**BJ\_PUM\_Lab-5\_ Regression tasks in Orange**

1. **Orange provides several datasets suitable for regression tasks, including:**
   * *Housing*: This dataset contains information about housing prices and features such as number of rooms, crime rate, etc., making it suitable for regression analysis.
   * Concrete: This dataset includes various properties of concrete mixtures and their compressive strength, making it suitable for regression analysis.
   * Energy efficiency: This dataset contains features related to building energy efficiency and their heating and cooling load, suitable for regression analysis.
2. **Orange offers various widgets for creating regression models:**

**2.1 *For pure linear regression***: - Linear Regression widget: This widget allows you to perform linear regression analysis. You can specify the target variable and select additional options such as normalization and regularization.

**2.2** [Unfort, in Orange-3 there is no direct(!) Widget with Polynomial regression!]

***For polynomial regression***: - use *Formula widget*: This widget extends linear regression to polynomial regression (or others) by allowing you to use the degree of the polynomial. You can select a variable and the degree of the polynomial to fit the data. Then connect the Formula Widgets with Lin Regr. one.

By manually creating polynomial features and then using linear regression, you can effectively perform polynomial regression in Orange-3..

1. **Other useful widgets for regression analysis in Orange include:**
   * *Support Vector Regression (SVR)* widget: This widget allows you to perform regression analysis using support vector machines. You can adjust parameters such as kernel type, regularization parameter, etc...
   * *Principal Component Analysis (PCA)* widget: PCA can be used for dimensionality reduction before regression analysis to improve model performance and interpretability.
   * *Rank* widget,
   * *Random Forest* widget,
   * *N-Nets* widgets

You can then compare the performance of different variants of regressions using cross-validation or by evaluating the models Indices in Test&Score Widget. Keep in mind that higher degrees of polynomials can lead to overfitting, so it's essential to choose the degree that balances model complexity and performance on unseen data.

Use Data Table widget to see coefficients of the regression model(s).

The intercept coefficient in linear regression represents the constant term in the regression equation.

In the equation:

𝑦=𝐶+𝑐1⋅𝑋1+𝑐2⋅𝑋2+…*y*=*C*+*c*1⋅*X*1+*c*2⋅*X*2+…

* *y* is the dependent variable (the variable you are trying to predict),
* 𝑋1,𝑋2,… are the independent variables (the features used to make predictions),
* 𝑐1,𝑐2,… are the coefficients of the independent variables, and
* 𝐶 is the intercept coefficient, also known as the constant or bias term.

**Indices in *Test and Score* Widget:**

1. **Train**: MSE, This refers to the training dataset used to train the regression model .
2. **Test**: MSE, This refers to the test dataset used to evaluate the performance of the regression model.
3. **Mean Squared Error (MSE)**: MSE measures the average of the squares of the errors, where the error is the difference between the actual value and the predicted value.
4. **Root Mean Squared Error (RMSE)**: RMSE is the square root of the MSE, providing a measure of the spread of the errors.
5. **Mean Absolute Error (MAE)**: MAE measures the average absolute difference between the actual values and the predicted values.
6. **Mean Absolute Percentage Error (MAPE)**: MAPE expresses the error as a percentage of the actual value, providing a relative measure of the accuracy of the model.
7. **R-squared (R2)**: R2 represents the proportion of the variance in the dependent variable that is predictable from the independent variables

Formula: 𝑅2=1−(∑(𝑦\_𝑖−𝑦^\_𝑖)^2)/( ∑(𝑦\_𝑖−E𝑦)^2), where: Ey is the mean of the actual value.

* An R2 value of 1 indicates that the regression model perfectly fits the data, meaning that all variability in the response variable (dependent variable) is explained by the predictor variables (independent variables).
* An R2 value of 0 means that the regression model does not explain any of the variability in the response variable, and it essentially predicts the mean of the response variable for all observations.
* R2 values close to 1 indicate that a high proportion of the variability in the response variable is explained by the predictor variables, suggesting a good fit of the model to the data.

So, the closer the R2 value is to 1, the better the model explains the variability in the data and the better its predictive performance.

When the R2 value is negative, it usually indicates that the model fits the data worse than a horizontal line.

This can happen if the model is overfitting the training data, or if the model is not appropriate for the data at hand. It's essential to investigate further to understand why the R2 is negative and take appropriate steps to improve the model's performance, such as reducing model complexity, regularization, or using a different model altogether.