Data Mining Group Project Report

Zilan Huang – [zh4u17@soton.ac.uk](mailto:zh4u17@soton.ac.uk); Bang Du – [bd1m17@soton.ac.uk](mailto:bd1m17@soton.ac.uk);

Zihan Wang - [zw1u17@soton.ac.uk](mailto:zw1u17@soton.ac.uk); Yang Zhou - [yz12a17@soton.ac.uk](mailto:yz12a17@soton.ac.uk);

Da An - [da1g17@soton.ac.uk](mailto:da1g17@soton.ac.uk); Mardhiyyah Rafrin - [mr1n17@soton.ac.uk](mailto:mr1n17@soton.ac.uk)

Table of Contents (For Convenience Use)

[ABSTRACT 1](#_Toc512955544)

[KEY TERMS 1](#_Toc512955545)

[1. INTRODUCTION 1](#_Toc512955546)

[2. UNDERSTANDING THE DATA 1](#_Toc512955547)

[2.1. COEFFICIENT CORELATION (TED) 1](#_Toc512955548)

[2.2. INFORMATION GAIN (ERNEST) 1](#_Toc512955549)

[2.3. MDS (AN DA) 1](#_Toc512955550)

[2.4. CLUSTERING (AN DA) 1](#_Toc512955551)

[3. DATA PROCESSING (TED) 1](#_Toc512955552)

[3.1. CLEAN DATASET 1](#_Toc512955553)

[3.2. DIMENSION REDUSION 1](#_Toc512955554)

[3.2.1. PCA 1](#_Toc512955555)

[3.2.2. STEPWISE 1](#_Toc512955556)

[4. PREDICTION MODEL (10-FOLDED CROSS VALIDATION) 2](#_Toc512955557)

[4.1. LOGISTIC REGRESSION (ERNEST) 2](#_Toc512955558)

[4.1.1. DESCRIPTION 2](#_Toc512955559)

[4.1.2. IMPLEMENTATION 2](#_Toc512955560)

[4.1.3. RESULT (ACCURACY, ROC, AIC, F1 SCORE, R2 SCORE) 2](#_Toc512955561)

[4.2. DECISION TREE (ERNEST) 2](#_Toc512955562)

[4.2.1. DESCRIPTION 2](#_Toc512955563)

[4.2.2. IMPLEMENTATION 2](#_Toc512955564)

[4.2.3. RESULT (ACCURACY, ROC, AIC, F1 SCORE, R2 SCORE) 2](#_Toc512955565)

[4.3. LDA (AN DA) 2](#_Toc512955566)

[4.3.1. DESCRIPTION 2](#_Toc512955567)

[4.3.2. IMPLEMENTATION 2](#_Toc512955568)

[4.3.3. RESULT (ACCURACY, ROC, AIC, F1 SCORE, R2 SCORE) 2](#_Toc512955569)

[4.4. SVM (Mardhiyyah) 2](#_Toc512955570)

[4.4.1. DESCRIPTION 2](#_Toc512955571)

[4.4.2. IMPLEMENTATION 2](#_Toc512955572)

[4.4.3. RESULT (ACCURACY, ROC, AIC, F1 SCORE, R2 SCORE) 2](#_Toc512955573)

[4.5. RANDOM FOREST (ZIHAN) 2](#_Toc512955574)

[4.5.1. DESCRIPTION 2](#_Toc512955575)

[4.5.2. IMPLEMENTATION 3](#_Toc512955576)

[4.5.3. RESULT (ACCURACY, ROC, AIC, F1 SCORE, R2 SCORE) 3](#_Toc512955577)

[4.6. COMPARATION & ANALYSIS 3](#_Toc512955578)

[5. DATA ANALYSIS (TED) 3](#_Toc512955579)

[6. CONCLUSION & FUTURE WORK 3](#_Toc512955580)

[GITHUB LINK 3](#_Toc512955581)

# ABSTRACT

# KEY TERMS

# INTRODUCTION

# UNDERSTANDING THE DATA

## COEFFICIENT CORELATION (TED)

## INFORMATION GAIN (ERNEST)

## MDS (AN DA)

## CLUSTERING (AN DA)

# DATA PROCESSING (TED)

## CLEAN DATASET

## DIMENSION REDUSION

### PCA

### STEPWISE

# PREDICTION MODEL (10-FOLDED CROSS VALIDATION)

## LOGISTIC REGRESSION (ERNEST)

### DESCRIPTION

### IMPLEMENTATION

### RESULT (ACCURACY, ROC, AIC, F1 SCORE, R2 SCORE)

## DECISION TREE (ERNEST)

### DESCRIPTION

### IMPLEMENTATION

### RESULT (ACCURACY, ROC, AIC, F1 SCORE, R2 SCORE)

## LDA (AN DA)

### DESCRIPTION

### IMPLEMENTATION

### RESULT (ACCURACY, ROC, AIC, F1 SCORE, R2 SCORE)

## SVM (Mardhiyyah)

### DESCRIPTION

### IMPLEMENTATION

### RESULT (ACCURACY, ROC, AIC, F1 SCORE, R2 SCORE)

## RANDOM FOREST (ZIHAN)

### DESCRIPTION

### IMPLEMENTATION

### RESULT (ACCURACY, ROC, AIC, F1 SCORE, R2 SCORE)

## COMPARATION & ANALYSIS

# DATA ANALYSIS (TED)

# CONCLUSION & FUTURE WORK

# GITHUB LINK

<https://github.com/MarioBrosECS/predicting-employee-attrition-Soton-ECS>