Predicting Diagnosis of Cervical Cancer Based on Risk Factors

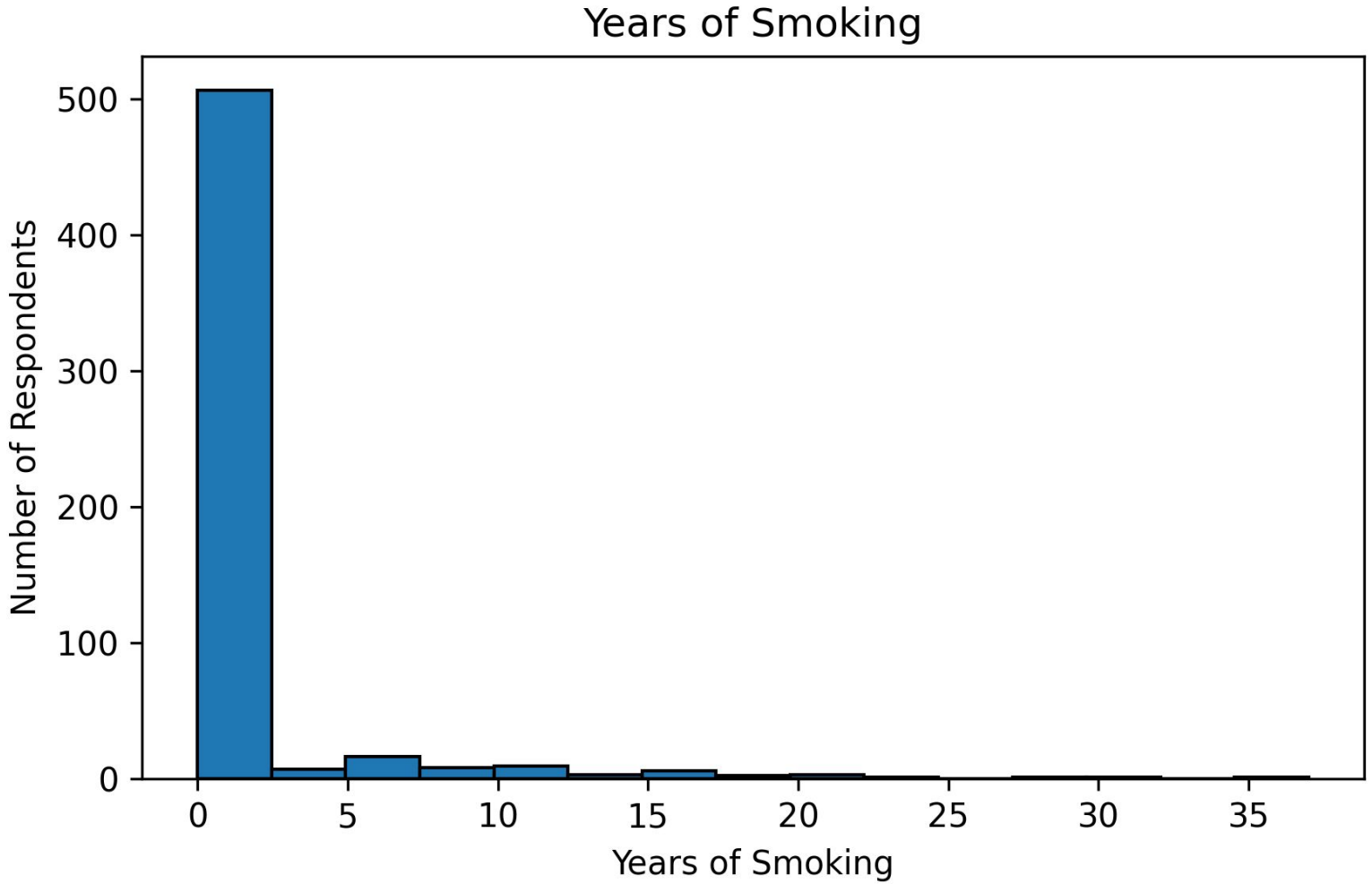
Dany Trainor

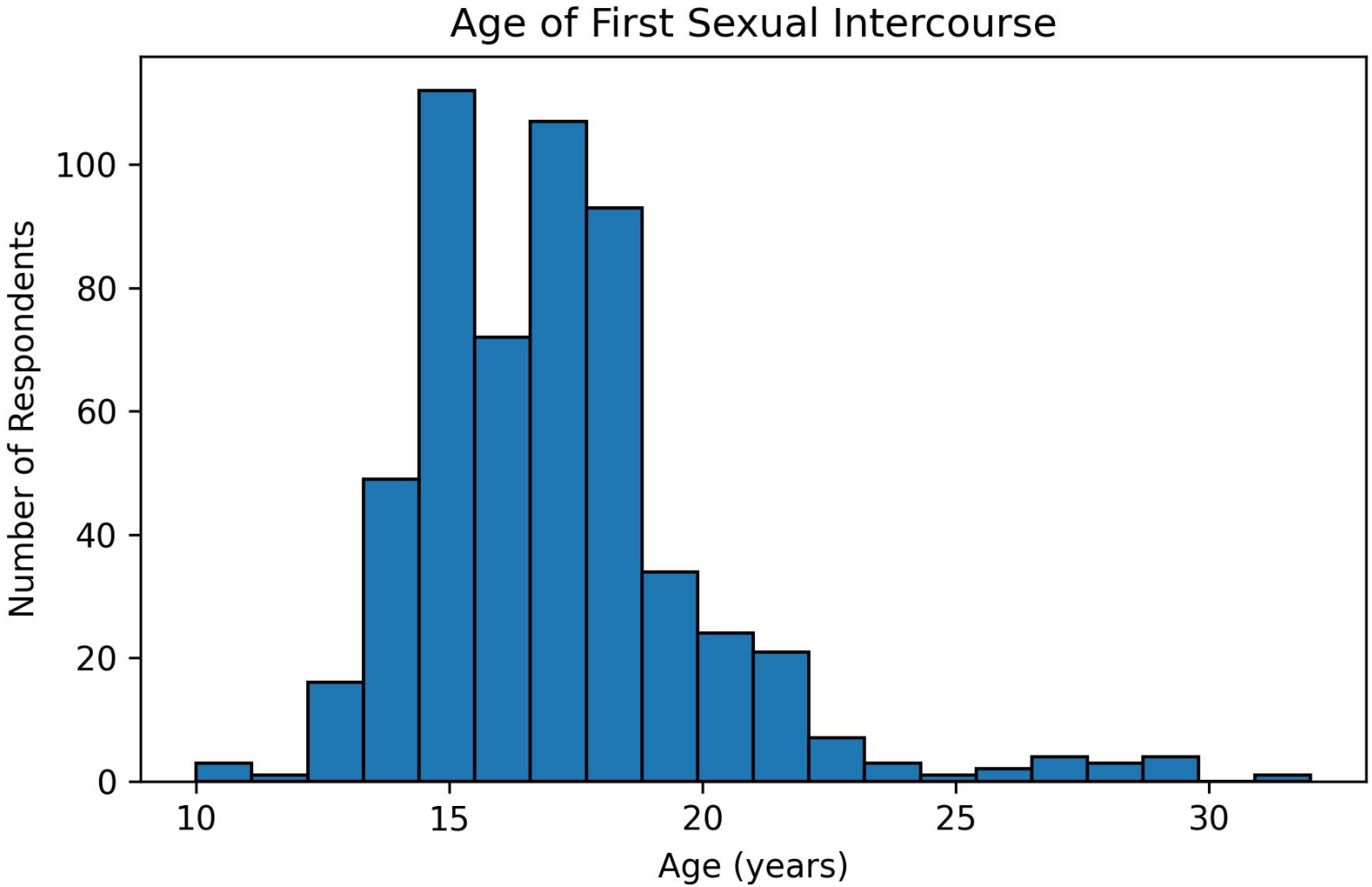
Data and Goals

- Data was collected from the UCI machine learning repository, originally collected via an anonymous survey conducted at Hospital Universitario de Caracas in Caracas, Venezuela

- Data includes 32 risk factors and results of 4 diagnostic tests for cervical cancer for each survey respondent

- Goal is to develop a machine learning model that can learn from this data and predict positive biopsy results for cervical cancer


Years of Smoking
Number of Respondents
500
400
300
200
100
0
0
5
10
15
20
25
30
35
Years of Smoking


Age of First Sexual Intercourse
Number of Respondents
0
20
40
60
80
100
10
15
20
25
30
Age (years)

Model: Neural Network

- Several models were built, but imbalanced data caused prediction issues

- Neural networks make predictions, analyze their mistakes, and then learn from them the next time around

- Neural network was able to achieve 93% accuracy

Recommendations

- More data: Larger sample with varied demographics, ideally data from multiple international hospitals and multiple survey takers

- Modify survey: more risk factors including a larger variety of health issues that can impact cervical cancer development and lifestyle information

- Analyze differences in diagnostic techniques