02450: INTRODUCTION TO MACHINE LEARNING AND DATA MINING

Project 1

	Section 1	Section 2	Section 3	Section 4	Exam Questions
Hugh Fulton s236798	20%	60%	35%	33.33%	33.33%
Dylan Myers s237017	30%	20%	35%	33.33%	33.33%
Simon Ritter s237138	50%	20%	30%	33.33%	33.33%

Technical University of Denmark

29.02.2024

Table of Contents

List of Figures

List of Tables

1 Description of the Data Set

Concrete is building material that is used all over the world. It consists of a mixture of a binder (usually cement), water, aggregates (ranging in size from sand to stones) and admixtures. The composition of a concrete sample determines its material properties. The most important one is its strength, which is of high interest when a new concrete structure is designed. The composition also affects the workability when the concrete is poured on the construction site, for which the requirements vary for each given situation.

Cement production generates a high amount of carbon dioxide emissions, which in the past has led to a lot of discussion, trying to either replace concrete elements completely with different and more sustainable materials, or to make concrete itself more environmentally friendly. This can be achieved mainly by reducing the amount of cement of a concrete element. Waste materials from industry, such as fly ash or burnt furnace slag provide a suitable alternative, although only up to a limited amount. There is also a limit to the amount of aggregate that can be used, as the cement is the element that binds the heterogeneous mixture together.

This report examines the "Concrete Compressive Strength" data set[1]. It has eight input attributes and one output attribute, which is the concrete compressive strength. A detailed description of the dataset can be found in the following chapter.

The entire dataset is a collection of the results of 17 different experiments from various places and scientists, that were gathered for the purpose of the above study. For this reason, the conditions, under which the experiments were conducted, are not known entirely. For example, there are various uncertainties related to the class of fly ash or the type of superplasticizers that was used. The dataset was, however, put together in such a way to guarantee a satisfactory amount of accuracy, by keeping all the known circumstances similar. They did this by removing observation with comparably lager size aggregates and observations with special curing conditions from the dataset.

The original paper[2] examines the material behaviour of high-performance concrete (HPC), which differs from normal concrete in its high amount of cement. HPC is a complex material for which it is difficult to predict its mechanical properties. The goal of the paper was to set up an artificial neural network to predict the concrete compressive strength from the eight input attributes. The study led to the following two conclusions *(citing):*

1. "The strength model based on the artificial neural network is more accurate than the model based on regression analysis."
2. "The compressive strength can be calculated using the models built with this methodology. It is convenient and easy to use these models for numerical experiments to review the effects of each variable on the mix proportions. For example, the strength model can be used to study the strength effects of age or water-to-binder ratio."

The lecture "02450 Introduction to Machine Learning and Data Mining" discusses artificial neural networks in week 8. Up until now, the focus lay on machine learning techniques, such as regression and classification. The two conclusions will therefore not be discussed any further. However, the idea of predicting the concrete compressive strength for any arbitrary concrete composition with regression is of high interest.

More specifically, if values of a to the model before unseen concrete sample are provided for some or all the eight attributes, how well does the model perform? Can a model be created to deliver a plausible estimate of the compressive strength, especially if the not all eight attributes are provided as an input? These questions shall be addressed in a regression task.

---

[1] Yeh,I-Cheng. (2007). Concrete Compressive Strength. UCI Machine Learning Repository. https://doi.org/10.24432/C5PK67.

[2] I.-C. Yeh, Modeling of strength of high-performance concrete using artificial neural networks, Cement and Concrete Research, Volume 28, Issue 12, 1998, Pages 1797-1808, ISSN 0008-8846, https://doi.org/10.1016/S0008-8846(98)00165-3.
(https://www.sciencedirect.com/science/article/pii/S0008884698001653)

The dataset does not provide any nominal attributes, which are needed for a classification task. However, it is always possible to create a nominal attribute from other, continuous attributes. For that purpose, let us define a categorization for the age of concrete:

- "fresh" for age < 27.99 days
- "mid-age" concrete for 27.99 days < age < 99.99 days
- "mature" concrete for age > 99.99 days

In the next report, the classification task will be about predicting one of those three age categories based on the other attributes, including the strength, which is greatly influenced by the age of the concrete. It needs to be investigated, which of the attributes are the most helpful in doing so. The starting assumption is the following: The content of cement and the strength, as well as the ratio of the first to the latter, might be used to predict the age category.

2 Detailed Explanation of the Attributes of the data

2.1 Description of Data Attributes

The selected dataset includes 9 features, 8 input, and 1 output. The ultimate output value is the compressive strength of concrete.

- **Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate (Continuous and Ratio):** These attributes quantify the components of concrete mixtures in kilograms per cubic meter or similar units, showcasing continuous data that falls on a ratio scale. They have a true zero point and allow for meaningful comparisons through division, which is characteristic of ratio data.
- **Age (Continuous and Ratio):** This attribute, representing the curing time of concrete in days, is also continuous and on a ratio scale.
- **Concrete Compressive Strength (Continuous and Ratio):** The target variable measures the concrete's strength in megapascals (MPa).

This analysis sets a strong foundation for exploring the relationships between concrete components and its compressive strength, applying machine learning models for prediction, and investigating the environmental impact of different mix compositions. The transformation of the 'Age' attribute into nominal categories will enable the identification of distinct patterns across different stages of concrete maturation, which could be crucial for both quality assurance and environmental studies.

2.2 Missing Values

There are no missing values in the dataset. Each attribute has values for all entries, indicating a complete dataset without any immediate signs of missing or null entries.

2.3 Summary Statistics and Potential Data Issues:

A table of summary statistics is included in the appendix 7.2 Summary Statistics.

- The summary statistics provide an overview of each attribute, including count, mean, standard deviation, minimum, 25th percentile, median (50th percentile), 75th percentile, and maximum values.
- All attributes show a wide range of values, appropriate to their nature. For example, Cement component ranges from 102 to 540 kg/m³, and compressive strength varies from 2.33 to 82.6 MPa, indicating diverse mix compositions and strengths.
- The Age attribute shows a broad range, from 1 to 365 days, which is expected given the varying curing times for different concrete mixes. This is prior to the manipulated from the original dataset.

- There are no immediately apparent signs of corrupted data. The continuous nature of the variables and their expected ranges suggest the data is in a usable state for analysis.

2.4 Considerations for Data Analysis

- The continuous and ratio scale of the attributes allows for various forms of statistical analysis and modelling.
- The absence of missing values simplifies the initial data preparation phase, allowing for direct application of machine learning techniques.
- The variability in attributes, especially in components like Fly Ash, Slag, and Superplasticizer, might require normalization or standardization before applying machine learning models to account for different scales and distributions.
- Given the wide age range from 1 to 365 days, the 'Age' attribute provides a valuable dimension of analysis. However, its continuous nature could obscure meaningful patterns within different age categories that affect concrete properties. To address this, it is necessary to convert the 'Age' feature into a nominal data feature, categorizing it into discrete bins such as 'fresh', 'mid-age', and 'mature'. This conversion facilitates the use of age as a categorical variable in machine learning models, allowing for more nuanced analysis and interpretation of its impact on concrete compressive strength. It is especially important when considering the practical aspects of concrete use and the potential for age to interact with other features in complex ways.

This analysis sets a strong foundation for exploring the relationships between concrete components and its compressive strength, applying machine learning models for prediction, and investigating the environmental impact of different mix compositions.

3 Data Visualization

3.1 Primary Visualizations of the Dataset

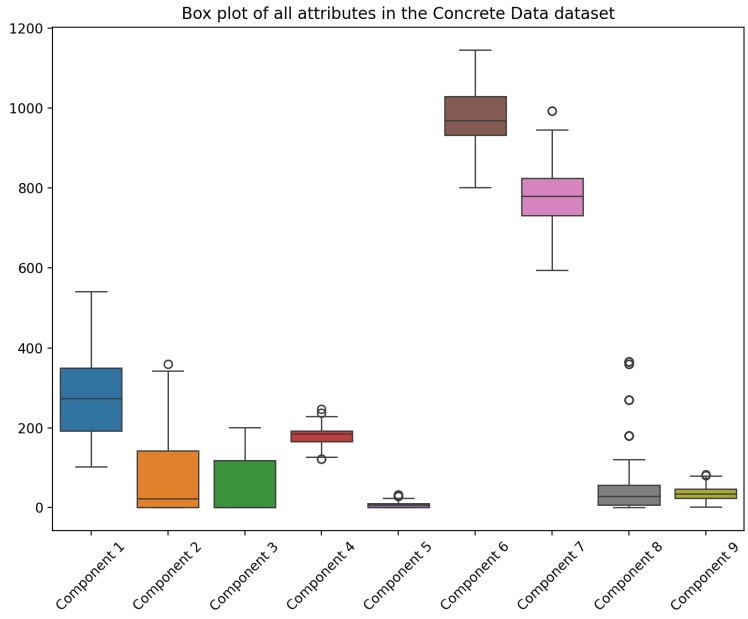


Figure 1 - Box Plot of data features

There are no problems with outliers in the data set, although the box plot shows some extreme values, marked with circles outside of the 25th – 75th percent quantile for components 2 (slag), 4 (water), 5 (superplasticizer), 7 (fine

aggregate), 8 (age) and 9 (strength). These data points, however, cannot be described as outliers, as their values were selected deliberately for the purpose of the various experiments. Especially for component 8 (age), the high values are important observations. For that, a concrete sample had to be created and stored for about 1 year before testing. As this is a lot more expensive than creating a sample and testing it 3 days later, it is obvious that there are fewer old samples available in the dataset. The small number of mature samples compared to the large number of old samples is the reason why the box plot describes these points as potential outliers.

Figure 2 shows the histogram of every attribute, with an appropriately chosen number of 20 bins. A normal distribution curve with the computed mean and empirical standard deviation (according to Table 2) is plotted on top of the histogram.

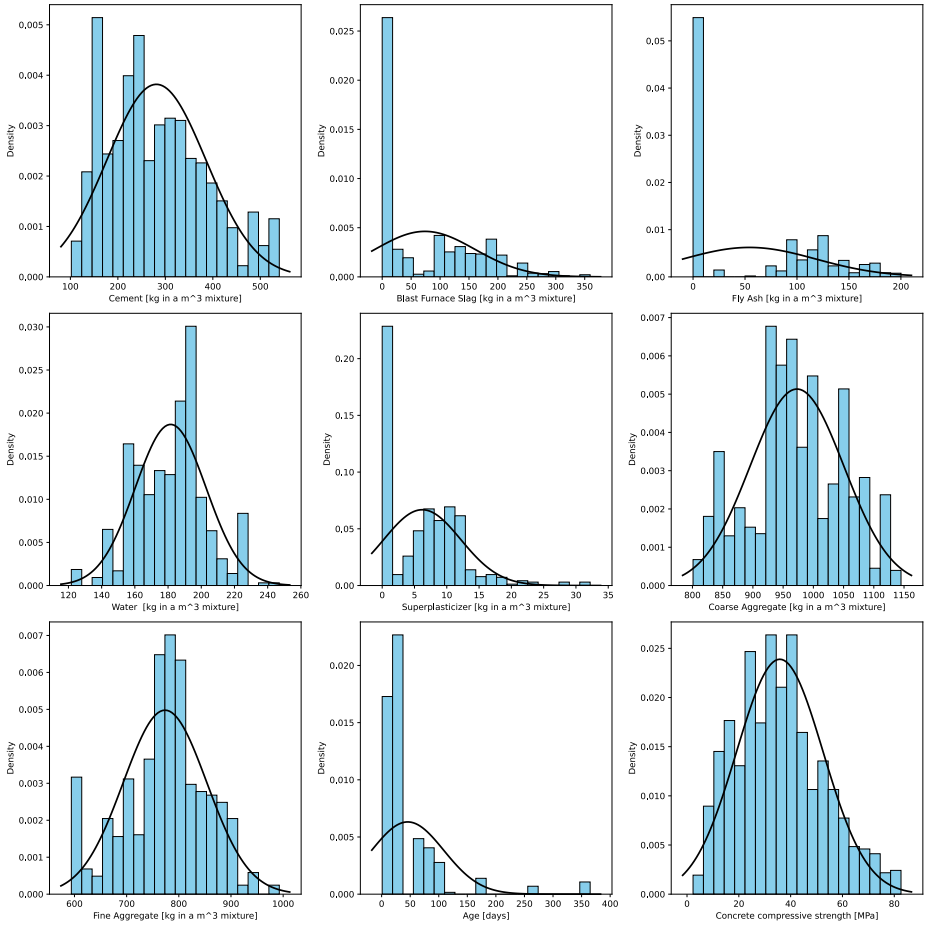


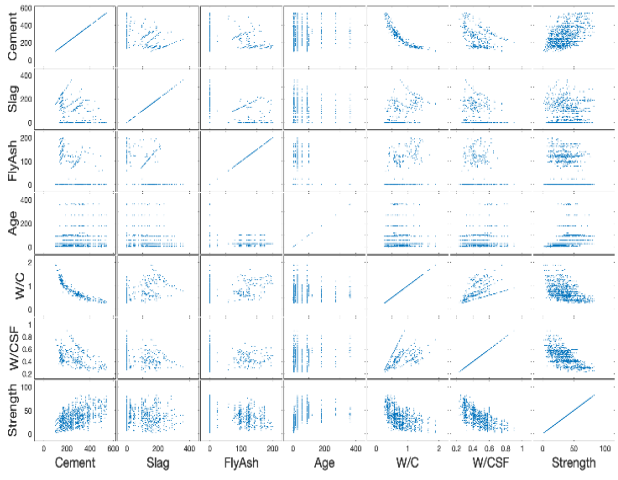
Figure 2 – histograms of every data feature (blue) with fitted normal distribution curves (black)

The illustration of the observations with histogram shows in most of the cases a very similar result as in Figure 1, but not in all of them. The histograms reveal more information about where the data lies. For example, the amount of fly ash in a sample is zero in around 50% of all observations, while approximately the other 50% of observations show a content of fly ash in the range between 70 – 200 kg/m^3. This information is not visible in the box plot illustration, which smears the real situation by taking averages over the two separate regions, where the data effectively lies.

From a visual inspection, none of the input features seem to be normal distributed, as the bell curves does not fit the histogram well enough. This makes sense, as in each original experiment the values of the features were selected with the goal of examining the effect of certain attributes on the concrete strength. The selection was therefore neither subject to random selection, nor to coincidence.

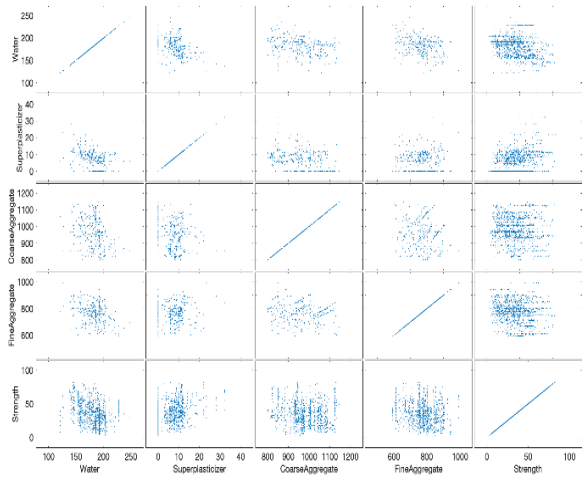
Furthermore, the normal distribution curves reach into the negative range on the x-axis, which is physically not possible. For example, it is not possible to have a negative amount of slag, fly ash or superplasticizer. Neither is a negative age (< 0 days) possible. This underlines the statement that those input features do not follow a normal distribution.

For the output feature "concrete compressive strength", the normal distribution curve fits the histogram best, compared to all the other features, which is an observation worth mentioning. This is not surprising, as the strength is the result of all the input features, for which a lot of different combinations are available with the thousand observations. The fit of the normal distribution to the histogram is, however, not at all perfect but rather vague and does not really seem helpful for task of predicting the strength.



1	-0.2752	-0.3975	0.0819	-0.8791	-0.4750	0.4978
-0.2752	1	-0.3236	-0.0442	0.3573	-0.2791	0.1348
-0.3975	-0.3236	1	-0.1544	0.2460	-0.1358	-0.1058
0.0819	-0.0442	-0.1544	1	-0.0293	0.1499	0.3289
-0.8791	0.3573	0.2460	-0.0293	1	0.5128	-0.5007
-0.4750	-0.2791	-0.1358	0.1499	0.5128	1	-0.6231
0.4978	0.1348	-0.1058	0.3289	-0.5007	-0.6231	1

Figure 3 - scatter matrix, 7x7. W/C is the ratio of Water/Concrete and W/CSF is the ratio of Water/ (Cement + Slag + Fly Ash). The corresponding correlation matrix is shown.



1	-0.6575	-0.1823	-0.4506	-0.2896
-0.6575	1	-0.2663	0.2225	0.3661
-0.1823	-0.2663	1	-0.1785	-0.1649
-0.4506	0.2225	-0.1785	1	-0.1672
-0.2896	0.3661	-0.1649	-0.1672	1

Figure 4 - scatter matrix, 5x5. The corresponding correlation matrix is shown.

Fehler! Verweisquelle konnte nicht gefunden werden. is a 7x7 Scatter plot matrix with two attributes created for data analysis – W/C and W/CSF. W/C is the ratio of water to cement, and W/CSF is the ratio of Water to Cement + Slag + Fly Ash. Both ratios are a measurement that we created. It is widely known that these ratios show a strong correlation to strength.

Fehler! Verweisquelle konnte nicht gefunden werden. shows the strength of concrete is positively corelated to the amount of cement used. Furthermore, the plots of W/C and W/CSF show negative correlation, meaning dryer concrete is stronger. It is also worth noting that age and strength have some positive correlation (0.3289), though there is a point of marginal returns, where further aging does not yield stronger concrete. There is also some variation within each age bin, however this most likely due to the many different combinations that are used (and measured) at each age, these deviations should not take away from the relationship between increased age and increased strength.

Fehler! Verweisquelle konnte nicht gefunden werden. is a 5x5 scatter matrix of the remaining attributes, containing three main relationships. Firstly, there is a negative correlation of 0.6575 between water and superplasticizer, as well as a negative correlation of 0.4506 between water and fine aggregate, suggesting that the ratios are inversely related to each other when forming the recipe for concrete examined in the study. Furthermore,

there is a slightly positive correlation of 0.3661 between strength and superplasticizer, suggesting some increased strength with more superplasticizer[3].

In the appendix you can find a 9x9 scatter plot of all the attributes from the data set plotted against each other, for further rigour.

The primary machine learning modelling aim of using regression for strength prediction seems feasible, as there are attributes within the dataset that directly appear correlated. From the correlation between the age and the strength, the classification task appears to be feasible as well. This feasibility of it will be discussed further in section 4.

3.2 Principal Component Analysis

To understand the variation explained by PCA components, we conducted an analysis that showcases how the cumulative variance in the data changes as we increase the number of principal components. This helps to determining the minimum number of components required to capture the essence of the dataset without significant information loss. Figure 5 plots the principal components against the percentage of variance explained. The plot typically exhibits a point at which the level of variance significantly flattens which may offer a point at which to decide how many principal components should be analysed. For our analysis we are happy to include all Principal Components in our analysis.

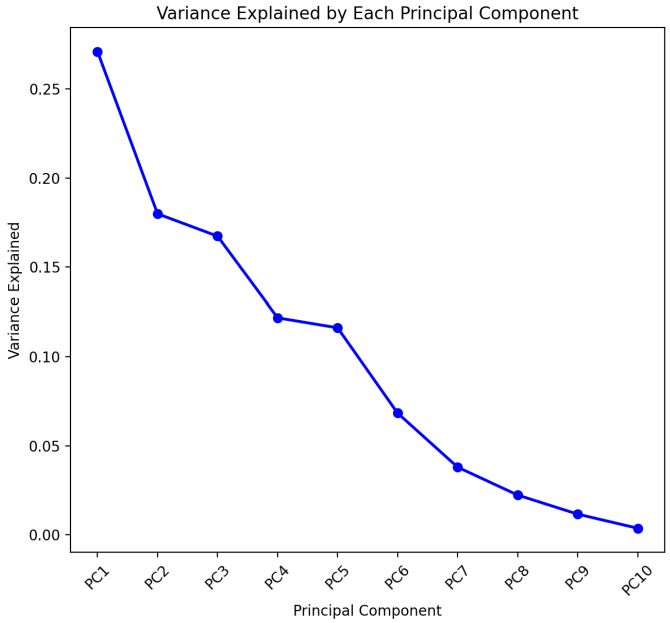


Figure 5- Variance explained by each principal component (eigenvalues)

Principal Directions of PCA Components

Table 1 - V1, and V2 values

Feature Number	Feature Description	V1 (PC1)	V2 (PC2)
1	Cement (component 1) (kg in a m^3 mixture)	-0.08592707	0.08494342
2	Blast Furnace Slag (component 2) (kg in a m^3 mixture)	-0.21609323	-0.64359181

[3] For a curious reader, there is no correlation between cement and superplasticizer (Section 7.1 b).

3	Fly Ash (component 3) (kg in a m^3 mixture)	0.4141312	0.11190844
4	Water (component 4) (kg in a m^3 mixture)	-0.56080941	-0.03816912
5	Superplasticizer (component 5) (kg in a m^3 mixture)	0.5286937	-0.32462798
6	Coarse Aggregate (component 6) (kg in a m^3 mixture)	-0.043138	0.65399826
7	Fine Aggregate (component 7) (kg in a m^3 mixture)	0.42128533	0.01228277
8	Age Category: Fresh (Age < 28 days)	0.00566866	0.06455277
9	Age Category: Mid-age (28 days ≤ Age < 100 days)	0.02935853	-0.15783181
10	Age Category: Mature (Age ≥ 100 days)	-0.0124539	0.04762321

First Principal Component (PC1)

The first principal component (V1) demonstrates the highest weight in absolute terms for the Water (component 4) with a negative coefficient, suggesting that water content inversely influences this component the most. The Superplasticizer (component 5) and Fine Aggregate (component 7) follow with positive weights, indicating their significant contributions to the variance captured by PC1. The negative weight for Cement (component 1) and Blast Furnace Slag (component 2) implies an inverse relationship with PC1.

Second Principal Component (PC2)

The second principal component (V2) is most heavily influenced by the Coarse Aggregate (component 6) with a large positive weight, indicating its strong contribution to the variance along PC2. In contrast, Blast Furnace Slag (component 2) has a significant negative weight, showing an opposite effect on PC2. Other features like Fly Ash (component 3) and the age categories show smaller contributions.

Implications

These principal components underline the complex interplay between the concrete's components and its curing time in relation to the variance observed in the dataset. PC1 and PC2, with their distinct sets of feature weights, highlight different aspects of the data's variability. This nuanced understanding facilitates more focused analyses, such as feature selection for predictive modelling or identifying patterns relevant to the quality and durability of concrete.

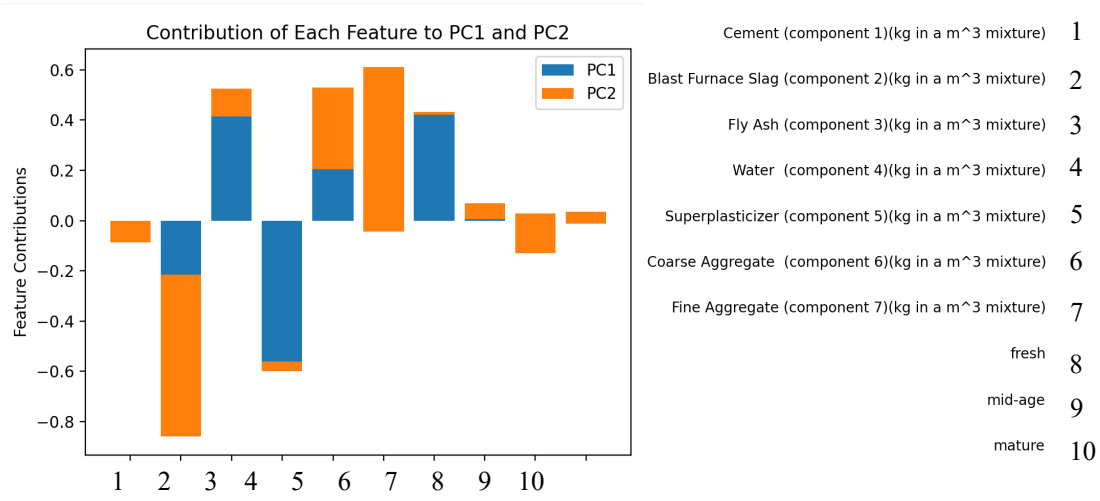


Figure 6 - Contribution of each feature to PC1 and PC2

For a more intuitive understanding of the principal directions in a 3D space, we can visualize the first three principal components using a 3D scatter plot shown in Figure 7. This visualization not only demonstrates the separation of data points in the space defined by these components but also provides insights into the dataset's inherent structure. The axes in this 3D representation correspond to the principal directions, with each axis representing a combination of the original features that maximize variance.

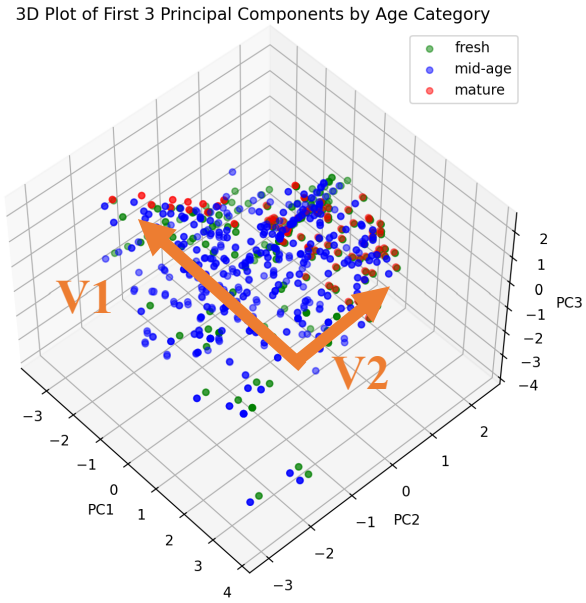


Figure 7 - 3D plot of first 3 principal components.

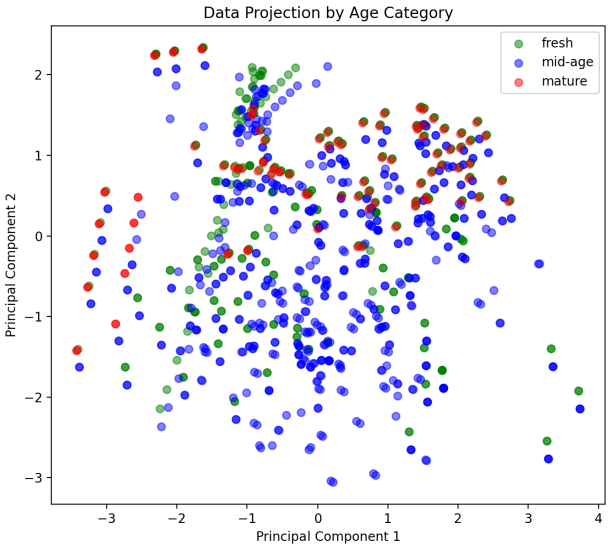


Figure 8 - Data projection of PC1 and PC2 separated by high and low compressive strength.

Data Projected onto Principal Components

Before performing PCA, as the data contained varying scales, we needed to standardize the data for analysis. This involved scaling each feature to have a mean of zero and a standard deviation of one. The standardized data was then projected onto the selected principal components. This projection transforms the data into a new space defined by the principal directions, facilitating the identification of patterns and relationships that were not apparent in the original space. The resulting projection can be visualized through a scatter plot of the data points in the space of the first two principal components, offering insights into the clustering and separation of the data.

Fehler! Verweisquelle konnte nicht gefunden werden. shows a scatter plot of the dataset projected onto the first two principal components (PC1 and PC2). The data points are color-coded according to the age of the concrete, with 'fresh' concrete shown in green, 'mid-age' in blue, and 'mature' in red. From this graph, we can determine:

1. **Age Category Clustering**: There is some visible clustering based on the age of the concrete. 'Fresh' concrete samples (in green) appear to be distributed differently from 'mid-age' (in blue) and 'mature' (in red) samples. This may indicate that the age of the concrete has a discernible relationship with the variance captured by the first two principal components.
2. **Principal Component 1 (PC1) Influence**: PC1, which represents the direction of the greatest variance, appears to show a spread of points that could be related to the concrete age categories. This suggests that certain material compositions or processes associated with different ages might influence the variability along PC1.
3. **Principal Component 2 (PC2) Influence**: The distribution of points along PC2 also seems to be influenced by the age categories, although the separation is not as distinct as with PC1. This indicates that the second principal component captures additional aspects of variance that could be related to concrete age, but perhaps intertwined with other features.
4. **Correlation with Concrete Age**: While the graph does not directly measure compressive strength, the distribution of colors might infer how the age of concrete could be related to other properties, possibly including strength. For example, if older concrete ('mature') is known to be stronger due to continued curing, we might expect to see those samples positioned differently in the plot compared to 'fresh' samples.
5. **Overlap and Diversity**: Despite some clustering, there is considerable overlap among the age categories, suggesting that age alone does not entirely explain the variance within the dataset. This overlap indicates that the differentiation in the properties of concrete is influenced by a complex interaction of its components and age, not solely by the age.

From this visualization, it is evident that PCA captures certain trends related to the age of concrete, but the overlap suggests that a combination of various factors contributes to the overall properties of the concrete. The analysis provides a foundation for more detailed exploration, potentially guiding further investigations into how concrete composition and age interplay to affect its characteristics.

4 Discussion

4.1 Summary

The dataset has a high number of observations of high quality. There are neither missing values nor outliers. None of the attributes are normal distributed, as the composition of each concrete sample was deliberately selected.

For the attributes “fly ash”, “burnt furnace slag” and “superplasticizer”, the histograms show two peaks, one of them lying at zero. This means that for some of the samples, these components were zero and therefore not contained.

The correlation analysis led us to two main takeaways. The first takeaway has to do with the correlation between strength and components in the concrete. The strength of the concrete will increase if it: 1) contains more cement 2) is drier (i.e, less water) 3) is aged for longer, however there is marginal returns after a certain amount and 4) contains more superplasticizer. It was also observed that the amount of water is inversely related to both the fine aggregate as well as superplasticizer concentrations in the concrete “recipes” used in this study.

PCA analysis on the concrete dataset uncovers that its variance is largely captured by the initial principal components, highlighting underlying patterns amidst its complexity. The first component reveals a negative correlation between water content and strength, suggesting drier mixtures are stronger, while emphasizing the roles of superplasticizer and fine aggregates. The second component points to coarse aggregate's significant impact on strength. This analysis indicates that concrete strength is influenced by a complex interplay of components, beyond simple univariate analysis, underscoring the importance of their interactions for accurate modelling.

These findings underscore the multidimensional nature of concrete strength determinants, indicating that simple univariate analyses may not fully capture the nuances of what affects concrete's compressive strength. The PCA also suggests that while individual components like cement content, water, superplasticizer, and aggregates play critical roles, their interactions and the ratios between them (such as water to cement ratio) are fundamental in determining concrete strength.

4.2 Feasibility of the machine learning aim

As mentioned above, using regression for the prediction of the concrete compressive strength appears to be a feasible task. The cement content is a good example for an attribute, that directly relates to the strength. However, other attributes such as age and the attribute ratios W/C and W/CSF show a good correlation to strength. These relationships shall be exploited in the next report.

PCA analysis on the concrete dataset reveals age significantly influences property variance, essential for classification tasks. However, overlaps in age categories present challenges, suggesting complexity beyond age alone affects concrete strength. This complexity requires models to consider multiple attributes and their interactions. The PCA underscores the need for sophisticated models to accurately classify concrete, highlighting the intricate relationships between components and their collective impact on properties. These insights are crucial for developing models that effectively predict concrete strength and categorize it by age.

5 Exam Questions

5.1 Question 1: Spring 2019 question 1

~~A~~: wrong, as "time of day" can be ordered and therefore is not nominal
~~B~~: wrong, x_4 is not nominal, as the "number of immobile buses" is ratio. 0 means no immobile bus, and 10 immobile buses are twice as much as 5 immobile busses.
~~C~~: wrong, as each step in x_1 "time of day" has a physical meaning of being 30' later and therefore is not ordinal
D: **Correct**. x_1 "time of day" is indeed interval and not ratio, as $x_1 = 0$ does not have a physical meaning (and on top of that $x_1 = 1$ is the lowest possible value)

5.2 Question 2: Spring 2019 question 2

Observation: only the first and the third component of the vector has a delta that is not 0. $x_1 - y_1 = 7$ and $x_3 - y_3 = 2$.

~~C~~: Wrong. The distance is 2 + 7 = 9
~~B~~: Wrong. (7^3 + 2^3)^(1/3) = 7.05
~~D~~: Wrong. (7^4 + 2^4)^(1/4) = 7.01
A: **Correct**. With p -> inf, the distance will approach 7.0.

5.3 Question 3: Spring 2019 question 3

Using this formula and the sigma_i are the diagonal components of the **S**-matrix

$$\text{Variance Explained} = \frac{\sum_{i=1}^{n} \sigma_i^2}{\sum_{i=1}^{M} \sigma_i^2}.$$

A: **Correct**. Variance Explained = **0.87 > 0.8**
B: Wrong. Variance Explained = 0.48 < 0.51
C: Wrong. Variance Explained = 0.52 > 0.5
D: Wrong. Variance Explained = 0.72 > 0.7

5.4 Question 4: Spring 2019 question 4

An observation with a low value of Time of day, a high value of Broken Truck, a low value of Accident victim, and a low value of Defects will typically have a negative value of the projection onto principal component number 4.

by interpreting the direction and magnitude of the coefficients (loadings) for the fourth principal component. A negative projection implies that the observation moves in the opposite direction of the component's vector in the multidimensional space, which is consistent with the given attribute values and their expected effects on the component.

5.5 Question 5: Spring 2019 question 14

Observation: no stemming, no stopword filtering: there are 13 different words, only "the" and "words" are contained in both documents.
f11 = 2
f10 = 6

f01 = 5
f00 = 20’000 – 13 = 19’987

J = 1/(7+6+19’000) = 0.00001

~~A~~: wrong
~~B~~: wrong
C: **correct**
~~D~~: wrong

5.6 Question 6: Spring 2019 question 27

Answer: B - $p(\hat{x}_2 = 0 \mid y = 2) = 0.84$

Explanation: This probability is calculated by adding the probabilities of $\hat{x}_2$=0 for both $\hat{x}_7$=0 and $\hat{x}_7$=1 given *y*=2, which are 0.81 and 0.03, respectively, leading to a total probability of 0.84. This indicates the likelihood of observing $\hat{x}_2$=0 (no or one broken truck) given light congestion.

6 References

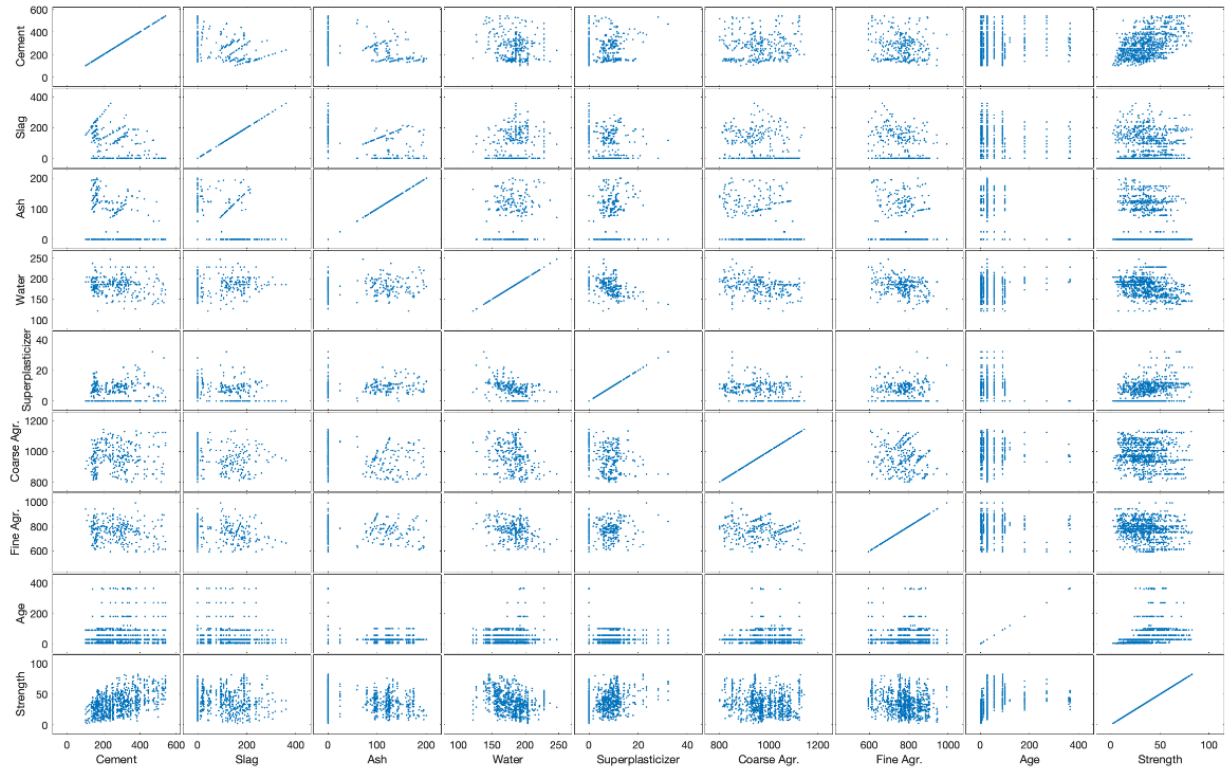
Dataset: I.-C. Yeh, "Concrete Compressive Strength," UCI Machine Learning Repository, 2007. [Online]. Available: https://doi.org/10.24432/C5PK67. [Accessed: 18-02-2024].

Original Paper: I.-C. Yeh, "Modeling of strength of high-performance concrete using artificial neural networks," Cement and Concrete Research, vol. 28, no. 12, pp. 1797-1808, 1998. [Online]. Available: https://doi.org/10.1016/S0008-8846(98)00165-3. [Accessed: 18-02-2024].

7 Appendix

7.1 All Variables Visualized

A) Shows each of the variables plotted against each other in a 9x9 scatter plot matrix. This scatter plot matrix can be used to visually see the correlation between each of the different variables.



B) Correlation between cement and superplasticizer is 0.0928

7.2 Summary Statistics

Table 2 - Summary Statistics of Data Set

	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)
count	1030	1030	1030	1030	1030	1030	1030	1030	1030
mean	281.17	73.90	54.19	181.57	6.20	972.92	773.58	45.66	35.82
std	104.51	86.28	64.00	21.36	5.97	77.75	80.18	63.17	16.71
min	102	0	0	121.75	0	801	594	1	2.33

25%	192.375	0	0	164.9	0	932	730.95	7	23.71
50%	272.9	22	0	185	6.35	968	779.51	28	34.44
75%	350	142.95	118.27	192	10.16	1029.4	824	56	46.14
max	540	359.4	200.1	247	32.2	1145	992.6	365	82.60