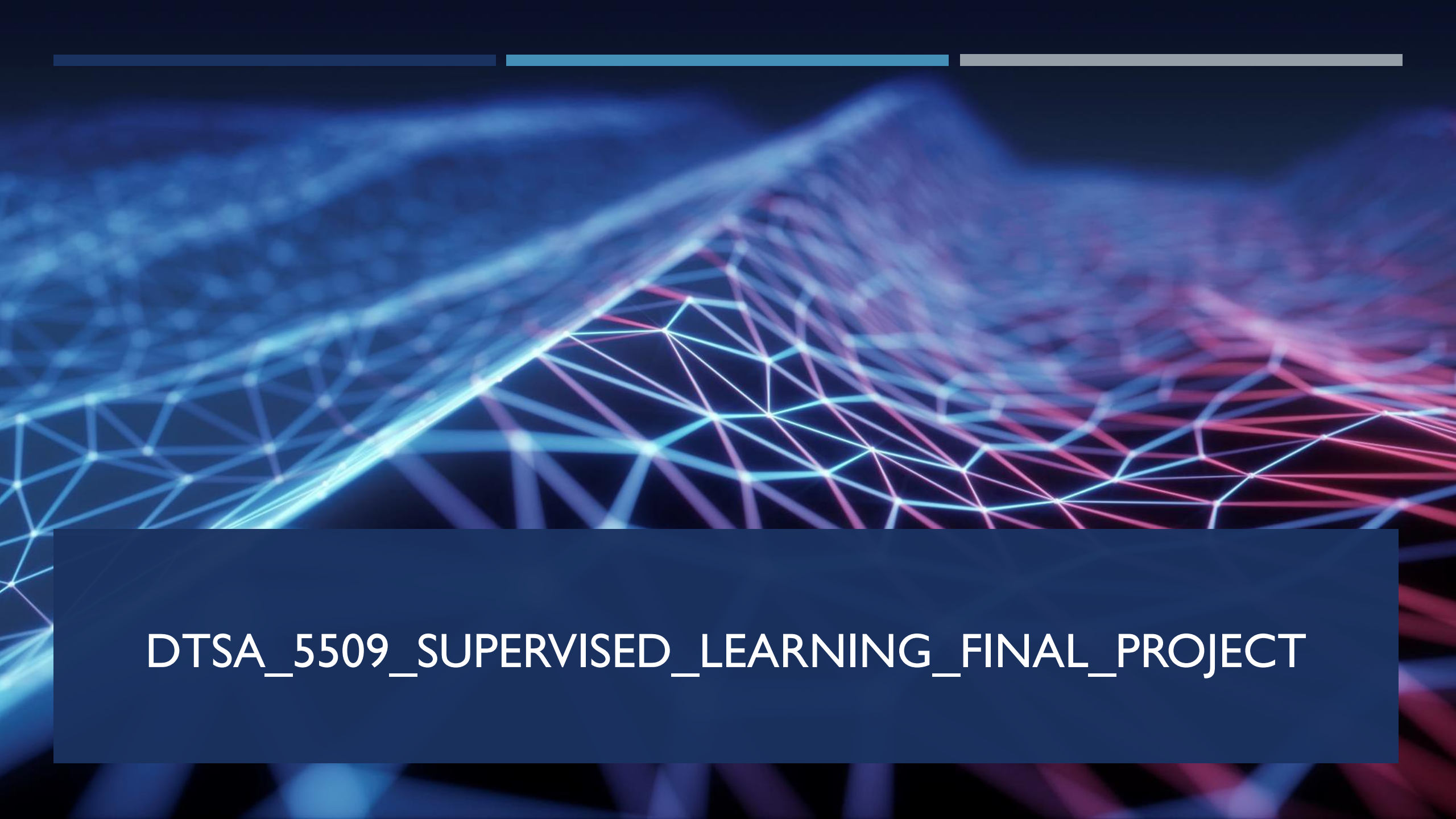


DTSA_5509_SUPERVISED_LEARNING_FINAL_PROJECT

PRESENTATION HEADLINES

- Problem Definition
- ML Approach
- Conclusions & Results

Problem Definition

The problem definition revolves around the Civil Engineering: Cement Manufacturing Dataset available on Kaggle. This dataset contains information about cement manufacturing and concrete properties. The goal is to predict the compressive strength of concrete using machine learning techniques applied to the dataset's features. The dataset consists of 1030 observations and 9 columns, with 8 columns representing the features used to predict the compressive strength. The features include cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and age.

Problem Definition

The compressive strength is the target variable of interest, representing the concrete's ability to withstand compressive forces. The dataset provides numerical values for all features, and there are no missing values.

During the exploratory data analysis (EDA) stage, it was discovered that there are 25 duplicated rows in the dataset, which were subsequently dropped to ensure distinct concrete mixtures are considered ending up with 1005 observations.

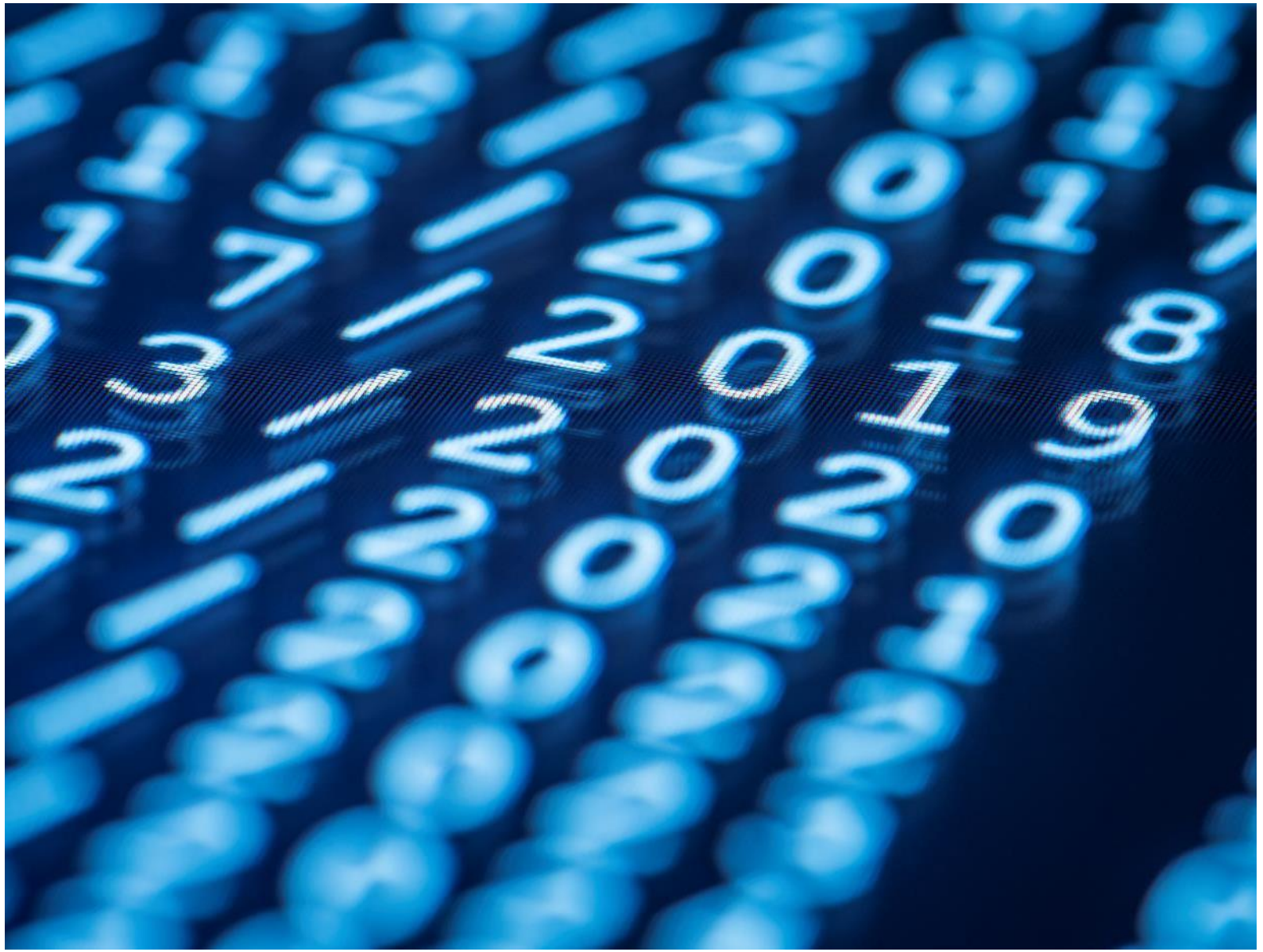
ML Approach

```
Model Comparison:
XGBoost Regression: 0.921205229611639 1
Random Forest Regression: 0.9172339433343911
Gradient Boosting Regression: 0.896468088755232
Decision Tree Regression (Cross-Validation): 0.8284440187640069
Neural Network Regression (Cross-Validation): 0.7295898277295405
Support Vector Regression: 0.6262211729783689
Linear Regression: 0.60070066074050 85
Lasso Regression: 0.5550728149310743
Ridge Regression: 0.5528291756490149

Best Model: XGBoost Regression
```

ML Approach

In the model comparison for ML approaches, several regression models were evaluated. The results showed that XGBoost Regression achieved the highest performance with a score of 0.921. It outperformed other models such as Random Forest Regression (0.917), Gradient Boosting Regression (0.896), Decision Tree Regression with cross-validation (0.828), Neural Network Regression with cross-validation (0.729), Support Vector Regression (0.626), Linear Regression (0.601), Lasso Regression (0.555), and Ridge Regression (0.553). Based on these results, XGBoost Regression was determined to be the best model for the given task.



THANK YOU