Math 402 Final Project Due date: May 8, 2025

The due date of the final project is May 8, 2023. Include your Python code and a short report of your data analysis in your submission. We meet in class 9am-10am Friday May 9, 2023 for presenting your findings. Each of you can have a 5–8-minute presentation.

For the project, you need to manually split the data as training data and test data. **Build your best model on the training data** (using SelectKBest or another method you choose and k-fold cross validations) and **test their performances on the test data**. Does your best model on the training data still works best for the test data? To make your research reproducible, use the following code to split the data.

```
from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler

X = df.drop("response", axis=1)

y = df["response"] #df is your data and response is the name of the
response variable

seed=100

X_train, X_test, y_train, y_test = train_test_split(X,y,
test_size=0.3, random_state=seed, shuffle=True)

scaler = StandardScaler()

X_train = scaler.fit_transform(X_train)

X_test = scaler.transform(X_test)
```

If you want to check interaction effects, limit the interactions effects to two variables only and use a smaller significance level like 0.001 for t-tests.

Question. Classification using the HCV data set:

https://archive.ics.uci.edu/ml/datasets/HCV+data

The target attribute for classification is Category (blood donors vs. Hepatitis C (including its progress ('just' Hepatitis C, Fibrosis, Cirrhosis).

Methods to be considered include K-Nearest neighbors, multinomial logistic regression (LogisticRegression(multi_class='multinomial'), linear discriminant analysis, quadratic discriminant analysis and naive Bayes.