IITM-IAARC Joint Hackathon

Predicting Compressive Strength of Concrete with Recycled Aggregates using Machine Learning

Harnessing the Power of Machine Learning to Build Sustainable Infrastructure

By- Lenin Kennedy

Team 4

About Myself

- My name is Lenin Kennedy
- I'm a 3rd year student from Vellore Institute of technology, Chennai.
- I am pursuing B. Tech in Electronics and Computer Engineering

VIT®

Vellore Institute of Technology

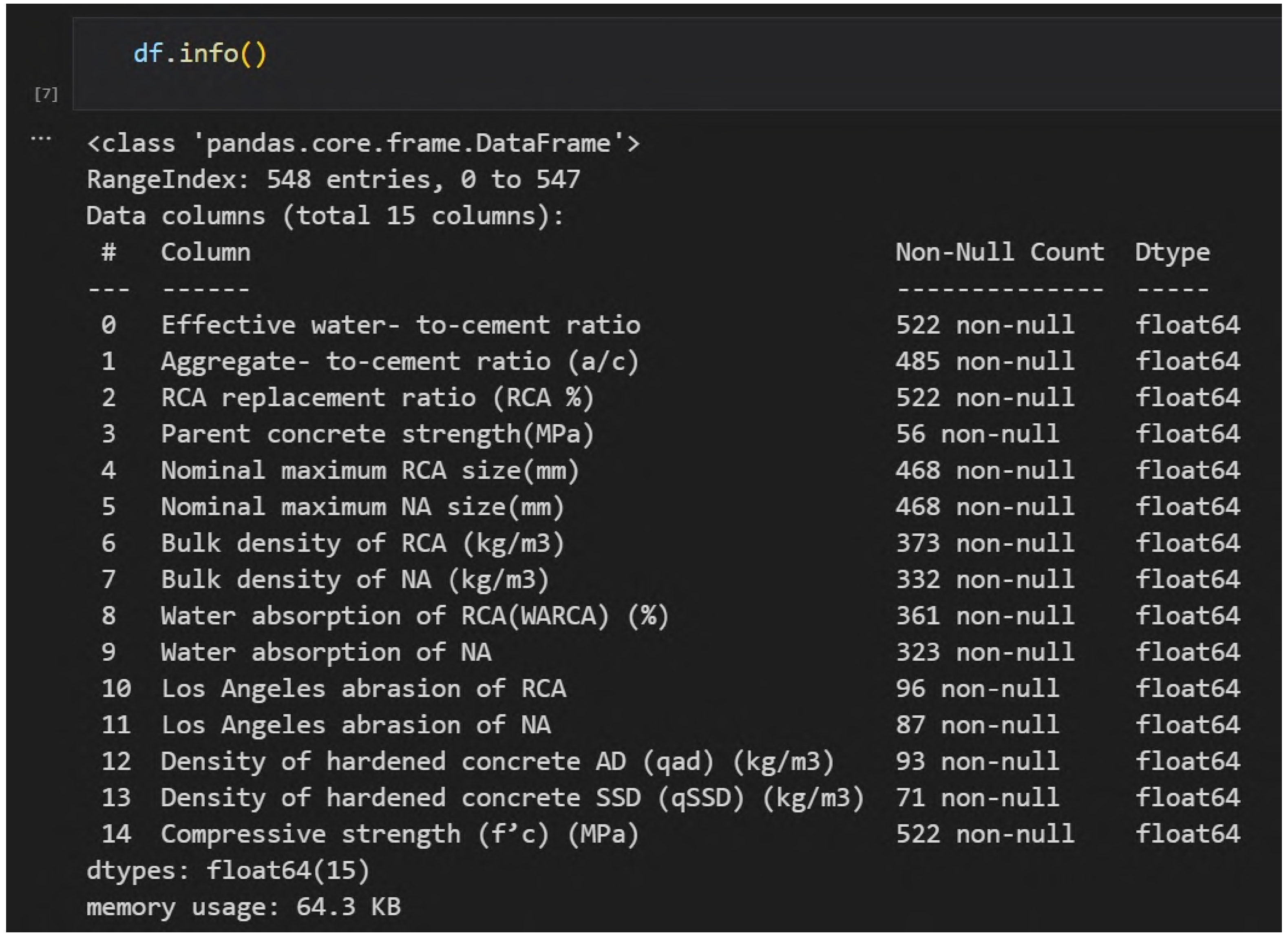
(Deemed to be University under section 3 of UGC Act, 1956)

Problem Statement

- **Goal:** Predict the compressive strength of concrete made using recycled aggregates.
- **Background:** Concrete production using recycled aggregates presents challenges in achieving sufficient compressive strength, which is crucial for structural integrity and durability.
- **Significance:** Accurately predicting the compressive strength of concrete with recycled aggregates can promote sustainable construction practices and reduce environmental impact.

Dataset Overview

The dataset comprises 19 features related to the compressive strength of concrete made using recycled aggregates. These features provide valuable insights into the composition and properties of the concrete.



```
df.info()
```

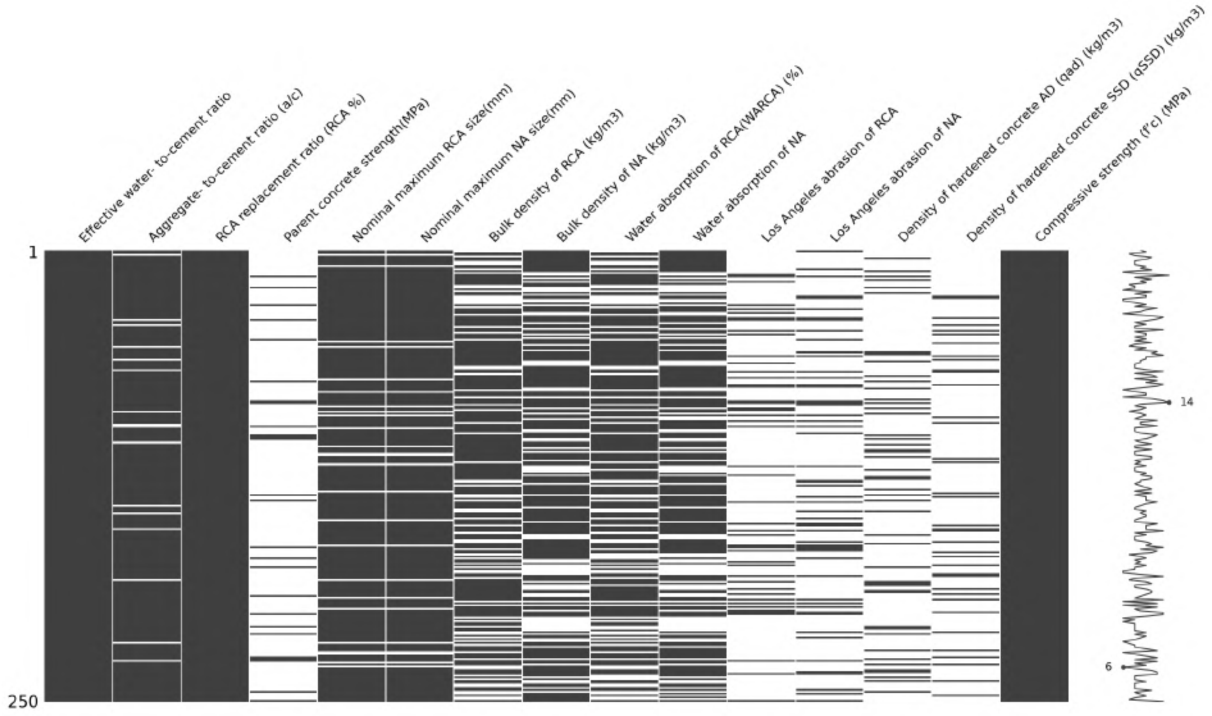
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 548 entries, 0 to 547
Data columns (total 15 columns):
 #   Column                                          Non-Null Count  Dtype  
---  ------                                          --------------  -----  
 0   Effective water- to-cement ratio                522 non-null    float64
 1   Aggregate- to-cement ratio (a/c)                485 non-null    float64
 2   RCA replacement ratio (RCA %)                   522 non-null    float64
 3   Parent concrete strength(MPa)                   56 non-null     float64
 4   Nominal maximum RCA size(mm)                    468 non-null    float64
 5   Nominal maximum NA size(mm)                     468 non-null    float64
 6   Bulk density of RCA (kg/m3)                     373 non-null    float64
 7   Bulk density of NA (kg/m3)                      332 non-null    float64
 8   Water absorption of RCA(WARCA) (%)              361 non-null    float64
 9   Water absorption of NA                          323 non-null    float64
 10  Los Angeles abrasion of RCA                     96 non-null     float64
 11  Los Angeles abrasion of NA                      87 non-null     float64
 12  Density of hardened concrete AD (qad) (kg/m3)   93 non-null     float64
 13  Density of hardened concrete SSD (qSSD) (kg/m3) 71 non-null     float64
 14  Compressive strength (f’c) (MPa)                522 non-null    float64
dtypes: float64(15)
memory usage: 64.3 KB
```

Note: Description for each of the feature is documented in the Jupyter notebook

Data Exploration and Preprocessing

During the data exploration phase, several techniques were employed to gain insights and understand the provided dataset. The following data exploration techniques were used:

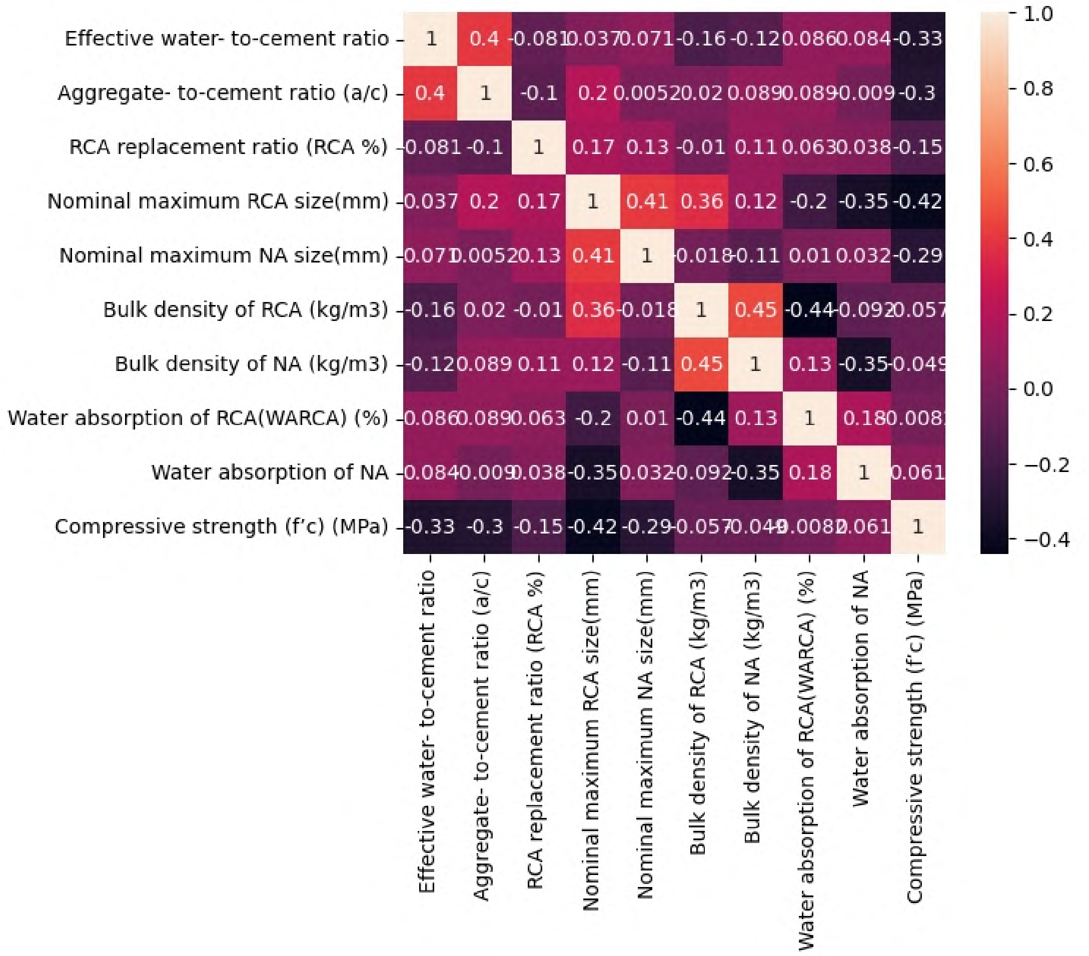
1. **Exploring the Data:** Data exploration is crucial in understanding the dataset, as it allows us to gain insights into the characteristics and patterns within the data.
2. **Visualizing Missing Values:** A sample of 250 data points revealed any gaps in the dataset caused by missing values.



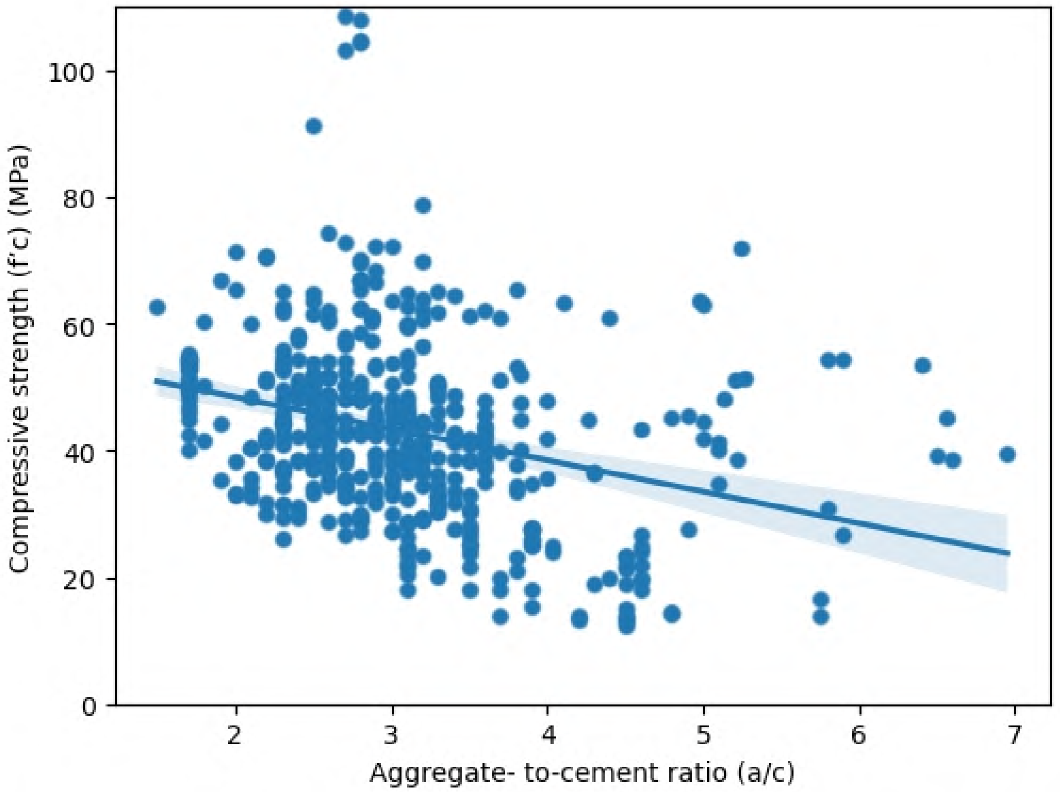
Missing data

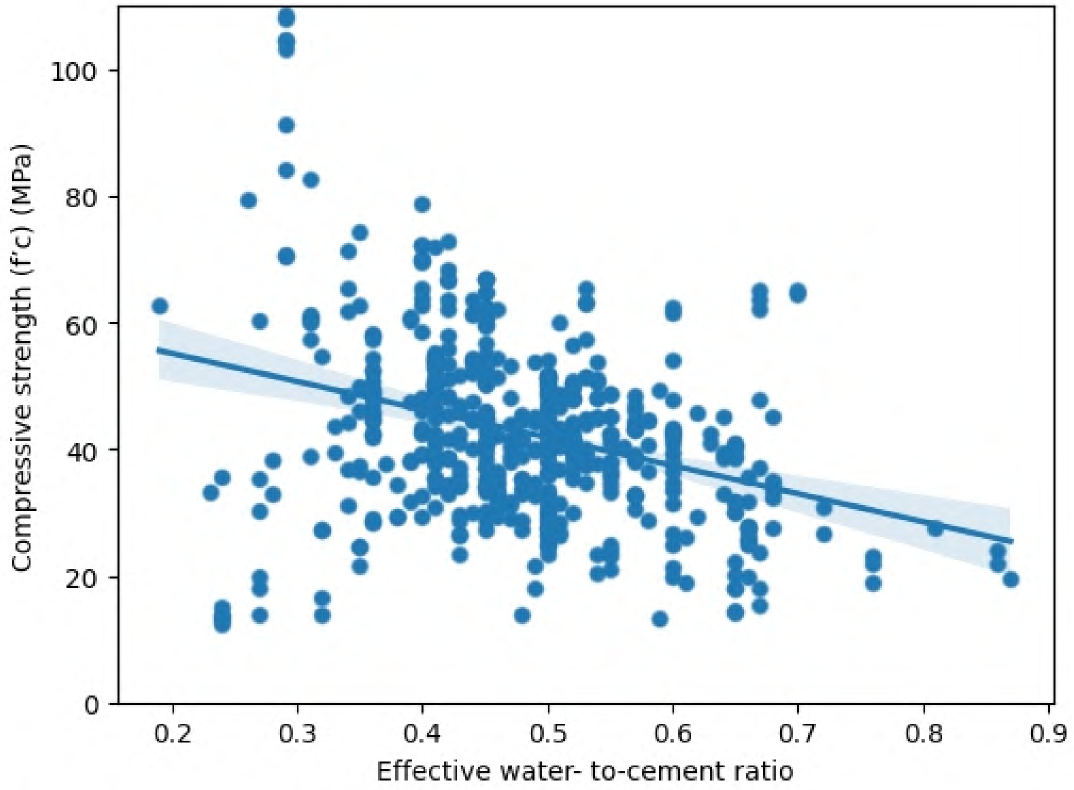
Data Exploration and Preprocessing

3. **Correlation Matrix:** Correlation matrix helped identify strong correlations among variables for feature selection.
4. **Relationship between Compressive Strength and Variables:** Scatter plots examined the relationship between compressive strength and relevant variables, aiding feature selection and model development.



	Effective water- to-cement ratio	Aggregate- to-cement ratio (a/c)	RCA replacement ratio (RCA %)	Nominal maximum RCA size(mm)	Nominal maximum NA size(mm)	Bulk density of RCA (kg/m3)	Bulk density of NA (kg/m3)	Water absorption of RCA(WARCA) (%)	Water absorption of NA	Compressive strength (f'c) (MPa)
Effective water- to-cement ratio	1	0.4	-0.081	0.037	0.071	-0.16	-0.12	0.086	0.084	-0.33
Aggregate- to-cement ratio (a/c)	0.4	1	-0.1	0.2	0.0052	0.02	0.089	0.089	0.009	-0.3
RCA replacement ratio (RCA %)	-0.081	-0.1	1	0.17	0.13	-0.01	0.11	0.063	0.038	-0.15
Nominal maximum RCA size(mm)	0.037	0.2	0.17	1	0.41	0.36	0.12	-0.2	-0.35	-0.42
Nominal maximum NA size(mm)	0.071	0.0052	0.13	0.41	1	-0.018	-0.11	0.01	0.032	-0.29
Bulk density of RCA (kg/m3)	-0.16	0.02	-0.01	0.36	-0.018	1	0.45	-0.44	-0.092	0.057
Bulk density of NA (kg/m3)	-0.12	0.089	0.11	0.12	-0.11	0.45	1	0.13	-0.35	-0.049
Water absorption of RCA(WARCA) (%)	0.086	0.089	0.063	-0.2	0.01	-0.44	0.13	1	0.18	0.0082
Water absorption of NA	0.084	0.009	0.038	-0.35	0.032	-0.092	-0.35	0.18	1	0.061
Compressive strength (f'c) (MPa)	-0.33	-0.3	-0.15	-0.42	-0.29	-0.057	-0.049	0.0082	0.061	1





Data Cleaning

During the data cleaning phase, several steps were taken to ensure the quality and reliability of the dataset. The following actions were performed:

1. **Dropping Irrelevant Features:**
   a. Water absorption of NA
   b. Bulk density of NA (kg/m3)
   c. Water absorption of RCA (WARCA) (%)
   d. Bulk density of RCA (kg/m3)

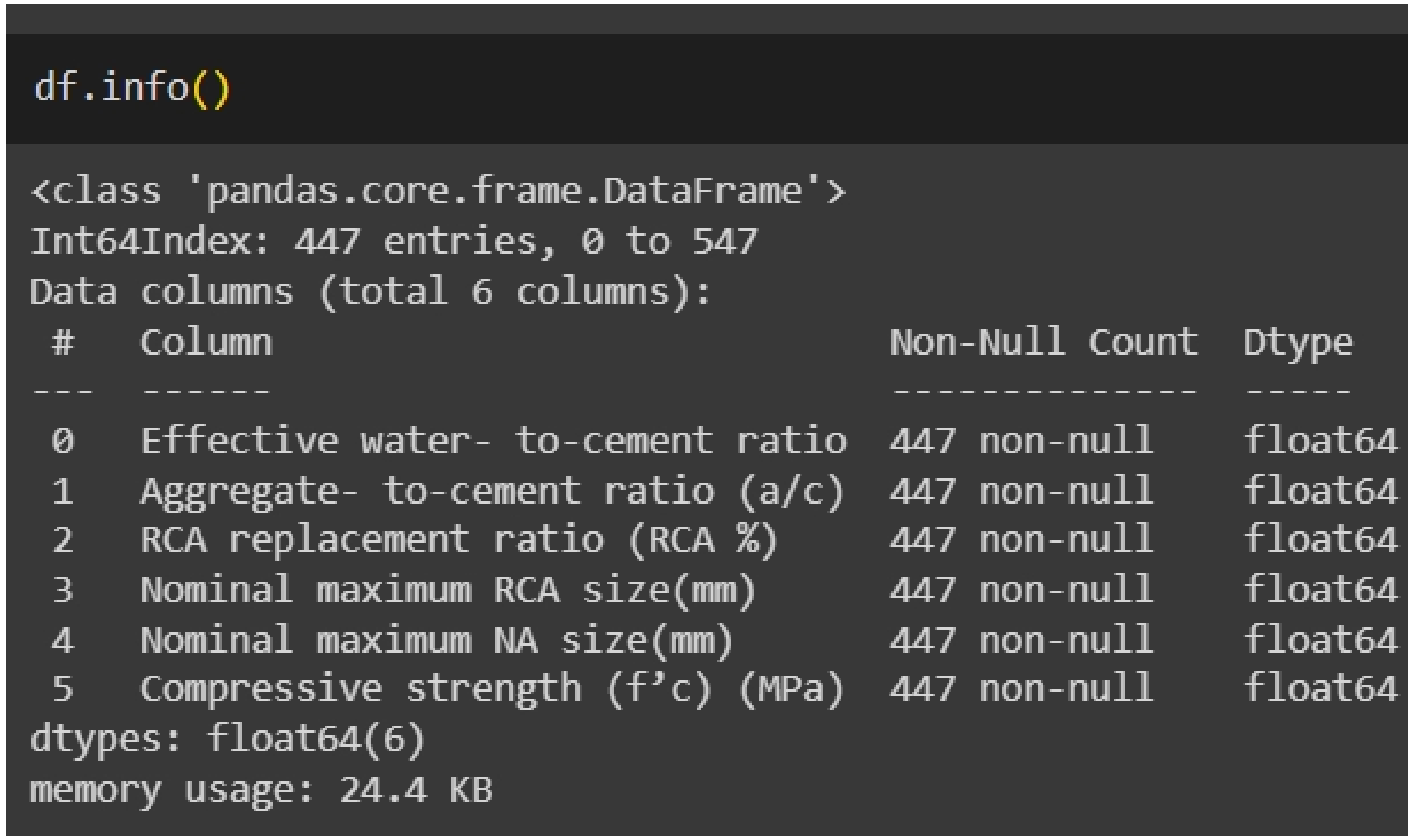
   These features had **more than 15% missing data** and were found to have weak correlation with the target variable, "Compressive strength." Hence, they were deemed irrelevant for the prediction task and were dropped from the dataset. guided by the dendrogram analysis.

Data Cleaning

2. **Handling Missing Values:**

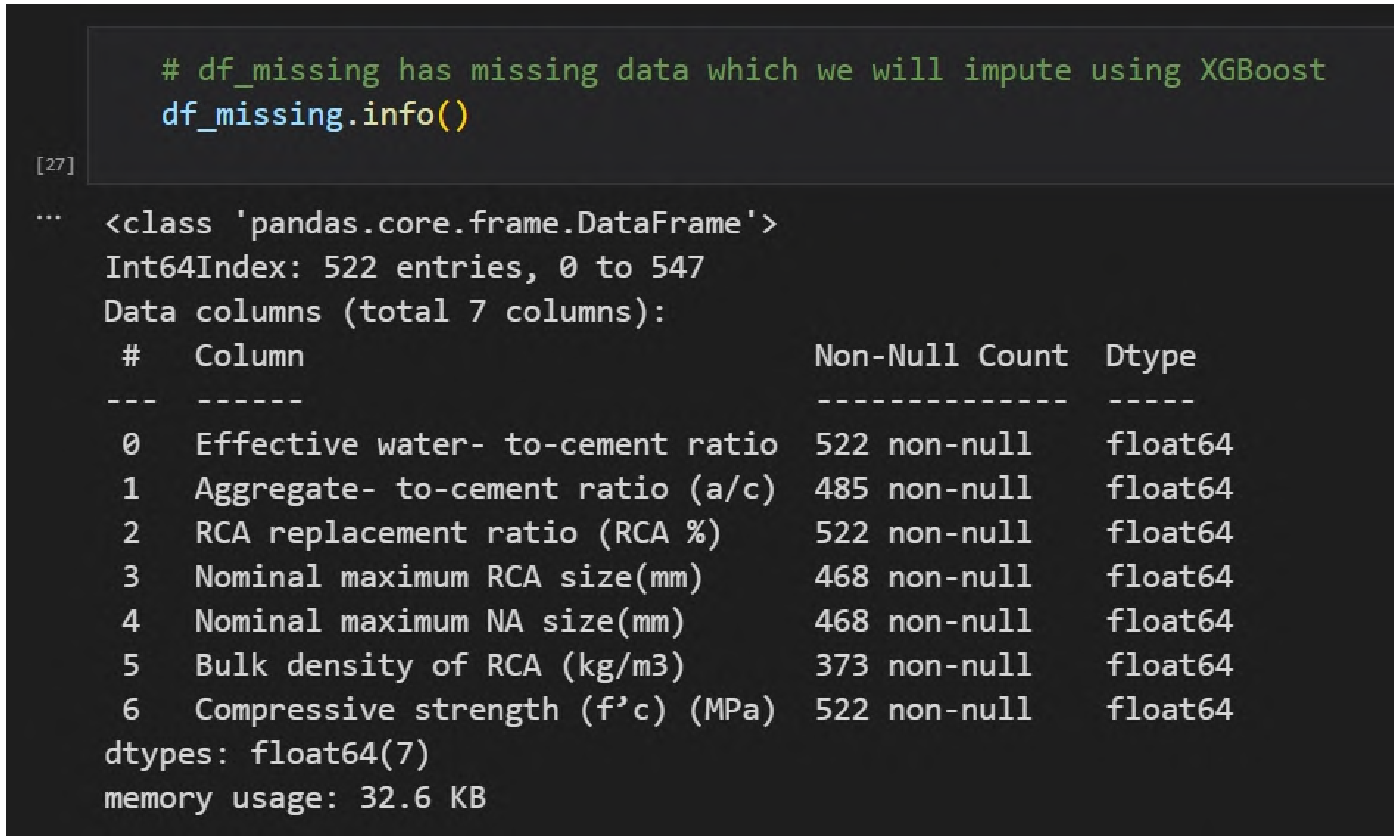
Additional rows containing null values were identified and subsequently dropped from the dataset. This ensured that the remaining data used for model training and testing was complete and free from missing values.

By the end of data cleaning, This is how the dataset looks like



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 447 entries, 0 to 547
Data columns (total 6 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Effective water- to-cement ratio  447 non-null    float64
 1   Aggregate- to-cement ratio (a/c)  447 non-null    float64
 2   RCA replacement ratio (RCA %)     447 non-null    float64
 3   Nominal maximum RCA size(mm)      447 non-null    float64
 4   Nominal maximum NA size(mm)       447 non-null    float64
 5   Compressive strength (f’c) (MPa)  447 non-null    float64
dtypes: float64(6)
memory usage: 24.4 KB
```



```
# df_missing has missing data which we will impute using XGBoost
df_missing.info()
```

```
[27]
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 522 entries, 0 to 547
Data columns (total 7 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Effective water- to-cement ratio  522 non-null    float64
 1   Aggregate- to-cement ratio (a/c)  485 non-null    float64
 2   RCA replacement ratio (RCA %)     522 non-null    float64
 3   Nominal maximum RCA size(mm)      468 non-null    float64
 4   Nominal maximum NA size(mm)       468 non-null    float64
 5   Bulk density of RCA (kg/m3)       373 non-null    float64
 6   Compressive strength (f’c) (MPa)  522 non-null    float64
dtypes: float64(7)
memory usage: 32.6 KB
```

Machine Learning Models

In our project, I employed various machine learning models for prediction. The following models were used:

- **Linear Regression**
- **Random Forest Regressor**
- **Gradient Boosting Regressor**
- **XGBoost:** A powerful gradient boosting framework known for its efficiency and accuracy in handling complex data and achieving high predictive performance.

These models were utilized to predict the compressive strength of concrete based on the available features.

Note: XGBoost is used to impute missing values, thus giving better perfomance

Results and Model Performance

Training data

Model Name	Accuracy (R2)	RMSE
Linear Regression	0.31	9.62
Random Forest Regressor	0.31	9.62
Gradient Boosting Regressor	0.63	7.09
XGBoost	0.86	5.65

XGBoost performs the best with 86% accuracy

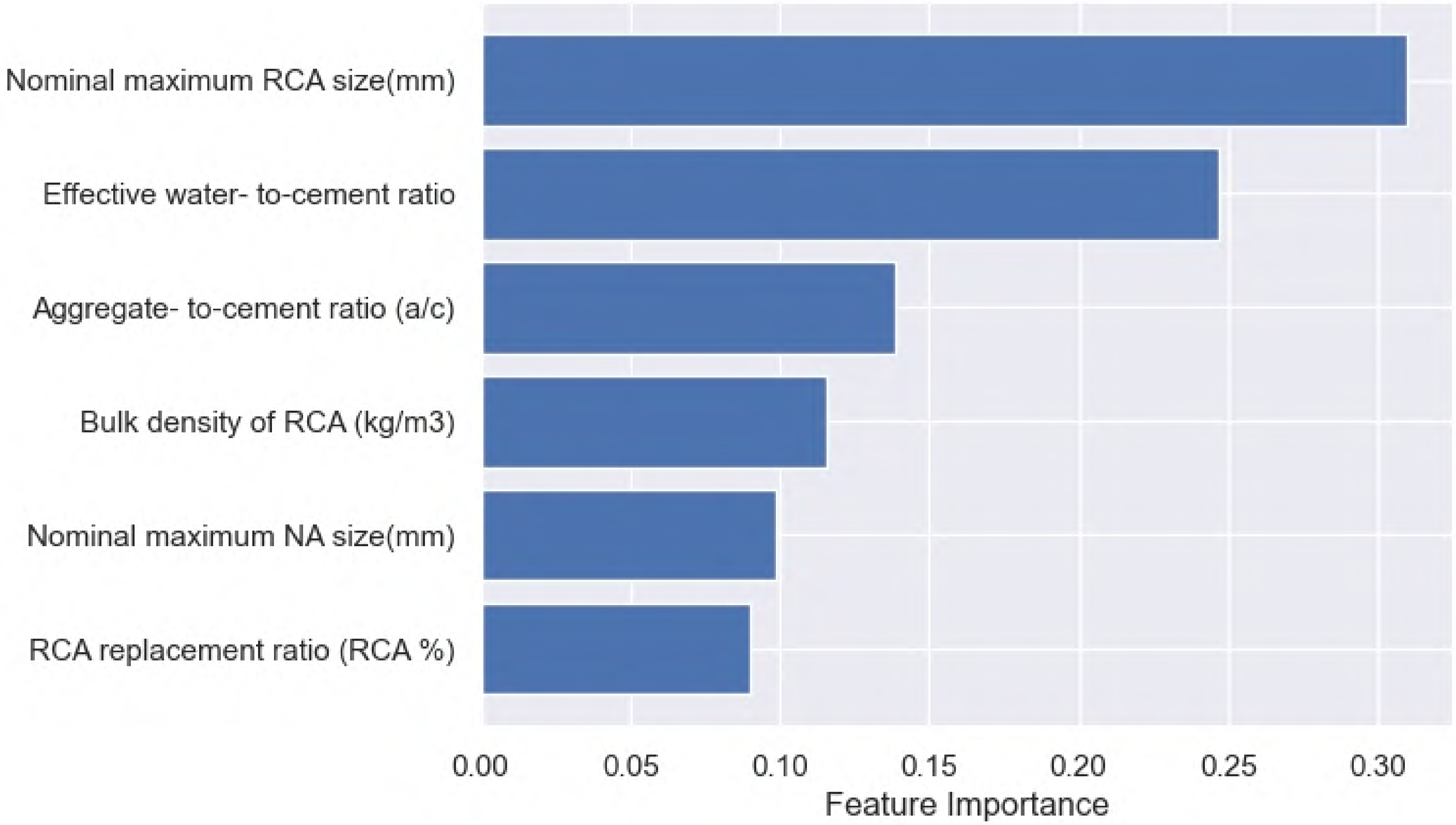
Results and Model Performance

Test data

Model Name	Accuracy (R2)	RMSE
Linear Regression	0.42	11.66
Random Forest Regressor	0.69	8.51
Gradient Boosting Regressor	0.66	893
XGBoost	0.73	7.97

XGBoost performs the best with 73% accuracy

Feature Importance



The most influential features on compressive strength are **Nominal maximum RCA size** and **Effective water-to-cement ratio**, followed by **Aggregate-to-cement ratio** and **Bulk density of RCA**.

Conclusion

- **Predicted compressive strength** of concrete with recycled aggregates using machine learning.
- Achieved **accurate predictions** with regression models (Random Forest, Gradient Boosting, XGBoost).
- Contributed to sustainable construction practices by reducing waste and minimizing costs.

The project makes a positive impact on the construction industry by promoting **sustainability** and **cost-effective** solutions.

Future Scope of the project

- **Handle missing values:** Implement data imputation techniques for missing values in relevant features.
- **Remove outliers:** Identify and address outliers to improve model performance.
- **Explore deep learning:** Investigate the use of deep learning models like Neural Networks for enhanced predictions.

These future steps will enhance the project's predictive capabilities, address missing data, and explore advanced modeling techniques for improved accuracy.

Thank you!