Cairo university

Faculty of Computers and Artificial Intelligence

Theoretical Foundations of Machine Learning

Medical Cost Personal-Project

By

Adham Adel Mohammed. ID:20210052

Ziad Ahmed Eliwa. ID:20210143

Abdelrahman Mohammed Ali. ID:20210518

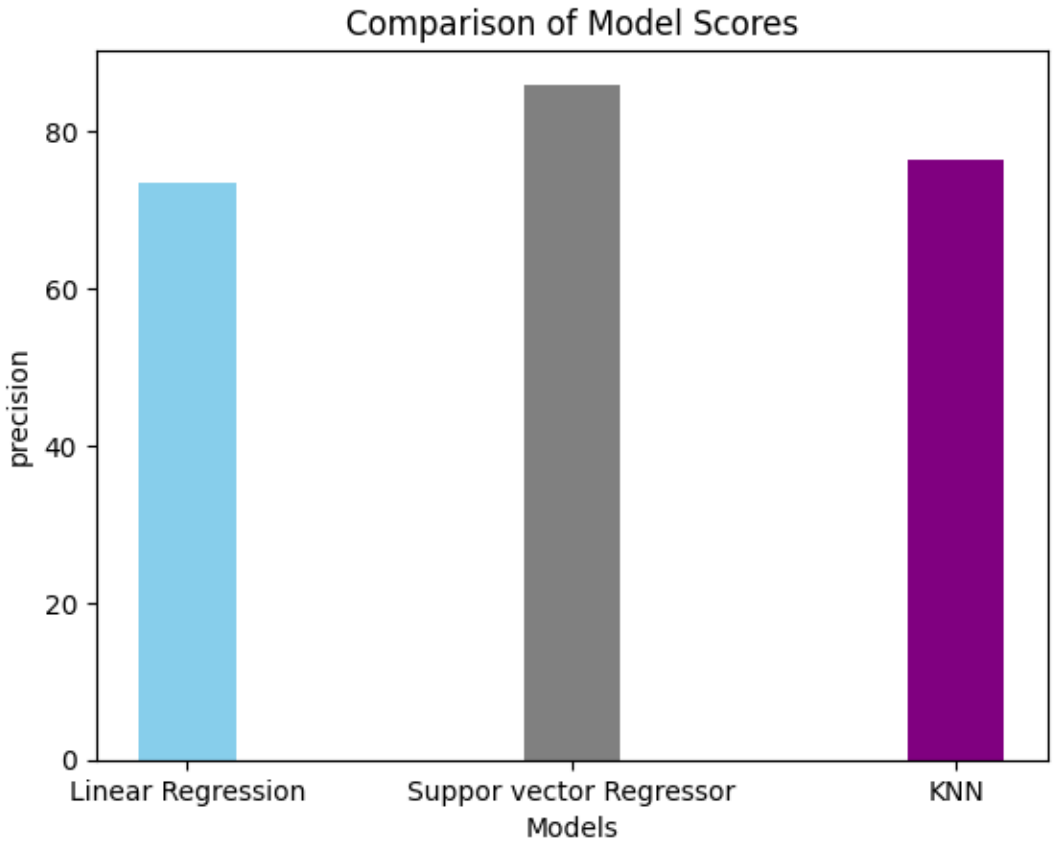
Muhannad Ibrahim Yassin. ID:20210411

Nourhan Mahmoud Eldesouky. ID:20211107

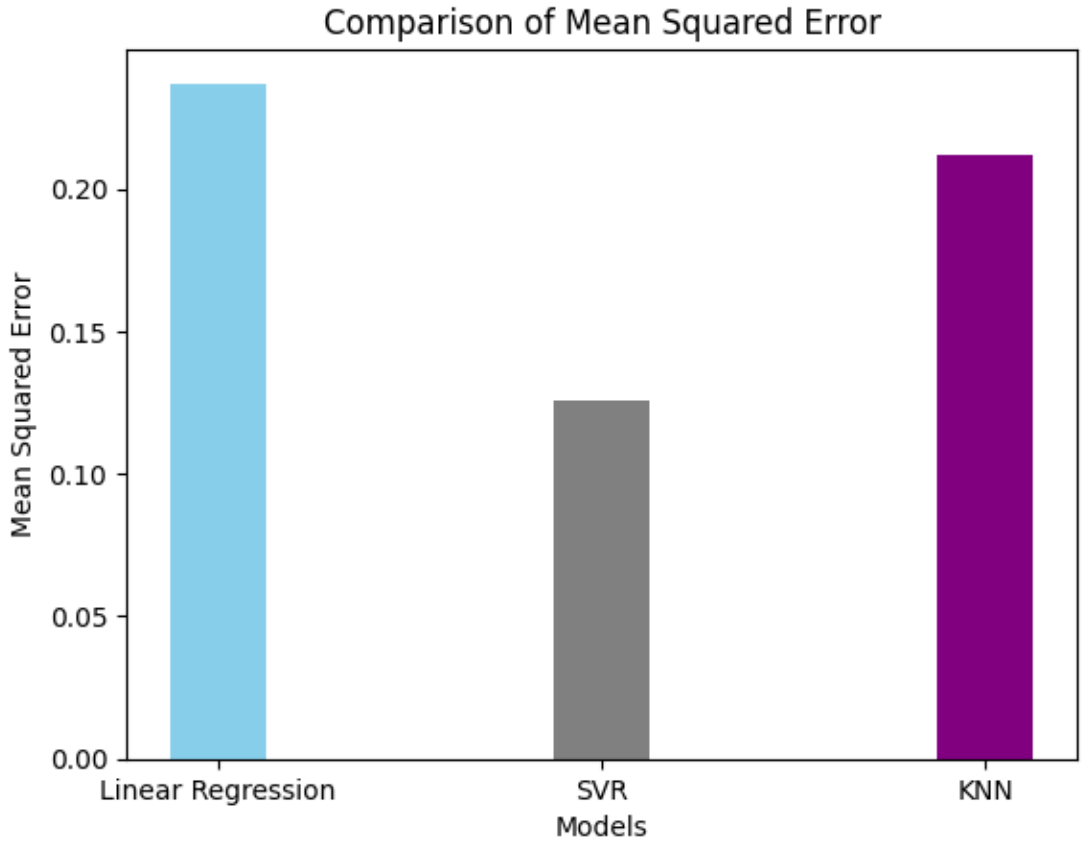
Model	***Linear Regression***	***Support Vector Regressor:***	***KNN (K-Nearest Neighbors***
Result	Almost: 0.73489 **73.48%**	Almost: 0.8595 **85.95%**	Almost: 0.76296 **76.29%**
Comment	1. Linear Regression is a simple and interpretable model that assumes a linear relationship between the features and the target variable, which might be limiting in capturing complex patterns in the data. 2. It performed the worst (oversimplify the underlying relationships) which means this model may not be the best choice for this dataset.	1. The support Vector Regressor model performed the best. 2. It uses a kernel function to transform the data into a higher-dimensional space, allowing for more flexible modeling. In this case, the 'rbf' (Radial Basis Function) kernel was used.	1. K-Nearest Neighbors is a non-parametric model that makes predictions based on the average of the target values of its neighboring data points. 2. It tends to work well when the data has clear clusters or local patterns. 3. KNN relies on the similarity of instances, and its performance may be sensitive to the choice of the number of neighbors (k).
Error analysis	**Errors:** Linear regression may struggle to capture non-linear patterns in the data, and it assumes a constant relationship between predictors and the target. It also assumes that the errors are normally distributed, which might not be true. **Improvements:** Consider feature engineering to	**Errors:** Despite its high precision, SVR can still make errors, especially if the choice of the kernel or hyperparameters is not optimal. **Improvements:** Perform hyperparameter tuning to find the optimal kernel parameters and	**Errors:** KNN may be sensitive to outliers and the choice of the number of neighbors (k). The assumption of local similarity may not always hold, as KNN doesn't perform well when there are a lot of features because it treats each feature equally. **Improvements:** Experiment with different values of k, consider data preprocessing

	capture non-linear relationships (polynomial regression or regularization techniques) or explore more sophisticated regression techniques that can handle non-linearity.	consider experimenting with different kernel functions to enhance performance.	to handle outliers, and explore other distance metrics to enhance the model's robustness.

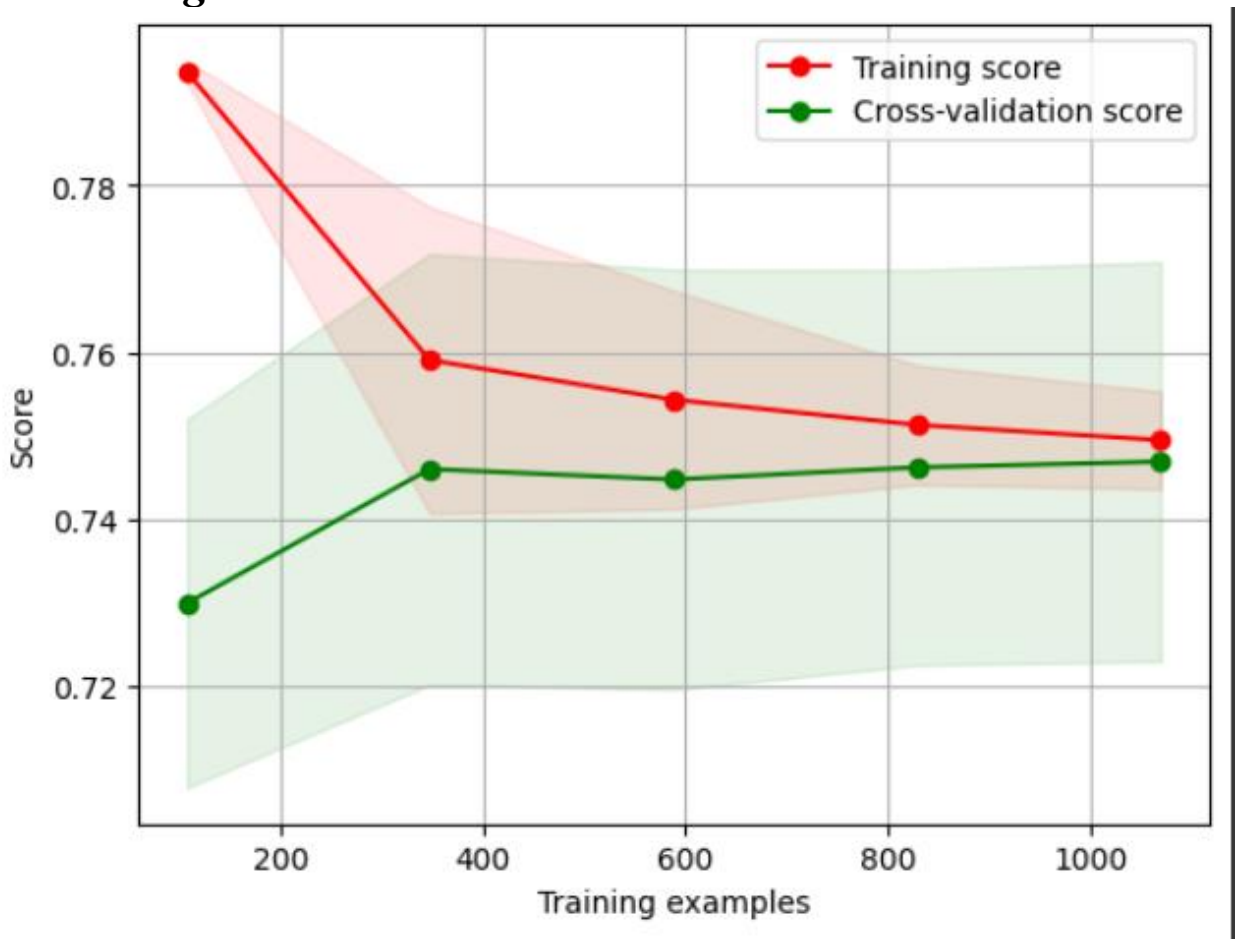
Visualization of the models and comparison between them:



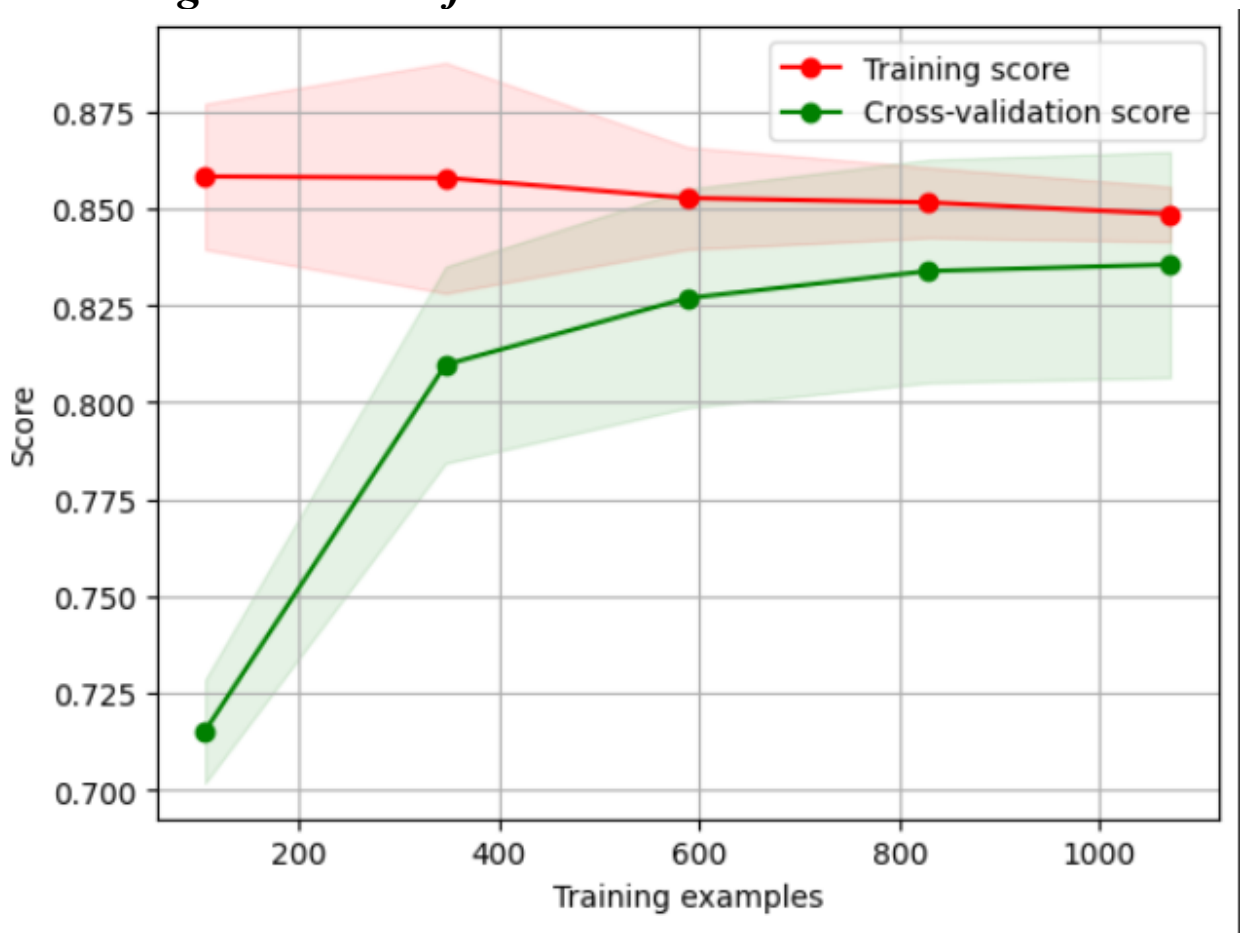
Comparison of Mean Squared Error



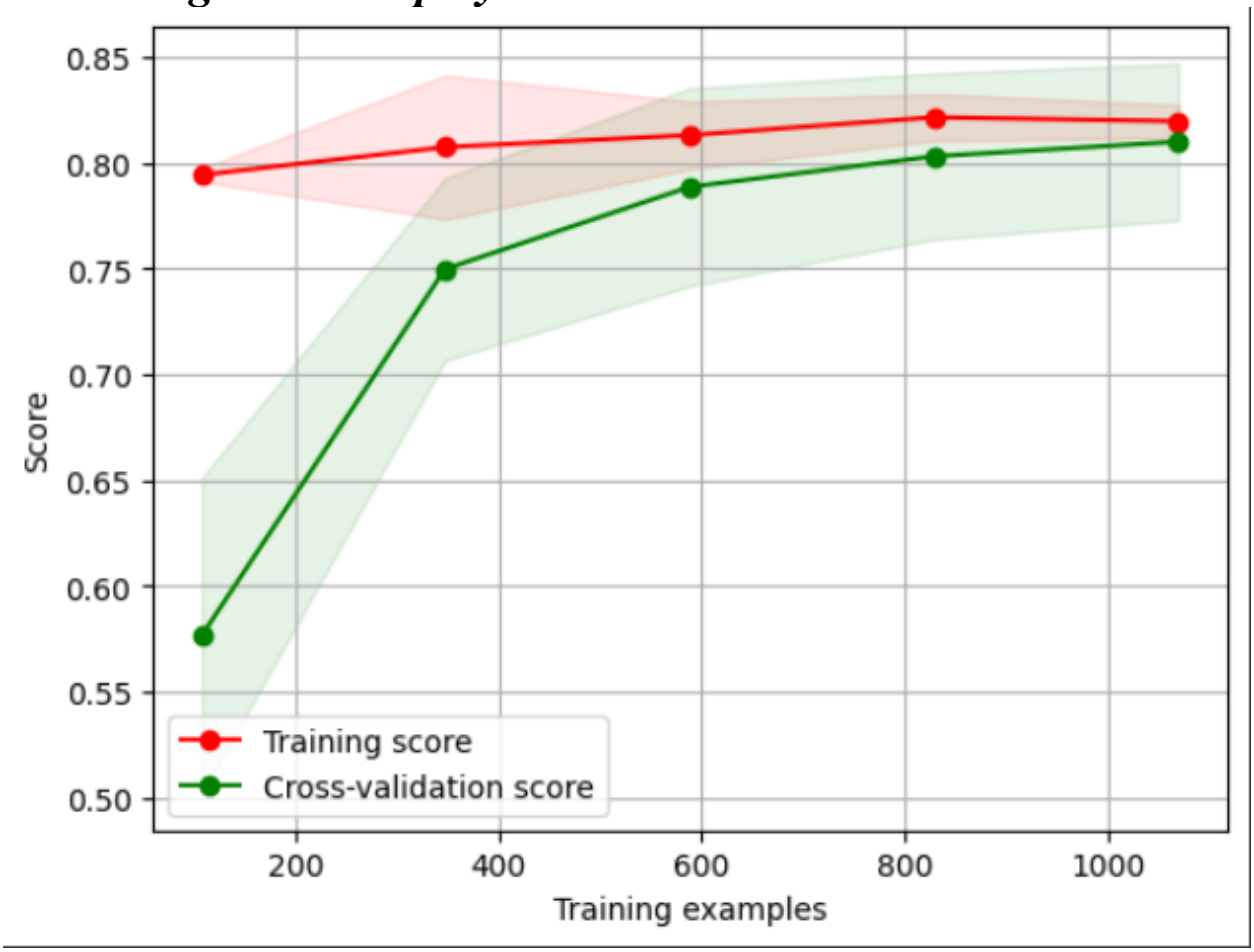
Models and their scores while training:
Linear Regressor



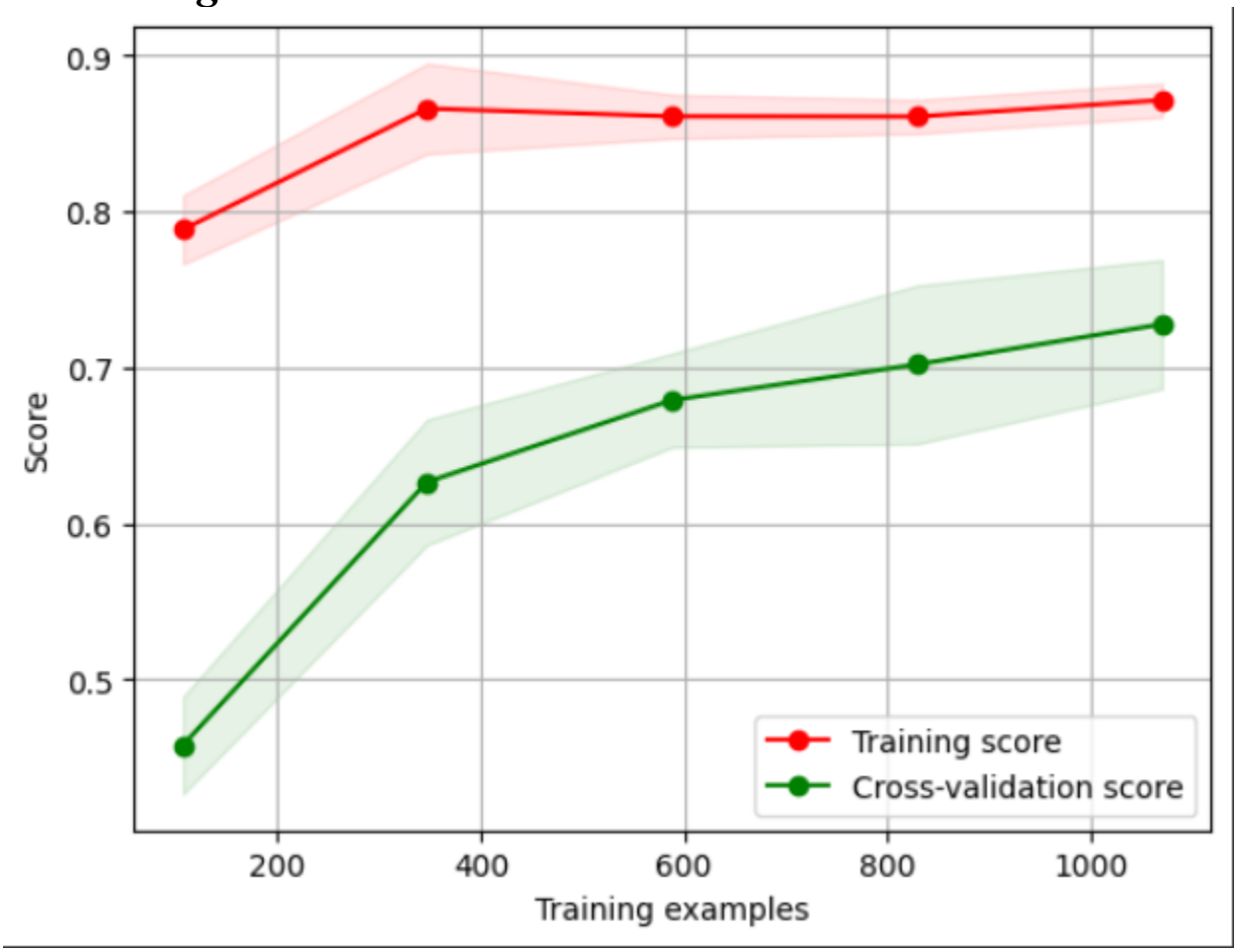
SVR using kernel = 'rbf'



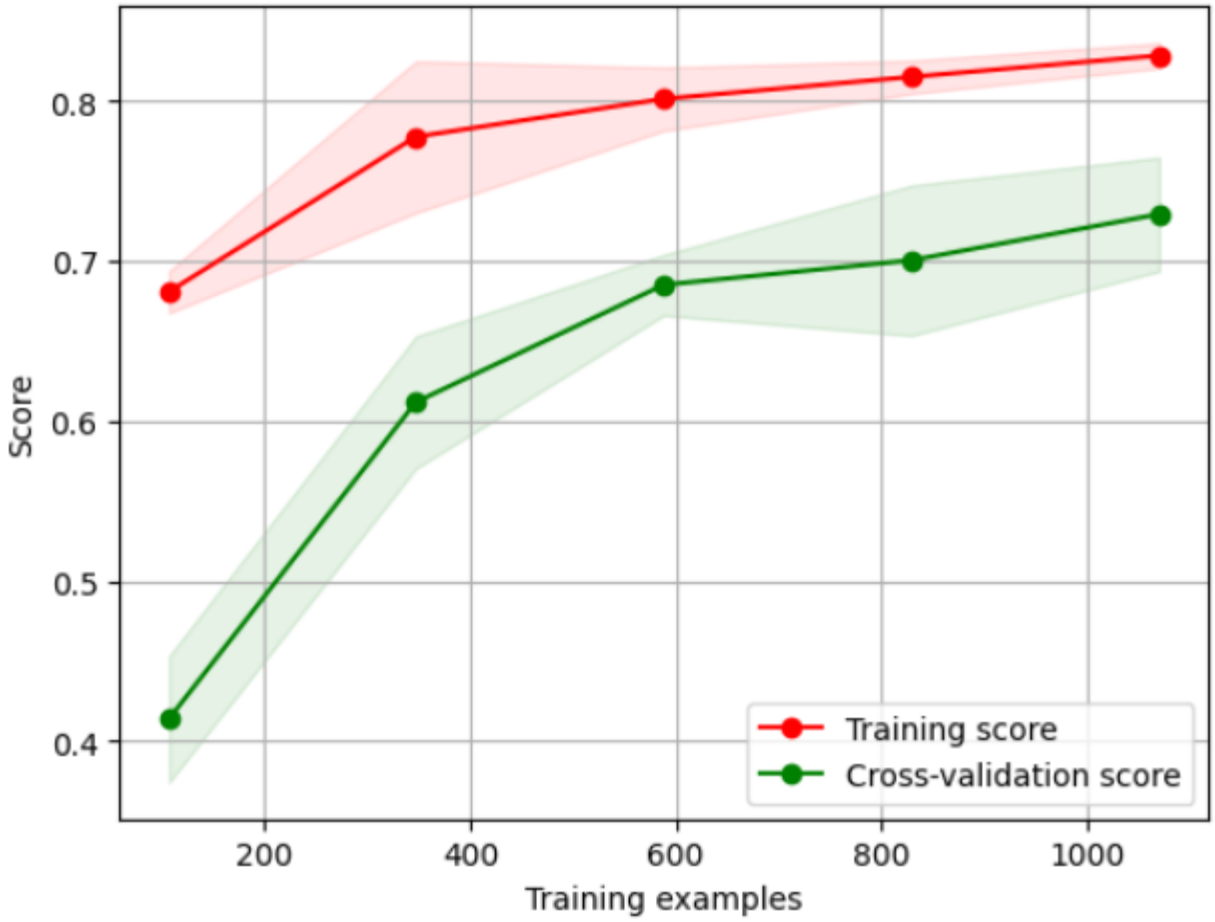
SVR using kernel = 'poly'



KNN using k = 3


0.9
0.8
0.7
0.6
0.5
Score
200
400
600
800
1000
Training examples
Training score
Cross-validation score

KNN using k = 5


0.8
0.7
0.6
0.5
0.4
Score
200
400
600
800
1000
Training examples
Training score
Cross-validation score