Concrete is an essential material in civil engineering. Its strength as a building material (typically measured by compressive strength, or ability to withstand compressive force) is a complex function of its ingredients and age after mixing. The purpose of the proposed project is to create a model that predicts concrete compressive strength based on the quantity of ingredients and age in days. The dataset was obtained from the UCI ML repository. The ingredients included in the dataset are cement, slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate each measured in kilograms in a cubic meter mixture. The target variable, concrete compressive strength, is measured in mega pascals (MPa).

**Problem statement:** Can the compressive strength of concrete be predicted and maximized based on the amounts of ingredients and age after mixing?

**Criteria for success:** Prediction of concrete compressive strength with an R-squared value of .8 or higher

**Scope of solution space:** All we have to work with in terms of features are the amounts of individual ingredients per cubic meter of mixture and the time aged. Some features may need to be engineered, possibly including ratios of ingredients, or binary features of whether or not an ingredient is present.

**Constraints:** All features and target variable must always be positive to be physically meaningful. A viable mixture must contain at least some cement, water, and other ingredients. A couple of features (ingredients) have many zero entries in the dataset. It must be determined whether these are valid zeros representing mixtures not containing those ingredients, or if they are missing data.

**General Approach:** After some exploratory data analysis to determine the relative importance of features, several regression techniques will be used to predict compressive strength. The data should be scaled or perhaps normalized to a percentage composition for the ingredients. A linear regression model will be compared to a random forest model, and the best model will be decided based on maximum R-squared value and minimum mean absolute error obtained from cross-validation.

**Deliverables:** A GitHub repo will be created containing clearly annotated notebooks for each step of the project. A slide deck and project report will also be included to summarize the results of the project.