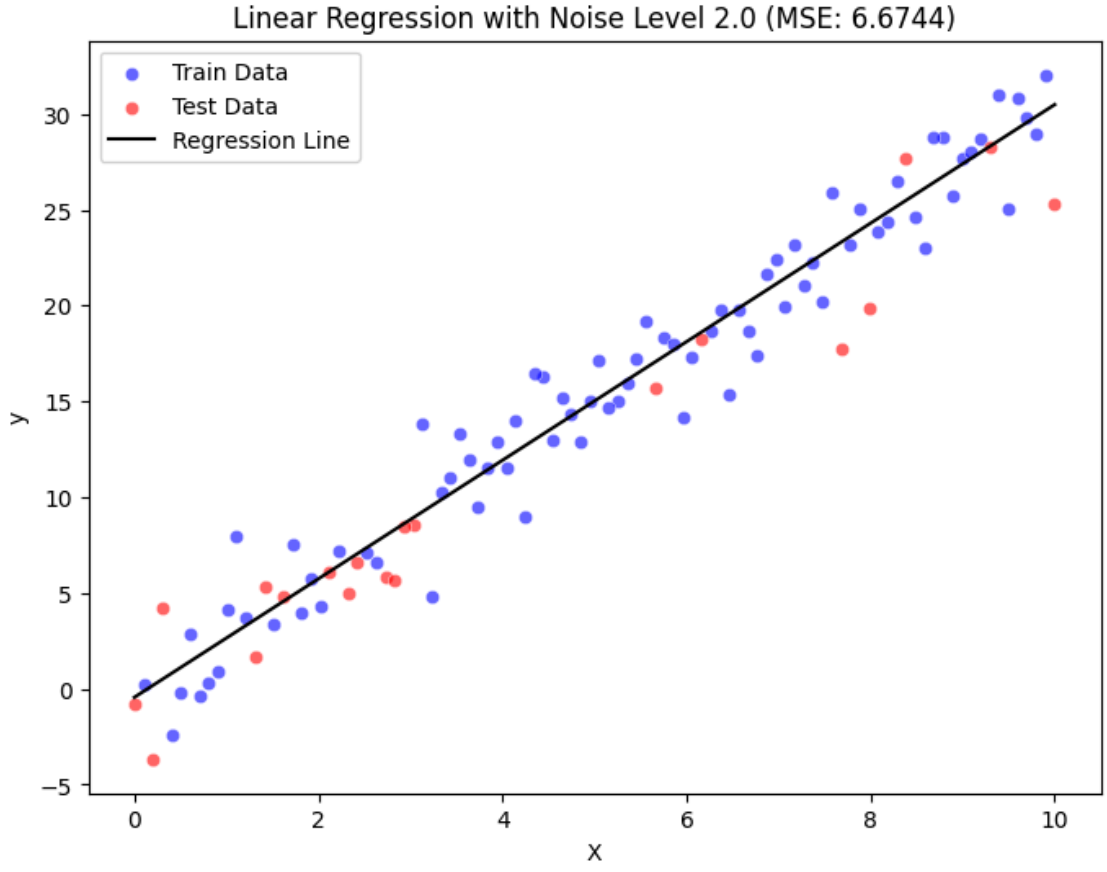
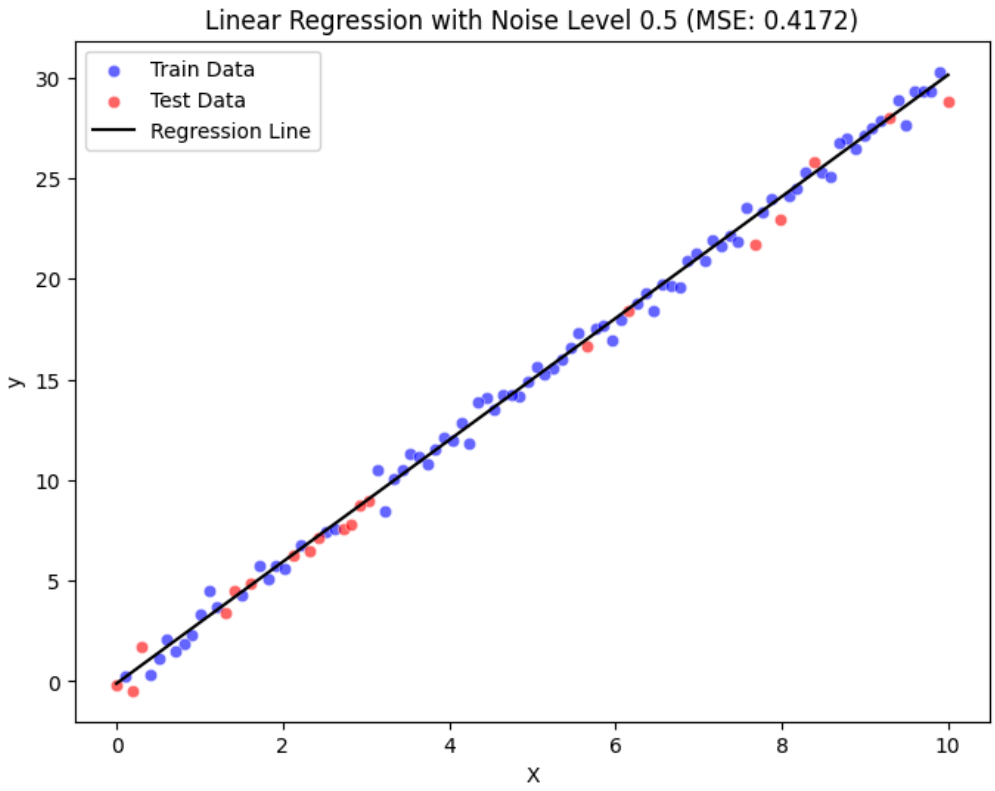
[Link to Code]

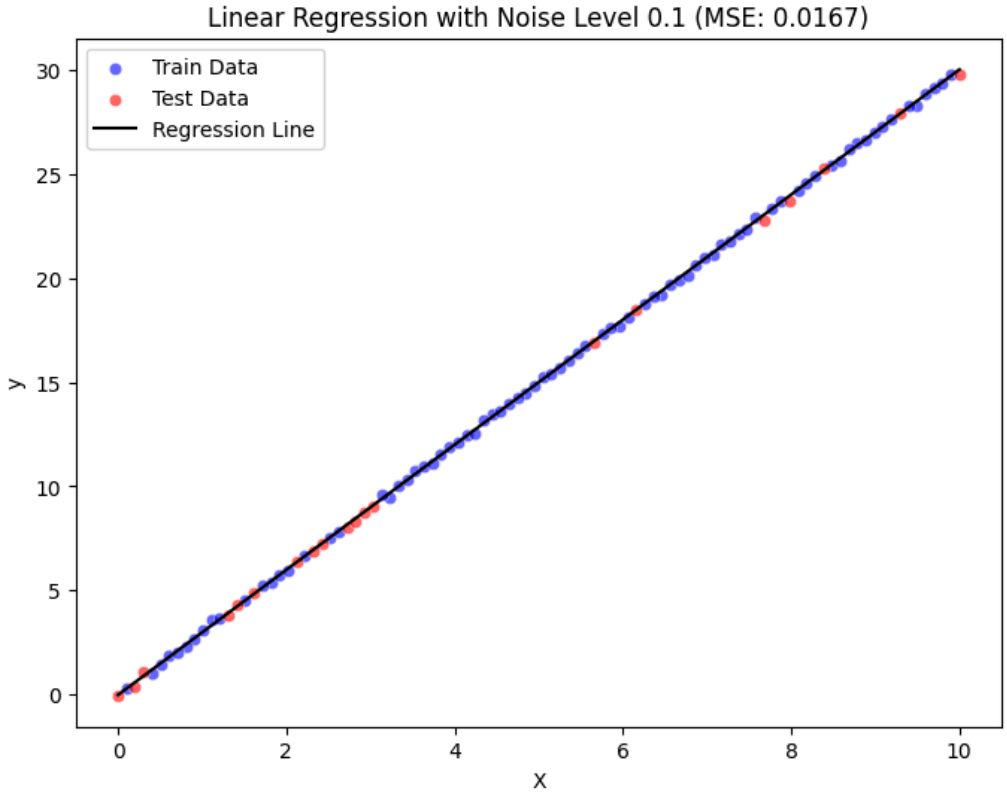
Task 1: Simple Linear Regression on Simulated Data

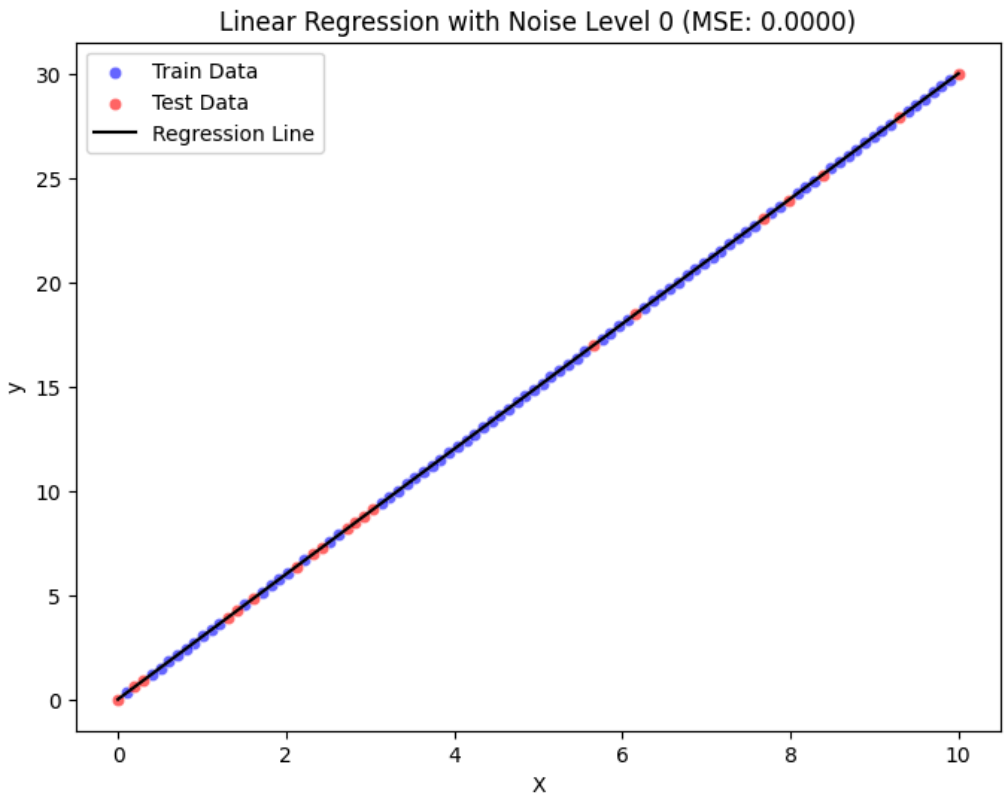
We can see that as the noise increases, the points deviate more from the regression line, leading to the MSE to rise.

Noise Level	MSE
0	0.00
0.1	0.0167
0.5	0.4172
2.0	6.6744




Linear Regression with Noise Level 0.5 (MSE: 0.4172)
Train Data
Test Data
Regression Line
y
X
0
5
10
15
20
25
30
0
2
4
6
8
10

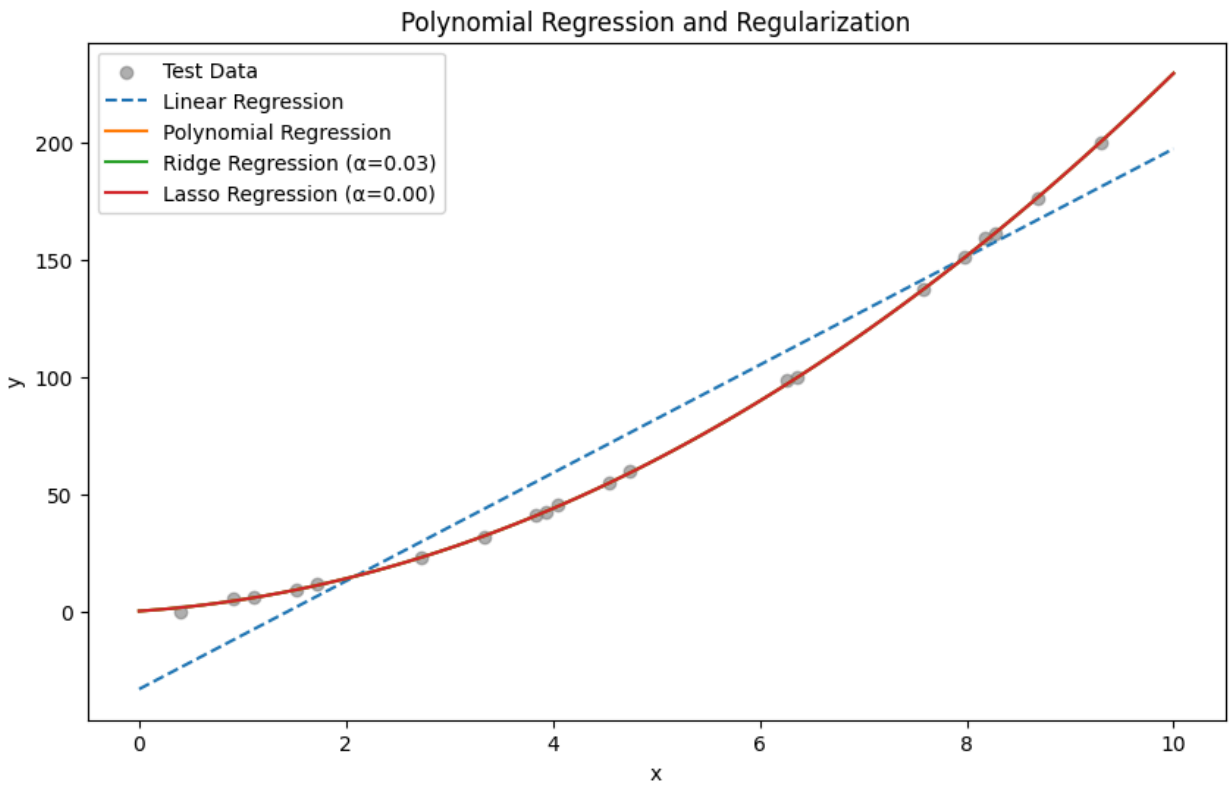

Linear Regression with Noise Level 0.1 (MSE: 0.0167)
Train Data
Test Data
Regression Line
y
X
0
5
10
15
20
25
30
0
2
4
6
8
10



Task 2: Polynomial Regression and Regularization

I determined the best alpha value using Cross Validation with 25 folds. I uniformly sampled 100 alpha values in [-3,3]. Here are the results:
Best Ridge Alpha: 0.02848035868435802
Best Lasso Alpha: 0.001

Linear Regression MSE: 166.9013
Polynomial Regression MSE: 0.5615
Ridge Regression MSE (α=0.028480): 0.5622
Lasso Regression MSE (α=0.001000): 0.5604



In the above graph the polynomial, ridge, and lasso regression all overlap each other. The best model is the Lasso Regression since it has the lowest MSE.
Ridge regression reduces the magnitude of coefficients but does not set them to zero. Lasso regression shrinks some coefficients to zero, performing feature selection. In general, Regularization prevents overfitting, making the model generalize better to the test data split. When a large alpha value is picked, Ridge regression will shrink coefficients too much, leading to underfitting. Meanwhile, Lasso regression will drive most coefficients to zero, making the model behave like a linear regression model.

Task 3: Classification using real data

The Breast Cancer dataset from sklearn.datasets contains 569 samples with 30 numerical features, used to classify tumors as malignant (0) or benign (1). The dataset is well-structured and does not contain missing values.

Since SVM is sensitive to feature magnitudes, we standardized all features using StandardScaler. We do not need to encode categorical variables as it has already been encoded: 0 for Malignant and 1 for benign. We split the dataset into 80% training and 20% testing using train_test_split.

The rbf(Radial Basis Function) kernel has the highest accuracy on the dataset of 96.5% compared to poly and linear. This suggests that the decision boundary is nonlinear, making rbf a better choice for capturing complex relationships in the data.

Kernel	Accuracy
Linear	0.956140350877193
Poly	0.8947368421052632
RBF	0.9649122807017544