ML Algorithms with Python Assignment

General Information:

This file introduces the topic for ML Algorithms in Python assignment. In this assignment, you will be required to write a report on the given topic.

Deadline:

ML Algorithms in Python Assignment **due date is November 24th @ 11:59 PM**.

Requirements:

- **Please attach the link to your Colab notebook's Python file that demonstrates all the procedures you used to extract the given information in the first page of your Word document. Include brief explanations for each task using comments within the Python file. Please also include the screenshots of your Python codes at the end of your assignment's Word document.**
- Please follow APA style in your paper for all the parts of your essay. You can find APA information on your course page.
- Please support all the provided information from other sources with appropriated references.
- Search for best practices to write different part of your paper. For example, for the Introduction part, you can click on the following link and watch that short video: https://www.youtube.com/watch?v=FTC-5P1VFFU
- The report **should not exceed 1500 words** (not including the title page or reference page). There is no minimum word count. The key is whether the essay thoroughly covers the provided topic.
- Turnitin score of less than 10% is required. Note that Turnitin does not count quoted material and References section.
- Providing the proof of AIgarism at the end of your assignment file is required (for each page of your paper separately). You can find more information about how to provide the proof of AIgarism in "AI Policy" file on Brightspace.
- Please only use Python for this assignment.

Research Topic:

Suppose you are a data analyst at an insurance company. Your role involves estimating the premium each customer should be charged based on their smoking habits, personal information (age, gender, BMI), family circumstances (number of children), and geographic location.

Drawing from the knowledge you have accumulated, please analyze the data provided and write a report that is both descriptive and predictive to reflect your findings.

Start by conducting a concise descriptive analysis of the data. The goal of this descriptive analysis is to understand the data provided and achieve greater accuracy in the predictive analysis phase. Next, clean the data using the techniques we've covered. Use your creativity and knowledge to identify the most accurate regression model for forecasting insurance charges. You can use various models such as multiple linear and non-linear regression, k-nearest neighbors, decision tree, random forest, piecewise regression, and piecewise regression tree. You also have several techniques at your disposal to enhance the model's accuracy, including feature selection, dummy variables, and data normalization.

When setting up your regression model, ensure that you shuffle the data first, then divide it into training data (80%) and testing data (20%). Train your regression model with the training data and subsequently evaluate it using the testing data. It is essential to assess your regression model with the evaluation metrics discussed in class. Also, work to prevent overfitting in your analysis by comparing the prediction accuracy on both the training and testing data sets.

Guidance for Improving Regression Analysis Performance:

For your reference, I conducted a regression analysis on the entire dataset (without dividing it into training and testing sets) and achieved an R^2 value of 0.87 and an RMSE of 4352. In my analysis, I used dummy variables to account for categorical variables and applied piecewise regression. I divided the data into eight distinct categories and developed a linear regression model for each. These individual linear regressions were combined into one overarching linear regression model using the IF condition.

I allocated one hour to work on this file. Notably, my analysis only used multiple linear regression. I did not consider nonlinear regression or other regression models. By creating better categorizations for your regression models, incorporating other regression techniques, or improving data cleaning, you might further enhance the performance.

Topics	**Max. Grade**
Assignment	
Quality of Writing	**40**
• Key elements of assignments are covered. • Clarity and Cohesion: Report is organized, well-written, and easy to understand. Ideas flow smoothly. • Conducts appropriate analyses. • Provides insightful commentary on the findings, demonstrating a clear understanding of the data. • Follows the appropriate structure/template for the report. • Uses relevant and reliable references to support the provided information. • Provides appropriated graphs to illustrate the findings.	
Use of Python for Data Analysis	**10**
• Appropriate Python commands are used effectively for data extraction and analysis. • Clear and concise comments are provided for each Python task performed.	
Descriptive Analysis	**15**
• Handling Missing Values & Outliers: Effectively handles missing values and outliers in the data. Justification is provided for the methods used. • Descriptive analysis should help to achieve better results in the predictive analysis. • Additional Analysis: Demonstrates creativity by investigating beyond the basic assignment requirements. Asks and answers interesting questions based on the data.	
Predictive Analysis	**25**
• Find the appropriate (and accurate) regression model for the provided case study (using Python). • Creativity in the modeling procedure. • Comparing the results of the considered models. • Investigating the overfitting.	
APA 7th	**10**
• Title page is present and properly formatted – separate page (the discussion titles are not required to appear on a separate page). • Reference page is present and properly formatted – separate page. • Citations/reference page follow APA guidelines. • Properly cites ideas from other sources.	
Note	
Please note that the final score will be reduced by 10 percent if the assignment is not submitted by the due date or if proof of AIgarism is not provided as required. Scoring for each section will be based on the quality and depth of the work in that area, with highest marks reserved for exceptional performance that goes beyond the basic requirements.	