**Executive Summary and Implications**

A.  Develop an executive summary using your data and results from task 2. The summary should be written for a technical audience of your data-analytics peers or members of your team and should include each of the following:

•   a statement of the problem and the hypothesis

In this analysis the problem that we are having is how much does customer cost for health insurance affect them based off different variables. The hypothesis is that the age of customers is statistically significant correlation with the cost of health insurance.

•   a summary of the data-analysis process

The data analysis process will start with checking for any missing or null values and displaying the sum for each variable. I then checked the data types for each column in the dataset, which is then followed by getting the statistics. These statistics will show the quartile, min/max, standard deviation, mean, and median of the numerical values. Next, I performed EDA on the dataset, where I looped through the columns to show all the unique values or unique count. Also, created a new data frame that holds only numerical values to loop through and plot histograms with seaborn. Furthermore, for categorical columns I used count plots in seaborn to show the counts for each value.

Once all exploratory data analysis was done, I checked if the p-values and significance for all columns using chi-testing. I also set an initial p-value at 0.05 and compared all variables p-values to the initial to check if a column was dependent “to reject the null hypothesis” or independent “fail to reject the null hypothesis”. Conclusion of the tests shows that all variables are independent, and we accept H0 “variables do not have a significant relation”.

Next, I started converting the categorical values to 1,0 for the regression test. Then split the data for 70% train, 30% test, and setting a random state. I then set up 4 different regression model that included linear, random forest, gradient boosting, SVR. This allowed me to check which model was the best model for the analysis. Furthermore, I checked which model was the best by looking at the actual prediction of the y test vs the y predictions, I then visualized the actuals vs the predictions with a plot graph. Those methods were still kind of hard to distinguish the best model, so I looked at the r2 score for each model, as well as seeing which model produced the lowest mean absolute error.

Finally, after getting the best model, gradient boosting, I then tested the model on unseen data from a data frame I created. Once, I got the prediction I then trained the whole dataset and stored the model using joblib. Lastly, I recalled the model that was previously saved in joblib, predicted the data frame that I had created before on the unseen data to compare the results, which resulted in the same value.

•   an outline of the findings

* Median age is 39
* Average charge is 13,270
* Max charge is 63,770
* Average BMI is 30.6
* Median children is 1
* Data contains less smokers than non-smokers
* More male smokers that female smokers
* The southeast region has the most smokers

•   an explanation of the limitations of the techniques and tools used

The limitations on the tools used was the choice of environment that I used. Since most of the analysis was done in Google Colab because its simplicity of setting up, importing libraries, ease of use, and ability to save the notebook in Google Drive, also, presented some drawbacks. The main limitation was that Google Colab does not allow GUI in the application because it is a cloud base environment. This limitation stopped the workflow in Google Colab, which then I had to use PyCharm ide to run Tkinter and finish the work.

•   a summary of proposed actions

After this analysis stakeholders can trust this model with an r2 score of 86.9% and a MAE 2428.6. The proposed actions is that this product/model is ready from production and can give its customers the best estimates of cost of health insurance.

•   expected benefits of the study (be as specific and quantitative as possible)

The expected benefits are to help new and existing customers determine what cost they will

be paying for health insurance. This information can help customers plan financially,

budgeting, or plan accordingly for health insurance “Job Health Insurance”. The hypothesis for this analysis is that a customer’s age is statistically significant correlation with the cost of health insurance.

B.  Present your Capstone project to a non-technical audience. Use Panopto to record yourself delivering a presentation that includes slides displayed by a multimedia tool (e.g., PowerPoint, Keynote, Prezi). Your recording should capture your multimedia presentation. Your presentation should also demonstrate appropriate communication skills for the audience.

Note: appearing on camera is optional.

1.  Cover in your presentation each of the following points:

•   a brief introduction of yourself and your background

•   a statement of the problem and the hypothesis

•   a summary of the data-analysis process

•   an outline of the findings

•   an explanation of the limitations of the techniques and tools used

•   a summary of proposed actions

•   expected benefits of the study (be as specific and quantitative as possible)

Link to Tableau workbook:

<https://public.tableau.com/views/Insurance_16464209849310/Dashboard1?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link>

C.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

None used.

D.  Demonstrate professional communication in the content and presentation of your submission.