A DATA CURATION SYSTEM AND PREDICTIVE MODEL FOR UNCONFINED CONCRETE STRENGTH

Submitted by:

MSDS 23.3

Prateek Kakkar, Ramgopal Reddy Putta, and Xuanzhi Li

Presented to:

Department of Mathematics

College of Science and Engineering

Seattle University, Seattle

Under the Supervision of:

Dr. Ariana Mendible

Sponsor Organization:

DeSimone Consulting Engineering (DCE)

Liaison:

Jeff Dragovich, PHD, PE, SE, F.ACI

Matthew Cummins, P.E., S.E.

May 22nd, 2023

TABLE OF CONTENTS

Acknowledgement

We would like to express our gratitude to our sponsor *DeSimone Consulting Engineers* for providing us with the opportunity to work with them on designing an industry solution. It has been an incredible learning experience, and we are grateful for the technical growth that we have achieved during our time with the company. We deeply appreciate the trust that *DeSimone Consulting Engineers* placed in our team.

We would like to extend our sincere appreciation to our project liaisons Jeff Dragovich, Ph.D, PE, SE, F.ACI and Matthew Cummins, P.E., S.E., for their support and guidance throughout our project. Their willingness to clarify our doubts and provide expert insights was invaluable to the success of our project. We deeply appreciate their dedication to our success and their commitment to ensuring that we had a meaningful learning experience.

We would like to express our deep and sincere gratitude to Dr. Ariana Mendible, Department of Mathematics, Seattle University, for her unwavering support, expert suggestions, and constant encouragement throughout the project. Her guidance was pivotal in enabling us to surmount the various challenges and obstacles that we encountered along the way.

Last but not the least, we would like to express our heartfelt appreciation to the Seattle University Project Center, College of Science and Engineering, for offering us an outstanding project and sponsor, and for cultivating an atmosphere that promotes academic excellence, practical learning, and professional development. The opportunity to work with the Project Center has been incredibly valuable to us, and we are grateful for its dedication towards our education and advancement. Their mission and values of promoting excellence, fostering collaboration, and providing practical learning experiences have been reflected in every aspect of our project.

Abstract

This project aims to improve data management in the construction industry by providing a comprehensive solution that delivers clean data and the development of a machine learning model to predict concrete strength and understand the key factors affecting it. The team, consisting of Data Science graduate students at Seattle University and liaising with a leading structural engineering firm, DeSimone Consulting Engineering, establishes a feasible approach demonstrating the potential for using technology to improve data management processes in the construction industry and will provide a comprehensive solution for the company's data analysis challenges. The solution extracts data from PDF files and transforms it into a structured format. The team then develops a machine learning model to predict concrete strength based on the extracted data. This project enables the construction company to have better insights into concrete strength and its influential variables, make better business decisions, and improve the overall efficiency of their operations.

1. Introduction

DeSimone Consulting Engineers (DCE) is a leading firm in the structural engineering, facade consulting, and forensic services industries. DCE provides a comprehensive range of services (Desimone, n.d.) for all types of buildings, with 14 offices in 4 countries and headquartered in New York City. DCE is highly regarded for its expertise in concrete engineering and for delivering high-quality engineering services for projects of varying sizes and complexity.

Concrete is a vital composite material used in construction that consists of cement, water, and aggregates (such as sand, gravel, or crushed stone). Its importance lies in its strength, durability, and ability to withstand various environmental conditions. To test this, wet concrete samples are collected from construction sites, molded into a cylindrical shape (Fig. 1(a)), and are then subjected to strength testing using a specialized machine that can apply a compressive load to the sample (Fig. 1(b)). The machine measures the force required to compress the sample, allowing engineers to determine the compression strength of the material. What sets this measurement apart is that it is not restricted or confined, allowing for lateral expansion while undergoing compression. As a result, this particular strength is referred to as unconfined compression strength. The reason for collecting these wet samples is to test their strength in the lab based on their location of placement, which is crucial for ensuring the structural integrity of the concrete.

Fig.1 (a) Wet concrete samples molded into cylinders

Fig.1 (b) Tested for Unconfined Concrete Strength

The challenge DCE faces is that it collects an abundance of field concrete samples for determining unconfined compressive strength, creating a vast dataset. However, currently, DCE lacks an efficient system to effectively store, analyze and utilize this data. This data includes details

about the site, including city, project type, building height or number of floors, ambient temperature; details about the concrete used in the building, like concrete Mix ID, concrete temperature, structural element, location of placement, and water added; and details about the testing sample like unit weight, air content, slump, specimen size, and date tested. The contractor usually presents this data to DCE in PDF format, but there is currently no system in place to effectively process and make use of this information. This makes it challenging for a practicing structural engineer to find the predominant mix design for elements in a geographic area for specific building types and to understand the performance of high-strength mixes in hot and cold cycles. This information could help identify the most successful mixes and highlight any issues with mixes that have strength problems.

Currently, DCE's approach for collecting and analyzing this data includes obtaining ready-mix concrete loads from the contractor, collecting cylindrical modeled specimens, and sending them to the labs for testing the unconfined concrete strength for 56 days. The results of these tests are then sent to the DCE's shared drive, where the engineers review them manually, and make decisions accordingly. Despite collecting a vast dataset of field concrete samples for unconfined compressive strength, DCE currently lacks an efficient system for effectively analyzing and utilizing this data. This poses several challenges for engineers, such as difficulties in identifying suitable mix designs for specific building types and elements or determining which high-strength mixes are performing well or have strength issues. The current network-attached storage exacerbates the problem, making it challenging to locate and retrieve specific reports efficiently. Therefore, a formalized framework for data storage, retrieval, and analysis is urgently needed. Furthermore, DCE requires a machine learning model to predict concrete strength and understand how the key influencing factors affect it under real-world conditions, such as local temperature variations. A systematic approach to data collection, cleaning, preprocessing, and modeling is necessary to uncover meaningful insights and patterns.

2. Methodology

2.1. About the data

The dataset used in this project comprises of Unconfined Concrete Strength test results for each set of specimens from various construction sites, produced by testing agencies. The data is currently in PDF format and needs to be extracted. Reports produced by two independent testing agencies were used for this project, which contain essential details related to Site, Concrete and Test specimens collected from wet concrete mixtures. Since the current storage system hinders data analysis due to its nested directories on a network-attached storage system, there is a need for a structured data such as a clean Excel file that categorizes and minimizes data redundancy, facilitating easy retrieval and analysis for engineers. This structured data enables accurate predictions of *Unconfined Concrete Strength.*

2.2. Designed solution

To address these problems, a robust framework is implemented, consisting of two major components – The Data Curation System and The Predictive System.

The *Data Curation System* is responsible for collecting, cleaning, and organizing the data for analysis. It involves scraping the data from PDF files, storing them in a structured tabular format, and integrating the data into a unified format before transferring them to Excel files. This ensures data integrity by eliminating redundancy and inconsistency. The *Predictive System* then utilizes the curated data to develop predictive models using advanced machine learning algorithms, enabling DCE to accurately predict mix designs for engineering structures in specific temperature ranges and geographic areas.

2.3. Data curation system

The data curation system was developed in Python to extract and analyze the Unconfined Concrete Strength test data from PDF lab reports. This is accomplished through the utilization of text scraping techniques, using various Python libraries like Pypdfium2, Pdfplumber, Camelot and Tabula. These python libraries are implemented to scrape the data from different sections of the PDF document, including the header, footer, and body. This ensures that all relevant information is captured accurately and efficiently.

After extracting the necessary data from the PDF files, the next step is data cleaning. The Unconfined Concrete Strength test results are manually entered into the PDF format by the testing agencies, increasing the likelihood of errors such as incorrect field names and naming formats. To ensure data accuracy and integrity, the cleaning process includes the identification and correction of such errors. This is done using a series of conditions and codes that can also be used for future corrections. The cleaning process involves several stages, such as removing redundant data, filling in missing values, and correcting inconsistencies.

After completing the data cleaning process, the data is organized into multiple tables with predefined data schemas. Data schemas provide a structured framework for managing and organizing data within a data management system. They define field data types and establish relationships between tables by assigning primary keys and foreign keys. A primary key serves as a unique identifier for a record within a table, while a foreign key is a field in one table that references the primary key in another table, establishing connections and relationships between them. This ensures that the data is stored in a structured and organized format across different tables, making it easier for DCE engineers and analysts to extract valuable insights and draw conclusions from the data. The data curation system exports data into three clean data files:

a. *Project Information:* This file is created by DCE analysts and contains information related to a project, such as the project name, building type, number of levels, height, address, site latitude and longitude etc.
b. *Specimen Field Report:* This file includes field or site data, such as the site weather, concrete mix ID, concrete supplier, time of specimen molding, specimen temperature, specimen density, air content, required compression strength of the specimen, and other related data points.
c. *Specimen Compression Test Report:* This file includes data related to the compression tests conducted by a testing agency, such as date of the test, specimen age in days, load at which the specimen broke, and other relevant data points.

This structured data is then utilized in the designing of a predictive system, which uses sophisticated machine learning techniques to create models that can accurately predict the Unconfined Concrete Strength of a concrete specimen.

2.4. Predictive system

Choosing a suitable machine learning algorithm for a given problem depends on several factors, such as the type of data, and the desired outcome. In this case, the obtained data is labeled, and therefore a supervised learning method is appropriate. Additionally, since the target attribute to be predicted is a continuous numerical value, the problem falls into the category of regression.

After considering all the above factors, an artificial neural network model was chosen to predict the Unconfined concrete strength with high accuracy and identify the key influencing factors that affect the concrete strength in real-world conditions such as variations in local temperature. This helps an engineer to make even better-informed decisions. The subsequent section provides a detailed description of the model and its specifications.

2.5 Artificial neural network

Neural networks are models inspired by the biological structure of the human brain, composed of processing units known as neurons. The strength of neural networks lies in their numerous interconnections and their ability to learn from data, which makes them powerful tools for prediction and classification. In this case, a neural network approach is chosen to address the complex relationships observed in the data between various variables such as Required Strength and Age-days, and their impact on the unconfined compressive strength (UCS).

In this project, a neural network was used, which consists of layers of neurons that are interconnected through weighted connections and biases. As shown in Fig 3, the first layer takes in the input data, and the last layer produces the output. The intermediate layers, known as hidden layers, perform computations on the data as it passes through the network. Each neuron has a weight that decides how important it is. In a feedforward neural network, the input data is passed through different layers of neurons, which extract features and produce a final output. The weights and biases of these connections are adjusted through the backpropagation algorithm, which involves computing the error between the predicted and actual output and propagating this error backwards through the network to adjust the weights and biases. This technique is particularly useful when the output is dependent on past inputs and when there is a feedback loop between the output and input layers.

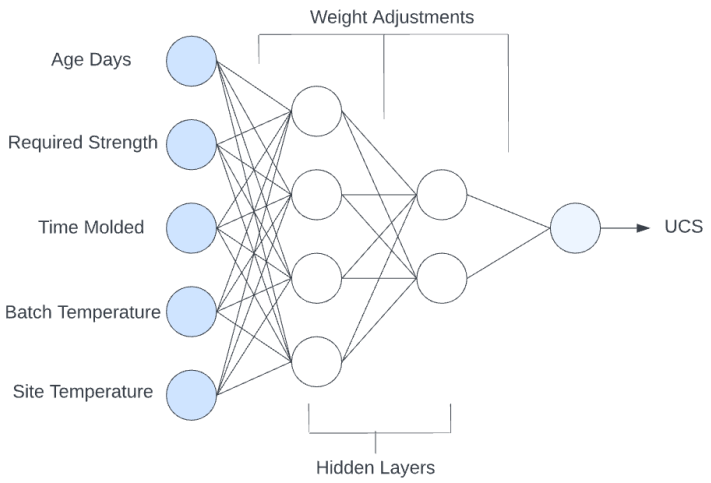

Fig 3. Artificial Neural Network

Given the number of available data points, the optimal performance was observed in the neural network architecture with 4 hidden layers having 300, 200, 100, and 50 neurons respectively. In order to prevent over-fitting, a simpler architecture was chosen, with regularization and dropout techniques implemented to further improve the model. Regularization, for instance, involves adding a penalty term to the loss function to limit model complexity, while dropout randomly removes neurons during training to avoid over-reliance on specific neurons. Specifically, the complete dataset was split into an 80% training set and a 20% testing set. Cross-validation was then used to evaluate the performance of the model by training it on different combinations of data subsets. The best epoch was chosen based on low validation loss.

To investigate the factors influencing unconfined concrete strength, permutation importance was used to determine the most significant variables in a neural network model. Permutation importance is a method that measures the importance of each feature in a model by shuffling its values and observing the effect on the model's performance. It works by measuring how much the model's accuracy changes when each variable is randomly shuffled. The greater the change in accuracy, the more important the variable is considered to be in predicting the target variable. By analyzing these scores, the most significant variables in predicting unconfined concrete strength were identified as Specimen Age Days and Specimen Required Strength.

After training the neural network model with the chosen architecture, and hyperparameters, its performance was evaluated by measuring the RMSE (Root Mean Squared Error) between predicted and actual values of the unconfined concrete strength.

3. Result

3.1. Data Visualization and Interpretation

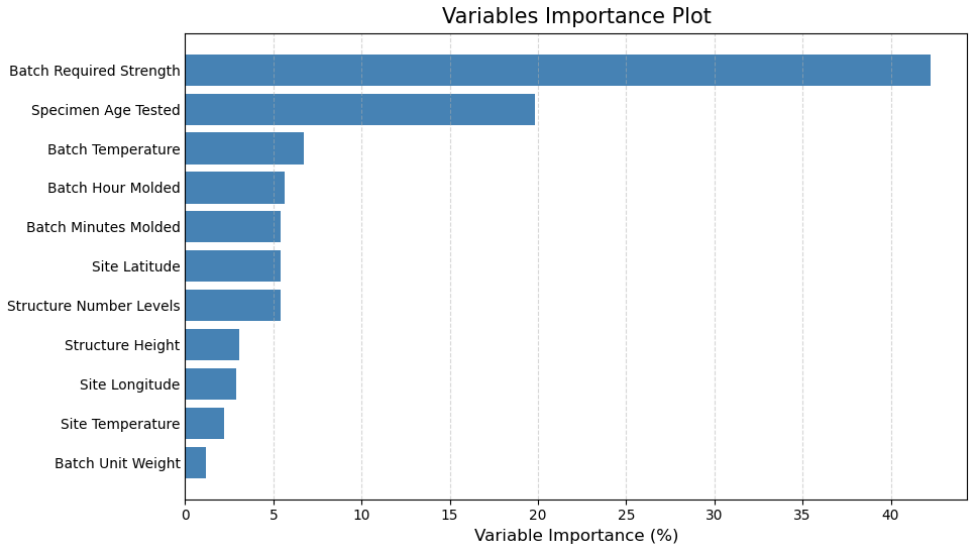

Figure 4: Percentage of Importance for Top 11 Variables in influencing Concrete Strength.

The model results in the RMSE value of 991, which means the predicted concrete strength value is off by +/-991 PSI.

The key influential variables are identified by the model. As illustrated in Figure 4, the top three influential features for predicting the measured concrete strength were found to be the batch required strength, the specimen age tested, and the batch temperature. This finding offers valuable insights into the key factors that impact concrete strength and can aid in enhancing quality control and optimizing the concrete production process.

In our study, we observed that the majority of specimens were tested on the 7th, 28th, and 56th days. To investigate the relationship between the required strength and the measured strength on these days, we plotted the data and visualized the results as shown in Figure 5. The linear regression lines indicate the trend, while the blue ticks represent the mean ratio with each required strength, and the bars show one standard deviation from the means.

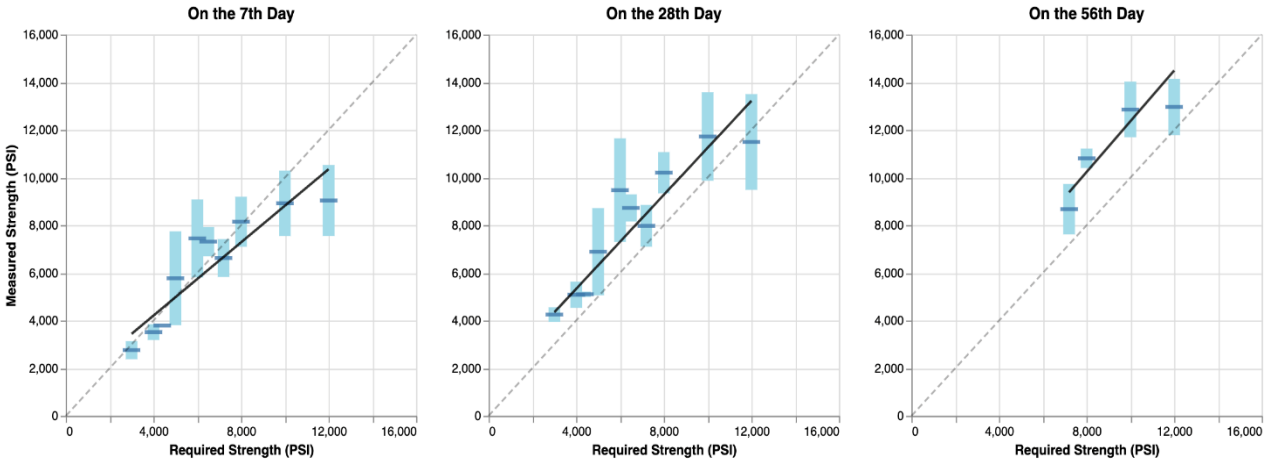

Figure 5: Measured Strength vs Required Strength on 7th, 28th, and 56th Day.

As depicted in Figure 5, on the seventh day, most specimens exhibited measured strengths that were close to but slightly lower than the required strength, except for a few with a very low required strength. By the twenty-eighth day, the majority of specimens had exceeded their required strength. Moreover, the concrete specimens continued to cure and exceed the required strength after the first twenty-eight days. Notably, we observed that for specimens with the largest required strength (12,000 PSI), the diagonal line passed through the mean of the measured strength on the 28th-day plot. Furthermore, any specimen with a lower measured strength than the mean was not qualified. This indicates that even though the majority of the concrete specimens exceeded the required strength in the first twenty-eight days, specimens with higher required strength need to be allowed to continue curing beyond the twenty-eighth day to ensure their quality.

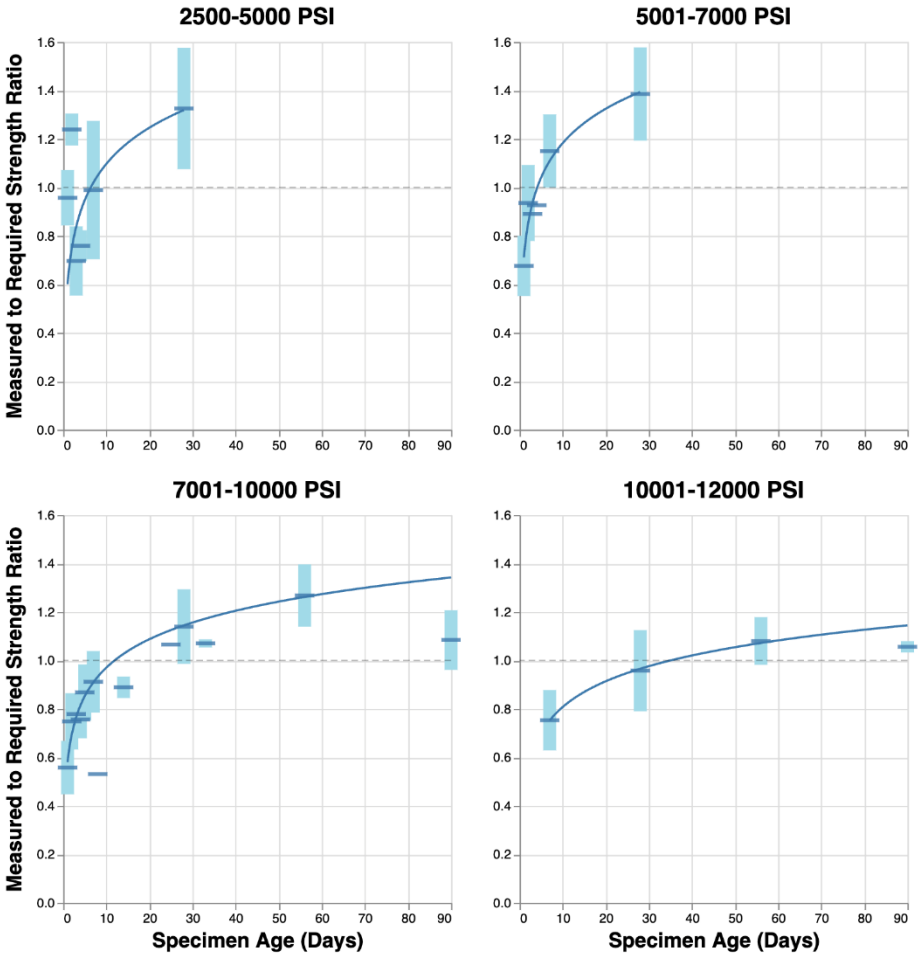

Figure 6: Measured to Required Strength Ratio vs Specimen Age for each Required Strength Range

In order to gain deeper insights into the performance of specimens with varying required strengths over time, we categorized the required strength into four bins. Subsequently, we plotted the ratio of Measured Strength to Required Strength against the Specimen Age for each bin. Considering that concrete samples tend to exhibit faster initial curing and slower later-stage curing, we employed logarithmic regression. To depict the trend, logarithmic regression lines were incorporated into the plot. As shown in Figure 6, specimens with required strength less than 10,000 PSI exhibit similar behavior, with those in the 5001 to 7,000 PSI range curing the fastest. On the other hand, specimens with required strength higher than 10,000 PSI cure much slower and their measured-to-required strength ratios do not rise as high as the others. In the lower right plot, we observe that the majority of specimens in this required strength range do not reach the required strength until the fifth or sixth day (one standard deviation away from the mean). Therefore,

concrete samples with a required strength higher than 10,000 PSI should be cured for at least 56 days, or potentially longer, to ensure their quality before their application in construction.

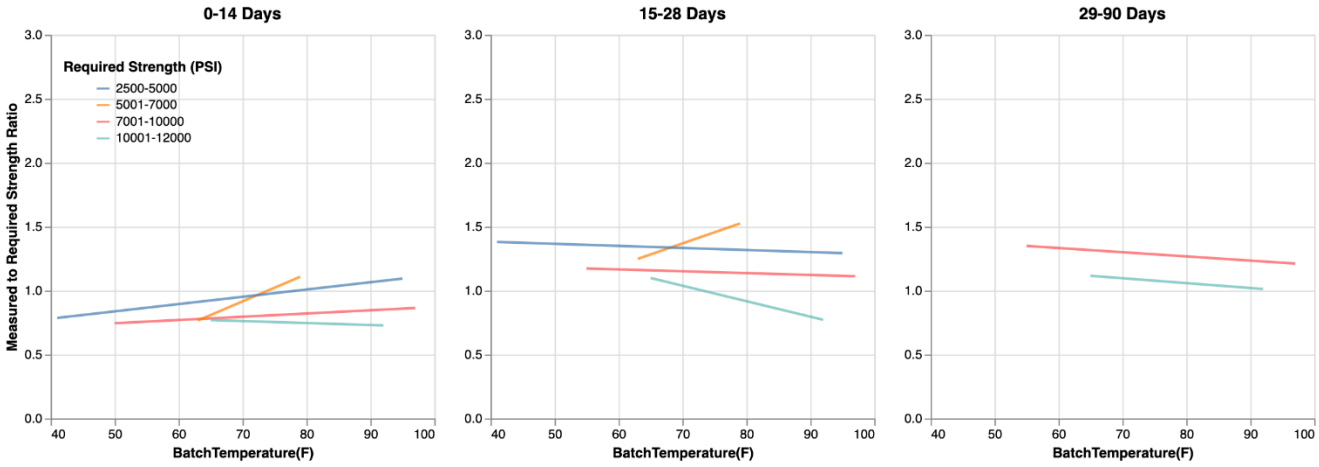

Figure 7: Measured to Required Strength Ratio vs Batch Temperature

Lastly, our analysis focused on the third most influential feature for predicting concrete strength, namely batch temperature. This feature denotes the temperature of the concrete specimen during the curing process. Figure 7 illustrates our findings, indicating that a higher batch temperature during the initial 14-day period is correlated with improved curing, as evidenced by the measured strength surpassing the required threshold.

However, beyond the first 14 days, this correlation ceases to hold true, with one exception: specimens with a required strength falling within the range of 5,001 to 7,000 PSI. Consequently, we recommend reducing the batch temperature for all specimens whose required strength falls outside of this range after the initial 14-day period, as this adjustment has the potential to enhance the measured strength. Currently, batch temperature manipulation involves the addition of hot water or ice, both of which may have a detrimental effect on the measured strength. Unfortunately, due to missing data regarding the amount of water added, our project was unable to explore the relationship between water added, batch temperature, and measured strength. To facilitate further investigation, we strongly recommend collecting information on water added, enabling a more comprehensive analysis of these interrelated factors in the future. Such an exploration will yield a more precise recommendation regarding batch temperature manipulation.

4. Conclusion

This project has helped DCE overcome their data storage, retrieval, and analysis challenges. Initially, DCE stored their concrete strength test reports stored in PDF files on a shared drive, which made it difficult for their engineers to retrieve the data and impossible to conduct data analysis. To address these issues, our team developed text scraper tools to extract the data from the PDF files and established a data schema to store the data in a structured format, for improved data storage and retrieval.

In addition, we have developed a predictive model for concrete strength using Artificial Neural Networks that can potentially save DCE's engineers time, money, and efforts by estimating concrete strength values without the need of conducting several strength measurement tests in the laboratories by various testing agencies. The model can be used to predict concrete strength for different batches with varying parameters, empowering engineers to adjust the curing process and enhance the overall quality of the concrete.

Finally, the plots we created provide key insights and advice on concrete curing. We analyzed the relationship between variables such as batch temperature, required strength, and specimen age and their impact on measured strength. Our findings suggest that reducing the batch temperature after the first 14 days for specimens with required strength above 5,000 PSI can potentially increase measured strength. This information can help DCE's engineers to adjust the curing process to improve the overall quality of their concrete.

Overall, the improved data management processes and machine learning models we developed have the potential to revolutionize the construction industry by providing engineers with accurate predictions and insights on concrete strength and curing processes.

5. Appendix

References:

1. Desimone. (n.d.). *Firm Profile*. Retrieved January 23, 2023, from https://www.de-simone.com/assets/DeSimone-Brochure-New-York-2022.pdf
2. Salvador, M. (2022, May 18). *Break tests vs the maturity method*. Kryton. https://blog.kryton.com/2021/12/break-tests-vs-the-maturity-method/
3. IS Code. (2022, November 22). *Unconfined compressive strength test of soil, UCS Test*. Civil Allied Gyan - No.1 Civil Engineering Blog. http://www.civilalliedgyan.com/2020/07/unconfined-compressive-strength-test-of-soil.html