Interactive Machine Learning Model

Chronic Kidney Disease

Nicholas Buse, Mrunmai Gadbail, Spencer Garrett, William Gray Renton

Dataset Introduction:

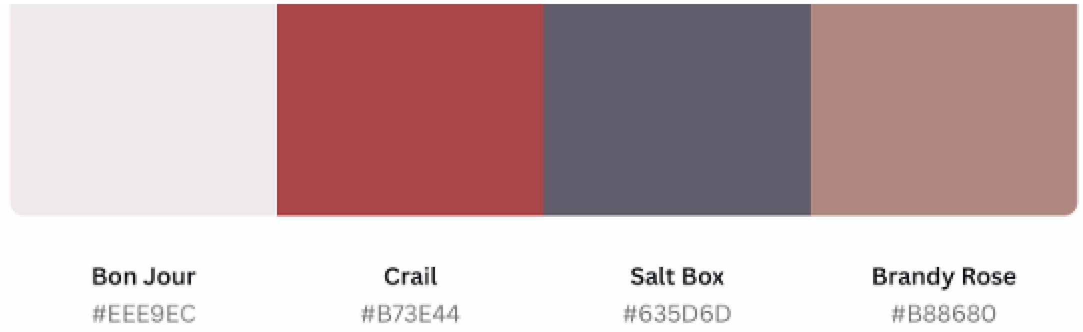
The dataset that we chose was the Chronic Kidney Disease Dataset from Kaggle.com. The dataset contains information from 1,659 patients diagnosed with Chronic Kidney Disease (CKD). It contains a number of factors involving the patients' family history, lifestyle, and environment. Using this dataset, we built a prediction model that would take in information from a user and return the chances of their information matching a positive diagnosis. We then included this model in an interactive dashboard that we published online so users could enter their information and get a prediction with the click of a button.

Data Cleaning:

Our data didn't require a lot of cleaning, but it did require some preparation work to be ready for our machine learning. The columns we dropped were the DoctorInCharge and PatientID columns, since their values weren't influential to our model and also dealing with data in this area, patients' confidential information should be protected. Some other work we did in preparation was scaling some columns and changing some categorical columns to numeric. This work will be expanded on in the Machine Learning part of this document.

Color Design:

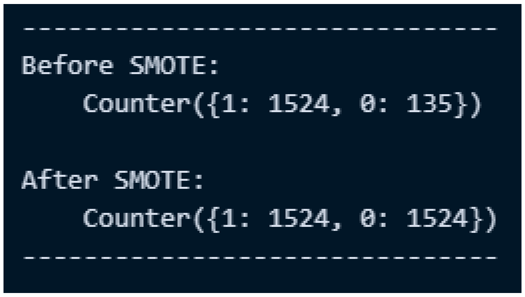
Our color palette was inspired by other medical related sites and works, such as the Red Cross and CKD Patients EDA, a dashboard published on Tableau Public. We chose:



Machine Learning Experiment:

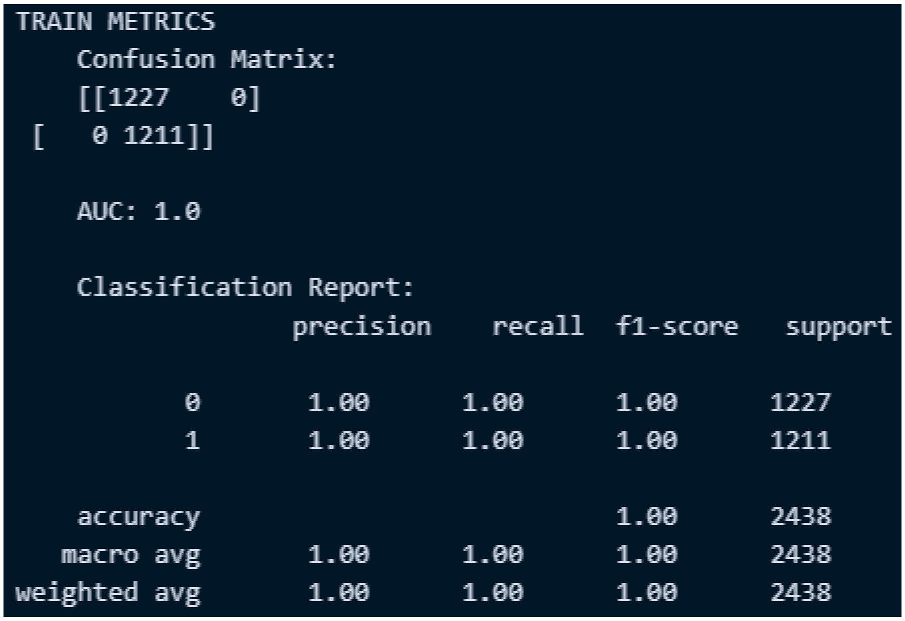
After initial cleaning and examination of our dataset, we first checked to see if our dataset was in balance in relation to the number of positive to negative diagnoses. This was unfortunately not the case. Our positives greatly out-numbered our negatives.

To solve this we implemented the balancing technique SMOTE, synthetic minority over-sampling technique. This method creates duplication of the minority group until balance is achieved with the majority group. This was our method in combating imbalance and over-fitting that some other methods produce.

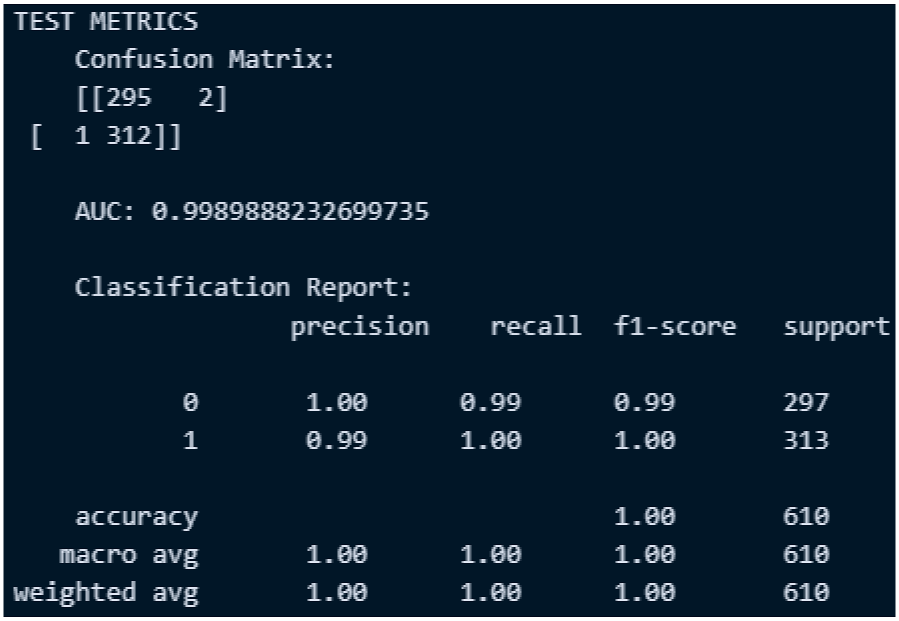


```
--------------------------------
Before SMOTE:
    Counter({1: 1524, 0: 135})

After SMOTE:
    Counter({1: 1524, 0: 1524})
--------------------------------
```

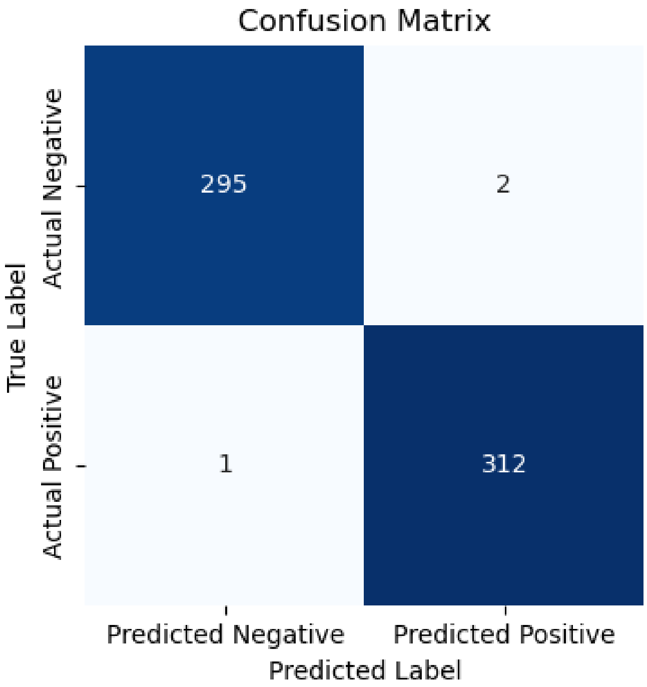
After running several models, including lgbm and xgboost, we compared their results with our end product model, SVC (Support Vector Classifier), and chose this as our prediction model.

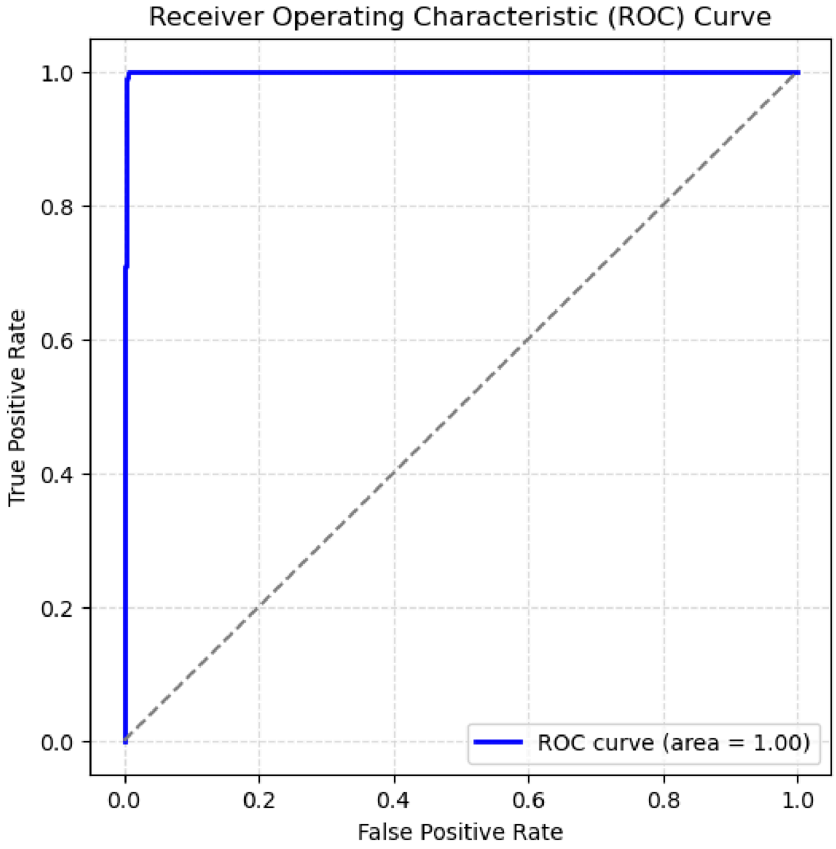


```
TRAIN METRICS
    Confusion Matrix:
    [[1227    0]
 [   0 1211]]

    AUC: 1.0

    Classification Report:
                  precision    recall  f1-score   support

           0       1.00      1.00      1.00      1227
           1       1.00      1.00      1.00      1211

    accuracy                           1.00      2438
   macro avg       1.00      1.00      1.00      2438
weighted avg       1.00      1.00      1.00      2438
```



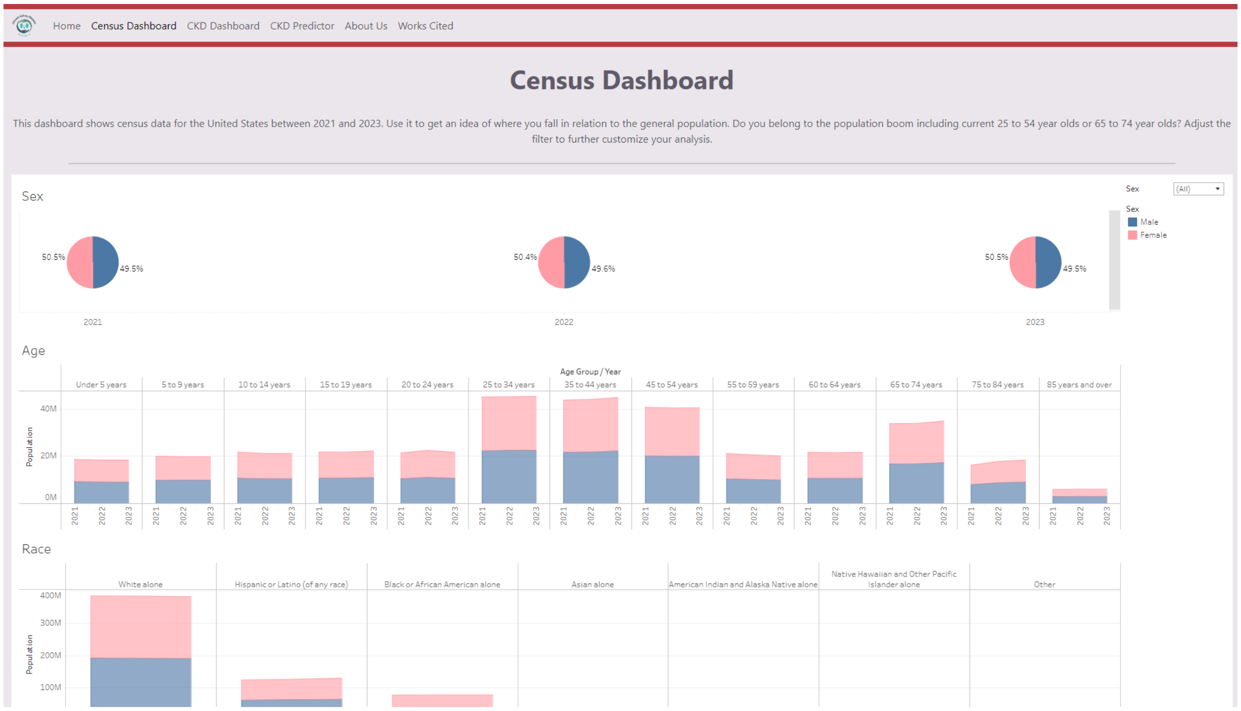
```
TEST METRICS
    Confusion Matrix:
    [[295   2]
 [  1 312]]

    AUC: 0.9989888232699735

    Classification Report:
                  precision    recall  f1-score   support

           0       1.00      0.99      0.99       297
           1       0.99      1.00      1.00       313

    accuracy                           1.00       610
   macro avg       1.00      1.00      1.00       610
weighted avg       1.00      1.00      1.00       610
```


Confusion Matrix
True Label
Actual Negative
Actual Positive
295
2
1
312
Predicted Negative
Predicted Positive
Predicted Label


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (area = 1.00)

Dashboard Design:

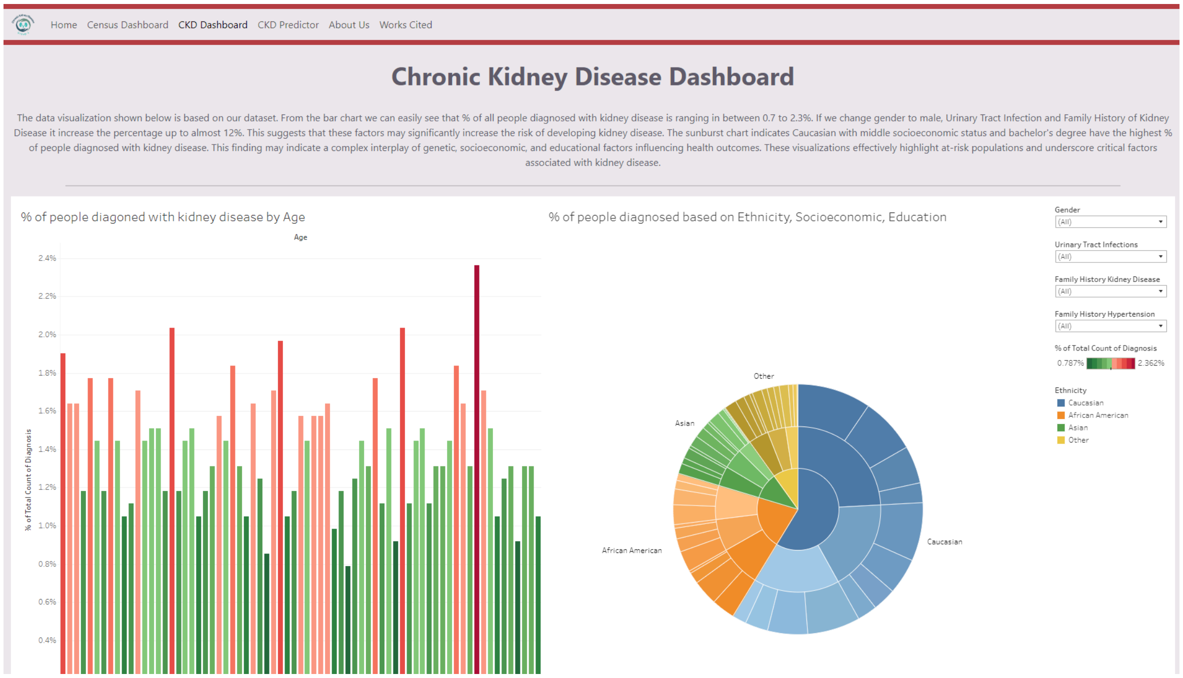
Our dashboard design folds into our website design. The user starts out on the landing page, where it asks the user if they are ready to calculate their chances of having a positive diagnosis of Chronic Kidney Disease. The user then moves to the first of our dashboards, a breakdown of where they stand in relation to the general population of the United States. Here they can see a breakdown of sex, age, and race of the United States.



Census Dashboard

This dashboard shows census data for the United States between 2021 and 2023. Use it to get an idea of where you fall in relation to the general population. Do you belong to the population boom including current 25 to 54 year olds or 65 to 74 year olds? Adjust the filter to further customize your analysis.

Our second dashboard compares data based on our dataset. Here, the user can see the percentage of people diagnosed with CKD by age and also, in a sunburst chart, the user can see the percentage of people diagnosed with CKD based on ethnicity, socioeconomic, and education.



Chronic Kidney Disease Dashboard

The data visualization shown below is based on our dataset. From the bar chart we can easily see that % of all people diagnosed with kidney disease is ranging in between 0.7 to 2.3%. If we change gender to male, Urinary Tract Infection and Family History of Kidney Disease it increase the percentage up to almost 12%. This suggests that these factors may significantly increase the risk of developing kidney disease. The sunburst chart indicates Caucasian with middle socioeconomic status and bachelor's degree have the highest % of people diagnosed with kidney disease. This finding may indicate a complex interplay of genetic, socioeconomic, and educational factors influencing health outcomes. These visualizations effectively highlight at-risk populations and underscore critical factors associated with kidney disease.

% of people diagoned with kidney disease by Age

% of people diagnosed based on Ethnicity, Socioeconomic, Education

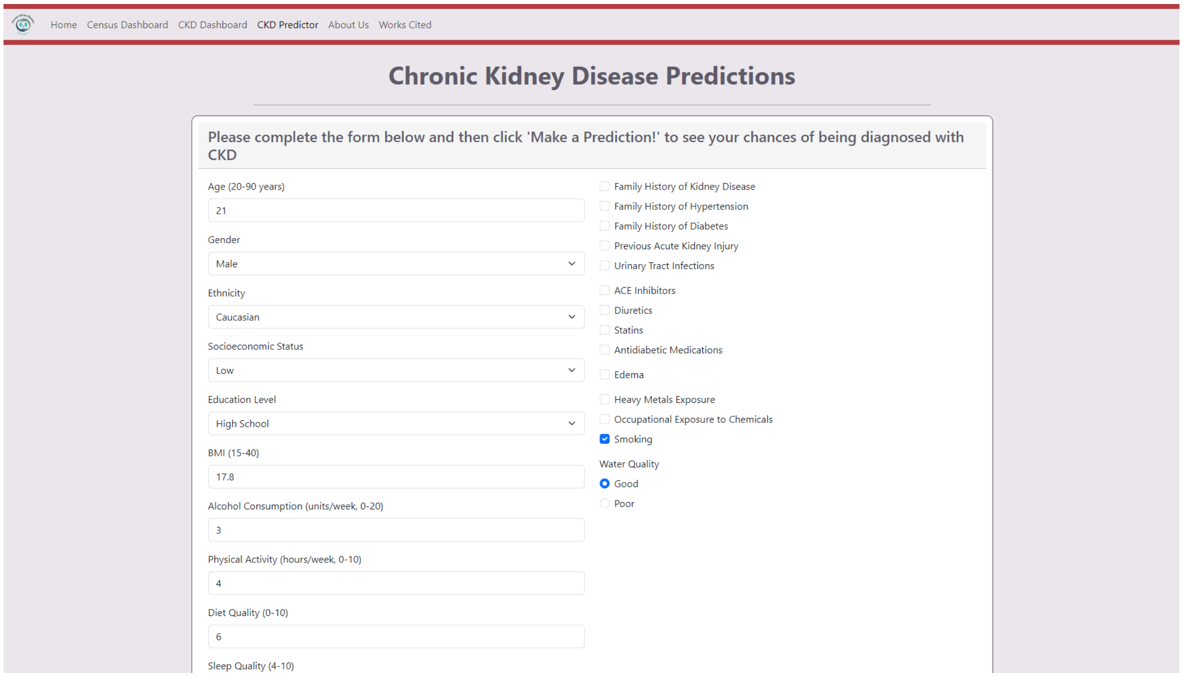
Gender: (All)

Urinary Tract Infections: (All)

Family History Kidney Disease: (All)

Family History Hypertension: (All)

The next page is our Machine Learning page. Here, the user will be able to input their own data and have it fed into our prediction model. They are able to input values such as age, sex, whether they have a family history of kidney disease or hypertension, and other values they may see in a blood test from their doctor. After clicking "make prediction," the model will return their percentage chance of having a positive or negative diagnosis according to our dataset.



Chronic Kidney Disease Predictions

Please complete the form below and then click 'Make a Prediction!' to see your chances of being diagnosed with CKD

Age (20-90 years)
21

Gender
Male

Ethnicity
Caucasian

Socioeconomic Status
Low

Education Level
High School

BMI (15-40)
17.8

Alcohol Consumption (units/week, 0-20)
3

Physical Activity (hours/week, 0-10)
4

Diet Quality (0-10)
6

Sleep Quality (4-10)

□ Family History of Kidney Disease
□ Family History of Hypertension
□ Family History of Diabetes
□ Previous Acute Kidney Injury
□ Urinary Tract Infections

□ ACE Inhibitors
□ Diuretics
□ Statins
□ Antidiabetic Medications

□ Edema

□ Heavy Metals Exposure
□ Occupational Exposure to Chemicals
☑ Smoking

Water Quality
● Good
○ Poor

The final pages are our About Us and Work Cited. Users can find out more about us as creators of the site, see other projects we've worked on following our github and linked in links, and see what references we have used for this project.

Call To Action:

As with any medical or information site, users should always consult their physician with questions regarding their health. Our site can however, give users some small insight into how a value changing drastically could potentially affect a positive or negative diagnosis.

Bias/Limitations:

Dataset quality is important to any machine learning project. One limitation of our dataset that causes potential bias involved the percentage of positive diagnosed out-numbering the amount of negative, skewing our model in the direction of predicting positive. This was in part intentional due to the nature of our subject in that we believe it would be better to produce a false positive result than a false negative. Another limitation of our dataset is the size of the set in general compared to the percentage of the population that suffer from CKD. A larger dataset would be better in creating more accurate predictions.

Conclusion/Reflection:

On reflection, our group would like to see our model perform on a larger dataset to increase accuracy and overcome potential bias. We feel, however, that our methods dealing with oversampling were effective in predicting with a high percentage, lending users greater insight into their own health and how factors affect diagnosis.

Works Cited:

Bootstrap. (n.d.) *Bootstrap - Open-source toolkit for developing with HTML, CSS, and JS.* Retrieved October 1, 2024, from https://getbootstrap.com/

Diamandis, D. (2023). **CKD Patients EDA**. Tableau Public. Available at: https://public.tableau.com/app/profile/dimitris.diamandis/viz/CKDPatientsEDA/Dashboard1

El Kharoua, R. (2023). **Chronic Kidney Disease Dataset Analysis.** Kaggle. Available at: https://www.kaggle.com/datasets/rabieelkharoua/chronic-kidney-disease-dataset-analysis/data

U.S. Census Bureau. (2023). **Data Tables**. Available at: https://data.census.gov/table