the company, including their age, employees’ number, education status, positions held in the company and number of years they have worked etc.

There are 1470 observations in this dataset and 35 attributes of which nine of those are object values meaning they are text values for example the department in which the employees work cannot be expressed as an integer but as text values. The remaining attributes have float64 values which means they are in decimal form. Float64 values require more computer memory to process and hence will make machine learning models slower and will require more time to process. To handle this, I rounded off the attributes: ‘distance travelled’, ‘hourly rate’ and ‘percentage hike’ to two decimal places and the remainder rounded them off to the nearest whole number/ integer. This is because for the three attributes ‘distance travelled’, ‘hourly rate’ and ‘percentage hike’ precision is key factor whereas the others are not that precise.

### Missing Values

Every attribute in this dataset has 147 observations missing at random. One of the easiest ways and common ways to handle missing data is drop the values from the dataset, this way doesn’t really introduce bias or skewedness to dataset which might be introduced by the imputation methods. I dropped the missing values and lost almost 95% of the data which means dropping the missing values is not applicable in this dataset.

In this analysis employee ‘Job Satisfaction’ is going to be my target variable, so for this attribute I did not use imputation methods to handle missing data as this will affect the machine learning process of training data, as using estimated data can train the model falsely. I used the median to impute on missing values for float attributes because of the presence of outliers in the dataset. For object attributes I used the mode to impute the missing values.

## Data Preparation and Evaluation

### 1.2.1. Data Cleaning

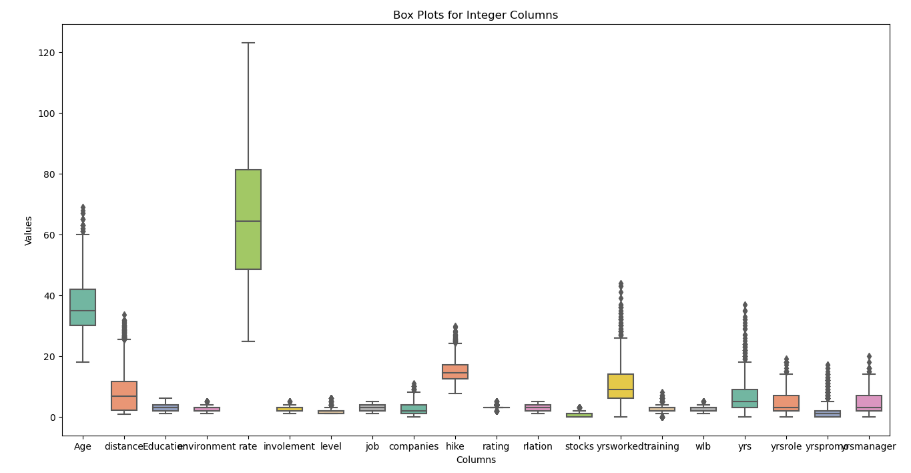
Data cleaning is defined as the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

* Removing unnecessary columns: for this dataset first I removed some of the columns as they were not adding any benefit to the analysis as the value was constant across the dataset, these included employee count, over 18 and standard hours. The other columns I removed were the daily rate, employee number, monthly income, monthly rate because these contains almost the same information as the hourly rate only that these are calculated for a day or month respectively.
* Renaming of columns: renaming of columns is done to improve readability of the columns and clarity of the code during the analysis.
* Duplicate Values: the dataset have no duplicated rows or columns.

### 1.2.2. Exploratory Data Analysis

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. (Patil, 2018). EDA also work is detecting relationships within the dataset and to identify outliers or any anomalies in the dataset. For this analysis I did a few visualisation to explore the dataset and these are explained below:

Outlier Detection: an outlier is defined by the Oxford Dictionary as a person or thing situated away or detached from the main body or system in other words a person or thing differing from all other members of a particular group or set. Outliers are common findings in data analytics the same in the case of my analysis of Job Satisfaction. To detect the outliers I used boxplots (*figure 1* below) to plot the numerical data indicating the presence of outliers and central tendency of the data.



*Figure 1*

The boxplot above indicates that for variables education, rate, job satisfaction and relation have no outliers. Methods to handle outliers includes, removing them from the dataset, I tried that but removing them from the datasets results in almost deletion of the whole dataset because of the way the outliers are distributed in the dataset.

So I had to use the capping/winsorisation method to dela with the outliers. Winsorisation involves the replacement of the smallest and biggest values in a dataset or variable with observations close to them thus mitigating the effects caused by outliers. The results are indicated in the figure below:

A graph of different colored boxes

Description automatically generated

*Figure 2*

Bar plots: I plotted several bar plots to answer to certain questions which can be derived from just looking at the dataset.

* What age group is most or least satisfied with their job? From the bar plot there is even distribution of sentiments on job satisfaction on all ages
* Is there a certain department which employees are not happy with? The research and development department has a greater percentage of employees compared to the other departments in the company. Also, of these a sizeable percentage are moderately satisfied to very satisfied with their jobs compared to other departments.
* What does the dataset show about the company gender equality policy? Is there gender imbalance? From the count plot it doesn’t really indicate any bias against a specific gender. Also the data indicates that a greater percentage of the women employed by the company kept their job compared by the percentage attrition of their male counterparts.
* How many of the employees are married and if this has any effect on the attrition.
* Are there specific roles that people like to work in and are not stressful to work in?
* Are the people who are not satisfied with their jobs the ones leaving the company?
* Which department in the company and roles have the high earners in the company?
* Which department in the company and roles have the most people who are happy with their jobs?

Correlation Matrix: this is a table which shows correlation coefficients between variables in a dataset indicating the strengths and direction of their relationships. From the correlation matrix plotted for this dataset, there some variable with strong positive correlation among themselves hence I remove some of those variables as keeping them will distort the machine learning modelling process, keeping in mind that correlation doesn’t mean causation.

### 1.2.3. Encoding

For encoding I used the ‘one hot encoding method’ which works by converting the categorical labels to binary which is easy to incorporate into machine learning model without affecting the models. I avoided using the label encoder because it has a disadvantage that it assigns a unique number from 0 to each class of data, hence a label with a high value maybe considered of more priority than one with a lower value. as an example, for the Attribute, 'Role' a 'sales representative’ is now represented by value '8' while a 'Healthcare representative' is now 0. this might look like the sales representative is more important than the healthcare representative.

### 1.2.4. Scaling:

Data scaling is the process of transforming the values of the features of a dataset till they are within a specific range, e.g. 0 to 1 or -1 to 1. This is to ensure that no single feature dominates the distance calculations in an algorithm and can help to improve the performance of the algorithm (Khoong, 2023). In my dataset there are attributes like age etc which have high values, these need to be scaled down so that they don’t seem to have a higher importance than the other variables.

### 1.2.5. Dimensional reduction:

With the use on the one hot encoding methods there is an increase in the dimensions of the dataset, hence there is need to employ dimensional reduction methods like PCA or LDA.

### PCA:

This is a dimensional reduction method used in data analytics to simplify the complexity of high dimensional data while retaining the patterns and meaning of the data. I used PCA because with the one-hot encoding there was introduction of more variables to my dataset hence an increase in the dimensionality of the dataset. This dimensionality reduction also helps in reducing the risk of data overfitting in machine learning models. *Figure 3* below is a scatter plot demonstrating the distribution of the principal components from the principal component analysis.

PCA is primarily concerned with capturing the maximum variance in the dataset. The principal components (PC1, PC2) are derived to maximize the variance of the data along these components. Points that are spread out in the PCA plot indicate high variability in the data along those directions.

PCA is an unsupervised technique that doesn't take into account any information about the target variable. The separation observed in a PCA plot is based solely on the overall variance in the dataset. The PCA plot helps identify which features contribute the most to the overall variance in the data. It doesn't specifically focus on how well the data separates based on the target variable.

A diagram of a graph

Description automatically generated with medium confidence

*Figure 3*

### LDA

LDA is a supervised technique that aims to find the linear combinations of features that best separate different classes (in this case, levels of job satisfaction). LDA explicitly considers information about the target variable which is not considered by a PCA

The linear discriminants (LD1, LD2) are derived to maximize the separation between different classes while minimizing the variance within each class.

The LDA plot provides insights into how well the features discriminate between different levels of job satisfaction.

Points that are close together in the LDA plot are expected to be more similar in terms of job satisfaction, and the plot shows how well the algorithm separates the classes.

A group of blue and yellow dots

Description automatically generated

# STATISTICAL TECHNIQUES:

Statistical analysis is an integral part of data analysis and machine learning as it gives meaning to meaningless numbers. Statistical methods involved in this study include descriptive analysis, hypothesis tests among others:

## Descriptive Analysis

A descriptive analysis of the dataset was performed:

A table with numbers and letters

Description automatically generated

*Figure 4*

The youngest employee of the company recorded was 14 years old which below the age of 18 where you assume someone can be fully employed. Hence an assumption was made that those below the age of 18 were obviously temporary workers or students working part time in the company and would eventually leave anyways, so keeping this data would definitely add bias to the analysis or machine learning models. The age groups below the age of 18 were then removed from the dataset for analysis.

Also noted is the steep difference between the lowest earner and the highest earner, with the latter earning 20 times more than the amount earned by the former. This might be due to the difference in their roles within the company, one might be in the executive and the other just a general worker.

The analysis also indicates that all the employees in the company live in vicinity of a 30km radius from the are of work and don’t have to travel long distances to get to and from work. One would assume that distance could not be a factor affecting whether one is satisfied with working with company or not.

The longest serving member of the company has been the for nearly 37 years and also the longest time period with an employee under the same manager if close to 20 years which indicate a good staff retention programme. But this has a negative impact in that some employees went as much as 17years without being promoted in their positions, and we can only assume they would get frustrated and ready to move given an opportunity and definitely would not be satisfied with their jobs.

Also, the descriptive analysis indicates that most of the employees about 40% work in the research and development department, and the same percentage of employees are males.

The most popular job role is the Sales Executive among nine different roles involved in this study.

Boxplots were plotted and outliers were also noted in the dataset, and these were winsorized to reduce their negative input on the dataset during analysis and machine learning model.

## Inferential Statistics

Inferential statistics is a branch of statistics that makes the use of various analytical tools to draw inferences about the population data from sample data (Gornick, 1993). To be able to do inference of the sample data onto the population I used the null hypothesis (when there is no effect caused by the variable), the null hypothesis assumes a status quo where everything is normal and does not cause anything. The significance level for the null hypothesis is 5%.

The Shapiro Wilk test was done for hypothesis testing of the numerical attributes, this indicated there wasn’t variable which is normally distributed. Sometimes data just needs to be transformed to follow a normal distribution, I used the boxcox method to transform the data to be normally distributed, but this also did not bring any good results.

Even after using the Boxcox method, normality is not satisfied for this analysis so I used the Kruskal-Wallis Test as a non-parametrics alternative for the ANOVA test, this also failed to reject the null hypothesis.

# MACHINE LEARNING

Machine learning is defined as the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyse and draw inferences from patterns in data (Oxford, 2023). There are two methods used in machine learning and these are, supervised and unsupervised learning:

**Supervised learning** is defined by its use of labelled datasets to train algorithms to classify data or predict outcomes accurately (IBM, 2023). This type of learning is also most useful if there is a known outcome. In the case of this study the original dataset, after all the data preparation has been done is used for machine learning modelling as a supervised dataset. This helps the business solve real-world problems, in this case to find if employees are satisfied with their jobs or not and what to do to improve their welfare.

Evaluation metrics used in supervised learning are clear, well defined learning objectives, and also the ability of the model to handle complex relationships between the input and the output these are some of the advantages of using a supervised learning model. Super vised learning can fall short in that it requires labelled data which might be time consuming and expensive to the company to get and use.

On the other hand, **unsupervised learning** uses algorithms to analyse and cluster unlabelled data. When I passed the dataset through the Principal Component Analysis and the Linear Discriminant Analysis, the data was grouped as components, but they are not labelled as attributes, and we can’t change the values hence modelling that dataset with components of after LDA I think qualifies as being an unsupervised learning.

Advantages of unsupervised learning include: labelled data is not required, hidden patterns in the data can easily be revealed and also exploratory analysis will be made easier by unsupervised learning.

## Prediction

In this analysis, I used the machine learning models for prediction, these are the KNN clustering method, SVM model and the logistic regression just to try and find the best model for my dataset.

On all the models I used my accuracy was really low on average around 30% for the dataset and also after performing PCA. After PCA the accuracy would go up a bit but with LDA it went all the way up to 100%accurate.

For data splitting I used mainly 80%training data and the remainder for testing purposes. I also tried decreasing the training data to 70% and so on and this improved the accuracy of the model but with continued decrease of the training data the accuracy was shown to start going down as well.

With KNN modelling I tried to use the GridSearchCV hyperparameter tuning to try and develop a model which would be more accurate. Thid improved the accurate yes but with a very small margin hence I can conclude that hyperparameter tuning does not work for this dataset

## Comparison

Comparing the metrics from the machine learning model predictions made indicated that on accuracy the logistic regression model has a slight edge above the other models. But in terms of box precession testing the KNN model surpasses the three models employed in this study

In terms of accuracy, precision, recall, and F1-score, the supervised model performs better than the unsupervised model. Given that supervised learning models are trained on labeledlabelled data with specific goals, this is to be expected. Figure 5 below illustrates the differences between the models.

A graph of different colored bars

Description automatically generated*figure 5: illustration of model comparisons.*

# REFERENCES

Hastie, Friedman, and Tibshirani, The Elements of Statistical Learning, 2001

Bishop, Pattern Recognition and Machine Learning, 2006

Davenport, T. H., Ronanki, R., “Artificial Intelligence for the Real World,” 2018. Accessed July 21, 2020.