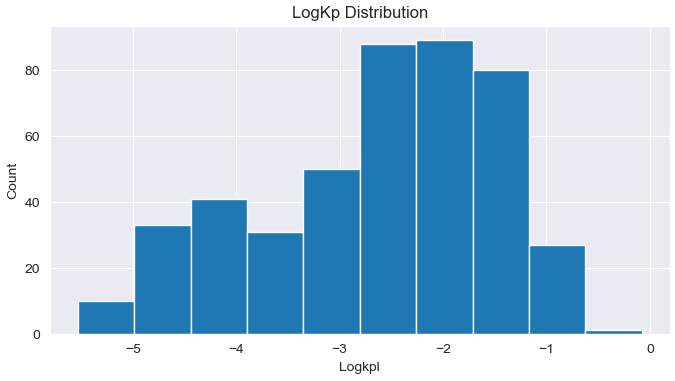
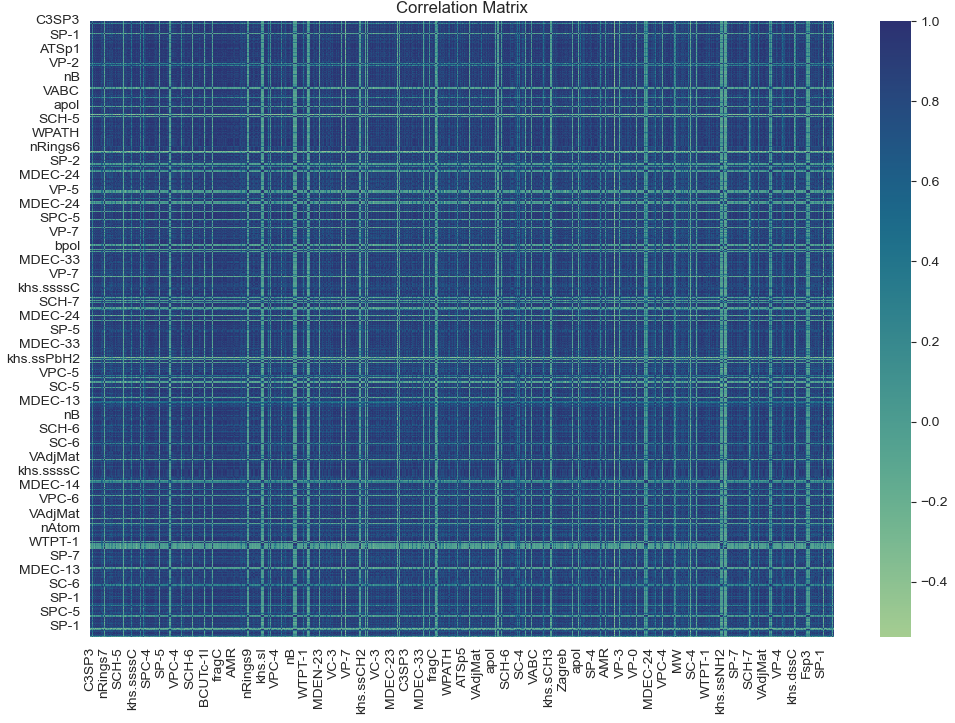
# Machine Learning

## Data

We used CDK descriptors 2.8 to generate descriptors for our compounds, salts are removed and molecules are neutralized. All the molecules are used excluding salts and water.

## Descriptive Analysis







This figure shows strong positive correlation between these two variables, meaning if one increases the other does.

Chart, scatter chart

Description automatically generated

This figure shows strong negative correlation between these two variables, meaning if one increases the other decreases.

## Data Cleaning

We dropped duplicated values, imputed missing values with the mean, and we have removed columns where correlation value more than 95% or less than -95% since this indicates in a multicollinearity problem we don’t want this in our features, so the processed data is 417 records with 151 features including our target variable “logkpl” with We split the data into train and test, where is test size is 20%.

## Models Comparison

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **MSE** | **MAE** | **MAPE** | **RMSE** | **R2** |
| Linear Regression | 15.85 | 1.11 | 0.56 | 3.98 | -12.89 |
| Random Forest | 0.29 | 0.39 | 0.18 | 0.54 | 0.74 |
| **XGBoost** | **0.26** | **0.35** | **0.16** | **0.51** | **0.77** |
| Gradient Boosting | 0.29 | 0.39 | 0.18 | 0.54 | 0.75 |
| CatBoost | 0.28 | 0.37 | 0.18 | 0.53 | 0.76 |
| LGBM | 0.29 | 0.38 | 0.18 | 0.54 | 0.75 |
| ANN | 0.48 | 0.48 | 0.24 | 0.69 | 0.58 |
| GNN | 0.55 | 0.66 | 0.34 | 0.74 | 0.07 |

## Feature Importance

