# Group Project ML

## Some thoughts:

1. What should we do with the missing data?

* Remove columns
* Remove rows
* Impute
  + KNN Imputes
  + Simple imputer (which metric?)
  + ….

2. Should we binarize data and medical expenses (depending on the algorithms)?

3. Perform feature selection?

* SelectKbest with Random Forest
* Ridge
* Lasso
* ….

4. Should we do feature engineering?

5. Try on different features on different algorithms? - Do a pipeline

6. Do we need to standardize?

* MinMax
* RobustScaler
* StandardScaler

7. Which models?

* Decision Trees
* SVM
* KNN
* NN (does it make sense on such a small and simple dataset?)
* Logistic Regression
* Naive Bayes

8. Test different measurements? Accuracy and F1score

9. Can we consider the dataset as imbalanced? (It’s a 61/39 ratio)

If yes, then:

* Random Undersampling
* SMOTE Oversampling

10. Ways to improve our work?

* Pipeline
* GridSearch (and there is another one I don’t remember but will search)