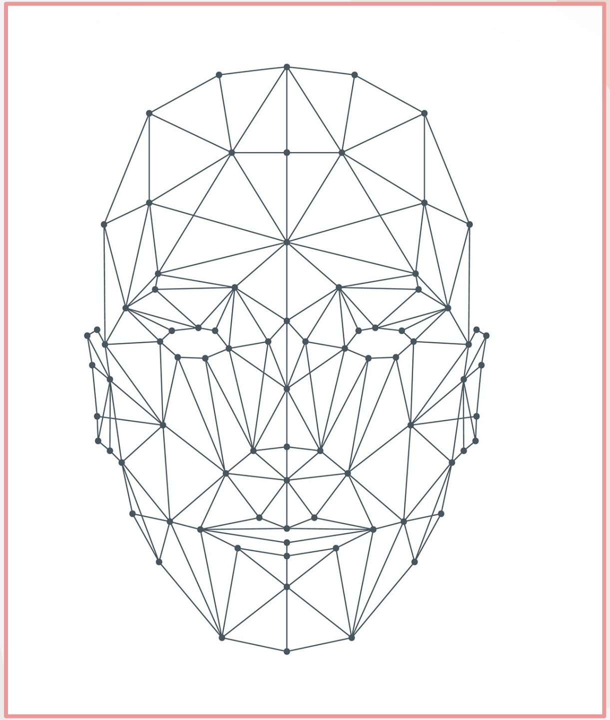
STAT 207: Final Project

~ Sarah Michalec - Jada Giddens - Armeen Sultan - Trish Qiu ~

Motivation and Introduction

- Research Introduction
  - Identity identification and facial recognition
  - Equally high accuracy for positives and negative
- Research Goals
  - **Primary**
    - Build predictive model to predict gender
  - **Secondary**
    - Yield reliable interpretive insights
    - Describe relationship of variables



Dataset Discussion

Dataset Source

- Retrieved from Kaggle on April 17th, 2024
- Author Jifry Issadeen
- Explanatory variables
  - 2 Numerical:
    - forehead_width_cm
    - forehead_height_cm
  - 5 Categorical:
    - long_hair
    - nose_wide
    - nose_long
    - lips_thin,
    - distance_nose_to_lip_long

Data Cleaning

- No implicit or explicit missing values
- No outliers
- Categorical variable counts all >50

- Overall, no rows dropped
  - Model representative of entire dataset

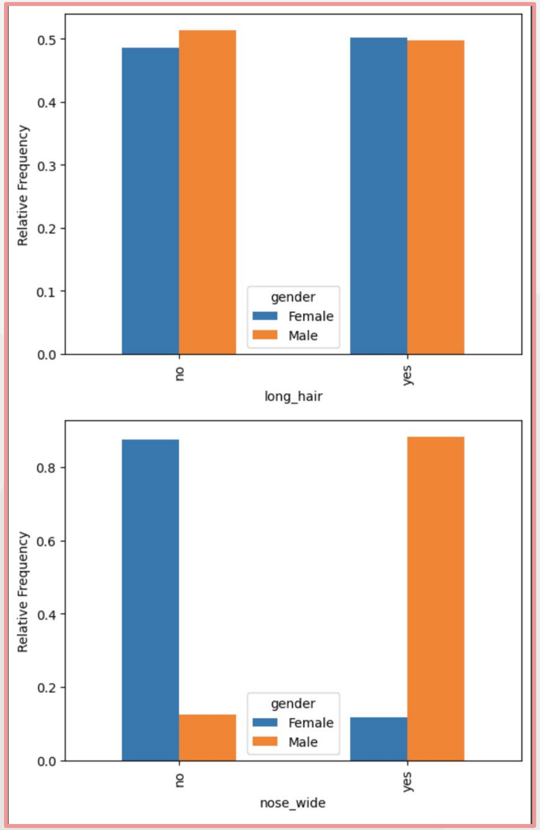
https://www.kaggle.com/datasets/elakiri coder/gender-classification-dataset?res ource=download.

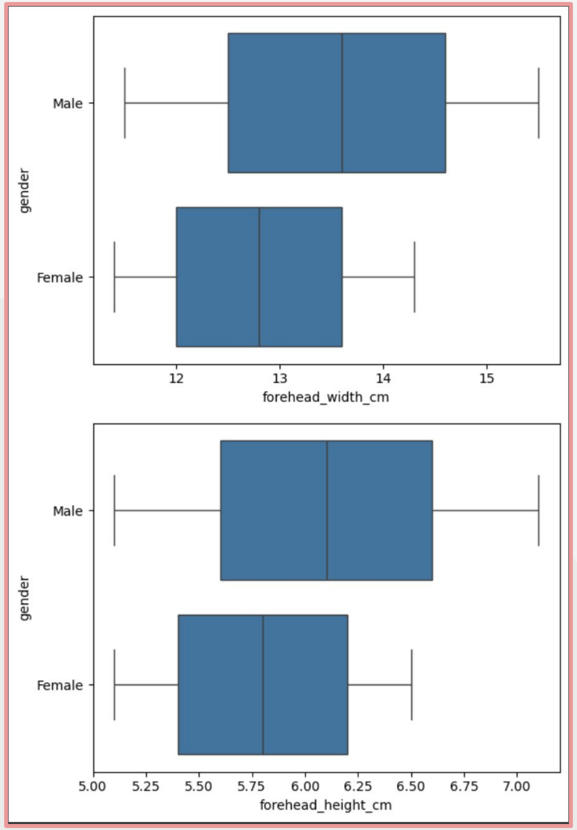
Dataset

	long_hair	forehead_width_cm	forehead_height_cm	nose_wide	nose_long	lips_thin	distance_nose_to_lip_long	gender
0	1	11.8	6.1	1	0	1	1	Male
1	0	14.0	5.4	0	0	1	0	Female
2	0	11.8	6.3	1	1	1	1	Male
3	0	14.4	6.1	0	1	1	1	Male
4	1	13.5	5.9	0	0	0	0	Female
5	1	13.0	6.8	1	1	1	1	Male
6	1	15.3	6.2	1	1	1	0	Male
7	0	13.0	5.2	0	0	0	0	Female
8	1	11.9	5.4	1	0	1	1	Female
9	1	12.1	5.4	0	0	0	0	Female

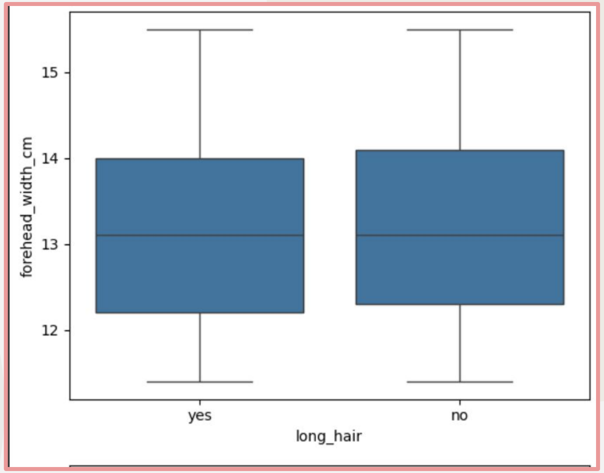
~ first 10 rows ~

Descriptive Analytics

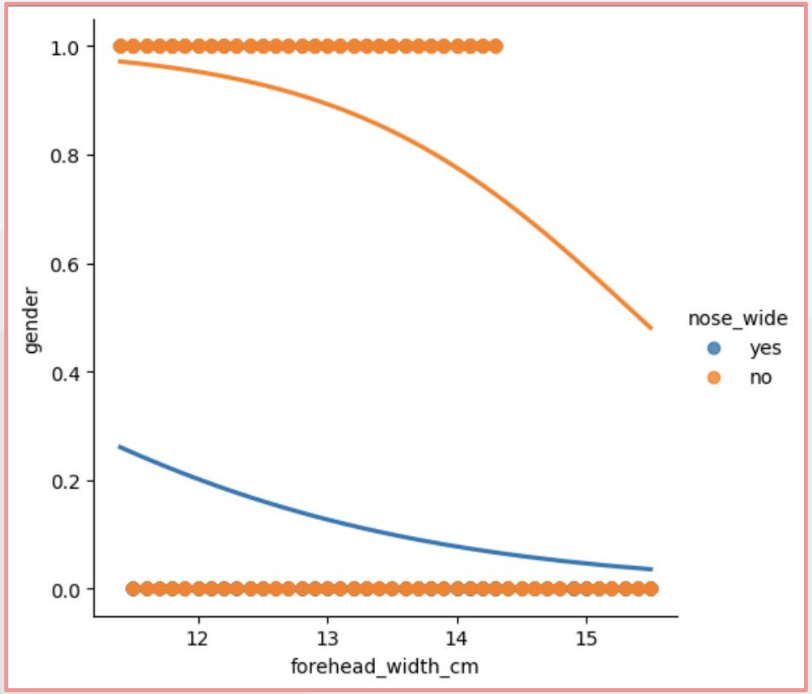




- Long_hair variable has weak association with response and other explanatory variables.
- Doesn't mean that the variable will under or overfit the model



Descriptive Analytics



- Slope or stretch of the simple logistic model differs
- An interaction term will be made for each pair with an interaction to effectively model the relation between these the explanatory variables and response variable.
- Log likelihood values
  - Full_model : -368.206
  - Full_model w/ interaction terms: -363.150

Best Model Discussion

- Tried Lasso, Ridge, Elastic Net Regression
  - 0.996231, 0.996396, 0.996231
  - Ridge model highest AUC

$$\hat{gender} = \frac{1}{1+\exp\left(-\begin{array}{l} -\,0.00073589 \\ +\,0.0041852(\text{long hair}) \\ -\,0.06137687(\text{forehead width cm}) \\ -\,0.04974995(\text{forehead height cm}) \\ -\,0.14924822(\text{nose wide}) \\ -\,0.14499015(\text{nose long}) \\ -\,0.14275644(\text{lips thin}) \\ -\,0.14630999(\text{distance nose to lip long}) \end{array}\right)}$$

~ equation ~

- No slopes zeroed out
  - long_hair lowest slope
- Probability Threshold: 0.48
  - FPR: 5%
  - TPR: 99%
- Multicollinearity
  - Numerical: none
  - Categorical: present
  - Numerical/Categorical: minor
- Cannot interpret slopes
  - Multicollinearity
  - Slope interactions

Conclusion

- Would recommend our model
  - High AUC → implies good fit
- Shortcomings
  - Multicollinearity
  - Slope interactions
- Future work
  - Eliminate multicollinearity and slope interactions
  - Test different datasets
    - More explanatory variables

References

Glover, E. (2024, February 23). Facial Recognition Technology, explained. Built In. https://builtin.com/articles/facial-recognition-technology-explained#:~:text=Facial%20recognition%20is%20a%20technology,of%20known%20faces%20or%20templates.

Najibi, A. (2020, October 26). Racial discrimination in face recognition technology. Science in the News. https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/