Data Science (Fall 2025)
Dr. Carlos Monroy
Due date: 11/5/25 @ 11:59PM

Assignment, comparing Weka and scikitlearn for building a Linear Regression model using the cars dataset and practice good formatting using Markdown. To be submitted via GitHub submission. Start early!

Include screen shots to support your answers and provide detailed responses with thoughtful insights.

To be done individually.

1. Create a Jupyter Notebook named linear_regression_weka_scikit.ipynb using as template the linear regression example found here:
   https://github.com/computingcelts/f25-ds-examples/blob/main/src/plot_ols.ipynb

   Modify the code to load the cars dataset, found here:
   https://github.com/computingcelts/f25-ds-examples/blob/main/datasets/cars.csv

2. Add a section called Part 1. Scikit Learn Multiple Linear Regression. Add your name on the first markdown cell.

3. Build a regression model to predict the "class" which is numeric. Remember that linear regression works only with features that are numeric, hence you have to think what to do with the categorical features.

4. Plot some descriptive stats of the dataset. Explain what the stats tell you about the dataset.

5. Run 10-fold CV and evaluate your model, how well it works.

6. What do the coefficients tell you about the features?

7. Add a section called Part 2. Multiple Linear Regression in Weka. For each of the next points, whenever there is a double asterisk that means you have to upload that image to your Jupyter Notebook along with some narrative explaining each point (this should be in a Markdown cell).

8. Now on your laptop Open Weka.

9. Load the cars dataset (import as csv). Make sure to select csv in the file type drop down (otherwise Weka complains). You might need to modify the header of the file to comply with Weka's format in case of issues. **

10. Remove the categorical features. Why do you have to do it? **

11. Change the class feature to be the one labeled class (which is numeric). **

12. Run the Multiple Linear Regression model. Doing a 10-fold CV, make sure to print the output of each of the 10 folds. **

13. Analyze the results. What do the results tell you? How well does the model perform? **

14. Describe in detail, what is the meaning of the coefficients, and how they explain the results? **

15. Since you are predicting numeric values, how evaluating the model compared with the Bayes Classifier you did in scikitlearn?

16. Compare your experience using Weka and scikitlearn. What are the pros and cons, which one is better and why?

Commit and push your JN to GitHub! Triple verify everything is fine.