PREDICTING CONCRETE MIX'S COMPRESSIVE STRENGTH COMPLIANCE FROM MIX COMPONENTS

Date: 02 June 2019

Meng Heng Ea
Student ID: s3716021
Email: s3716021@student.rmit.edu.au

Contents

Table of Figures

Table of Tables

Executive Summary

Concrete is one of the most commonly used construction materials in building structures in civil engineering. Research projects conducted in Australia aims to investigate the influence of different combinations of concrete mixture components on concrete properties. Researchers initially identify a pool of potential mixture designs and ideally make them into concrete specimens and conduct repeated experimental testings to answer research questions.

However, these mixture designs identified could be in large quantities, and due to budget and time constraints as well as limited resources, it is not possible to produce concrete specimens for all mixtures.

The goal of this project is to develop a model that can classify researchers' initial pool of mixture designs into compliant or non-compliant type based on components: cement, water, coarse aggregates, fine aggregates, fly ash, blast furnace slag and superplasticizers. 'Compliant' means the mix will achieve a compressive strength of 40 MPa or above at 28 days or 'non-compliant' as achieving a compressive strength less than 40 MPa at 28 days according to Australian Standards 3600, Concrete Structures.

Data of 425 concrete mix made up of varying levels of the components will be fit to two classification algorithms; namely, k-nearest neighbours and decision tree algorithms at following training/test sets split ratio: 50/50, 60/40 and 80/20 percentages. The evaluation metric used to compare the models is the precision score.

Consequently, the best model results from fitting data at 80% to 20% training/testing split, which achieves mean precision score of 93.653% with standard deviation 5.340% on the testing set after 10-fold cross-validation after hyperparameter tuning.

1. Introduction

1.1. Problem Statement

Concrete is one of the most commonly used construction materials in building structures in civil engineering. For many years, there have been many research projects conducted. These aims to investigate how different combinations of concrete mixture components and additions of new ones such as minerals and superplasticizers can improve concrete's properties.

In research projects, any potential mixture designs that could answer research questions are identified based on the researchers' expertise and previous investigations. Ideally, all mixture designs would be produced into specimens, and these specimens would then be subjected to repeated testings before reaching conclusions on which mixture designs to adopt.

However, the potential mixture designs identified could be in large quantities, and due to budget and time constraints as well as limited resources, it is not always possible to produce concrete specimens for all of them.

Therefore, the goal of this project is to develop a model that can assist researchers to narrow down which mixture design to be made into specimens for further testings.

1.2. Project Scope and Research Question

First and foremost, the concrete produced based on these mixture designs from the experiments must meet the required standards locally in order to ensure trust in their applicability in real life. For this project, compressive strength is considered.

The model will be limited to be used by those researchers working to improve concrete structures in Australia. As majority of Australia's cities are located along the coastline, concrete members situated in coastal areas are in the B2 exposure classification according to Table 4.3 of Australian Standards 3600 (Standards Australia 2009, p. 51). For these structures, concrete strength of 40 MPa at 28 days is commonly used to place and thus deemed compliant (Standards Australia 2009, p.54).

The essential components of concrete are cement, water, coarse aggregates and fine aggregates. In addition to this, fly ash and blast furnace slag were the popular minerals used to design mixture and have long been researched on. Further, superplasticizer is also commonly used as an additive to the mixture to improve concrete workability during placing.

As a result, the question this project aims to answer is to develop a model can classify mixture design based on components: cement, water, coarse aggregates, fine aggregates, fly ash, blast furnace slag and superplasticizers. The predicted output from the model would be "compliant", meaning concrete mix will achieve a compressive strength of 40 MPa or above at 28 days or "non-compliant" otherwise, meaning achieving a compressive strength of less than 40 MPa at 28 days.

2. Methodology

2.1. Data Description

The model will be developed based on data from Yeh (1998). Below is the preview of the first ten observations of the dataset:

Cement (component 1)(kg in a m^3 mixture)	Blast Furnace Slag (component 2)(kg in a m^3 mixture)	Fly Ash (component 3)(kg in a m^3 mixture)	Water (component 4)(kg in a m^3 mixture)	Superplasticizer (component 5)(kg in a m^3 mixture)	Coarse Aggregate (component 6)(kg in a m^3 mixture)	Fine Aggregate (component 7)(kg in a m^3 mixture)	Age (day)	Concrete compressive strength(MPa, megapascals)
540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	79.99
540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	61.89
332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	40.27
332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	41.05
198.6	132.4	0.0	192.0	0.0	978.4	825.5	360	44.30
266.0	114.0	0.0	228.0	0.0	932.0	670.0	90	47.03
380.0	95.0	0.0	228.0	0.0	932.0	594.0	365	43.70
380.0	95.0	0.0	228.0	0.0	932.0	594.0	28	36.45
266.0	114.0	0.0	228.0	0.0	932.0	670.0	28	45.85
475.0	0.0	0.0	228.0	0.0	932.0	594.0	28	39.29

There is a total of 8 quantitative independent variables: cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate and age. The dependent variable is the concrete compressive strength, which is also quantitative. According to the read_me file of the dataset folder from the source and upon inspection in python, there are no missing values. There are a total of 1030 rows observations. The followings are ranges of values for each attribute:

- Cement: 281.168 kg/m^3 to 540 kg/m^3
- Blast Furnace Slag: 0.000 kg/m^3 to 359.400 kg/m^3
- Fly Ash: 0.000 kg/m^3 to 200.100 kg/m^3
- Water: 121.800 kg/m^3 to 247.000 kg/m^3
- Superplasticizer: 0.000 kg/m^3 to 32.2000 kg/m^3
- Coarse Aggregates: 801.000 kg/m^3 to 1145.000 kg/m^3
- Fine Aggregates: 594.000 kg/m^3 to 992.600 kg/m^3
- Age: 1 day to 365 days
- Compressive Strengths: 2.330 MPa to 82.600 MPa

2.2. Data Retrieving

Python pandas package is used to import the source CSV file into python environment as a data frame. Upon reading the file into python, the data frame was checked against the source file to determine their obvious discrepancies as followings:

- The top and last ten observations were printed to check against the source file to make sure the data was imported as expected.
- The columns' names were then checked if they were assigned to appropriate corresponding columns.
- The number of rows and columns were then checked against the information given by data description such as there are a total of 1030 rows and nine columns expected.

- The data types for each column were then checked to see if python type each column meaningfully according to data description; otherwise, coercion needs to be done accordingly.

Summaries statistics of each numerical columns were printed to give their ranges of values. By inspecting each attribute's minimum and maximum statistics and checking them against the range of values expected, this can ensure that the data imported is meaningful numerically.

Finally, python function .info () was used to print out the number of non-missing values in each column. In this case, there are not any.

2.3. Data Preprocessing:

Now, we proceed to pre-process the dataset into proper shape for modelling algorithms. Because the scope of the project only concerns those compressive strengths achieved at 28 days, the data will be subsetted using the 'age' variable to include those observations at 28 days only and then drop the variable 'age' from the dataset. Further, the quantitative variable 'compressive strengths' will be converted to a categorical variable, where:

- Observations with compressive strength less than 40 MPa labelled as "non-compliant".
- Also, those with compressive strengths more than or equal to 40 MPa labelled as "compliant".

2.4. Data Exploration:

2.4.1. Single Variables:

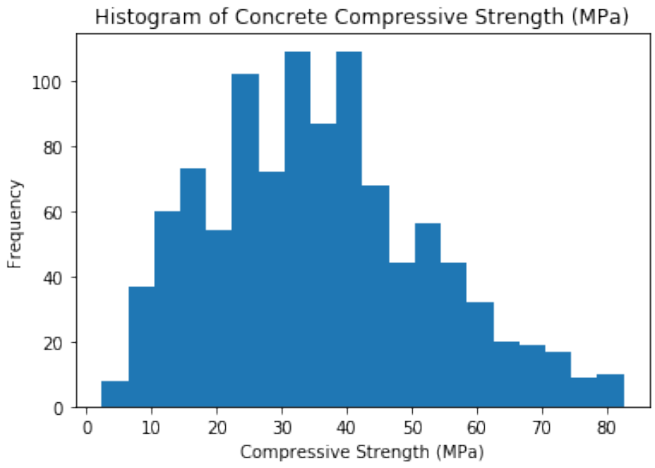

Figure 1 Histogram of Concrete Compressive Strength (MPa) at 28 Days

Figure 1 illustrates the histogram of concrete compressive strengths of mixtures in megapascals at 28 days from the dataset before it was converted to a categorical variable. It is observed that the distribution is approximately normally distributed and slightly right-skewed.

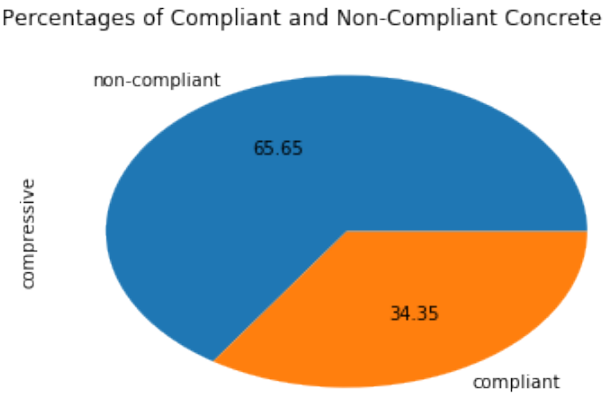

Figure 2 Percentages of Compliant and Non-Compliant Concrete Mixtures

Figure 2 is a pie chart illustrating the proportions of compliant and non-compliant concrete mixtures after the concrete compressive strengths have been converted to a categorical variable. It is seen that non-compliant mixtures account for two-thirds of total mixtures present in the dataset, while compliant mixtures account for only around one-third of total mixtures.

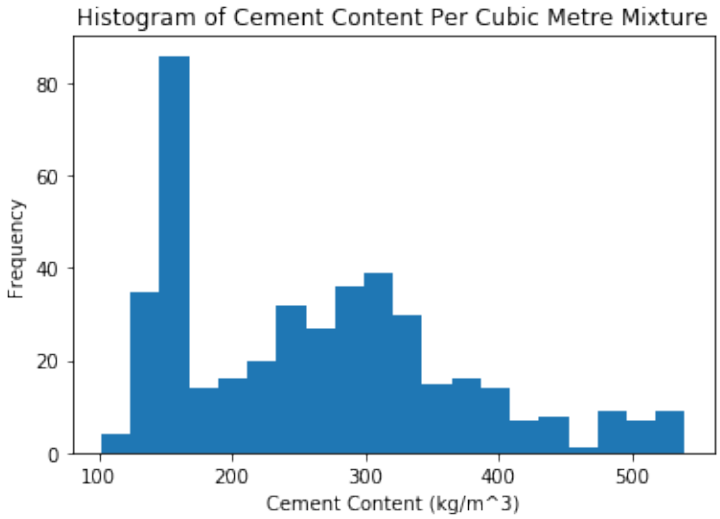

Figure 3 Histogram of Cement Content (kg/m^3)

Figure 3 shows the histogram of cement content in kilograms per cubic metre mixture. The distribution is not normally distributed, and the highest mode occurs at around 150 kg/m^3.

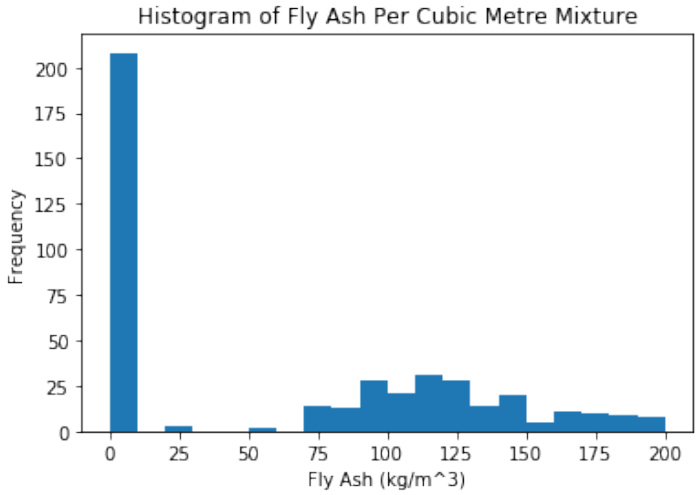

Figure 4 Histogram of Fly Ash Content (kg/m³)

Figure 4 shows the histogram of fly ash content in kilograms per cubic metre mixture. The highest mode of its distribution is at 0 kg/m^3. This is expected as this dataset is from a designed experiment and this is to account for when the mixtures contain no fly ash as it is a mineral addition and not an essential component of concrete mix.

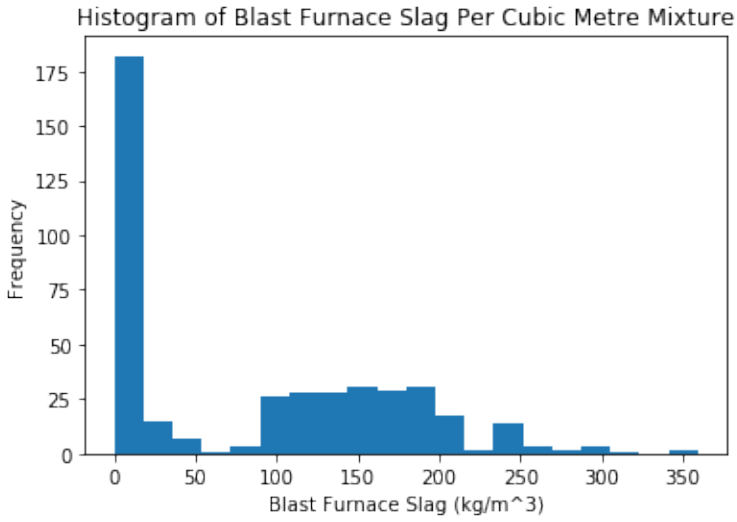

Figure 5 Histogram of Blast Furnace Slag Content (kg/m³)

Figure 5 shows the histogram of blast furnace slag content in kilograms per cubic metre mixture. Similar to fly ash, the highest mode of the distribution is at 0 kg/m^3. This is expected as this dataset is from designed experiment and this is to account for when the mixtures contain no blast furnace slag as it is a mineral addition and not an essential component of concrete mix.

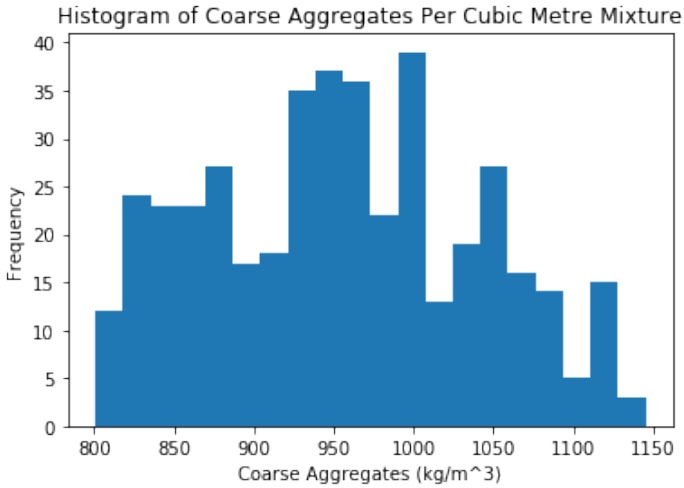

Figure 6 Histogram of Coarse Aggregates Content (kg/m³)

Figure 6 illustrates the histogram of coarse aggregates content in kilograms per cubic metre mixture. There is no discernable distribution pattern for this attribute.

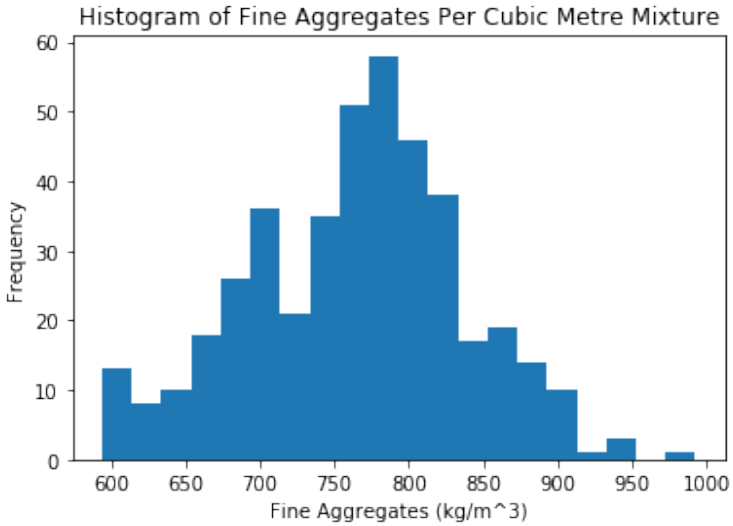

Figure 7 Histogram of Fine Aggregates Content (kg/m3)

Figure 7 illustrates the histogram of fine aggregates content in kilograms per cubic metre mixture. It is observed that its distribution is approximately normally distributed with a relatively heavy left tail.

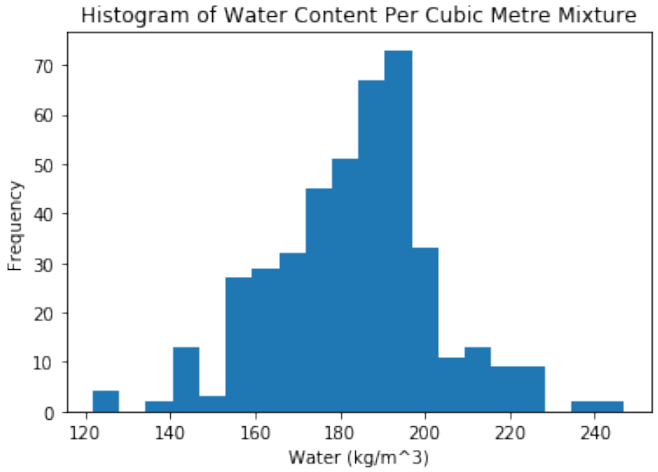

Figure 8 Histogram of Water Content (kg/m3)

Figure 8 illustrates the histogram of water content in kilograms per cubic metre mixture. It is also relatively normally distributed.

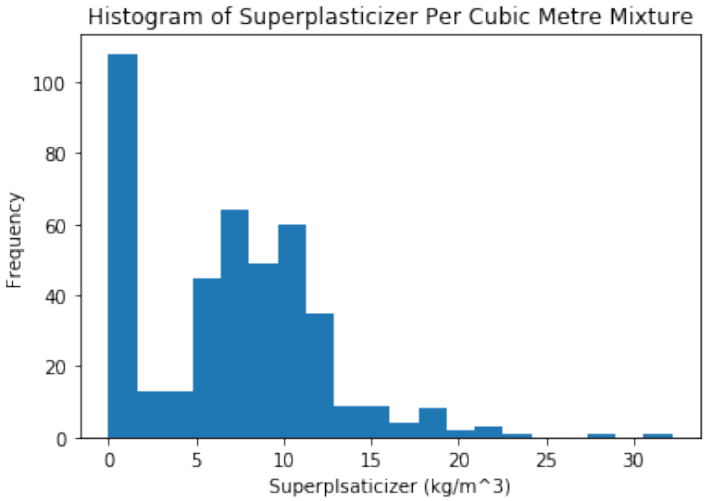

Figure 9 Histogram Superplasticizer Content (kg/m3)

Figure 9 illustrates the histogram of superplasticizer in kilograms per cubic metre mixture. There is no discernable distribution pattern for this particular attribute. The highest mode of the distribution is at 0 kg/m^3. This is expected as this dataset is from designed experiment and this is to account for when the mixtures contain no superplasticizer as it is an additive to improve concrete's workability and not an essential component of concrete mix.

2.4.2. Relationship Between Pairs of Attributes

Hypothesis: the figures of two horizontal box plots for each concrete mix components by compliance types shown below are to investigate whether it can be observed that any components have influence on categorizing compliance type.

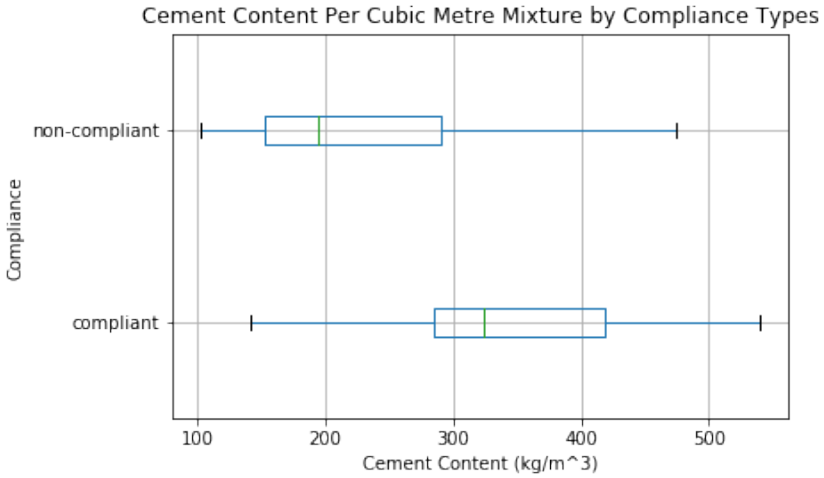

Figure 10 Cement Content by Compliance Types

Figure 10 shows horizontal box plots of cement content in kilograms per cubic metre mixture for both compliant and non-compliant types. Not surprisingly, the compliant mixtures have higher cement content as cement is an essential component for concrete to achieve higher compressive strengths.

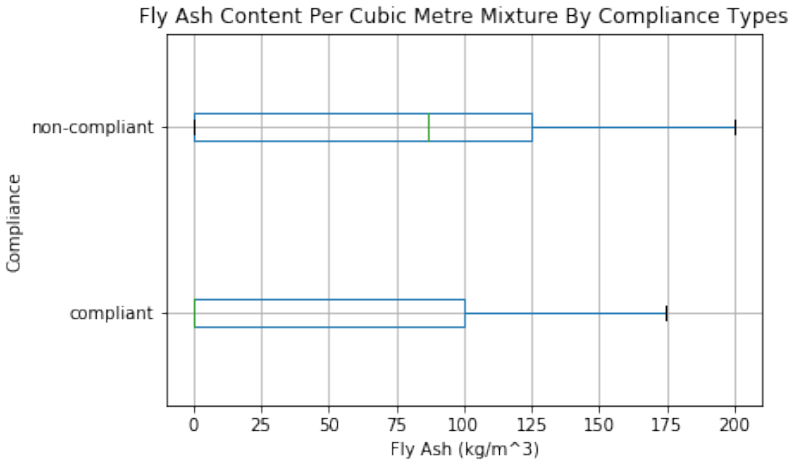

Figure 11 Fly Ash Content by Compliance Types

Figure 11 shows horizontal box plots of fly ash in kilograms per cubic metre mixtures for both compliant and non-compliant types. The same relationship of higher fly ash content present in compliant mixtures cannot be observed from this figure as it was for figure 9 with cement content.

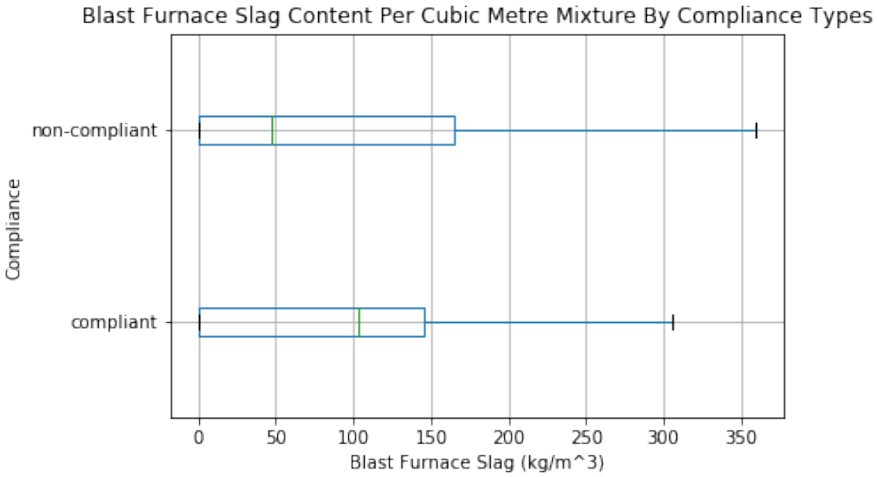

Figure 12 Blast Furnace Slag Content by Compliance Types

Figure 12 shows horizontal box plots of blast furnace slag content in kilograms per cubic metre mixtures for both compliant and non-compliant types. Similarly, the relationship of higher blast furnace slag content present in compliant mixtures cannot be observed from this figure as it was for figure 9 with cement content.

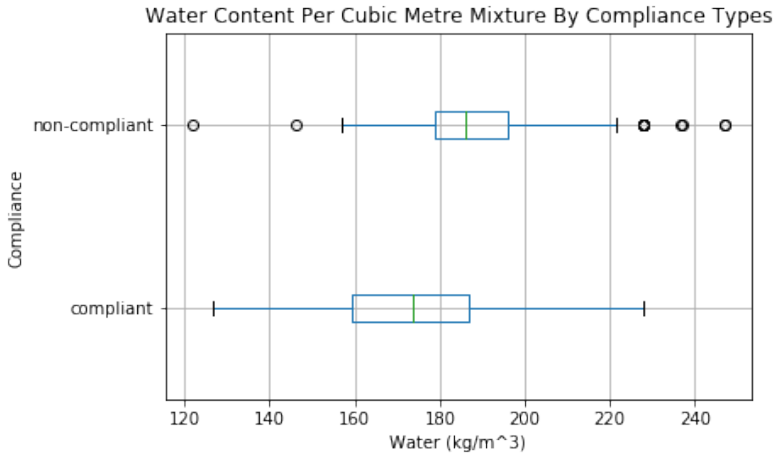

Figure 13 Water Content by Compliance Types

Figure 13 illustrates horizontal box plots of water content in kilograms per cubic metre mixtures for both compliant and non-compliant types. It is anticipated that mixture with higher water content would be weaker in strength than a mixture with lower water content. This is evident as can be observed from figure 13 that ‘non-compliant' mixture with concrete compressive strengths less than 40 MPa contain higher water in kg per cubic metre mixture.

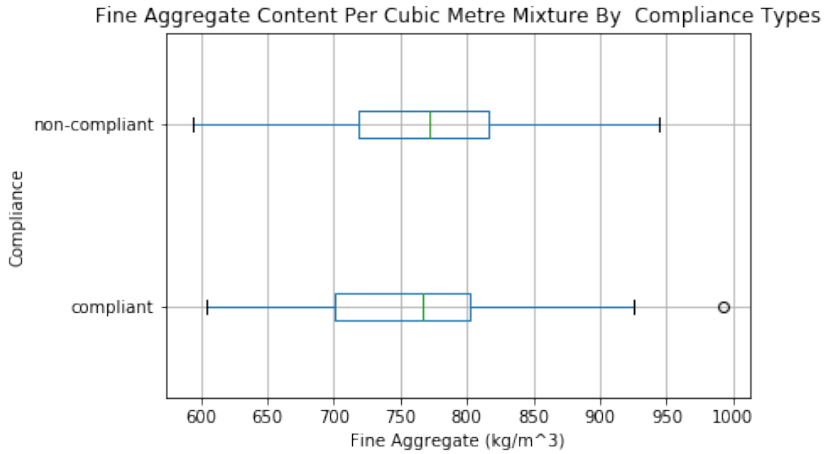

Figure 14 Fine Aggregates by Compliance Types

Figure 14 illustrates horizontal box plots of the amount of fine aggregates in kilograms per cubic metre mixtures for both compliant and non-compliant types. There seems to be no relationship suggesting that higher amount of fine aggregates present in the mix would lead to having the mixes deemed compliant or in other words, having higher compressive strengths.

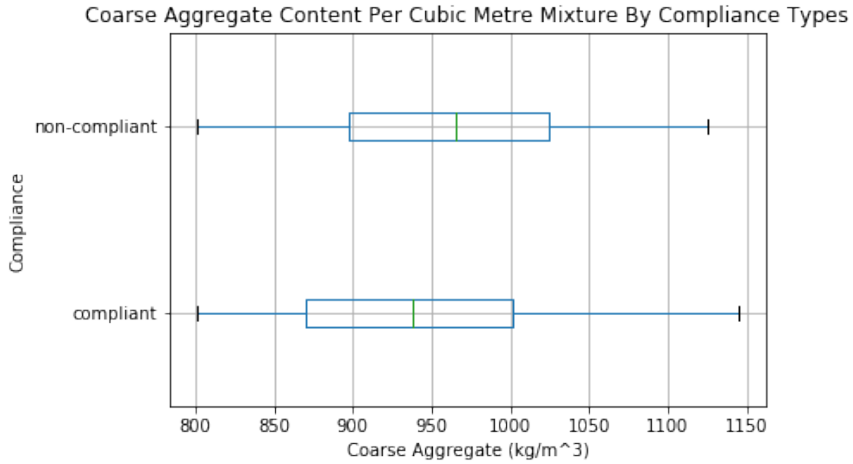

Figure 15 Coarse Aggregates Content Per Cubic Metre Mixture by Compliance Types

Similarly, figure 15 illustrates horizontal box plots of the amount of coarse aggregates in kilograms per cubic metre mixtures for both compliant and non-compliant types. There seems to be no relationship either suggesting that higher amount of coarse aggregates present in the mix would lead to having the mixes deemed compliant or in other words, having higher compressive strengths.

Hypothesis: the figures of scatter plot between one component of concrete mix against another illustrated below are to investigate if there are any relationship or lack thereof amongst components.

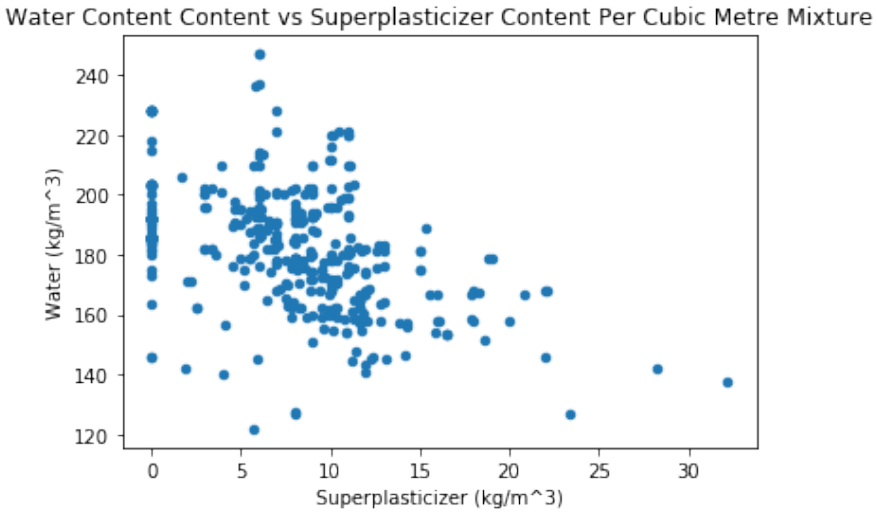

Figure 16 Scatterplot of Water Content vs Superplasticizer Content (kg/m³)

Figure 16 is a scatter plot of water content against superplasticizer content per cubic metre mixture. As superplasticizer is added to improve concrete's workability during placing and thus make it possible to reduce water content, it is anticipated that the more superplasticizer added in, the amount of water reduced. As shown in the plot, this proves to be the case as superplasticizer seems to be negatively correlated with water, and the linear relationship is moderately strong.

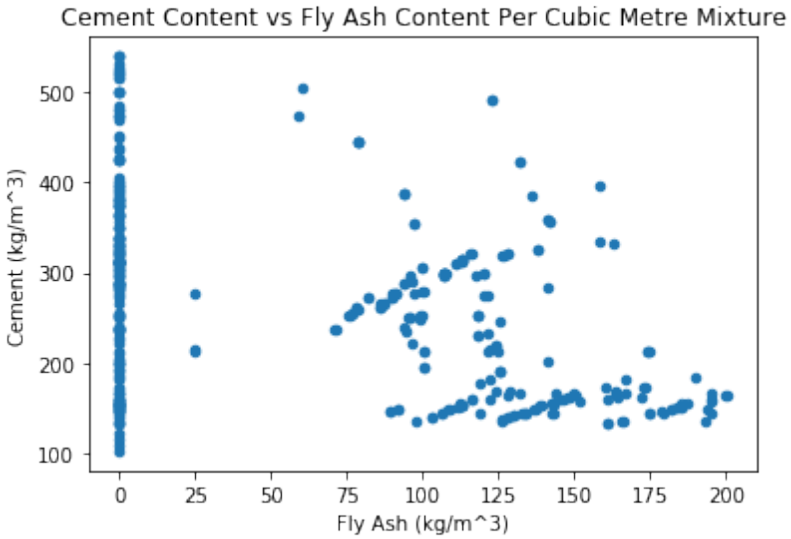

Figure 14 Scatter Plot of Cement Content vs Fly Ash Content (kg/m³)

Figure 17 is a scatter plot of cement content against fly ash content per cubic metre mixture. There appears to be no relationship between the two attributes.

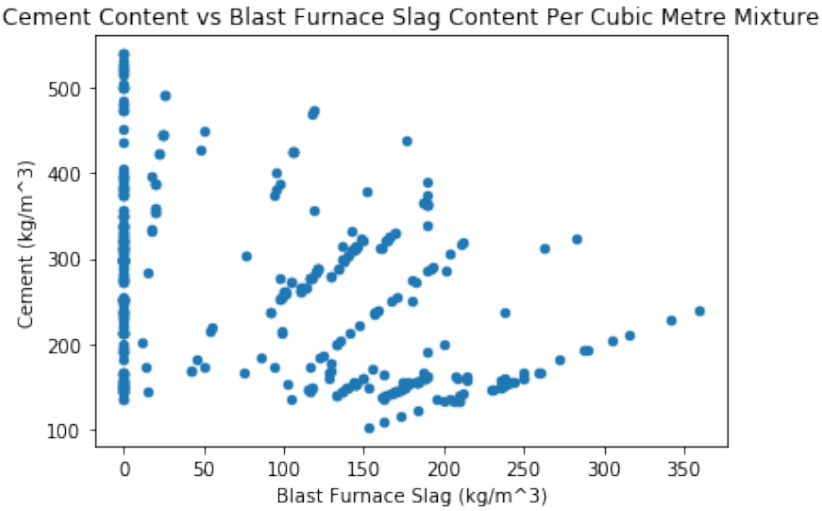

Figure 15 Scatter Plot of Cement Content vs Blast Furnace Slag Content (kg/m3)

Figure 18 is a scatter plot of cement content against blast furnace slag content per cubic metre mixture. There appears to be no relationship between the two attributes.

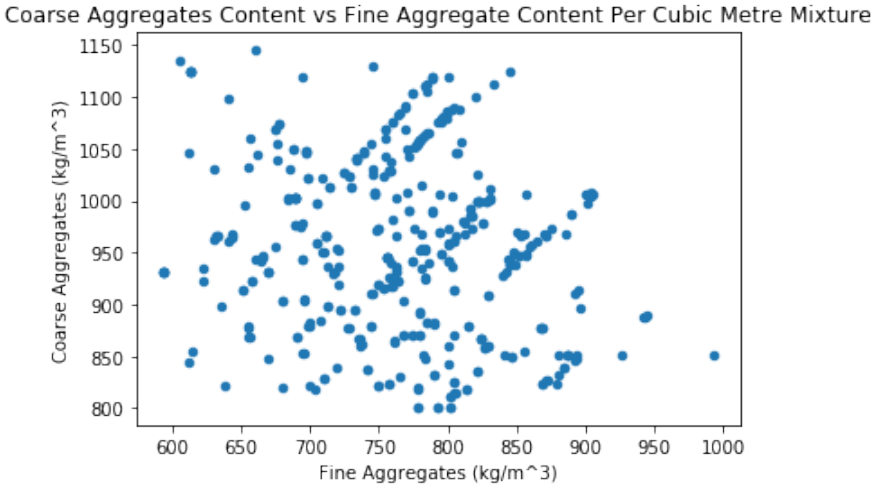

Figure 16 Scatter Plot of Coarse Aggregates vs Fine Aggregates Content (kg/m3)

Figure 19 is a scatter plot of coarse aggregates content against fine aggregates content per cubic metre mixture. There appears to be no relationship between the two attributes.

2.5. Data Modelling:

2.5.1. Evaluation Metric:

The definitions and decision paths of each confusion matrix components generated by the model are as followings:

- True Positive (TP): model predict the mix design as “compliant” and in reality is “compliant”.
    - Decision: produce the mix design’s concrete specimens and subject to property testings.
    - Meaning: testing useful mix design that in reality, meet Australian Standards.

- True Negative (TN): model predict the mix design as “non-compliant” and in reality, is "non-compliant."
    - Decision: do not produce specimens of the concrete mix design.
    - Meaning: saving money, time and effort on mix design that will eventually not meet Australian Standards.

- False Positive (FP): model predict the mix design as “compliant” but in reality, is "non-compliant."
    - Decision: produce the mix design’s concrete specimens and subject to property testings.
    - Meaning: spending money, time and resources of testing mix design that will eventually not meet Australian Standards.

- False Negative (FN): model predict the mix design as “non-compliant but in reality, is "compliant."
    - Decision: do not produce specimens of the concrete mix design.
    - Meaning: missing out on actually useful mix design.

The objective is to save money, effort and time by not testing mix designs that will eventually not meet Australian Standards while maximising the chance of selecting compliant mix designs. This means minimising false positive while maximising true positive. Therefore, the most crucial evaluation metric for this project is precision, which is TP / (TP + FP).

2.5.2. Modelling steps:

The modelling task is a supervised classification problem. Two supervised machine learning algorithms, k-nearest neighbour and decision tree algorithms, were used to fit the data.

For each algorithm, the data is split into three ratios of training and testing:

- 50% for training and 50% for testing
- 60% for training and 40% for testing
- 80% for training and 20% for testing, totalling to six fittings of the models.

2.5.2.1. K-nearest neighbour:

For the K-nearest neighbour algorithm, in each training and testing sets ratio, the steps of model building are outlined as followings:

- Since k-nearest neighbours is a non-parametric algorithm, there is no need to transform the attributes to normal distribution. However, because the range of values of all attributes varies significantly from one to another, MinMaxScaler is used to transform all attributes to the range [0, 1].
- The algorithm’s hyperparameters will be tuned using GridSearchCV algorithm to achieve the highest precision score. Followings are ranges of values chosen for hyperparameters:
    - parameter ‘n_neigbours' are chosen arbitrarily from 3 to 30 in an increment of 1.

    - ‘uniform' is chosen for ‘weight' parameter, meaning that all points in each neighbourhood are weighted equally. This is because all concrete components are in the same unit, kilograms per cubic metre of the mixture.
    - Power parameter 1 and 2 are chosen to tune, which will enforce Manhattan distance and Euclidean distance, respectively. ‘p = 1' is used here to take into account most of the dimensionality and measuring distance based on all attributes. However, ‘p=2' is also chosen because some particular components might influence concrete compressive strengths more than others.
- The GridSearchCV will try models from every possible combination of parameters specified. This algorithm splits the data into a training set and validation set where the training set is used to tune the parameter and validated against the validation set. This is done ten times as specified to ensure accuracy of the validated model and avoiding overfitting and underfitting.
- K-fold cross-validation is finally used to fit the original dataset. The algorithm split the dataset into a training set and testing set where the training set is fit to the validated model from hyperparameter tuning, and then validated against the testing set. This is done ten times as specified to ensure the accuracy of the score and avoid overfitting and underfitting.

2.5.2.2. Decision Tree

For the decision tree algorithm, in each training and testing sets ratio, the steps of model building are outlined as followings:

- For the decision tree algorithm, there is no need to normalise the data. Thus, inverse_transform was used to transform the features back to their original values.
- The algorithm’s hyperparameters will be tuned using GridSearchCV algorithm to achieve the highest precision score. Followings are ranges of values chosen for hyperparameters:
    - Max_features: is chosen from 4 and 7 to avoid fitting, because, from prior research, some components are most influential to concrete's compressive strength. Those are cement, fly ash, slag and superplasticizer/water. Consequently, when splitting each node, the algorithms would want to consider at least four features.
    - Min_samples_leaf is 10 to 55 in an increment of 5 chosen as ten observations at least needed to infer to avoid overfitting.
    - Max_depth is chosen arbitrarily from 3 to 7 in an increment of 1.
    - ‘gini’ and ‘entropy’ are both chosen as ‘criterion’ parameter as ways of calculating homogeneous set since there is no consensus on which one outperforms another.
    - Min_sample_splits is chosen arbitrarily as 10 to 100 in an increment of 5.
    - Class_weight is chosen to be ‘None’ or ‘balanced’ with the latter to consider avoiding bias on majority class.
- The GridSearchCV will try models from every possible combination of parameters specified. This algorithm splits the data into a training set and validation set where the training set is used to tune the parameter and validated against the validation set. This

is done ten times as specified to ensure accuracy of the validated model and avoiding overfitting and underfitting.

- K-fold cross-validation is finally used to fit the original dataset. The algorithm split the dataset into a training set and testing set where the training set is fit to the validated model from hyperparameter tuning, and then validated against the testing set. This is done ten times as specified to ensure the accuracy of the score and avoid overfitting and underfitting.

3. Results

3.1. K-nearest Neighbours Algorithm

3.1.1. 50/50 training and testing split

Table 1 Confusion Matrix of K-nearest Neighbour on Testing Set with 50/50 Split

Confusion Matrix		Predicted	
		Non-compliant	Compliant
Actual	Non-compliant	130	10
	Compliant	31	42

```
Classification Error Rate = 0.19

                precision    recall  f1-score   support
non-compliant        0.81      0.93      0.86       140
    compliant        0.81      0.58      0.67        73

    micro avg        0.81      0.81      0.81       213
    macro avg        0.81      0.75      0.77       213
 weighted avg        0.81      0.81      0.80       213
```

3.1.2. 60/40 training and testing split

Table 2 Confusion Matrix of K-nearest Neighbour on Testing Set with 60/40 Split

Confusion Matrix		Predicted	
		Non-compliant	Compliant
Actual	Non-compliant	107	5
	Compliant	15	43

```
Classification Error Rate = 0.12

                precision    recall  f1-score   support
non-compliant        0.81      0.98      0.89       112
    compliant        0.94      0.57      0.71        58

    micro avg        0.84      0.84      0.84       170
    macro avg        0.88      0.78      0.80       170
 weighted avg        0.86      0.84      0.83       170
```

3.1.3. 80/20 training and testing split

Table 3 Confusion Matrix of K-nearest Neighbour on Testing Set with 80/20 Split

Confusion Matrix		Predicted	
		Non-compliant	Compliant
Actual	Non-Compliant	53	3
	Compliant	12	17

```
Classification Error Rate = 0.18

                precision    recall  f1-score   support
non-compliant        0.82      0.95      0.88        56
    compliant        0.85      0.59      0.69        29

    micro avg        0.82      0.82      0.82        85
    macro avg        0.83      0.77      0.78        85
 weighted avg        0.83      0.82      0.81        85
```

3.2. Decision Tree Algorithm

3.2.1. 50/50 training and testing split

Table 4 Confusion Matrix of Decision Tree on Testing Set with 50/50 Split

Confusion Matrix		Predicted	
		Non-compliant	Compliant
Actual	Non-compliant	130	6
	Compliant	19	58

```
Classification Error Rate = 0.12

                precision    recall  f1-score   support
non-compliant        0.80      0.96      0.87       136
    compliant        0.88      0.57      0.69        77

    micro avg        0.82      0.82      0.82       213
    macro avg        0.84      0.76      0.78       213
 weighted avg        0.83      0.82      0.81       213
```

3.2.2. 60/40 training and testing split

Table 5 Confusion Matrix of Decision Tree on Testing Set with 60/40 Split

Confusion Matrix		Predicted	
		Non-compliant	Compliant
Actual	Non-compliant	95	16
	Compliant	8	51

```
Classification Error Rate = 0.14
                precision    recall  f1-score   support

non-compliant       0.92      0.86      0.89       111
    compliant       0.76      0.86      0.81        59

    micro avg       0.86      0.86      0.86       170
    macro avg       0.84      0.86      0.85       170
 weighted avg       0.87      0.86      0.86       170
```

3.2.3. 80/20 training and testing split

Table 6 Confusion Matrix of Decision Tree on Testing Set with 80/20 Split

Confusion Matrix		Actual	
		Non-compliant	Compliant
Actual	Non-compliant	48	9
	Compliant	5	23

```
Classification Error Rate = 0.16
                precision    recall  f1-score   support

non-compliant       0.91      0.84      0.87        57
    compliant       0.72      0.82      0.77        28

    micro avg       0.84      0.84      0.84        85
    macro avg       0.81      0.83      0.82        85
 weighted avg       0.84      0.84      0.84        85
```

4. Discussion

Table 7 Model Comparison by Precision Score

Summary statistics of the precision score on the 10-fold cross-validated final model		K-Nearest Neighbors		Decision Tree	
		Mean	Standard Deviation	Mean	Standard Deviation
training/testing split ratios	50%-50%	93.028%	5.281%	72.523%	10.172%
	60%-40%	93.653%	5.340%	71.1521%	11.4315%
	80%-20%	93.761%	4.389%	69.698%	10.774%

According to Table 7, model resulting from fitting data at 80% to 20% training/testing split achieves mean precision score of 93.653% with standard deviation 5.340% on the testing set after 10-fold cross-validation of the final model after hyperparameter tuning. This is the best precision score achieved among all six fittings of the model. Further, it can be observed that for the dataset with all numerical features, k-nearest neighbours algorithm outperforms decision tree on all training/testing splitting ratios. Moreover, K-nearest neighbour performance at all training/testing splitting ratios is relatively consistent compared to decision tree algorithm. As noted, decision tree performance decreases as testing size decreases.

5. Conclusion and Recommendation

In conclusion, k-nearest neighbour algorithm fitting at 80%/20% training/testing splits with hyperparameters shown as followings is recommended with features standardised using MinMaxScale prior to fitting:

```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski'
, metric_params=None, n_jobs=None, n_neighbors=22, p=1, weights='unifor
m').
```

6. References

Standards Australia 2009, *Concrete Structures*, AS 3600-2009, viewed 04 June 2019, SAI Global database.

Yeh, I.-C 1998, 'Modeling of strength of high-performance concrete using artificial neural networks', *Cement and Concrete Research*, Vol. 28, No. 12, pp. 1797-1808.