Insurance claim analysis: demographic and health

Authors: Maria M, Zhuoyi Zhao, Xuqing Wu

Summary of questions and results

Q1: What is the estimated non-linear regression of the insurance claim? (Use DecisionTreeRegressor) How does it compare to the linear model?

Using a non-linear regression model, this research question seeks to examine the link between the independent variables and the insurance claim. To identify the most important predictors of insurance claims and to forecast the anticipated claim amount for a given collection of predictor variables, the DecisionTreeRegressor model will be employed.

Result: We created a regression tree model and a linear regression model, and the LASSO regression under the best value of penalization using cross validation. Here are the MSEs we calculated for each model, and coefficients on features for the best LASSO model we could generate.

```
'''
Tree MSE test set: 62462237.38
Linear MSE training set: 43396839.03
Linear MSE test set: 41346465.04
best value of penalization: 40.48
[(220.67777789171802, 'age'), (1973.9114357246187, 'bmi'), (2578.4219804815343, 'bloodpressure'), (696.9285432747407, 'children'),
(-0.0, 'gender_female'), (15.758597983817875, 'diabetic_No'), (-20670.70474042754, 'smoker_No'), (1918.4386309856318, 'region_northeast'),
(0.0, 'region_northwest'), (-477.0371559518968, 'region_southeast')]
best lasso model mse: 41401657.36
'''
```

Q2: What is the average insurance claim of male and females in various regions? How do insurance claims vary with BMI for each gender?

This research question aims to explore the relationship between gender and insurance claims. The research question will investigate the average insurance amount of gender in different regions, as well as the BMI for each gender. The findings will help to identify any trends and correlations between health status and insurance claims in gender.

Result: From the output we got, we see that the average insurance claim of northeast is highest in the region for both male(30.1%) and female (31.9%). For the male, the average insurance claim of northwest is lowest. But for females, the average insurance claim of southwest is lowest. The scatter plot shows that gender has a minor effect on the relationship between BMI and insurance claims, as indicated by the slight difference in slope between the two OLS trendlines. For females, the OLS trendline equation is claim = 297.639 * bmi + 3527.38 (R-squared = 0.026152). For males, the OLS trendline equation is claim = 471.391 * bmi +- 670.397

(R-squared = 0.050066). The key finding is that the average insurance claim of northeast is highest in all of the regions.

Q3: Can we predict whether an insured person is diabetic or not based on their age, BMI, gender, blood pressure, and smoking status? (Use the ML)

The purpose of this research question is to determine whether diabetes status can be predicted using a number of independent factors. Based on factors including age, Weight, gender, blood pressure, and smoking history, the project will employ machine learning algorithms to create a prediction model that can determine whether an insured individual has diabetes or not. The findings will help to identify the most important predictors of diabetes and improve the accuracy of diabetes diagnosis.

Result: By using machine learning to perform the classification task, we find the accuracy score of our mode is 0.47. After we change the hyperparameter within the machine learning model such as “p_criterion”, “ p_splitter”, and “p_max_depth”, we generate a list of accuracy scores in the form of pandas dataframe and in a descending order. The highest accuracy score achieved was 0.524, which was obtained when using a Gini impurity criterion with random splitter and maximum depth of 1, and when using an entropy criterion with random splitter and maximum depth of 1.

Motivation and background

From the perspective of insurance buyers, it is important for them to know the criteria for setting the claim. Based on the buyer's personal information, they could estimate the insurance claim. We want to use machine learning to build a regression line.

Also, we want to use machine learning to predict whether an insured person is diabetic or not based on their age, BMI, blood pressure, and smoking status because we want to improve the accuracy of a diabetes diagnosis. Machine learning is a powerful tool for identifying relationships in a large dataset. By developing a predictive model, we can potentially identify high-risk individuals who may benefit from early intervention.

Diabetes and smoking are two major public health challenges that have significant implications for individual and societal health. By examining the prevalence of diabetics and smokers in different regions, we can gain insights into the distribution and patterns of these conditions, which can inform public health policy and interventions aimed at reducing the incidence of diabetes and smoking.

Healthcare costs are a significant concern for individuals, healthcare providers, and insurers, and understanding the relationship between the prevalence of diabetics and smokers and

the insurance amount can help us identify the factors that contribute to healthcare costs. By understanding these relationships, we can develop more effective healthcare policies and interventions that can improve healthcare access, affordability, and outcomes.

Overall, investigating the prevalence of diabetics and smokers and the insurance amount can provide important insights into public health and healthcare costs, which can inform policies and interventions aimed at improving health outcomes and reducing healthcare costs.

Dataset

The dataset has 11 columns and 1340 rows. The insurance_data.csv file contains detailed information about insurance claims, including age, gender, BMI, blood pressure, diabetic status, number of children, smoking status, and region of the insured person. The link is below here:

https://www.kaggle.com/datasets/thedevastator/insurance-claim-analysis-demographic-and-health

Method

Q1:

To answer the questions, What is the estimated non-linear regression of the insurance claim? How does it compare to the linear model?, we ended up:

1. Import necessary packages such as pandas to process the dataset.
2. Clean the dataset
    a. Drop NaN values
    b. Select columns of the dataset that we need to use and return a new dataset
3. Build the regression tree and calculate MSE
    a. Create features and labels dataframe by selecting all suitable columns.
    b. Convert categorical variables in features to numerical variables using one-hot encoding.
    c. Split the dataset into training and testing sets, where 20% of the data is used for testing, and the remaining 80% is used for training.
    d. Create a decision tree regressor using DecisionTreeRegressor().
    e. Train the model using the training data.
    f. Use the trained model to make predictions on the test data.
    g. Calculate the mean squared error (MSE) between the actual labels and predicted labels.
4. Build the linear regression model and apply LASSO regression(Challenge goal: machine learning multiple models)

a. Extract the categorical variables and convert them into numerical variables using one-hot encoding and store in dummies.
b. Create features and labels dataframe by selecting all suitable columns.
c. Extract the numerical variables and store it in X_numerical.
d. Store the names of the numerical variables in a list list_numerical.
e. Convert the X_numerical to a float data type.
f. Concatenate the X_numerical and dummies DataFrames and store it in X.
g. Split the dataset into training and testing sets, where 20% of the data is used for testing, and the remaining 80% is used for training.
h. Initialize the Lasso regression model with a regularization parameter (alpha) of 1.
i. Fit the Lasso regression model to the training set.
j. Predict the target variable for the training set and calculate the mean squared error (MSE) between the predicted values and the actual values.
k. Predict the target variable for the test set using and calculate the MSE between the predicted values and the actual values.

5. Cross validation(Challenge goal: result validity)
    a. Initialize a Lasso regression model with cross-validation (LassoCV) with the following parameters:
        i. cv: number of cross-validation folds (5 in this case).
        ii. random_state: random seed for reproducibility (0 in this case).
        iii. max_iter: maximum number of iterations (10,000 in this case).
    b. Fit the LassoCV model to the training set.
    c. Print the best value of penalization (alpha) selected by the LassoCV model.
    d. Initialize a new Lasso regression model using the selected alpha value.
    e. Fit the lasso_best model to the training set.
    f. Print the coefficients of the lasso_best model and their corresponding feature names.
    g. Calculate the mean squared error (MSE) between the predicted target variable and the actual target variable for the test set.

Q2: What is the average insurance claim of male and females in various regions? How do insurance claims vary with BMI for each gender?

Interactive plot 1: Mean Claim by Region and Gender Bar Plot (New library challenge goal by using dash to create interactive plot and dropdown menu)

- How mean insurance claims vary by region and gender?
- Are there significant differences in mean insurance claims between genders within each region?

a. Load the insurance data from a CSV file using Pandas.
b. Filter the data to remove any missing values or outliers using the filter_file function from the diabetic module.
c. Group the data by region and gender and calculate the mean claim for each group using the Pandas library.
d. Create a bar plot using Plotly Express that shows the mean claim for each region and gender group. The x-axis shows the regions, the y-axis shows the mean claim values, and the bars are color-coded based on regions.
e. Create a dash app with a dropdown menu to the plot that allows the user to select a gender and update the plot to show only the mean claim for that gender within each region.

Interactive plot 2: Insurance Claims by Region Pie Chart (New library challenge goal by using dash to create interactive plot and add hover text to the chart)

- What percentage of insurance claims come from each gender in each region?
- Are there significant differences in the distribution of insurance claims between genders within each region?
    a. Filter the data to show only the insurance claims for the selected region using the Pandas library.
    b. Create a pie chart using Plotly Express that shows the percentage of insurance claims by gender for the selected region. The slices are color-coded based on gender, and the percentage of insurance claims are displayed on each slice.
    c. Create a Dash app with a dropdown menu that allows the user to select the gender.
    d. Add hover text to the pie chart that shows the percentage and gender of each slice.

Interactive plot 3: Insurance Claims by BMI Scatter Plot (New library challenge goal by using dash to create interactive plot, dropdown menu, add a trendline using ols argument to trendline)

- How do insurance claims vary with BMI for each gender?
- Is there a significant difference in the relationship between BMI and insurance claims between genders?
    a. Filter the data to show only the insurance claims for the selected gender using the Pandas library.
    b. Create a Dash app with a dropdown menu that allows the user to select a gender.
    c. Create a scatter plot using Plotly Express library that shows the relationship between bmi and claim for the selected gender. The x-axis shows the BMI values, the y-axis shows the claim values, and the points are color-coded based on the gender (blue represents male, red represents female).
    d. Add a trendline to the scatter plot using the ols argument to trendline.

e. Update the scatter plot whenever the user selects a new gender from the dropdown menu. The x-axis shows the BMI values, the y-axis shows the claim values, and the points are color-coded based on the gender.

Q3:
Can we predict whether an insured person is diabetic or not based on their age, BMI, gender, blood pressure, and smoking status? (Use the ML)

In order to meet the challenge goal, we decided to train the model by changing its hyperparameter after we got the accuracy score of our model and find the best model to predict whether an insured person is diabetic or not based on their age, BMI, gender, blood pressure, and smoking status.

Since we are predicting whether the person is diabetic or not based on different criteria, this is a classification task.

a. We will start by separating our data into its features and its labels by filtering our dataset (focusing on diabetic, age, BMI, gender, blood pressure, and smoker columns)
b. Then we will import the class from sklearn (decision tree)
c. Train the classification model on the training set.
d. We will also use the accuracy_score function to evaluate how well our model is doing on our dataset. Evaluate the performance of the model on the testing set.
e. (Challenge goal) After getting the result, we choose to change the hyperparameters (p_criterion, p_splitter, p_max_depth) of our model to identify which model is “best” at estimating whether a person is diabetic or not.

Results

Result and Analysis for Q1:

Firstly, observing the coefficient(mean absolute error) for each feature in the cross validation model, we notice 2 zero values. One is on “gender_female” and the other is on “region_northwest”. This shows that under the best value of penalization, they have no impact on the insurance claim amount.

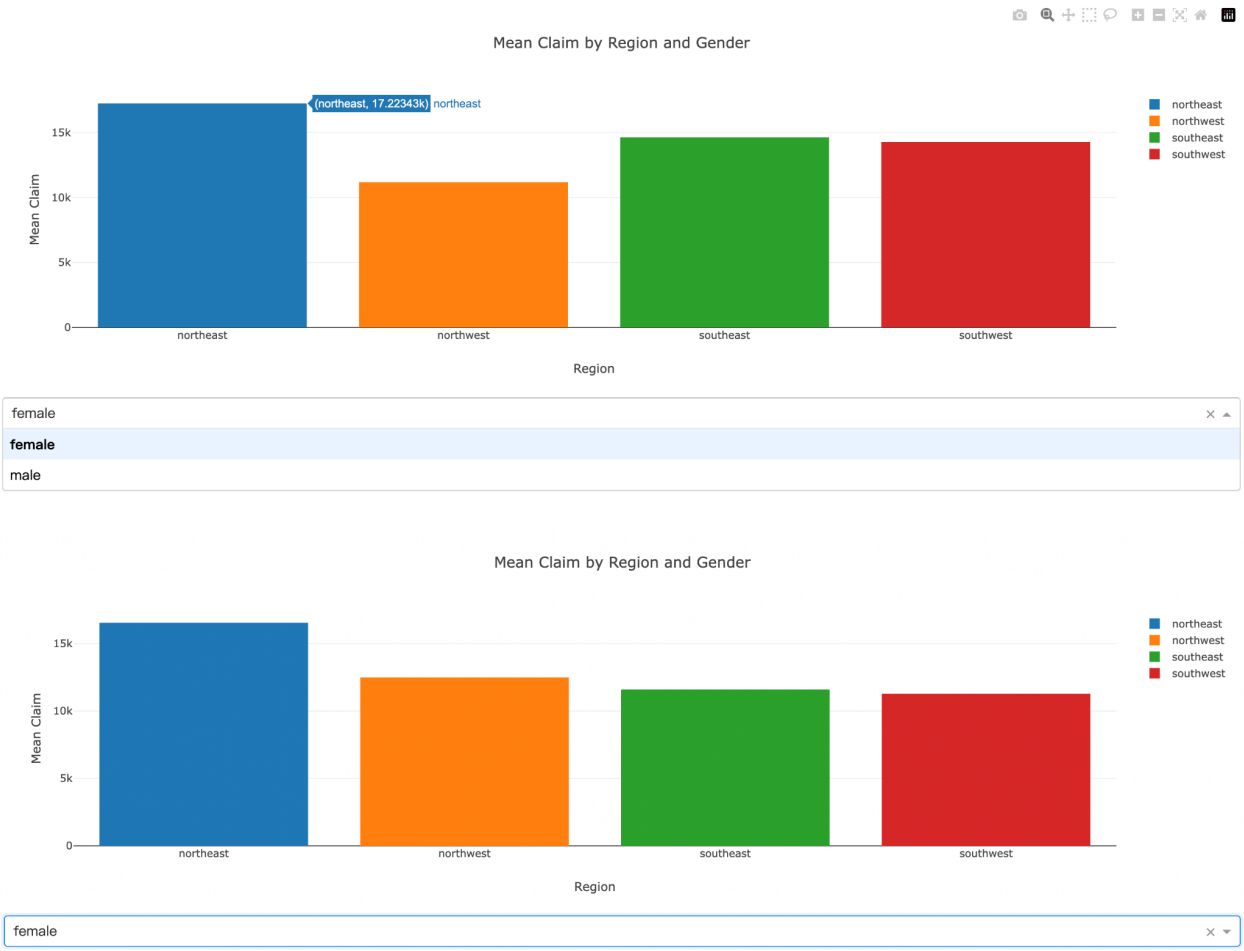
We can also see that the feature 'age' has the lowest MAE value of 220.67, which indicates that it is the most important feature for predicting the target variable. The next most important feature is 'bmi' with an MAE of 1973.91, followed by 'bloodpressure' with an MAE of 2578.42.

The feature 'smoker_No' has a negative MAE, which could indicate that it has a strong negative impact on the target variable.

Lastly, comparing the MSEs for all our models, it is interesting that the best LASSO model test set MSE is higher than the original linear regression model test set MSE, which is unexpected. Possible reasons for this might be the sample is not large enough, or the number of folds in cross validation is not large enough.

Result and Analysis for Q2 - interactive plot using dash library 1:

Interactive plot 1: Mean Claim by Region and Gender Bar Plot

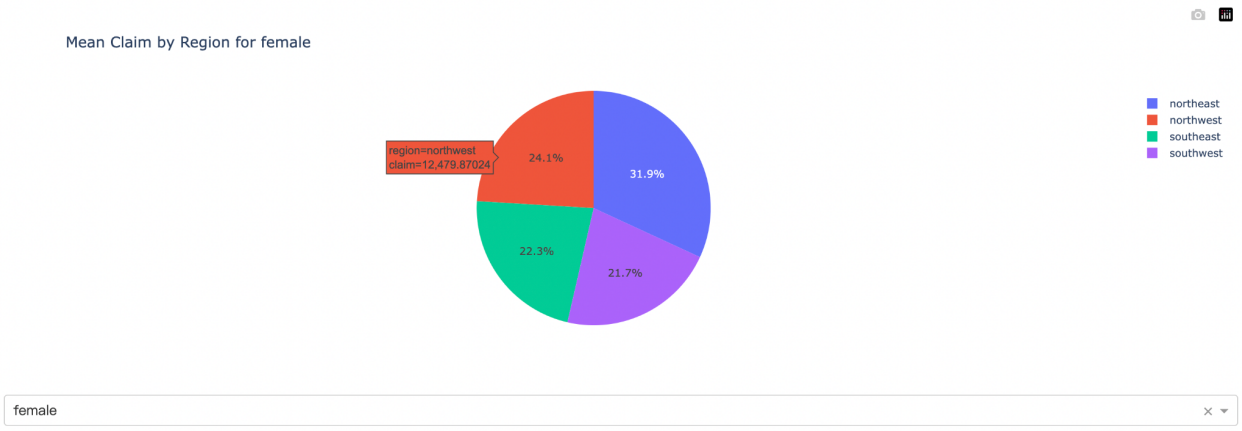


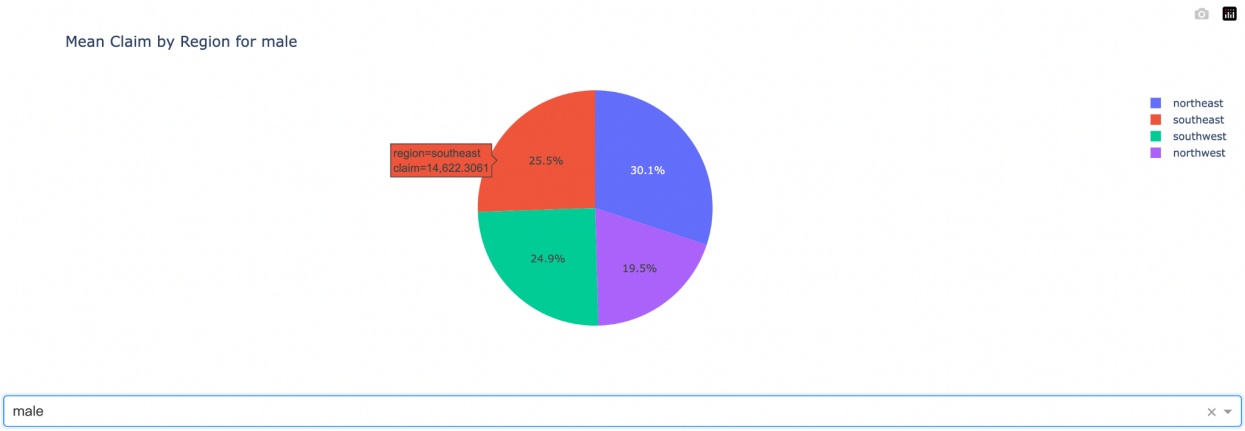
- How mean insurance claims vary by region and gender?
- Are there significant differences in mean insurance claims between genders within each region?

This plot shows the mean insurance claims by region and gender. The plot shows that the mean insurance claim varies across different regions and genders. For example, the highest mean claim is observed in the northeast region for both male and female genders, while the lowest mean claim is observed in the northeast region for males and southeast for females. The bar chart also shows that, on average, males have higher mean claims than females in all regions except for the Northeast, where females have slightly higher claims. Among males, the highest mean

claim amount is in the Northeast ($17,22343k), followed by the Southeast ($14,62231k), Southwest ($14,26716k), and Northwest ($11,17297k). Among females, the highest mean claim amount is in the Northeast ($16,53376k), followed by the Northwest ($12,47987k), Southeast ($11,58985k), and Southwest ($11,27441k).

Result and Analysis for Q2 - interactive plot using dash library 2:
Interactive plot 2: Insurance Claims by Region Pie Chart





The results of our analysis show that there are variations in the mean claim made by different genders across different regions. The following subsections discuss the findings:

- What is the mean claim made by females in different regions?
- What is the mean claim made by males in different regions?
- How does the mean claim made by females compare to that made by males across different regions?

Our analysis shows that the mean claim made by females varies across different regions. The highest mean claim was made in the northeast region, accounting for 31.9% ($16,534) of the total claims made by females. This was followed by the southeast region, where females made a

mean claim of 22.3% ($11,590). In the northwest and southwest regions, the mean claim made by females was 24.1% ($12,480) and 21.7% ($11,274), respectively. These results indicate that females in the northeast region tend to make higher claims compared to other regions.

Similar to females, our analysis shows that the mean claim made by males varies across different regions. The highest mean claim was made in the northeast region, accounting for 30.1% ($17,223) of the total claims made by males. This was followed by the southeast region, where males made a mean claim of 25.5% ($14,622). In the northwest and southwest regions, the mean claim made by males was 19.5% ($11,173) and 24.9% ($14,267), respectively. These results indicate that males in the northeast region also tend to make higher claims compared to other regions.

Our analysis shows that the mean claim made by males is generally higher than that made by females across different regions, except in the northwest region. In the northeast region, females made a mean claim of $16,534, which was lower than the mean claim made by males of $17,223. Similarly, in the southeast region, females made a mean claim of $11,590, which was lower than the mean claim made by males of $14,622. In the southwest region, females made a mean claim of $11,274, which was also lower than the mean claim made by males of $14,267. However, in the northwest region, females made a slightly higher mean claim of $12,480 compared to females who made a mean claim of $11,173.

Result and Analysis for Q2 - interactive plot using dash library 3:

Interactive plot 3: Insurance Claims by BMI Scatter Plot

- What is the relationship between BMI and insurance claims?
- Is there a difference in the relationship between BMI and insurance claims for males and females?
- How does gender affect the relationship between BMI and insurance claims?

The scatter plot shows a positive relationship between BMI and insurance claims, meaning as the BMI increases, so does the insurance claim amount. The OLS trendlines for both genders show a similar pattern, with a steeper slope for males than females. The R-squared values for both genders are low, indicating that BMI is not a strong predictor of insurance claims. For females, the OLS trendline equation is claim = 297.639 * bmi + 3527.38 (R-squared = 0.026152). For males, the OLS trendline equation is claim = 471.391 * bmi +- 670.397 (R-squared = 0.050066).

The scatter plot shows that the relationship between BMI and insurance claims is stronger for males than females, as indicated by the steeper slope of the OLS trendline for males. However, the R-squared values for both genders are

The scatter plot shows that gender has a minor effect on the relationship between BMI and insurance claims, as indicated by the slight difference in slope between the two OLS trendlines. However, the R-squared values for both genders are still low, indicating that gender does not significantly impact the relationship between BMI and insurance claims.

Result and Analysis for Q3 - Can we predict whether an insured person is diabetic or not based on their age, BMI, gender, blood pressure, and smoking status?

The initial model trained with default hyperparameters achieved an accuracy score of 0.468, indicating that the model could correctly classify 46.8% of the test data. However, the hyperparameter tuning results show that the accuracy score can be improved by changing hyperparameters. The highest accuracy score achieved was 0.524, which was obtained when using a Gini impurity criterion with random splitter and maximum depth of 1, and when using an entropy criterion with random splitter and maximum depth of 1.

Based on these results, we can conclude that the Decision Tree Classifier algorithm can be used to predict diabetes with moderate accuracy, but the model's performance can be improved by tuning hyperparameters. The results of this analysis can have significant implications for patient care and outcomes. For example, the model can be used to identify patients who are at high risk of developing diabetes, allowing healthcare providers to develop targeted prevention and intervention programs to improve patient outcomes. The model can also be used to optimize diabetes management by identifying patients who are at high risk of complications, allowing healthcare providers to develop personalized treatment plans that improve patient outcomes while minimizing the risk of adverse events.

```
Accuracy score: 0.4681647940074906
hyperparameter tuning reusults:     Accuracy P_Criterion P_Splitter  P_MaxDepth
4    0.52434        gini     random           1
12   0.52434     entropy     random           1
9    0.50936     entropy       best          10
14   0.50936     entropy     random         100
1    0.50562        gini       best          10
10   0.50187     entropy       best         100
7    0.49813        gini     random        1000
0    0.48689        gini       best           1
8    0.48689     entropy       best           1
11   0.48689     entropy       best        1000
13   0.47191     entropy     random          10
5    0.46442        gini     random          10
2    0.45318        gini       best         100
3    0.45318        gini       best        1000
15   0.44195     entropy     random        1000
6    0.41199        gini     random         100
```

Impact and Limitations

Q1 impact: Since we have established a linear regression model, if insurance customers have their own statistics about features, they could estimate their claim. Also, if someone wants to reduce his or her claim, he or she should first consider reducing BMI or blood pressure, because age couldn't be changed.

Q1 limitation:The feature 'smoker_No' has a negative MAE, but it is important to further investigate this result to ensure that it is not due to overfitting or other errors in the model.

Q2: The results presented in our graph analysis have several limitations that must be considered while interpreting them.

Firstly, the analysis only focuses on mean insurance claims and does not provide information on the distribution of claims. Therefore, the analysis does not account for any outliers or extreme values that may influence the results. Also, the data used in the analysis may be biased as it only considers insurance claims from a particular population. The analysis does not account for other factors that may affect the insurance claims, such as occupation, income, or health status, which may impact the results. Besides, the scatter plot analysis indicates a weak relationship between BMI and insurance claims with a low R-squared value for both genders. Therefore, the analysis may not be reliable for predicting insurance claims based solely on BMI. Lastly, the model trained for predicting diabetic status based on age, BMI, gender, blood pressure, and smoking status, achieved a low accuracy score with default hyperparameters. Therefore, the model's results may not be entirely reliable and must be further validated with a larger dataset or alternative models. In conclusion, while the results provide insights into the variations in mean insurance claims and the relationship between BMI and insurance claims, the limitations outlined above must be taken into account while interpreting the findings. The results

may not be generalizable or reliable in predicting the insurance claims or diabetic status of the broader population.

Q3: The results of our third research question model evaluation suggest that the decision tree models with the best performance had an accuracy score of around 0.52. The hyperparameter tuning results suggest that the criterion used to split the data and the method of splitting did not have a significant impact on the model's accuracy. However, the best-performing models had a maximum depth of 1, 10, or 100.

The implications of these results depend on the context of the dataset and the specific use case. If the dataset is used for predicting a patient's medical condition, an accuracy score of 0.52 might not be sufficient. In this case, other more complex models, such as neural networks, might be more appropriate. However, if the dataset is used for exploratory analysis or as a part of a larger decision-making process, the decision tree models might still provide useful insights. The potential beneficiaries of this analysis could include healthcare providers, researchers, and policymakers who use the dataset for various purposes. However, it is important to note that the dataset might have biases that could impact the model's performance and generalizability. For example, the dataset might be biased towards specific demographics, geographic regions, or healthcare providers. Therefore, caution should be exercised when applying the results of this analysis to other populations or settings.

The limitations of this analysis include the relatively low accuracy score of the best-performing models, which might limit their usefulness in some contexts. Additionally, the hyperparameter tuning results were based on a single train-test split of the dataset, which might not be representative of the model's performance on other splits. Therefore, practitioners should perform further validation and testing of the models before using them for decision-making. Finally, the results of our analysis should not be used to replace clinical judgment or medical expertise, and should be used only as a complementary tool for decision-making.

Challenge Goals

Our first question fulfills the machine learning challenge goal and result validity goal. We not only used the tree regression that we learned in classes, but also included another model type: the LASSO regression model. It aims to build a linear regression model, which has a very different machine learning pipeline. The cross validation is a way to validate (result validity) our result for the LASSO model and create the best linear model. It proves facts that gender doesn’t matter, and the MSEs are close to each other, leading to the claim that our result in LASSO regression is very likely to be correct.

Our second research question fulfills the New Library Challenge goal. In order to meet the New Library challenge goal, we made three interactive plots using the Dash library, which we haven't discussed in this course.

- **Interactive Data Visualization:** Our code creates three interactive plots using the Dash library - a bar chart, a pie chart, and a scatter plot. Users can interact with the plots by selecting options from the dropdown menus and the plots update dynamically. This fulfills the goal of creating interactive data visualizations that allow users to explore and analyze data in an interactive way.

Our third research question fulfills the machine learning challenge goal because we not only use a Decision Tree Classifier to make predictions about whether an individual is diabetic or not based on features such as age, gender, bmi, blood pressure, smoker status, and region but also performs hyperparameter tuning to optimize the decision tree classifier model by testing different values for criteria, splitter, and max_depth. Finally, we use the accuracy score to assess the performance of the model.

Work Plan Evaluation

For this project, our general planned tasks were:

- Data preparation: clean and prepare the data, performing exploratory data analysis
    - Estimated time: Around 2 hours.
    - Actual time: Less than half an hour. It turned out that our dataset was very complete, so it was easy to process.
- Model Selection
    - Estimated time: Around 2 hours. We decided to discuss the model first, then select and implement appropriate machine learning algorithms.
    - Actual time: More than 5 hours. There were so many models for machine learning, and we discussed the best models to use for a long time. We also spent a lot of time figuring out the algorithms behind the models in order to apply them.
- Model Evaluation
    - Estimated time: Around 2 hours. We decided to evaluate the performance of the models on the testing set, select the best model, and validate the model.
    - Actual time: More than 10 hours. We found it very hard to figure out the method to test machine learning and graphs. We talked to the TA several times about the testing methods and received some feedback. Writing the testing codes also took us a long time.

More specifically, we divided the project based on our 3 questions:

Research question 1:

Planned:

- About 6 hours
- Responsible for research question 2: Xuqing Wu
- Workflow:

  Week 7- Week 9: Cleaning and preparing the data set. Working on research question 1 and helping other group members to solve the problem. Xuqing will focus on developing code and testing the code, and we will schedule a Wednesday afternoon to debug. Attending weekly meetings on Friday 5 pm on zoom in order to solve the problems we met together.

  Week 10: Working on our final project report.

Actual:

- Took more than 10 hours to choose model, write code and debug
- Took about 4 hours to write the final report
- Actual time management is quite close to the planned workflow.

Research question 2:

Planned:

- About 4 hours
- Responsible for research question 2: Zhuoyi Zhao
- Workflow:

  Week 7- Week 9: Helping our group members to clean the dataset, and supporting our group members if they need help (Via zoom or discord). Start Working on Research question 2. Attending weekly meetings on Friday 5 pm on zoom in order to solve the problems we met together.

  Week10: Working on our final project proposal. Summarizing the results we find based on research question 3.

Actual:

- Took more than 24 hours to struggle with the dash plot in coding and debugging.
- Took about 2 hours to write the report and conclude the output results.

- Actual time management is not close to the planned workflow, because we decided to make a challenge problem in plot by using the Dash. Therefore, we spent lots of time on searching related knowledge to learn how to code and work on each plot.

Research question 3:

Planned:

- About 9 hours
- Responsible for research question3: Maria M
- Workflow:

  Week 7- Week 9: Helping our group members to clean the dataset, and supporting our group members if they need help (Via zoom or discord). Start working on research question 3. Attending weekly meetings on Friday 5 pm on zoom in order to solve the problems we met together.

  Week10: Working on our final project proposal. Responsible for making our project slides and summarizing the results we find based on research question 3.

Actual:

- The actual work time was roughly the same as the estimated work time. However, I spent much more time debugging the error message, which I had never encountered before, such as a deprecated error. Additionally, I spent a longer time writing the test code for my research questions since it is difficult to test the accuracy of the results generated by machine learning.

Testing

Testing for the graph:

In order to test the code, we created a smaller csv file “test_small_data.csv” that is structured similarly to the main dataset. We then ran our pre-processing methods on this smaller dataset to ensure that they were modifying the data as expected before being plotted.
We created a test function called test_filter_file() to test the filter_file() function. Within this function, we first loaded the smaller dataset using Pandas read_csv() method. We then passed this dataset to the filter_file() function to filter it based on certain conditions.
After filtering the dataset, we used assert statements to test if the filtered dataset had the expected number of rows, columns, and if it contained any NaN values. By using assert statements, we were able to automatically check if the expected results were obtained and the code was working correctly.

We ran this test function by calling it in the main function and executing it. By testing our pre-processing methods on smaller datasets, we can ensure that the methods are working correctly before using them on the main dataset. This helps to avoid any unexpected errors that may occur due to incorrect processing of data. The use of assert statements also helps to ensure that the expected results are obtained and the code is working correctly. Overall, this approach of testing helped us to build confidence in our code and ensure that the final results were accurate and reliable.

Testing for Machine Learning:

We implemented three different test functions for our third research question functions: filter_file, fit_and_predict_diabetic, and hyperparameter_tuning, all of which are located in the diabetic module.

The purpose of these test functions is to ensure that the code works correctly and to increase confidence in the results produced by our code.

The first test function, test_filter_file, tests whether the filter_file function returns a Pandas DataFrame with the expected columns. This is an important test because the code relies on this DataFrame to function correctly. This test uses an assert statement to check whether all of the expected columns are present in the DataFrame.

The second test function, test_fit_and_predict_diabetic, tests whether the fit_and_predict_diabetic function returns a float between 0 and 1. This function is intended to predict the likelihood of a person having diabetes, which is represented as a probability between 0 and 1. The test uses two assert statements to check whether the result is indeed a float and whether it falls within the expected range we assumed.

The third test function, test_hyperparameter_tuning, tests whether the hyperparameter_tuning function returns a Pandas DataFrame with the expected columns. This function is intended to optimize the performance of the model by tuning hyperparameters, and the resulting DataFrame is used to select the best hyperparameters for the model. This test uses an assert statement to check whether all of the expected columns are present in the DataFrame.

In the main function, the data variable is read in from a CSV file, and the filter_file function is used to extract the relevant columns. These filtered data are then used as inputs to the three test functions. By doing this, the code ensures that the functions are working correctly on a smaller subset of data before running them on the full dataset.

Overall, our test functions aim to provide a robust and comprehensive approach to testing the code, and increase the transparency and trustworthiness of the results we produced.

Collaboration

The resource we referred to:
https://deepchecks.com/how-to-check-the-accuracy-of-your-machine-learning-model/

The resource of LASSO regression we referred to:
https://www.kirenz.com/post/2019-08-12-python-lasso-regression-auto/

The resource we referred for dash library:
https://plotly.com/python/