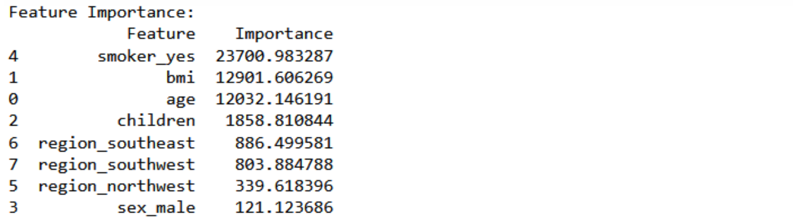
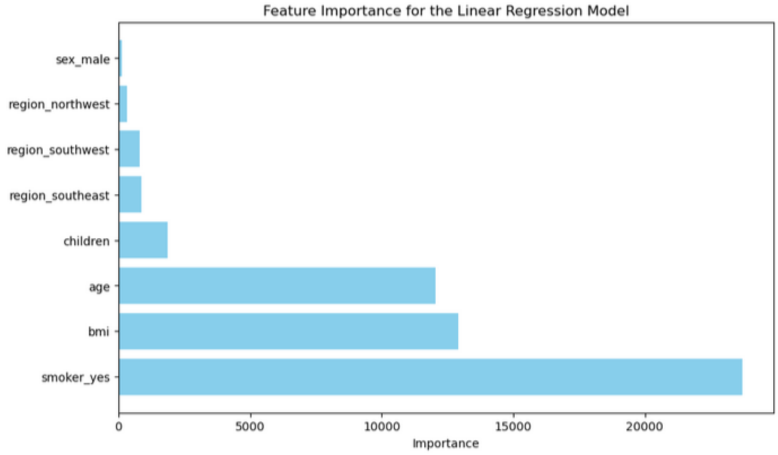
Interpretable models 1 Report

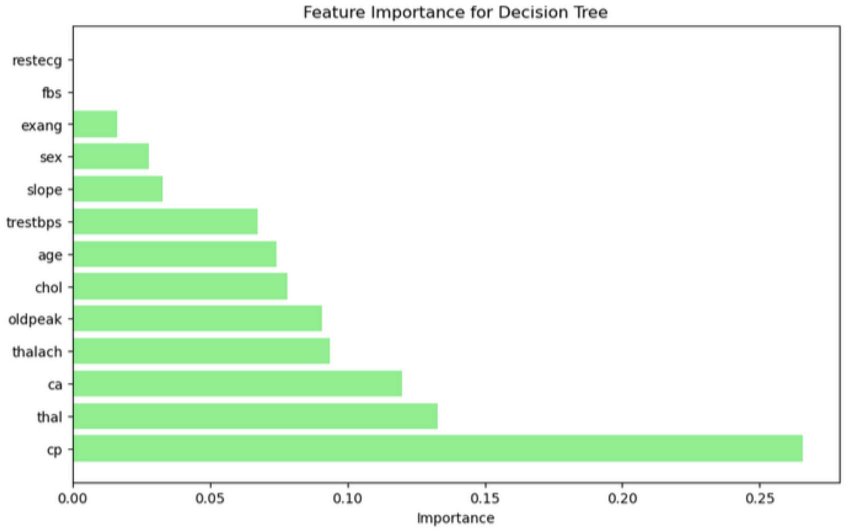
1. **Linear Regression**
   - R2 score: 0.7605
   - The following figure shows model coefficients and their names:



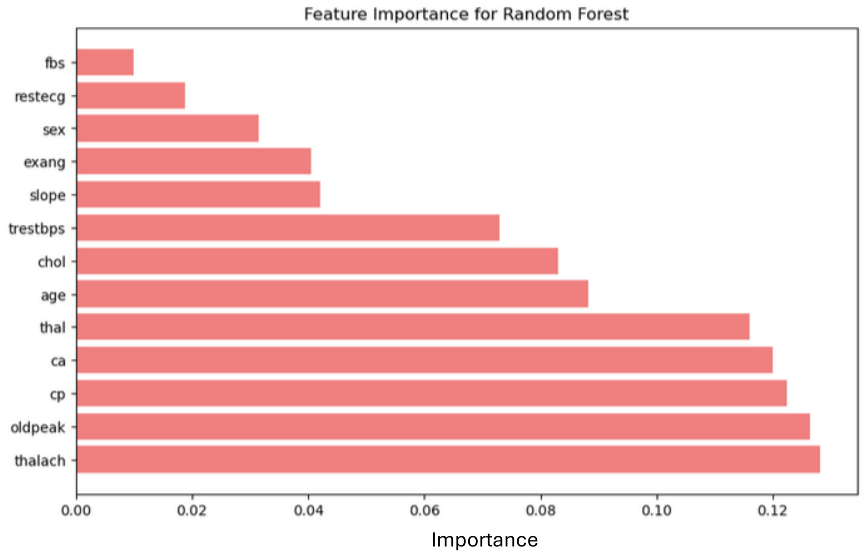
```
Feature Importance:
            Feature    Importance
4        smoker_yes  23700.983287
1               bmi  12901.606269
0               age  12032.146191
2          children   1858.810844
6  region_southeast    886.499581
7  region_southwest    803.884788
5  region_northwest    339.618396
3          sex_male    121.123686
```



2. **Decision tree**
   - Accuracy score: 0.9617
   - The following figure shows explanations corresponding to the feature importances in the tree:



3. **Random Forest**
   - Accuracy score: 0.9912
   - The following figure shows explanations corresponding to the feature importances in the tree:



4. **Simplifying the model**
    - They are the same, namely 'thal', 'ca', 'cp', 'oldpeak', 'thalach', but they are not in the same order of importance. In the first graph, 'cp' has a way higher importance than the other 4 features, while in the second graph the top 5 are more 'grouped' together separately from the other features.
    - After retraining a RF model using only the top 5 features the accuracy becomes: 0.9882. the accuracy therefore decreases (from 0.9912), but I would not classify it as decreasing "Significantly"
5. **General Questions**
    - Yes, I think these explanations can be useful, but they have limitations. Feature importance and coefficients provide information on how the model makes its predictions. With the feature importance and coefficients you can see the relationship between features and the outcome. For more complex models they are too oversimplified, but in this case they can provide the required insights.
    - In cases like these where the model is simple and the features have clear relationships with the target variable they can be trusted. If it were a DNN model or it was very overfit, the explanations that this method gives are not specific enough (that is: they don't capture nuances that well).
    - After searching in the textbook by Molnar (Interpretable Machine Learning, A guide for making black box models explainable), I found these two methods:
        - Method 1 (linear regression): PDP:
          Partial Dependence Plots visualize the relationship between features and the target variable, while keeping other features constant. It can show how a feature affects predictions across its entire range of values, which could be useful for understanding non-linear relationships.
        - Method 2 (Decision tree): Tree visualization:
          Flowcharts can be used to visualize a decision tree, where every feature is represented with a node. The branches represent decisions made based on the feature's value, this then provides a step by step explanation of how predictions are made.
    - Reference:

Molnar, C. (2024, July 31). *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/