**Machine Learning Regression Analysis for Cyanine Dye Absorption**

**Summary:**

In this experiment, you will learn how to use machine learning (ML) regression analysis to predict the wavelength of maximum absorption of cyanine dyes. You will train and test different types of regression models—including simple linear regression, multiple regression, and regularized regression—and will learn how to visualize the results. You will assess model performance using quantitative metrics and will explore the concept of overfitting data. You will also learn about the processes of feature selection and evaluating feature importance in regression models. Finally, you will use your regression analysis for hypothesis generation, gaining insight into features that affect cyanine dye absorption. This lab will also give you practice reading, modifying, and writing code in a Mathematica notebook.

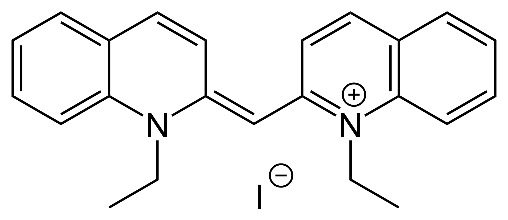
In an optional extension of this experiment, you will explore advanced topics including feature engineering, tree-based regression models, and feature importance evaluation in non-linear regression models.

**Theoretical Background:**

Absorption Spectra of Conjugated Cyanine Dyes:

*UV-visible spectroscopy* measures the absorption of light in the ultraviolet (UV) and visible portions of the electromagnetic (EM) spectrum. These absorption events arise from *electronic transitions* in molecules, in which an electron is excited from a lower-energy to a higher-energy state. These electronic states are described by *molecular orbitals (MOs)*, which are analogous to atomic orbitals but can span the entire molecule. The lowest energy, and thus longest wavelength, transition occurs when electrons are excited from the *highest occupied molecular orbital (HOMO)* to the *lowest unoccupied molecular orbital (LUMO)*. Changes in the molecular structure that raise the energy of the HOMO and/or lower the energy of the LUMO will cause the electronic absorption to shift to longer wavelengths.

*Cyanine dyes* are a class of molecules containing a conjugated *polymethine chain*, consisting of alternating single and double carbon-carbon bonds, bounded by two nitrogen atoms, one of which is an *iminium* cation. An example is 1,1ʹ-diethyl-2,2ʹ-cyanine (shown with an iodide counterion):



Cyanine dyes absorb strongly in the visible portion of the EM spectrum and thus are brightly colored. Due to their optical properties, cyanine dyes are widely used in the life sciences and biotechnology fields, including as labels in single-molecule fluorescence microscopy.

Machine Learning:

*Machine learning* algorithms use statistical tools to “learn” from an existing dataset in order to make predictions about new data. There are a variety of machine learning applications outside of chemistry, ranging from image analysis to language translation to medical diagnosis. Machine learning has also become increasingly prevalent in chemistry, with applications in spectroscopy, materials discovery and property prediction, calculation of potential energy surfaces and force fields, and microscopy, among other areas. (A recent overview of machine learning in chemistry is available: <https://pubs.acs.org/doi/book/10.1021/acs.infocus.7e4001>)

As the first step in the analysis workflow, a machine learning model must be *trained* using *training data*. This training dataset must be sufficiently large and diverse to allow the model to generalize well to new data. The performance of the model can then be assessed by the accuracy of its predictions for an independent set of *test* *data*. There are two broad categories of machine learning algorithms: *supervised* and *unsupervised*. In supervised learning, the training dataset contains both the underlying data and the known labels or values for each data instance. A supervised machine learning model can then be trained to predict the label or value for new data. In unsupervised learning, the goal is to identify the underlying structure in the dataset, which is useful for data that does not contain labels or for which the appropriate labels are unknown. The supervised learning category encompasses both *regression* and *classification* models. A classification model divides data into one or more distinct groups. In contrast, a regression model predicts a numeric target value for each data instance.

Regression analysis seeks to predict the value of a *dependent variable* (or *target*) using one or more *independent variables* (or features). The simplest type of regression model is a *simple linear regression* with a single independent variable, in which the relationship between the dependent variable and the independent variable is modeled as a line. In the *ordinary least squares* approach, the slope of this line is determined by minimizing a cost function, such as the *residual sum of squares (RSS)*, that measures the difference between the predicted and actual values. The resulting slope is referred to as a *coefficient* or *weight* of the independent variable. In a system with two or more independent variables, this approach can be generalized to a *multiple regression* model, with a distinct slope or coefficient for each feature.

Basic linear regression models can fail for datasets containing too few instances of training data and/or too many features. We say that a model is *overfitting* when it is too specific to the training data and fails to generalize well to new data; mathematically, this scenario corresponds to containing more parameters than justified by the data. One approach to improving model performance is to use *regularized* (or *penalized*) regression models, which have the effect of minimizing model complexity. Two common penalized regression models are *ridge regression* (*a.k.a.* L2 regularization) and *lasso regression* (*a.k.a.* L1 regularization). Other approaches for improving model performance are *feature selection*, in which only the features most correlated with the target variable are used in the analysis, and *feature engineering*, in which combinations of features are generated from the original list of independent variables. Finally, there are alternatives to linear regression models, including *tree-based* models in which predictions are structured as a tree with nodes, branches, and leaves. Two such models are the *Decision Tree* algorithm, which is simple but more prone to overfitting, and the *Random Forest* algorithm, which builds a collection of Decision Trees to improve performance and minimize overfitting.

The performance of regression models can be evaluated using qualitative and quantitative approaches. Qualitatively, the actual vs. predicted values of the dependent variable can be plotted to visualize model performance. In addition, performance can be evaluated quantitatively using different metrics. This experiment implements four different metrics to assess model performance: *mean absolute error (MAE)*, *mean squared error (MSE)*, *root-mean-square error (RMSE)*, and the *coefficient of determination* (*R*2). These metrics are described in more detail in the computational notebook.

In addition to predicting the value of the target variable for new instances of data, regression analysis can be used to gain insight into the system being studied. One approach is to assess the relative weight of the different features incorporated into the model, or the *feature importance*. In linear models trained on normalized data, the coefficient of each feature reports directly on how much that feature affects the final model. Evaluating feature importance can be more challenging in non-linear models. One model-agnostic approach for this task is *SHapley Additive exPlanations (SHAP)*, which can be used to interpret different types of regression models. Evaluating which features contribute most to the model, or which features are best correlated with the target variable, can ultimately be used in the process of *hypothesis generation*, coming up with explanations for the behavior observed in the system. The final goal of this experiment is to combine your chemistry intuition with interpretation of your regression models to come up with hypotheses for what factors govern cyanine dye absorption.

Another brief introduction to machine learning for chemistry students can be found here: <https://chem.libretexts.org/Courses/Intercollegiate_Courses/Cheminformatics_OLCC_(2019)/8%3A_Machine-learning_Basics/8.1%3A_Machine_Learning_Basics>

Dataset and Implementation:

The target variable in this regression analysis is the wavelength of maximum absorption of different cyanine dyes, referred to as *MaxAbsorbanceWavelength* in the dataset. The total dataset consists of 147 unique molecules. The wavelength of maximum absorption for each dye was determined computationally using a ML-based package called UVVisML (<https://doi.org/10.5281/ZENODO.5986671>). To predict cyanine dye absorption in the regression analysis, 13 different molecular features were computed; these features are described in the computational notebook. In this experiment, the 147 cyanine dyes are split into training (80%) and test (20%) datasets. Before being used to train and test regression models, outliers are removed from the dataset and the feature values are scaled; these *preprocessing* steps can improve model performance.

This experiment is implemented in the technical computing software *Mathematica*, which is a popular general-purpose programming language that contains built-in tools for machine learning and regression analysis. The computational notebook will introduce you some of the basic features of Mathematica as well as to some basic programming concepts, which are independent of programming language. Mathematica tutorials are readily available, including on the Wolfram website (<https://www.wolfram.com/>), and two recommendations are provided in the “Getting Started” section of the notebook.

**Protocol:**

Overview:

The protocol is divided into two sections, both of which are included in the “Cyanine Dye Regression Tutorial Student Handout” notebook:

1. *Part I* walks you through the workflow of performing different types of linear regressions, visualizing the results, implementing quantitative performance metrics, feature selection, and the evaluation of feature importance. Finally, you will use the results of your regression analyses to generate hypotheses about what factors govern cyanine dye absorption. Please complete this section during the lab period.
2. *Part II* introduces more advanced topics, including feature engineering, tree-based regression models, and the evaluation of feature importance in non-linear regression models. This section is optional.

Part I:

1. Open the “Cyanine Dye Regression Tutorial Student Handout” notebook in Mathematica
   1. If needed, first work through the “Getting Started” and “Basics of Mathematica” section to get familiar with Mathematica
   2. When prompted, allow Mathematica to run the initialization cells in the notebook
2. Work through the sections in Part I of the notebook, following the prompts to add new code blocks or modify the analysis:
   1. Run the “Get and Preprocess the Data” section to import, explore, and preprocess the dataset to be used in this analysis
   2. Run the “Building Regression Models” section to train and test simple linear regression and multiple regression models
   3. Run the “Model Evaluation” section to calculate different performance metrics for the regression models
   4. Run the “Avoiding Model Overfitting” and “Feature Selection” sections to implement Pearson Correlation analysis and feature selection
   5. Run the “Regularized (or Penalized) Regression” section to train regularized regression models
   6. Run the “Feature Importance” section to evaluate the contribution of different variables to linear regression models
3. When you reach the “Gaining Insight from Regression Models” section, look back through your analyses and try to formulate hypotheses for factors that might affect cyanine dye absorption. Discuss your hypotheses with your lab group and the instructor.

Part II:

1. Work through the sections in Part II of the notebook, following the prompts to add new code blocks or modify the analysis:
   1. Run the “Feature Engineering” section to generate new features and test their performance in the regression model
   2. Run the “Tree Models” section to train and test two tree-based regression models, Decision Tree and Random Forest
   3. Run the “Feature Importance using SHapley Additive exPlanations (SHAP)” section to implement SHAP analysis for feature importance evaluation of the tree-based regression models
2. Optional: Try something different! Test different preprocessing steps (such as outlier removal or data scaling) or test changes to the regression analysis (such as a different model or different hyperparameters). How does the performance change?

**Discussion Questions:**

Part I:

1. Compare the performance of all tested models with all tested parameters:
   1. Show a table with the performance metrics for all tested models and parameters
   2. Plot the actual vs. predicted wavelength for all tested models and parameters (you can use multiple plots if needed)
   3. Do the quantitative metrics generally agree with the “eye test” for model performance?
2. Assess model performance:
   1. Which model performed best? Which model performed worst?
   2. Which single-variable simple linear regression model performed best? Was this result expected or unexpected?
   3. How did the regularized regression models compare to the multiple regression model without regularization?
3. Discuss the feature selection and feature importance analysis:
   1. Which feature(s) were best correlated with the wavelength of maximum absorption?
   2. Which feature(s) had the greatest importance in the regression models?
   3. Did the best correlated and highest-impact features agree or did they differ?
4. Use your regression models for hypothesis generation:
   1. Which features in the regression analysis does your chemical intuition say could be responsible for differences in cyanine dye absorption? Which features seem less important?
   2. Does your chemical intuition agree with the features identified by the Pearson Correlation and feature importance analysis?
   3. Synthesizing all these factors, can you formulate a hypothesis for what governs cyanine dye absorption?

Part II:

1. Add the performance metrics and plotted performance of your new regression models to the table and plot(s) from Part I
2. Discuss the results of the feature engineering section:
   1. What new feature did you create? What was your reasoning for creating that feature?
   2. How did the simple linear regression using this new feature compare to other tested models?
3. Discuss the tree-based regression models:
4. How did the Decision Tree and Random Forest model performance compare to the linear regression models?
5. What features were important in the Decision Tree and Random Forest models according to SHAP analysis? How did these features compare to the ones that contributed most to the regression models?
6. Optional: If you tried something different in the analysis, discuss the results! What did you learn?