# Data Science Techniques and Applications - Dataset analysis

February 23, 2021

Orlando Taddeo, MSc Data Science, Student ID 13180720

## 1 Academic Declaration

"I have read and understood the sections of plagiarism in the College Policy on assessment offences and confirm that the work is my own, with the work of others clearly acknowledged. I give my permission to submit my report to the plagiarism testing database that the College is using and test it using plagiarism detection software, search engines or meta-searching software".

## 2 Introduction

The dataset I decided to use is '**Concrete Compressive Strength Data**' [1], that contains data about a sample of 1,030 structural concrete specimens. The dataset comprises different types of information about the concrete composition and maturing time. The target variable is the *concrete strength*, so the aim of the data collecting conducted is to try to deduce such strength from the other parameters. I decided to use this data because my background is in Structural Engineering, and I worked mainly with concrete in all the projects I took part in.

Structural concrete is one of the most common construction materials around the world. It is made by mixing cement, water, aggregates of various sizes and chemical additives with different functions. The obtained mix is fluid for a few hours and can then be poured in moulds of virtually any shape.

Fig. 1: Fluid concrete

As time passes, the cement hardens and becomes a strong matrix that holds together the aggregates, forming the actual structure. Embedded in hardened concrete are steel reinforcing bars, that are able to absorb tensile stresses. In fact, concrete has a very limited capacity of absorbing tensile forces: this can be seen from cracks appearing almost everywhere in building structures. Coupling with reinforcing steel solves the problem brilliantly.

Fig. 2: Reinforced concrete

As it can be seen, concrete is usually not produced in a strictly controlled factory environment, as it happens with steel, so there is a significant variability in the final product. This affects all the mechanical properties of the material, but we will now focus only on strength. Evaluating compression strength of the concrete starting from its composition (in terms of water, aggregate, etc…) has always been a challenge for Structural Engineers. Building this dataset is the initial part of an attempt to use Statistical Learning methods to derive the compression strength from the concrete composition.

## 3 Dataset description

The dataset comprises 1,030 rows (or instances) and 9 columns (or features). Each instance represents a concrete sample, which composition is indicated by the first 8 features. The last one is the compressive strength registered in a lab test, that represents the target value of possible Statistical

Learning procedures. The data has been previously cleaned and organised, and the dataset does not contain missing values. The first 5 rows of the dataframe are shown below (including the header).

[3]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

source_path = '/Users/orlandotaddeo/Desktop/concrete data/Concrete_Data.csv'
source_df = pd.read_csv(source_path)

print(source_df.head())
```

```
   Cement (component 1)(kg in a m^3 mixture)  \
0                                      540.0
1                                      540.0
2                                      332.5
3                                      332.5
4                                      198.6

   Blast Furnace Slag (component 2)(kg in a m^3 mixture)  \
0                                                0.0
1                                                0.0
2                                              142.5
3                                              142.5
4                                              132.4

   Fly Ash (component 3)(kg in a m^3 mixture)  \
0                                         0.0
1                                         0.0
2                                         0.0
3                                         0.0
4                                         0.0

   Water  (component 4)(kg in a m^3 mixture)  \
0                                      162.0
1                                      162.0
2                                      228.0
3                                      228.0
4                                      192.0

   Superplasticizer (component 5)(kg in a m^3 mixture)  \
0                                                2.5
1                                                2.5
2                                                0.0
3                                                0.0
4                                                0.0
```

```
   Coarse Aggregate  (component 6)(kg in a m^3 mixture)  \
0                                             1040.0
1                                             1055.0
2                                              932.0
3                                              932.0
4                                              978.4

   Fine Aggregate (component 7)(kg in a m^3 mixture)  Age (day)  \
0                                              676.0         28
1                                              676.0         28
2                                              594.0        270
3                                              594.0        365
4                                              825.5        360

   Concrete compressive strength(MPa)
0                               79.99
1                               61.89
2                               40.27
3                               41.05
4                               44.30
```

Below, we briefly describe the features used.

**Cement**: measured in kg per cubic metre of the final mixture, it is the basic component of concrete. It is originally a powder, that reacts with water and hardens to form a solid matrix in which the aggregate is embedded.

**Blast Furnace Slag**: measured in kg per cubic metre of the final mixture, it is a by-product of the iron production. Adding it to the concrete mix has shown to improve the mechanical properties of the latter, such as workability, durability, strength and resistance to environmental chemical agents (e.g., chlorides) [4].

**Fly Ash**: measured in kg per cubic metre of the final mixture, is a by-product of coal combustion. Its addition to the concrete mixture has shown to improve workability and reduce the cost [2].

**Water**: measured in kg per cubic metre of the final mixture, it is needed to hydrate the cement and making it become a strong matrix after its evaporation. High content of water improves workability, but reduces the strength of the final product.

**Superplasticizer**: measured in kg per cubic metre of the final mixture, allows to reduce the amount of water in it, then increasing concrete strength whilst keeping good workability.

**Coarse Aggregate**: measured in kg per cubic metre of the final mixture, is is basically a mix of small stones obtained by crushing natural rocks. It is the most important element that provides the concrete with its strength.

**Fine Aggregate**: measured in kg per cubic metre of the final mixture, it is similar to the coarse one, but it is made up if smaller grains (diameters ranging from a few millimetres to fractions of a millimetre).

**Age**: measured in days, it is the time passed from when the concrete is poured to when it is tested. It greatly affects the concrete strength, but after about 4 weeks (i.e. 28 days), the increase

in strength with time tends to be insignificant.

**Concrete compressive strength**: it is the target value of the data analysis. It is measured by crushing a concrete sample in a laboratory and measuring the maximum force that it is applied to crush the specimen. It is measured in MegaPascal (that is, Newton per squared millimetres), and represents the concrete strength per unit of area.

To allow for an easier visualisation of the data, we substitute the feature names with some shorter ones. The mapping from the old to the new feature names is contained in the dictionary shown below.

[4]:
```
new_names = {'Cement (component 1)(kg in a m^3 mixture)': 'cement',
             'Blast Furnace Slag (component 2)(kg in a m^3 mixture)': 'slag',
             'Fly Ash (component 3)(kg in a m^3 mixture)': 'ash',
             'Water  (component 4)(kg in a m^3 mixture)': 'water',
             'Superplasticizer (component 5)(kg in a m^3 mixture)':␣
 ↪'superplast',
             'Coarse Aggregate  (component 6)(kg in a m^3 mixture)':␣
 ↪'coarse_agg',
             'Fine Aggregate (component 7)(kg in a m^3 mixture)': 'fine_agg',
             'Age (day)': 'age',
             'Concrete compressive strength(MPa)': 'comp_str'}

source_df.rename(columns=new_names, inplace=True)
```

We can now show the main dataframe features using the `pandas DataFrame.info()` and `DataFrame.describe()` methods.

[5]:
```
print(source_df.info())
```

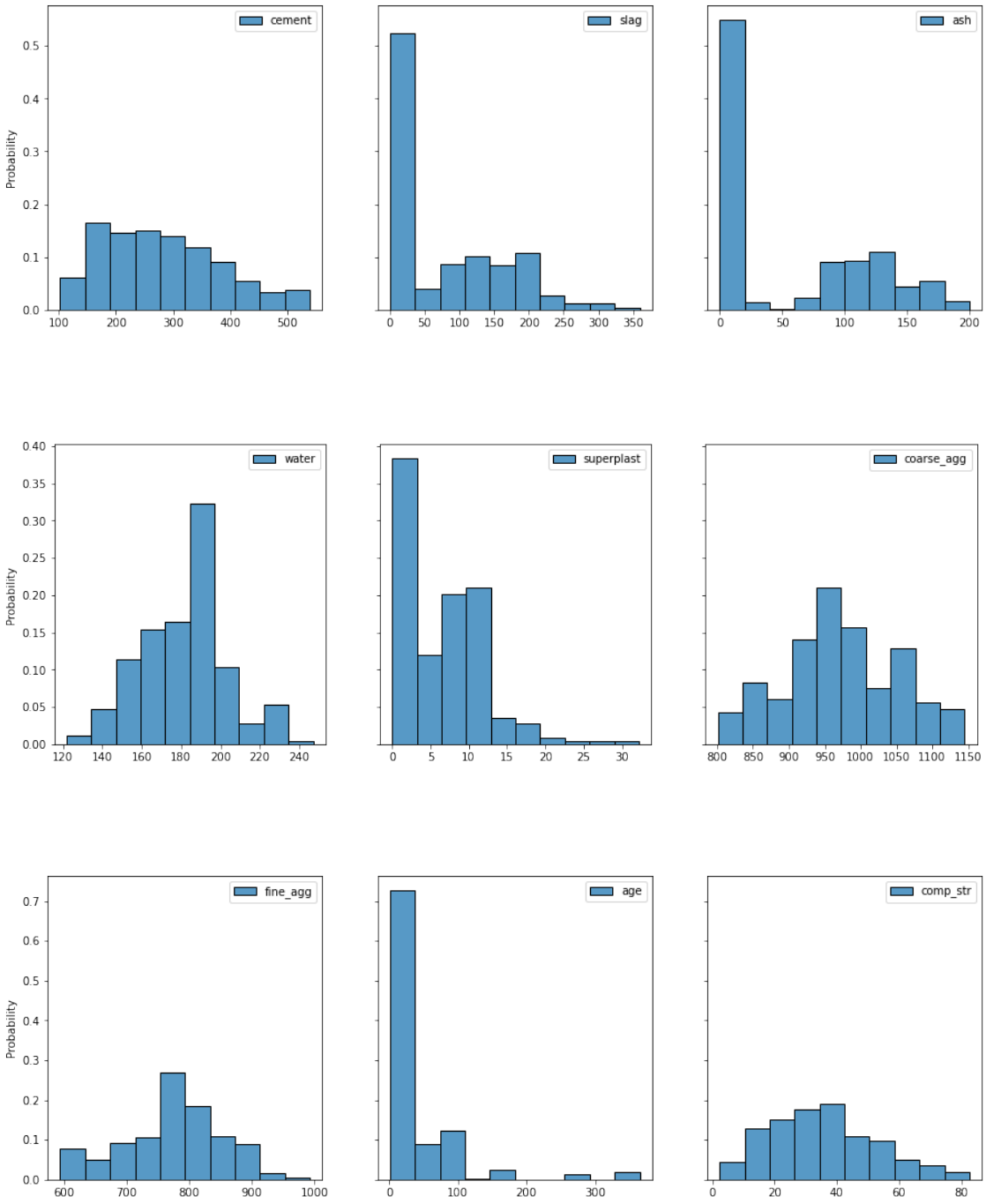
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1030 entries, 0 to 1029
Data columns (total 9 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   cement      1030 non-null   float64
 1   slag        1030 non-null   float64
 2   ash         1030 non-null   float64
 3   water       1030 non-null   float64
 4   superplast  1030 non-null   float64
 5   coarse_agg  1030 non-null   float64
 6   fine_agg    1030 non-null   float64
 7   age         1030 non-null   int64
 8   comp_str    1030 non-null   float64
dtypes: float64(8), int64(1)
memory usage: 72.5 KB
None
```

[6]:
```
print(source_df.describe())
```

```
              cement         slag          ash        water   superplast  \
count  1030.000000  1030.000000  1030.000000  1030.000000  1030.000000
mean    281.167864    73.895825    54.188350   181.567282     6.204660
std     104.506364    86.279342    63.997004    21.354219     5.973841
min     102.000000     0.000000     0.000000   121.800000     0.000000
25%     192.375000     0.000000     0.000000   164.900000     0.000000
50%     272.900000    22.000000     0.000000   185.000000     6.400000
75%     350.000000   142.950000   118.300000   192.000000    10.200000
max     540.000000   359.400000   200.100000   247.000000    32.200000

         coarse_agg     fine_agg          age     comp_str
count  1030.000000  1030.000000  1030.000000  1030.000000
mean    972.918932   773.580485    45.662136    35.817961
std      77.753954    80.175980    63.169912    16.705742
min     801.000000   594.000000     1.000000     2.330000
25%     932.000000   730.950000     7.000000    23.710000
50%     968.000000   779.500000    28.000000    34.445000
75%    1029.400000   824.000000    56.000000    46.135000
max    1145.000000   992.600000   365.000000    82.600000
```

As we can see, the dataset is free from missing values, and all of them are numeric (either floating-point or integer type). We can approach the data analysis, as suggested in [3], by considering each of the features as a sample of a random variable. From the `describe()` method's output, we can see the main statistics used to describe samples of random variables, related to the sample contained in the dataset: count, mean, standard deviation, main quartiles, minimum and maximum values. We can have a better idea of how the data is distributed, however, by plotting histograms of each feature, reported below.

[7]:
```python
def hist_plotter(data_frame, n_plots_per_line):
    plt.rcParams["figure.figsize"] = (15, 5)
    n_feat=data_frame.shape[1]
    indices = np.array(range(n_feat))
    ind_mat = np.reshape(indices, (-1, int(n_feat/n_plots_per_line)))

    for i in range(ind_mat.shape[0]):
        fig, axes = plt.subplots(nrows=1, ncols=n_plots_per_line, sharey=True)
        for j in ind_mat[i]:
            feature = data_frame.iloc[:, [j]]
            sns.histplot(data=feature,bins=10, stat='probability', ax=axes[j%3])
        plt.show()


hist_plotter(source_df, 3)
```



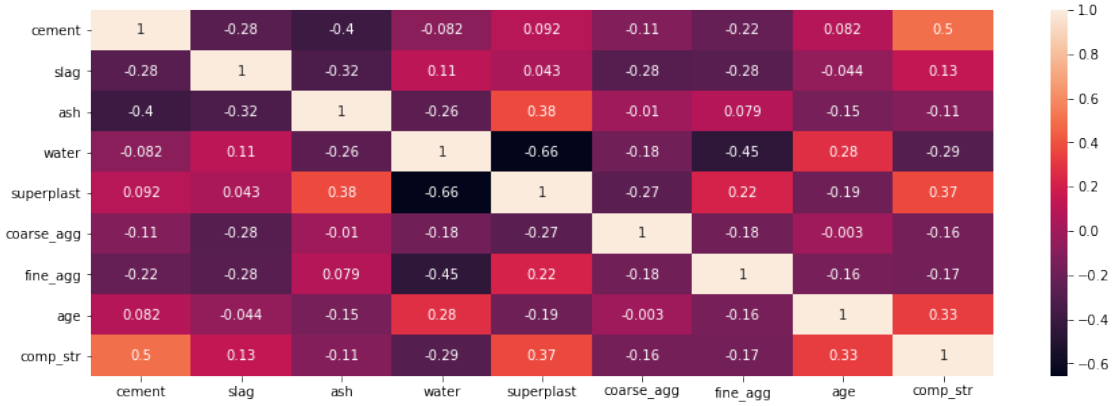

Some of these distributions are rather skewed. This is because they reflect the actual population of concrete specimens in the built environment. For example, as we said earlier, almost all the concrete tested has an Age of approximately 28 days, so the distribution of the variable **Age** is significantly skewed towards this value. The same happens with some components which presence in concrete is limited of a specific range in terms of percentage of the weight of the final mixture. Fly ashes, for example, don't usually exceed 15-20% of the total weight. With regard to the strength (the

target variable), the distribution is skewed as well. In fact, the vast majority of concrete structures is built using concrete with a compressive strength between 25MPa and 45MPa, even though it is always easier, nowadays, to produce concrete that exceeds 80MPa in strength.

In order to understand whether the features used are linearly correlated to each other, we can compute their correlation matrix, graphically shown below.

[8]:
```
corr_df = source_df.corr()
sns.heatmap(corr_df, annot=True)
```

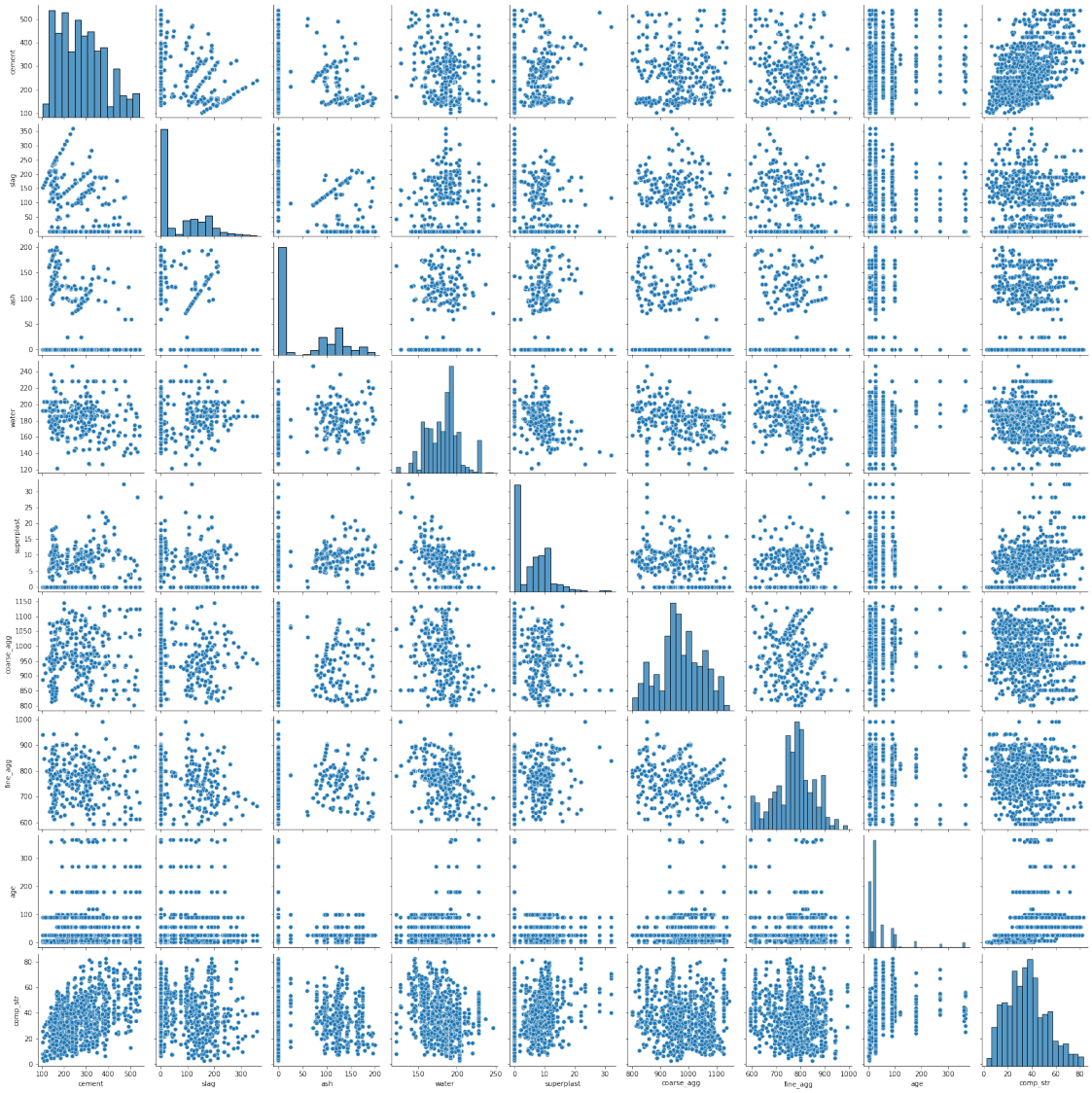
[8]: <AxesSubplot:>




As it can be seen, the correlation coefficients contained in the matrix are quite low, with a maximum absolute value of 0.66. The highest correlation coefficients are those between water and other components, such as superplasticizer, fine and coarse aggregate. This happens because the amount of such components is related to the quantity of water used by guidelines on concrete production. Weak linear relationships appear to exist between cement content and strength, as well as age, water, superplasticizer content and strength. It is widely recognised, in the scientific literature, that water content and age strongly affect the strength. However, these relationships may be nonlinear, so it is worth having a look at a graphical representation of the data, because the correlation coefficients might not catch this. We can do this using a pairplot, reported below.

[9]:
```
sns.pairplot(source_df)
```

[9]: <seaborn.axisgrid.PairGrid at 0x7fdaf6f89f70>




As we can see, there is a significant variability in the data.

The relationship between concrete age and compression strength is such that, as age increases between 0 and 28 days, the strength increases significantly as well. However, approximately after a month of maturing time, the strength stays more or less constant. This is a highly nonlinear relationship, so it is not captured by the correlation coefficients that we calculated before. Age is, therefore, not surprisingly, one of the best indicators of the value of the target variable. Finally, it is worth noting that values of age over 100 days are very rare. Limiting the age variable maximum to this value in the plot might reveal a clearer trend in the data.

At first glance, cement content seems to be linearly related to the compression strength, and this relationship becomes weaker as the cement content increases, likely because some asymptotic value is approached.

The relationship between water and compression strength is not very clear. Although it is well

known from practice that increasing the water content reduces the concrete strength, this correlation does not seem to be very strong in the dataset, given the variability present.

It can be seen that some features are somehow correlated to each other, at least at first glance. In particular, the presence of coarse aggregate, fine aggregate, ash and slag content appear to follow a linear relationship. This is mainly due to the way concrete is made: some specific ratios between these components are normally adopted, so it is not surprising that they are related to each other. They are possible candidates for an exclusion from the feature space, when aiming to dimensionality reduction.

Finally, it is worth noting that some kinds of concrete do not contain fly ash and/or blast furnace slag at all, and they are usually classified as different materials. In fact, their properties may differ significantly from concrete in which such components are present. These kinds of concretes are represented by points aligning vertically on the zero value in the plots in which the content of fly ashes and slag is on the horizontal axis. It could be useful to remove such points and create a separate dataset before applying Statistical Learning algorithms. If, for example, we would like to carry out a regression analysis, such points could affect the result without carrying any information about the actual relationship between ashes and slag and other components. This would happen because they represent instances in which ash and slug are not present at all, and their role is played by other components (e.g., fine aggregate).

# 4 Conclusions

The dataset under examination contains data about concrete components, together with values of its compression strength. The data were pulled together to investigate the relationships between such components and the material's resistance against compression loads. Currently, there is a significant difficulty in finding appropriate mathematical models to capture the influence of the quantity of such components in the mixture on the final material strength. Therefore, it seems appropriate to try to apply the methods of Statistical Learning to predict the concrete strength, without the use of physical models.

We carried out a first-level analysis of the dataset, using both numerical and visual data exploration. On the basis of the correlation coefficient values, we observed that just in very few cases some linear relationship between some predictors could exist. In particular, the biggest coefficients relate the quantity of water to that of aggregates and additives. However, this might derive from the "recipes" used in making the concrete mix, where some constant ratios are used when adding components to the mixture. An interesting information, but again not surprising, is that the cement content is related to the compression strength. Most likely, however, a best-fit function would be nonlinear. The subsequent visual exploration of the pairplot highlighted how, a few cases, different features could be correlated nonlinearly. For example, we found a confirmation that the maturing age has a significant effect on the final strength. The variability in the data, however, remains the greatest difficulty to tackle.

In the following coursework, we will try to apply a dimensionality reduction technique (i.e., PCA) to attempt to reduce the number of features that could be used to successfully predict the concrete strength.

## References


[1] Kaggle, Concrete Compressive Strength Data, available online at https://www.kaggle.com/vivekgediya/concrete-data. Last accessed 23/02/2021.

[2] Mehta, P.K., 2004, May. High-performance, high-volume fly ash concrete for sustainable development. In Proceedings of the international workshop on sustainable development and concrete technology (pp. 3-14). Ames, IA, USA: Iowa State University.

[3] Mohammed J. Zaki, Wagner Meira, Jr., Data Mining and Machine Learning: Fundamental Concepts and Algorithms, 2nd Edition, Cambridge University Press, March 2020. ISBN: 978-1108473989.

[4] Zulu, B.A., Miyazawa, S. and Nito, N., 2019. Properties of blast-furnace slag cement concrete subjected to accelerated curing. Infrastructures, 4(4), p.69.