Machine Learning Model Tracking Document

1. Dataset Information

Dataset Name:	Well 782
Number of Samples:	782
Number of Features:	4 (Tf, Rs, Gg, Api)
Target Variable:	Pb
Outlier Handling:	None
Feature Engineering Applied:	None
Scaling/Normalization Applied:	None
Encoding Applied:	None

2. Preprocessing Steps

Step	Description
Train-Test Split	70% - 30%
Shuffling	Yes, using random_state=42
Handling Missing Data	None
Feature Scaling	SVM only
Feature Selection	None

3. Models Used & Hyperparameters

Model	Hyperparameters	Training Time
Linear Regression	None	0.0007
Ridge Regression	alpha = 0.001	0.0011
Lasso Regression	alpha = 0.001	0.0021
Decision Tree	max_depth = 4	0.0011
KNN	n_neighbors = 5	0.0009
SVR	C = 100.0, kernel = linear	12.8890

4. Evaluation Metrics

Model	MSE	RMSE	MAE	R^2 Score	Adjusted R^2
Linear Regression	241250.0	491.17	384.37	0.8158	0.8126
Lasso Regression	241250.0	491.17	384.37	0.8158	0.8126
Ridge Regression	241250.0	491.17	384.37	0.8158	0.8126
Decision Tree	226862.4	476.30	356.14	0.8268	0.8238
KNN	222437.7	471.63	361.29	0.8302	0.8272
SVR	211659.1	460.06	352.83	0.8384	0.8356

5. Cross-Validation Summary (5-Fold)

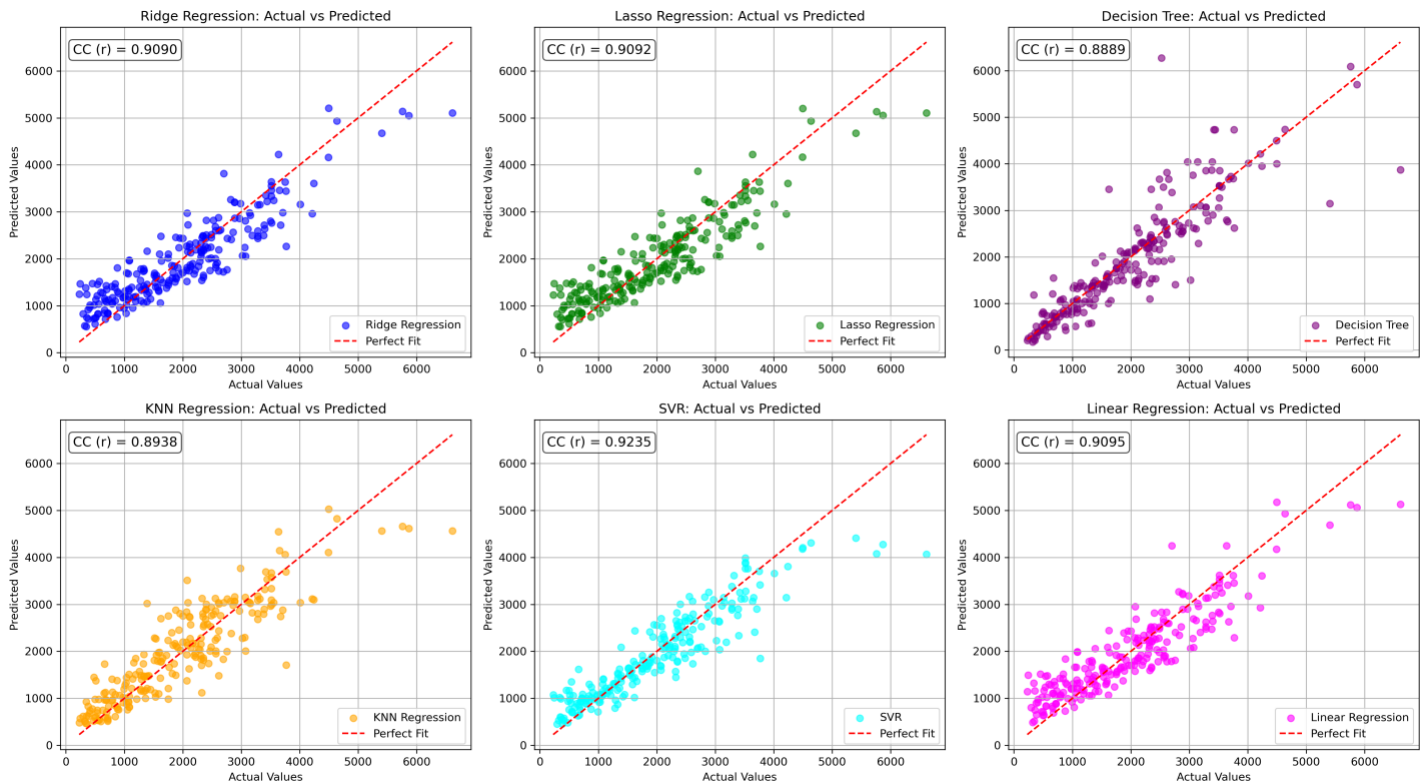| Model	RMSE Mean	RMSE Std	MAE Mean	MAE Std	R^2 Mean	R^2 Std
Decision Tree	517.10	67.93	321.83	44.24	0.8330	0.0360
KNN	553.23	71.48	393.79	39.81	0.8126	0.0083
Ridge Regression	586.66	106.67	405.94	34.80	0.7904	0.0316
Lasso Regression	586.66	106.67	405.94	34.80	0.7904	0.0316
Linear Regression	586.66	106.67	405.94	34.80	0.7904	0.0316
SVR	651.21	138.53	450.23	70.81	0.7413	0.0573

6. Visualizations

Figure 1: Actual vs. Predicted Values for Pb (Well 782)
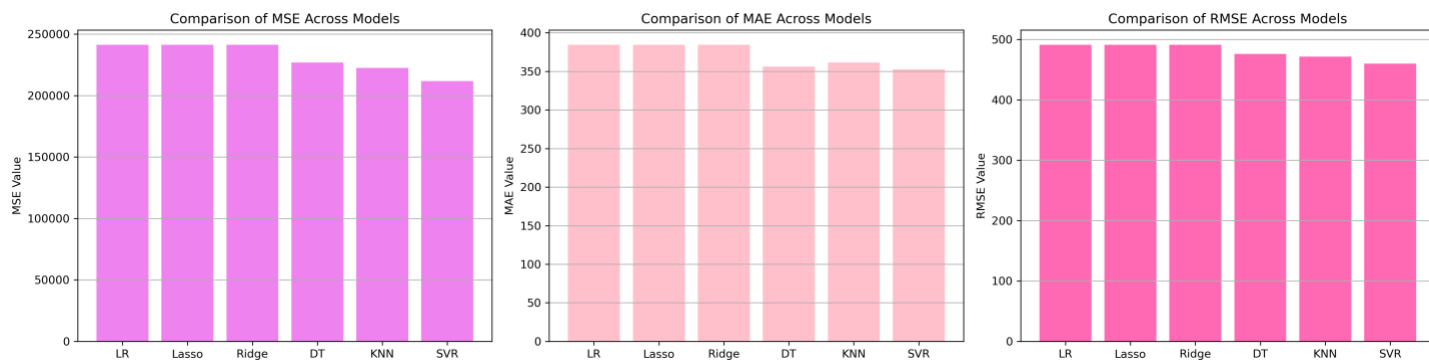
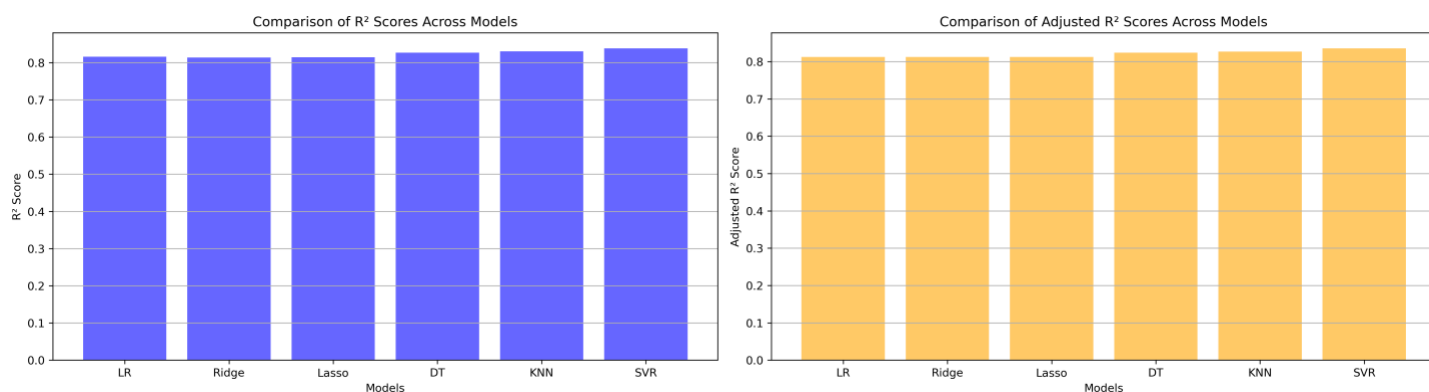*Figure 2: Bar Charts of MSE, RMSE, MAE for Pb (Well 782)*

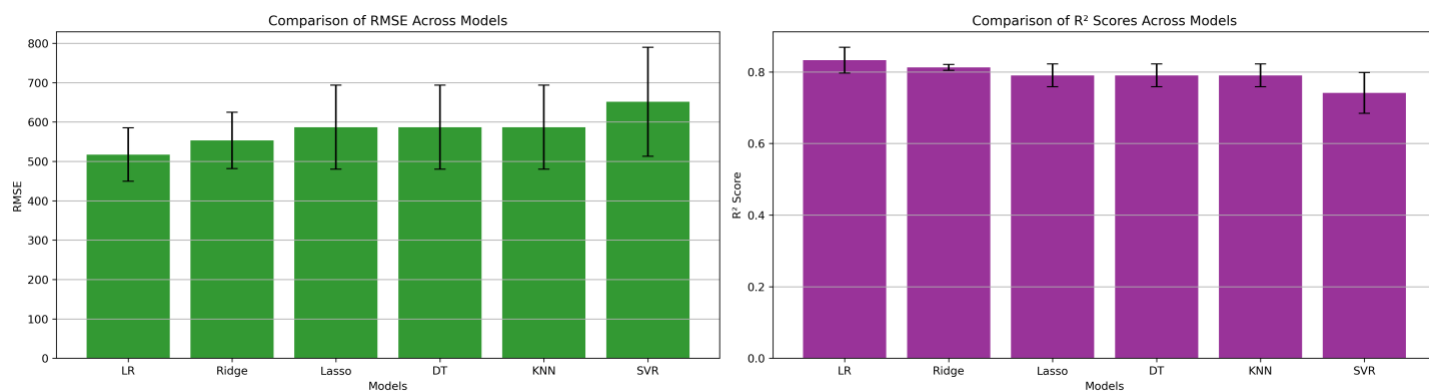*Figure 3: Bar Charts of R^2 and Adjusted R^2 for Pb (Well 782)*

Figure 4: Error Bars for RMSE and R^2 from CV for Pb (Well 782)

7. Observations & Next Steps

- **Best Performing Model**: SVR (RMSE: ~460.06, R^2: ~0.8384)
    - SVR performed better than other models on test set
    - All models scored R^2 between 0.74 and 0.84
    - Performance noticeably weaker than other wells (160, 283)

8. Code Access

The complete source code for data preprocessing, model training, evaluation, and visualization is available here. The repository includes organized Jupyter notebooks for each phase, dataset, and target, as well as requirements for reproducibility.