PREDICTING AUTISM WITH MACHINE LEARNING

BY OLIVER ZAGORIN

ABOUT ME

https://www.oliverzagorin.com/
www.linkedin.com/in/oliver-zagorin
https://github.com/NeonOstrich/

Educator
IO Psychologist & Consultant
Data Analyst

Seeking internships & part & full-time roles.

Special interests: Education, Health, Business

WHAT IS AUTISM?

Autism, or Autism Spectrum Disorder (ASD), is a neuro-developmental condition characterized by difficulties in social interaction, communication challenges, and repetitive behaviors. It is a spectrum disorder, meaning it varies in severity and symptoms from person to person by a wide margin. And it is about diagnosed four times as frequently in boys as girls.

1 in 36 children in the US are diagnosed with ASD in 2020.

This is an increase of 178% since 2000.

WHY IS AUTISM RATE INCREASING?

Environmental factors could be causing an increase in the prevalence of autism in the population.

- Advanced parental age
- Maternal health concerns
- Prenatal exposure to pollution
- Premature/Low birth weight

Improved assessment tools and screening among children who exhibit autistic traits in school.

- IDEA (Individuals with Disabilities Education Act)
- Standardized Screenings
- Improved Assessment
- Schoolwide Assessment
- Access to experienced professionals in schools

A greater understanding of the Spectrum of Autism. Many different presentations.

- Autistic Traits
- Tech & Entrepreneur Leaders
- Female Autistic Presentation
- Comorbid Conditions
- Sensory Challenges

WHY IS AUTISM TESTING CHALLENGING?

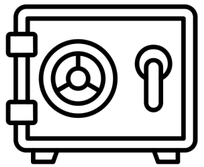
Ineffective

One study found 1 in 4 autistic children under 8 go undiagnosed.

Costly

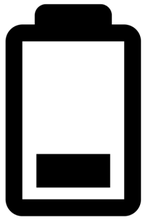
Testing can take months and cost thousands of dollars.

Inaccessible



Finding qualified specialists & traveling & hotel stays for testing.

Assessment Fatigue



The individual being assessed may become unable to continue.

WHY IS TESTING WITH MACHINE LEARNING USEFUL?

Inform Decisions

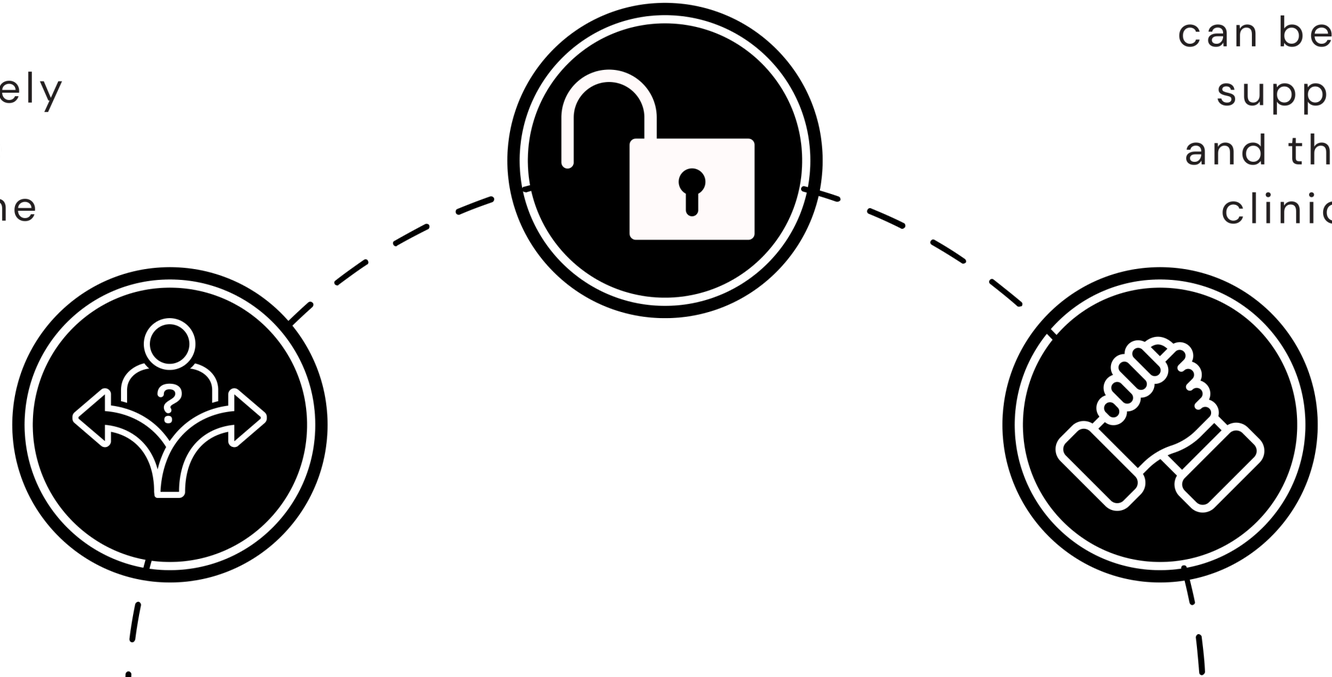
Decide whether to spend the money, time, and effort. Understand what your diagnosis is likely to be before investing in the process.

Cost and Access

Access this tool free of cost from the comfort of your home or classroom, and from regions where diagnostic professionals are not available.

Support

Empower adults and individuals who are not seeking accommodations. This can be a powerful tool to support introspection and therapy, and inform clinical interventions.



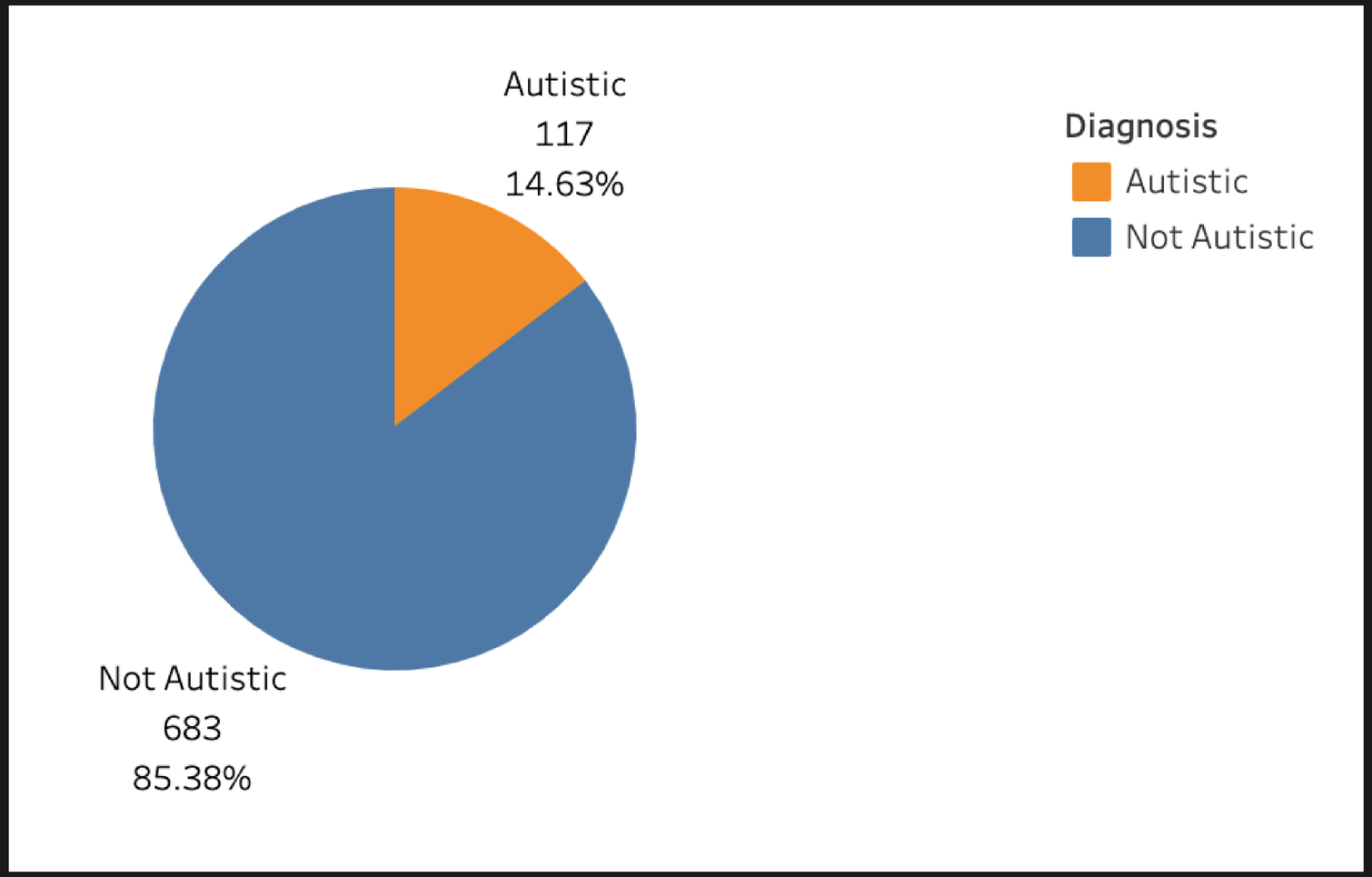
MODEL INTRODUCTION

- We utilized a dataset that includes birth information, demographic data, & responses to the Autism Spectrum Quotient (ASQ) test.

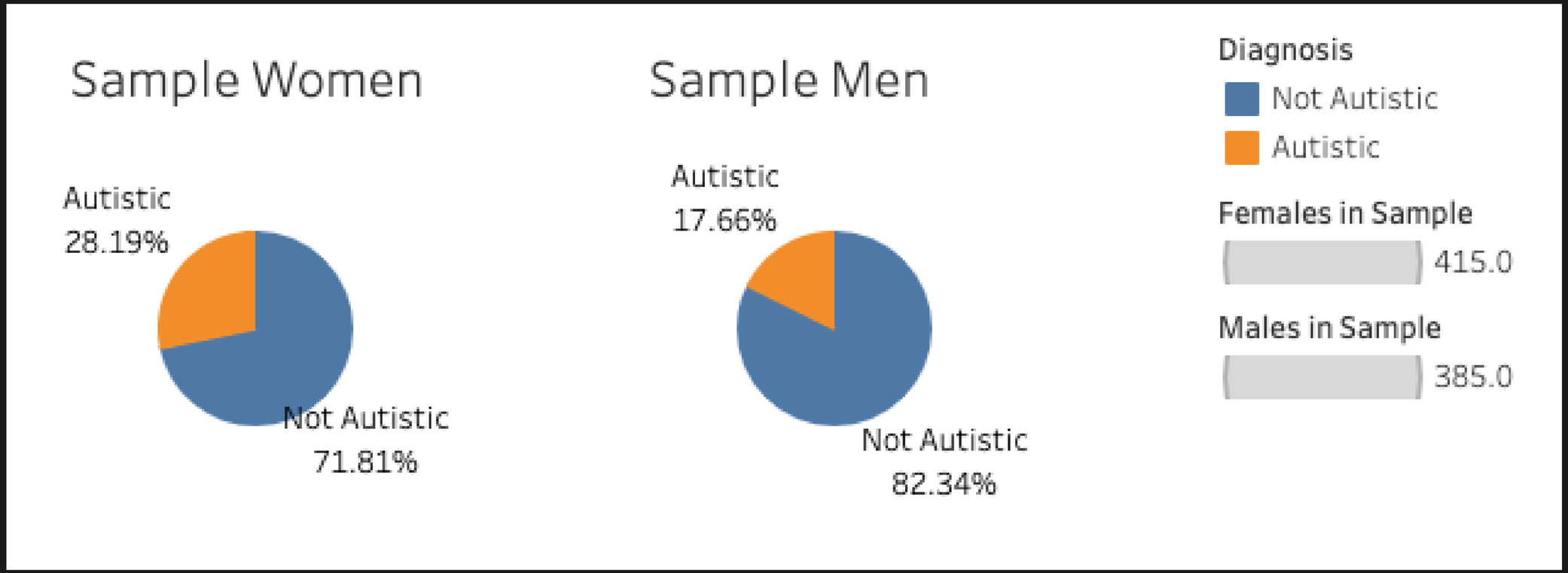
- We generated models using Random Forest, Decision Trees, Logistic Regression, Support Vector Classifier, Neural Network, and Keras Tuner Neural Network. We ran these on their own and in conjunction with Random Oversampling.

- We evaluated our models using F1 scores for each target designation as well as accuracy scores.

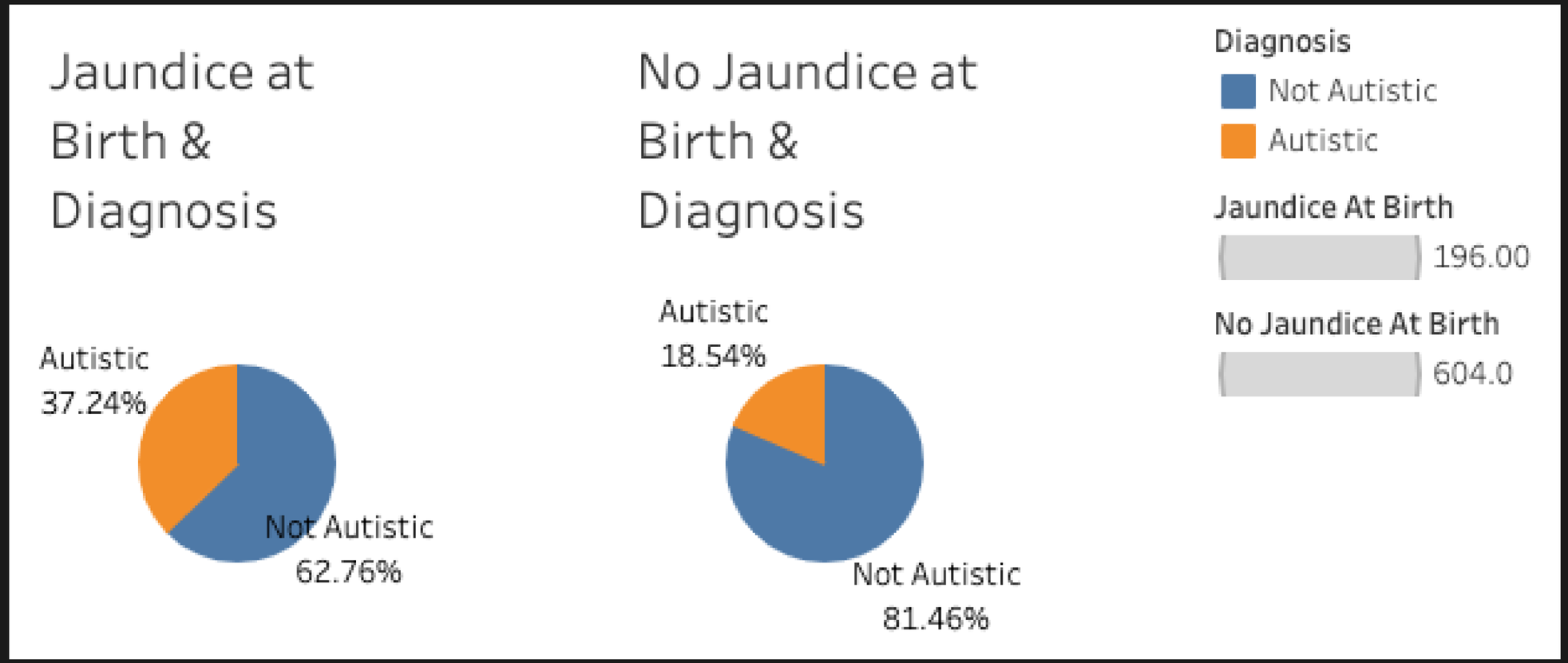
DIAGNOSIS IN SAMPLE



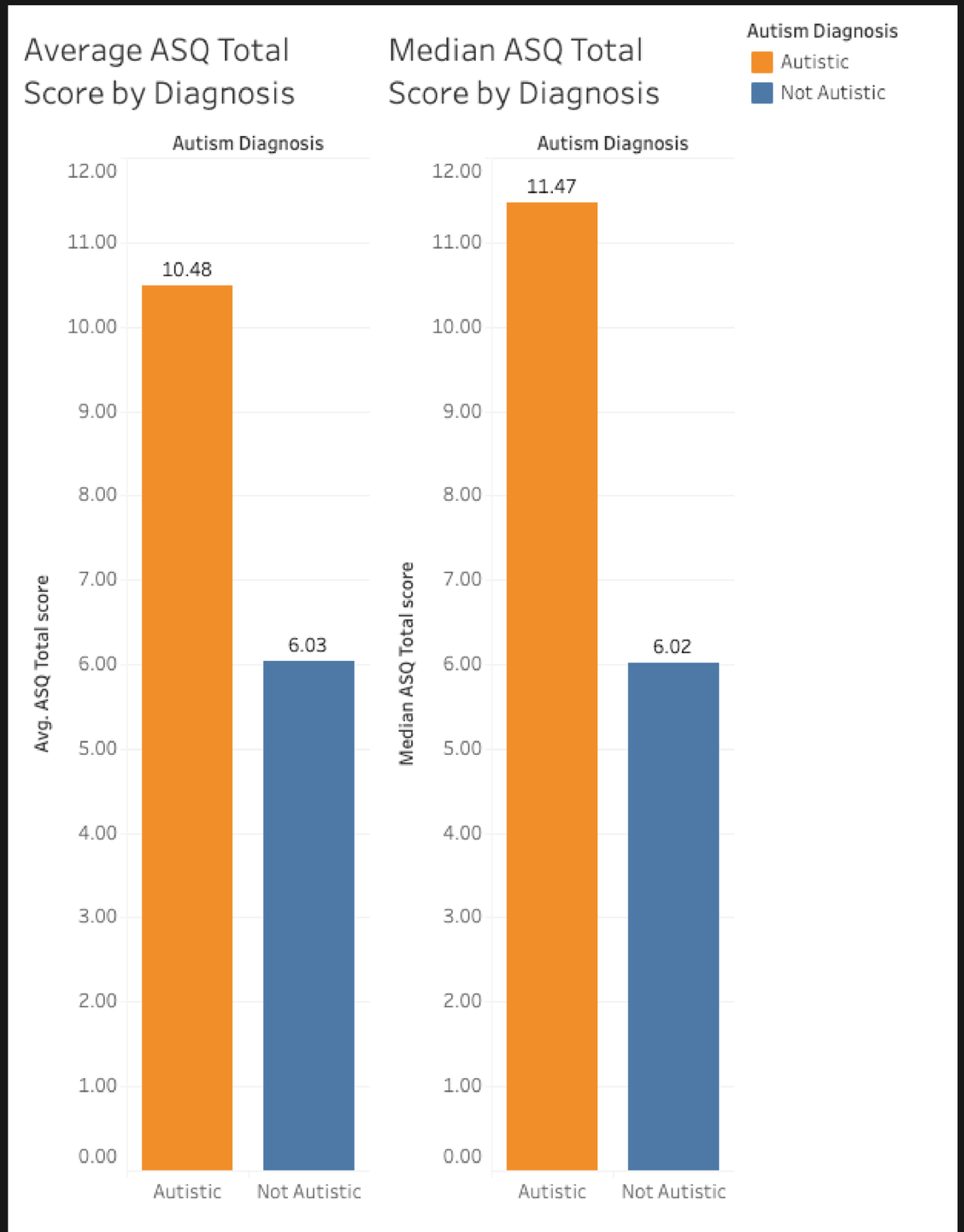
DIAGNOSIS BY GENDER



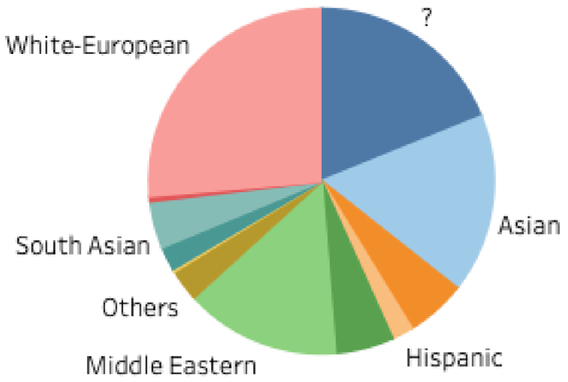
DIAGNOSIS BY JAUNDICE

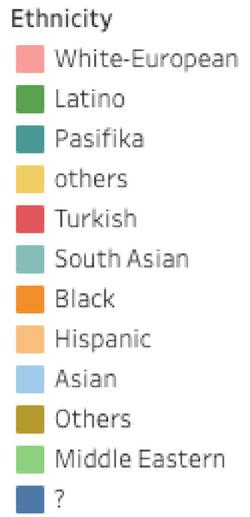


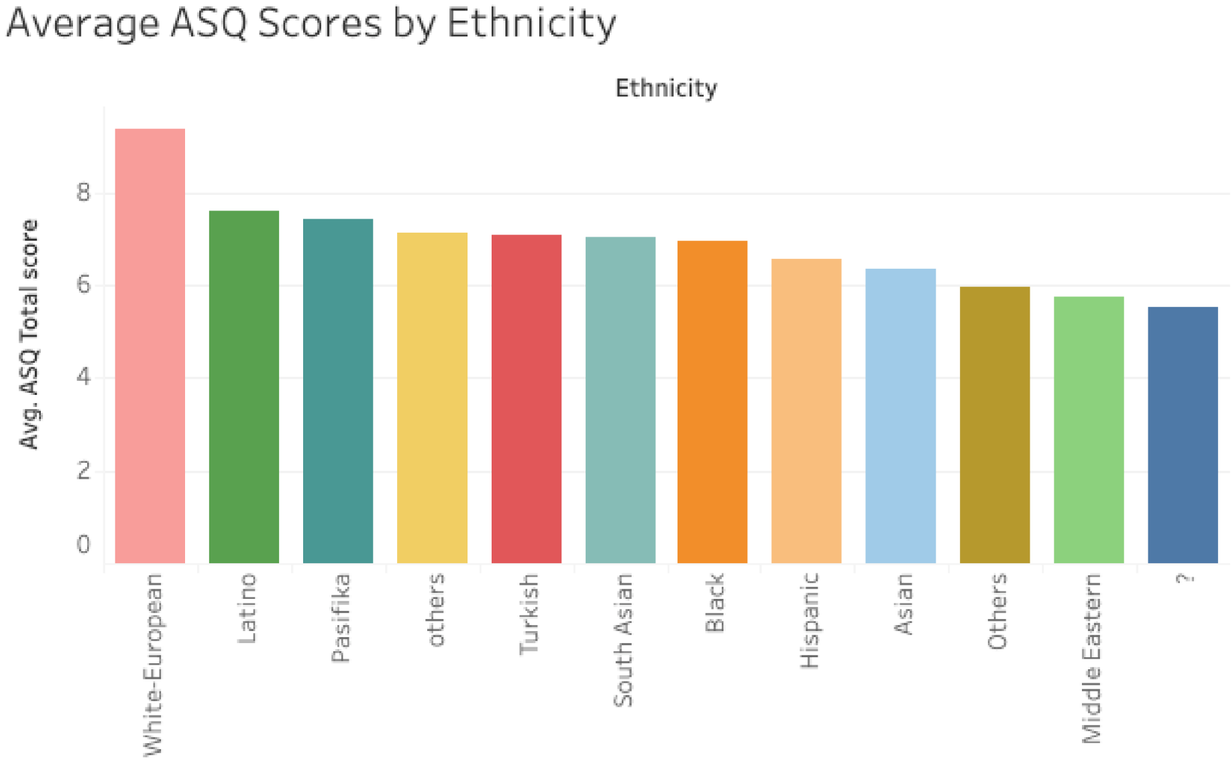
DIAGNOSIS BY AUTISM SPECTRUM QUOTIENT (ASQ)



ETHNICITY IN SAMPLE







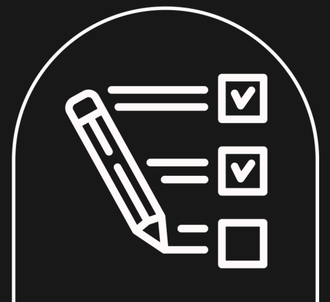
MODELS

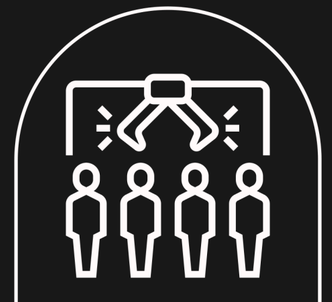
Standard

- Decision Tree
- Random Forest
- Logistic Regression
- Support Vector Classifier
- Neural Network
- Keras Tuner Neural Network

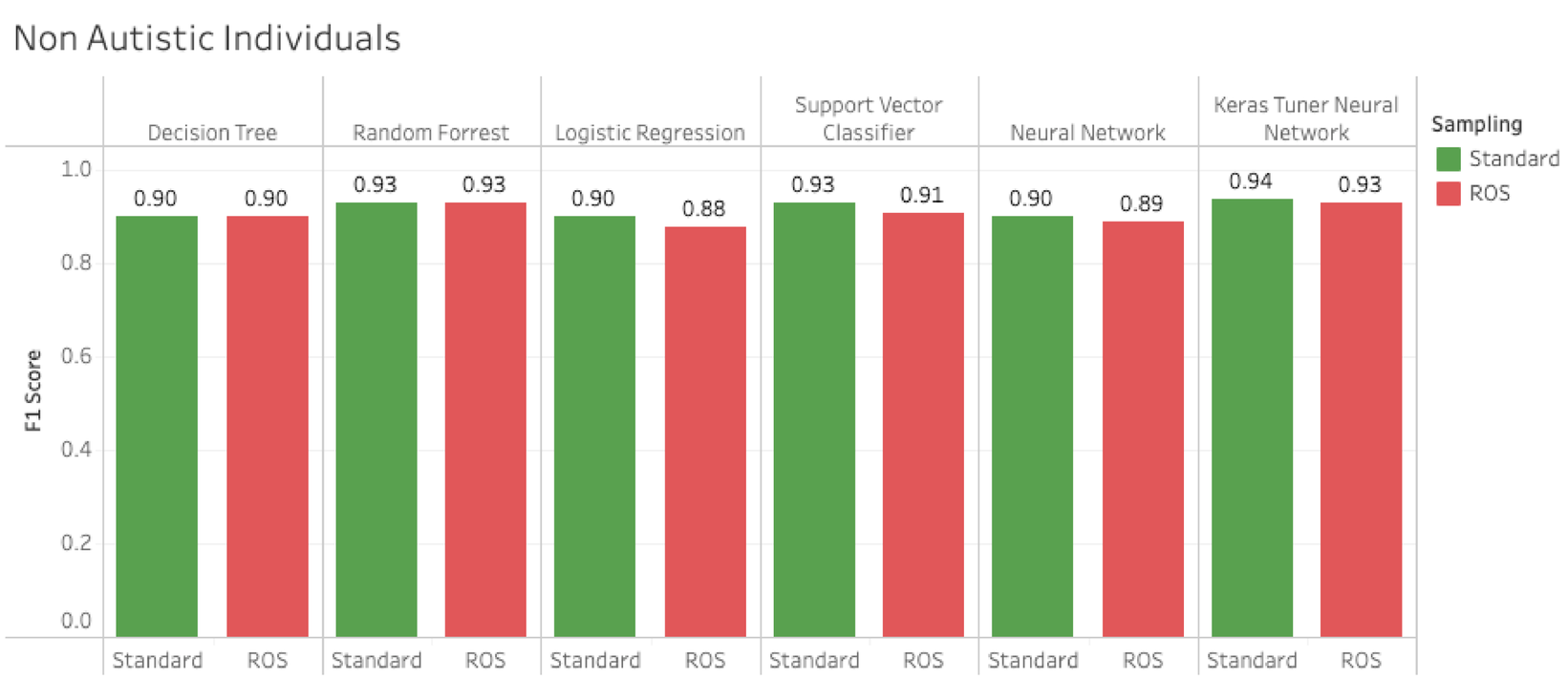
Random Oversampling(ROS)

- ROS Decision Tree
- ROS Random Forest
- ROS Logistic Regression
- ROS Support Vector Classifier
- ROS Neural Network
- ROS Keras Tuner Neural Network



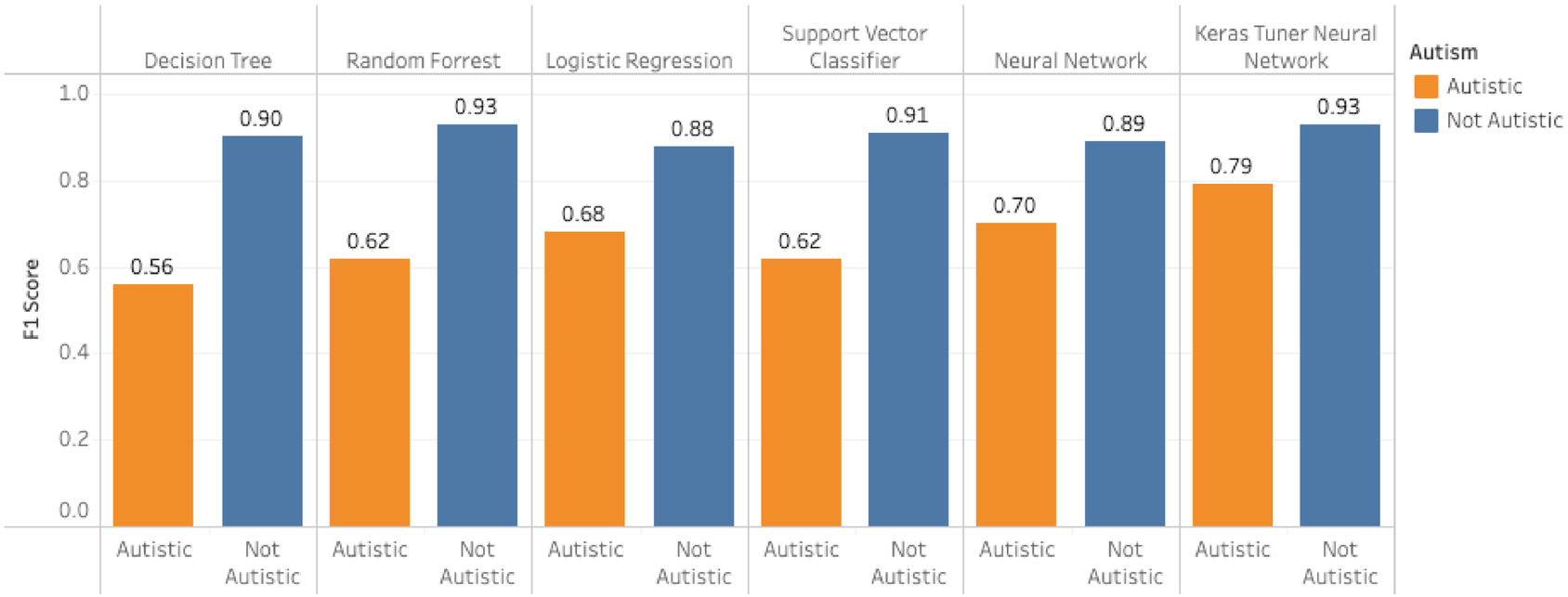


F1 SCORES BY ANALYSIS & DIAGNOSIS

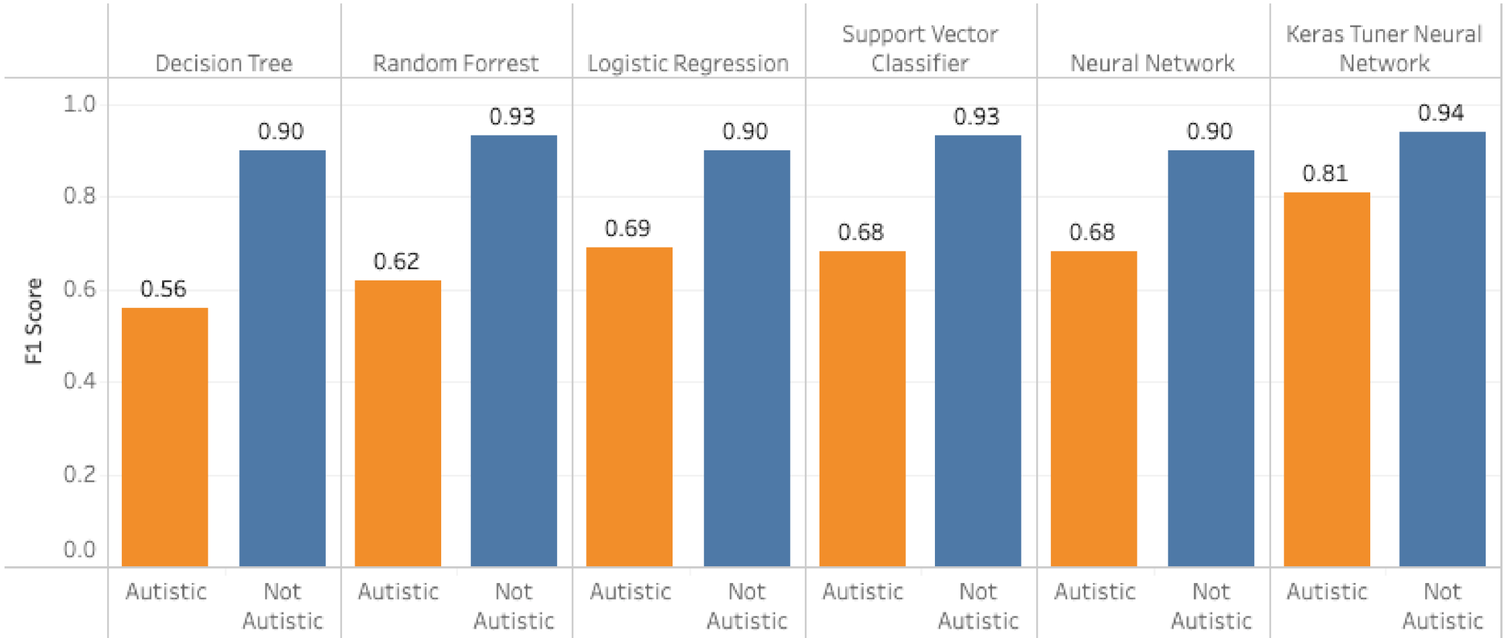


F1 SCORES BY ANALYSIS & SAMPLING

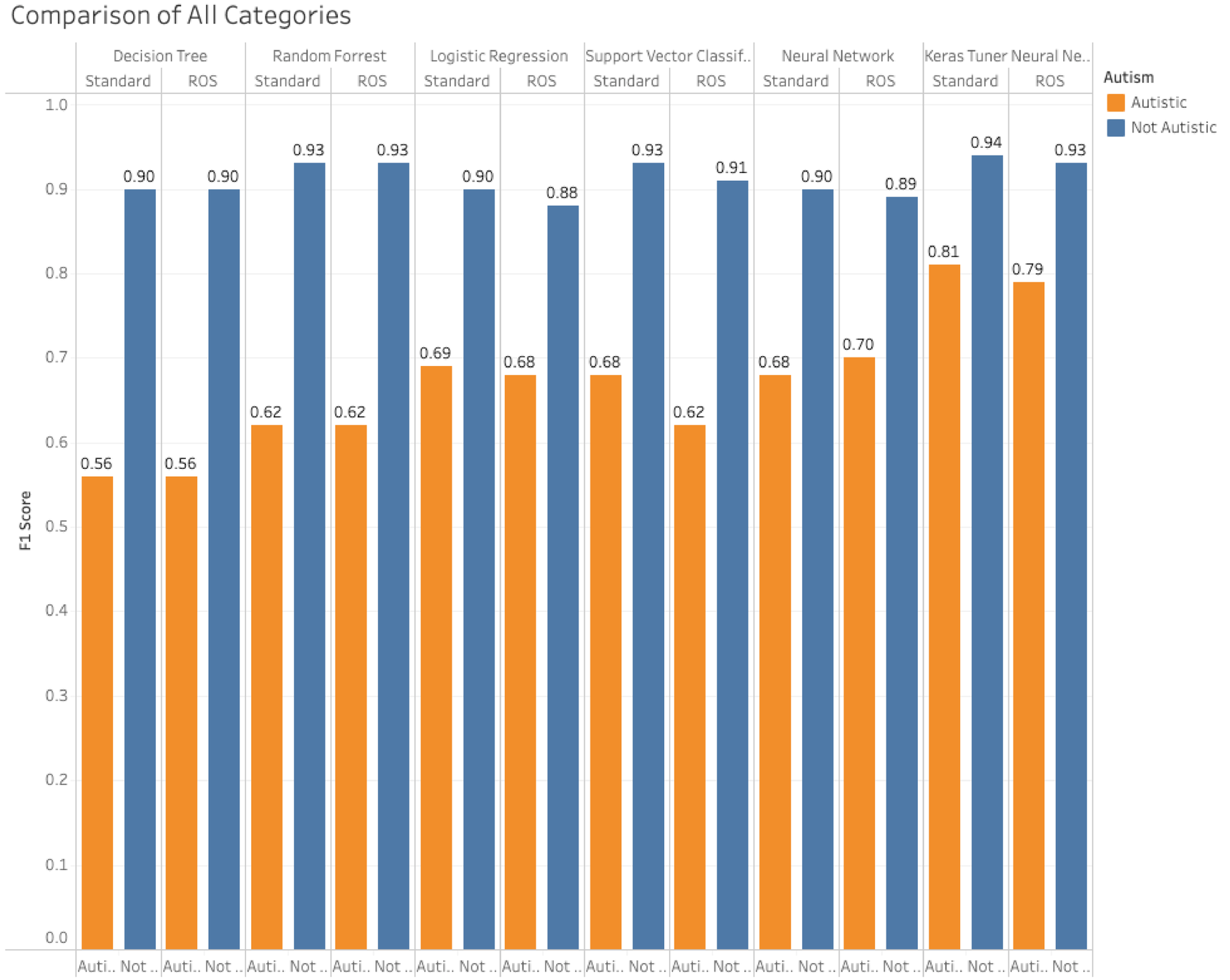
Random Oversampling



Standard Analysis



F1 SCORES BY ANALYSIS, SAMPLING & DIAGNOSIS



ACCURACY


Accuracy Score by Analysis
Decision Tree
Random Forrest
Logistic Regression
Support Vector Classifier
Neural Network
Keras Tuner Neural Network
Sampling
Standard
ROS
Accuracy
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.83
0.83
0.89
0.89
0.84
0.82
0.89
0.85
0.85
0.84
0.91
0.89
Standard
ROS

CONCLUSIONS

01

ROS

Random Oversampling can improve model performance in some cases. However the optimal overall model was generated with a standard analysis.

02

DIAGNOSIS

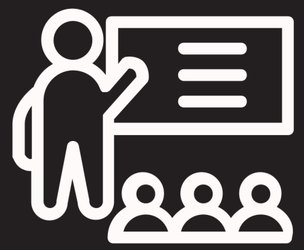
All models were consistently more effective in identifying non-autistic individuals. Emphasis should be made on generating models that are more capable of identifying autistic individuals.

03

F1 VS. ACCURACY

F1 scores should be used to compare models, as this measure is more sensitive to errors. Accuracy is a poor measure as it did not accurately portray all performance.

IMPACT OF A MODEL



Teachers

Striving to optimally support every student.

Parents

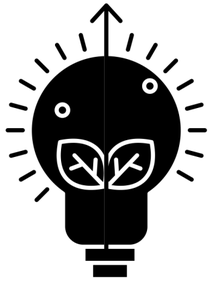
Desperate to understand and empower their children.

Autistic Individuals

Struggling with self-definition, acceptance, and integration.

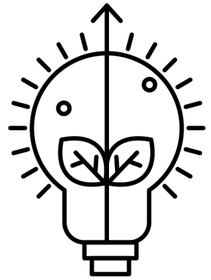
REFLECTIONS

Shortcomings



- Limited data
- Not very much data
- Accuracy only at 91%
- F1 Scores only at .81 & .94

Next Steps



- Wider dataset (more variables and info on each individual)
- More data (more people with and without autism in the dataset)
- More accurate model (using advances in machine learning, neural networks, and deep learning to achieve greater accuracy)

RESOURCES

This project can be accessed on github at: https://github.com/NeonOstrich/Predicting-Autism-with-Machine-Learning

The dataset we used is available at the following link: https://www.kaggle.com/competitions/autism-prediction/data.

Key Statistics obtained from:
The CDC
Rutgers University
Children's Hospital of Philadelphia