Caregivers and Machine Learning 2023
Homework assignment 3

Deadline: Monday April 10th at 11:59pm
Submission: Submit through the `https://learn.vectorinstitute.ai` course page.

- Submit your answers to questions 1 and 2 as a PDF file titled "`name_hw3_writeup.pdf`".
- Submit your code for question 3, as a Jupyter notebook file titled "`name_hw3_code.ipynb`"

Problem 1: [25 pts] Conceptual understanding

(A) **[5 pts]** What is the advantage of using a bagging algorithm like Random Forest over a single Decision Tree?

(B) **[5 pts]** What is the purpose of a loss function in machine learning?

(C) **[5 pts]** How does an outlier in a small data set affect a linear regression model? How about a large data set?

(D) **[5 pts]** What is an example of an operation where using a vector is faster than using for loops? Explain why.

(E) **[5 pts]** Why is the Sigmoid Function used in Logistic Regression?

Problem 2: (Bonus) [10 pts] Theory

(A) [**5 Pts**] Effects of correlation in bagging.

In the analysis of bagging, the m bootstrapped datasets drawn from p_D are not independent and we anticipated that we don't get the $1/m$ variance reduction in the predictions. Show that if the sampled individual predictions y_i have variance σ^2 (assuming $\sigma^2 > 0$) and correlation coefficient ρ among each other, then

$$\mathrm{Var}\left(\frac{1}{m}\sum_{i}^{m} y_i\right) = \frac{1}{m}(1-\rho)\sigma^2 + \rho\sigma^2 \tag{1}$$

Hints:

- The correlation coefficient is $\rho(x, y) = \frac{\mathrm{Cov}(x,y)}{\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}}$.
- For a sum of random variables: $\mathrm{Var}(\sum_i^m x_i) = \sum_i^m \mathrm{Var}(x_i) + 2\sum_{i<j} \mathrm{Cov}(x_i, x_j)$.
- The covariance is $\mathrm{Cov}(x, y) = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$.

(B) [**5 pts**] Entropies.

Let (X, Y) have the following joint distribution:

X Y	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

(a) Compute the marginal distributions $p(X)$ and $p(Y)$.

(b) Compute the entropies $H(X)$ and $H(Y)$.

(c) Evaluate the specific conditional entropies $H(X|Y = i)$. for all values of $i = 1, 2, 3, 4$.

(d) Compute the conditional entropies $H(X|Y)$ and $H(Y|X)$.

(e) Compute the joint entropy $H(X, Y)$.

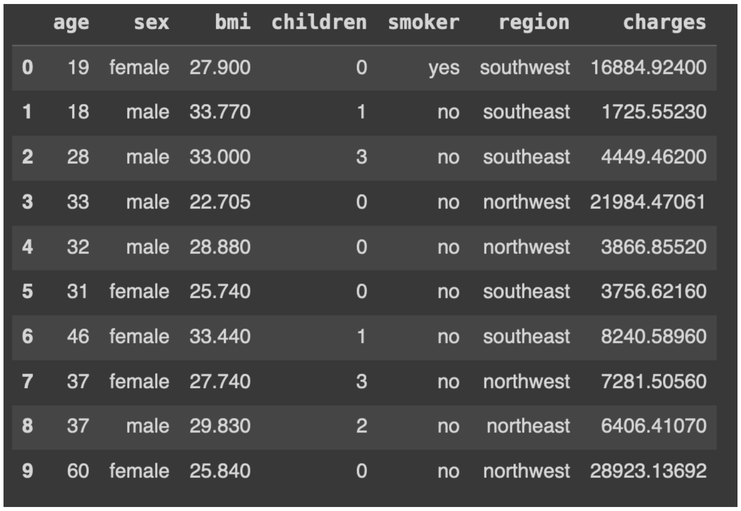
Problem 3: [25 pts] Linear Regression and Logistic Regression

This coding homework is adapted from an assignment by Alex Yun for the AI for Clinician Champions course.

In this coding homework, you will learn some basic data visualization and preprocessing techniques using Python and how to apply regression models to a medical cost data set in two contexts: (1) regression task and (2) classification task. The data set we consider consists of the following variables or features:

- `age`: age of the primary beneficiary
- `sex`: sex of the beneficiary (male or female)
- `BMI`: body mass index; a value derived from the mass and height of the beneficiary
- `children`: number of children covered by the insurance
- `smoker`: whether the beneficiary smokes or not (yes or no)
- `region`: Residential area of the beneficiary in the U.S.
- `charges`: individual medical costs billed by the insurance

Here is an example of the top 10 rows in the data set:



	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
5	31	female	25.740	0	no	southeast	3756.62160
6	46	female	33.440	1	no	southeast	8240.58960
7	37	female	27.740	3	no	northwest	7281.50560
8	37	male	29.830	2	no	northeast	6406.41070
9	60	female	25.840	0	no	northwest	28923.13692

Using the features in this data set, you will develop:

- A linear regression model to predict medical costs (“`charges`”)

- A logistic regression model to predict whether the individual is a smoker or a non-smoker ("`smoker`").
- Derive appropriate performance metrics to evaluate both the linear regression and logistic regression models.

Here is a link to the Google Colab notebook: `https://colab.research.google.com/drive/10hI27OZeaqP3gnjWR74C0jrgrSoN_TE4?usp=sharing`.

First, save a copy of the file in your own Google Drive (*File* → *Save a copy in Drive*), then run each cell from the top, read the explanations, and solve the following questions in the notebook:

(A) **[4 pts]** Data Visualization

(a) Plot the distribution of the independent variables: "`sex`", "`smoker`", and "`region`".

(b) Plot the relationship between the independent variables above and the dependent/target variable "`charges`".

(B) **[5 pts]** Metrics to evaluate a regression model.

Calculate the Mean Squared Error (MSE), Mean Absolute Error (MAE) and the coefficient of determination (R2) values using Python's math library, based on the equations shown below.

$$\text{Mean Squared Error } (MSE) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

$$\text{Mean Absolute Error } (MAE) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|^2 \tag{3}$$

$$R^2 = 1 - \frac{\text{RSS}}{TSS}, \quad \text{where} \tag{4}$$

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad TSS = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{5}$$

and where y_i is the true label, $\hat{y}_i$ is the prediction and $\bar{y}$ is the mean of the observed data.

(C) **[8 pts]** Build a Logistic Regression Model.

(a) Split the data set into train and test sets. Retain 10% of the data for the test set, and set the random state to '0'.

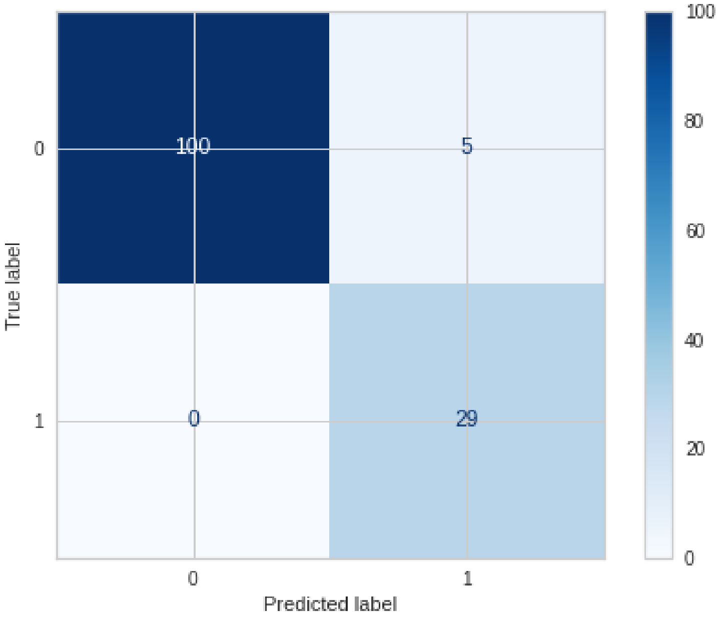
(b) Instantiate a Logistic Regression model and name it "`logit_model`"
.

(c) Fit the logistic regression model using the training data.

(D) **[8 pts]** Metrics to evaluate a classification model.

Given the information in the confusion matrix below, calculate the Precision, Recall and F1 score for this model using Python's math library.

A **Confusion Matrix** compares the actual target values (true label) with those predicted by the machine learning model (predicted label). Hint:



- True Positive (tp): Both the predicted label and true label are 1
- False Positive (fp): The predicted label is 1 but the true label is 0
- True Negative (tn): Both the predicted label and true label are 0
- False Negative (fn): The predicted label is 0 but the true label is 1

Common Evaluation Metrics for Classification Models:

$$\text{precision} = \frac{tp}{tp + fp} \tag{6}$$

$$\text{recall} = \frac{tp}{tp + fn} \tag{7}$$

$$F_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{8}$$

- **Precision**: "Precision" is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is: Of all individuals that were predicted to be smokers, how many are actually smokers?

- **Recall**: ”Recall” is the ratio of correctly predicted positive observations to all observations in the actual class. The question recall answers is: Of all the individuals that were actually smokers, how many did we predict to be smokers?
- **F1 score**: ”F1 Score” is the weighted average of Precision and Recall.