

**Postgraduate Certificate in Software Design with Artificial Intelligence**

**Advanced Machine Learning and Neural Networks - (AL\_KSAIG\_9\_1)**

**Minor Exercise 1**

Student ID: A00267948

Student Name: Daniel Foth

GIT: <https://github.com/DanielsHappyWorks/aml-data-discovery>

Brief Description: This assignment aims to outline data assumptions made before applying models to dataset (e.g. increase in rooms, increase in price), the strengths and weakness to each model, the accuracy of each model using cross validation and a conclusion which will outline the best model for the chosen dataset and why.

Contents

[Data 3](#_Toc34486703)

[Regression 3](#_Toc34486704)

[Clustering 4](#_Toc34486705)

[SVN 4](#_Toc34486706)

[Neural Network 4](#_Toc34486707)

[Conclusion 4](#_Toc34486708)

# Data

The data set utilised for this exercise is the “Adult” data set. The data was extracted from the census bureau database by Barry Becker and contributed to the UCI Repository by Ronny Kohavi and Barry Becker.

UCI Repository: <https://archive.ics.uci.edu/ml/datasets/adult>

The data contains 32561 rows and 15 features. Which include:

1. Age - Numeric
2. Workclass - Categorical
3. Fnlwgt - Numeric
4. Education - Categorical
5. education-num- Numeric
6. marital-status - Categorical
7. occupation - Categorical
8. relationship - Categorical
9. race - Categorical
10. sex - Categorical
11. capital-gain- Numeric
12. capital-loss- Numeric
13. hours-per-week- Numeric
14. native-country - Categorical
15. salary – Categorical (>50K or <=50K)

For this specific problem, we are trying to establish whether a person makes over 50K a year. The “Adult” data set has a few missing values. This data set has both numerical and categorical data. Below you can find a table describing each feature individually.



# Assumptions

With this much data available it’s hard to assume what will best affect the accuracy of a model. From looking at the fields available I expect that age, occupation, education and hours per week will affect the precision the most as those are always a good indicator as to how much a person may earn.

The data set is very much skewed in favour of people earning less than 50K as the ratio of data entries is 3.15:1. This could affect how well the models can deal with categorising over 50K pay.

# Regression

### Model Accuracy

### Strengths & Weaknesses of Models

# Clustering

### Model Accuracy

### Strengths & Weaknesses of Models

# SVN

### Model Accuracy

### Strengths & Weaknesses of Models

# Ensemble

### Model Accuracy

### Strengths & Weaknesses of Models

# Neural Network

### Model Accuracy

### Strengths & Weaknesses of Models

# Conclusion