# Model Card

This model is used to analyse the results of an experiment to determine the optimum ingredients for concrete. It actually consists of three models:

* K Nearest Neighbour Regressor
* Random Forest Regressor
* Neural Net with one hidden layer

# Model Description

The data is in the form of a CSV table consisting of 1030 rows and nine columns. Each row of data represents a sample of concrete which was crushed to find its compressive strength. The first seven columns list the ingredients used in each sample of concrete. The eighth column is the number of days the sample was allowed to cure, and the final column is the compressive strength of the sample.

The table is shuffled, then split into a training set (80%) and a test set (20%). The first eight columns are selected as the predictors and the last column, the compressive strength, forms the target. All the values are real and continuous. There are no classes. All three models are used for regression. The output is the predicted compressive strength.

* The K-Nearest Neighbour regressor architecture is based on sklearn’s KNeighborsRegressor. A grid search with five-fold cross-validation was used to tune the hyperparameters:
  + Algorithm: ball\_tree
  + Leaf\_size: 30
  + N\_neighbors: 5
  + p: 2
  + Weights: distance
* The Random Forrest architecture was based on sklearn’s RandomForestRegressor. A grid search was also used for tuning:
  + Criterion: squared\_error
  + Max\_features: None
  + N\_estimators: 100
* The Neural Net architecture was based on torch nn, with manually tuned hyperparameters:
  + One hidden layer with 64 hidden elements
  + Activation function: Sigmoid
  + Optimiser: stochastic gradient decent
  + Loss function: mean squared error

# Performance

The performance of each model was based on the mean absolute percentage error, (MAPE) measured on the test set.

* K-nearest neighbour: 20.0%
* Random Forest: 12.3%
* Neural Net: 7.7%

K-nearest neighbour:

Chart, scatter chart

Description automatically generated

Random Forest:

Chart, scatter chart

Description automatically generated

Random Forest Feature Importance:

Chart

Description automatically generated

Neural Net:

Chart, scatter chart

Description automatically generated

# Limitations & Trade-offs

The k-nearest neighbour and random forest models are limited by their accuracy. This is due to the highly non-linear nature of the input data. While the random forest outperforms the k nearest neighbour it is still not enough to adequately represent the data.

The neural net performs better on the non-linear data due to the use of the sigmoid activation function but serves as a ‘black box.’ Unlike the random forest it is unable to interpret the data.