TechStat Health Solutions

Final Report

Chronic Kidney Disease Detection and Prevention Plan

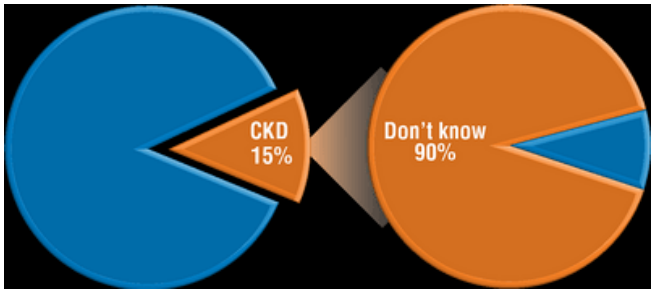
May 23, 2023

Team Members: Sylvia Chung, Kati LiPetri, Nick Malone, Spencer Puterbaugh, and Josh Squires

Table of Contents

Problem Overview

According to the United States Centers for Disease Control and Prevention (CDC), there are approximately 37 million adults in the US (15% of the adult population) who have chronic kidney disease (CKD), but 9 out of 10 of those adults are not aware that they have CKD[1].



Kidneys are an important organ of the human body that function to filter waste and toxins from blood, filtering all blood in a person's body every 30 minutes. CKD is a medical condition with varying levels of severity in which a person's kidneys are damaged and are unable to filter blood as well as they should. In a person with CKD, waste and toxins are not fully filtered and removed from the blood, meaning they remain in the body and cause additional health problems like heart disease, stroke, anemia, increased infections, mineral imbalance in the body, loss of appetite, and lower quality of life[2].

If left untreated, CKD eventually turns into kidney failure, also known as end-stage renal disease (ESRD). By this point dialysis or kidney transplant are needed for the person to survive. It is not possible to reverse damage already done to kidneys, but if caught earlier, steps can be implemented to slow the progression of CKD[2].

TechStat Health Solutions is a healthcare consulting company dedicated to improving the health and lives of people around the world. TechStat's goals are to develop a solution for earlier detection and prevention of CKD and ESRD in the US that is easy to access and use by healthcare providers and patients.

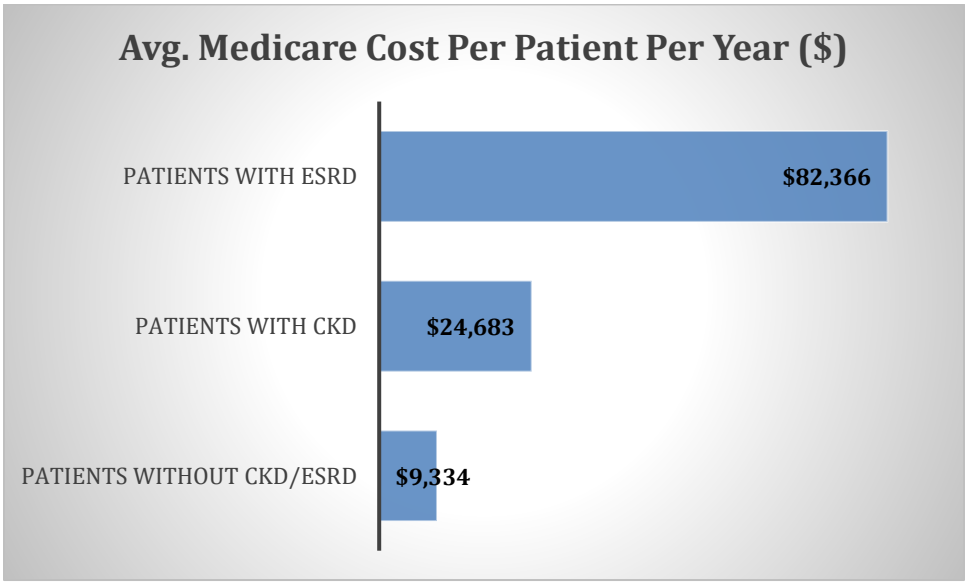
The following Initial Findings report outlines TechStat's initial findings to better detect and prevent CKD and ESRD. The most significant factors in detection of CKD from health will be determined along with non-health factors that could play a role in whether a person develops CKD, such as environmental pollution levels (air and water), food desert locations, and socioeconomic status throughout the US.

[1] "Using Electronic Health Records to Identify Patients with Chronic Kidney Disease." Centers for Disease Control and Prevention. Centers for Disease Control and Prevention, February 18, 2022. https://www.cdc.gov/kidneydisease/publications-resources/electronic-health-records.html

[2] "Chronic Kidney Disease Basics." Centers for Disease Control and Prevention. Centers for Disease Control and Prevention, February 28, 2022. https://www.cdc.gov/kidneydisease/basics.html

Financial Impact

Chronic Kidney Disease not only impacts people's health and livelihood individually, but the collective economic burden of CKD on the US healthcare system is substantial. According to data on Medicare spending in 2020, the average cost per patient per year was 2.6x more for patients with CKD and 8.8x more for patients with ESRD when compared to patients without CKD or ESRD[3].



In 2020, 13.9% of Medicare fee-for-service patients aged 66 or older were diagnosed with CKD, and this group accounted for about 25% of total Medicare fee for service spending (or $75 Billion). In addition, Medicare fee for service spending for patients of all ages diagnosed with CKD accounted for 23.5% of total Medicare fee for service spending (or $85.4 Billion)[3].

Implementation of early detection and prevention of CKD has the potential to save Medicare up to $53 Billion per year based on costs per patient per year and total spending on CKD in 2020.

[3] United States Renal Data System. 2022 *USRDS Annual Data Report: Epidemiology of kidney disease in the United States.* National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2022.

Project Goals

TechStat is dedicated to formulating a user-friendly platform and mobile application that not only educates the public but supplements the confidence of the next plan of action for the user with the goal of early detection and prevention or slowing the progression of CKD. Our application will be within reach of physicians, patients, and related organizations to maintain a current understanding of the state of CKD within relevant geographic locations throughout the US.

Objective	Action	Outcome
Determine variables that indicate high correlation of risk for development of CKD and ESRD.	Perform Exploratory Data Analysis of public health data, to maintain compliance with the Health Insurance Portability and Accountability Act.	Identification of specific locations, demographic, and health factors that are at high risk and indicate comorbid ailments.
Create a predictive model for CKD and ESRD.	Research and design several machine learning models (i.e., decision tree, random forest, logistic regression, etc.) using relevant health data.	Produce a predictive algorithm with optimal precision and accuracy for early detection of CKD.
Build a user-friendly application for public use, specifically healthcare professionals and patients.	Release application available in app stores online.	The public can use the educational and interactive application for detection of CKD and communication of next steps.
Outreach to those in high-risk areas of CKD and ESRD.	Utilize data analysis to determine the environmental factors and locations in the US that influence risk of CKD.	Advertise to those in high-risk areas the application to gain a user base.
Financial improvement.	Partner with insurance companies, Medicare, and educational institutions who can utilize our platform for detection of CKD.	Patients, insurance companies, and Medicare save money with the prevention and slowed progression of CKD and ESRD.
	Work with testing facilities that perform testing of relevant health data. Support expansion of organizations that advocate for healthy food sources and target environmental pollution.	Facilitate the exposure and accessibility of tests relating to CKD and ESRD and organizations in concurrence with increasing healthy eating and decreasing environmental pollution across the US to detect and prevent CKD.

Description of the Data

Our data sources are summarized in the following table:

Dataset Name	Description	Reference
Chronic Kidney Disease Dataset	Dataset with 25 features (numeric and categorical) that can be used to predict CKD	https://www.kaggle.com/datasets/mansoordaku/ckdisease
US Chronic Disease Indicators: Chronic Kidney Disease	Dataset with information on CKD prevalence in the individual states of the US	U.S. Chronic Disease Indicators: Chronic Kidney Disease | Chronic Disease and Health Promotion Data & Indicators (cdc.gov)
US Renal Data System (USRDS) 2022 Annual Data Report (ADR)	Multiple datasets with information on CKD prevalence and expenditures in the US	Annual Data Report | USRDS (nih.gov)
Kidney Disease Mortality by State	Dataset from the US Centers for Disease Control and Prevention (CDC) with information on CKD mortality in each US state	https://www.cdc.gov/nchs/pressroom/sosmap/kidney_disease_mortality/kidney_disease.htm
Prevalence of Diagnosed CKD by US State and County	Dataset from the CDC with % of the population of Medicare beneficiaries aged 65 years and up diagnosed with CKD in each US county	https://nccd.cdc.gov/ckd/detail.aspx?Qnum=Q705
Food Access Research Atlas	Dataset with supermarket access information in the US as well as demographic information from the US Department of Agriculture (USDA)	USDA ERS - Download the Data
Fast food Establishment Data	Dataset with locations of each fast food establishments in the United States	https://www.kaggle.com/datasets/datafiniti/fast-food-restaurants
Air Quality Data	Dataset with air contaminant data per monitoring location in the US	https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual
Water Quality Data	Dataset with water quality information in the US	Water Quality Data Home

Overview of the Data

CKD Dataset

The Chronic Kidney Disease (CKD) dataset was sourced from Kaggle and downloaded as a csv file[4]. This dataset consists of healthcare related data, collected from 400 different patients, and includes 24 independent variables, 1 dependent variable for classification of whether the patient has chronic kidney disease (i.e., ckd vs notckd), and an identifier column. The list of independent variables contains both numerical and categorical values, such as a patient's red blood cell count and whether the patient has diabetes. See the Appendix A1 for the data dictionary for this dataset. The dataset does contain missing values as well as erroneous values. The handling and correction of which will be addressed in the Description of Transformation of Data section of this document.

Water Quality Data

The water quality dataset was downloaded from the Water Quality Portal from the National Water Quality Monitoring Council and contains water quality data along with geographic location data of each tested water sample[5]. The dataset consists of test result data from water samples collected in the US during 2022. The original dataset consisted of 81 variables and up to 460,232 entries per variable. Of the original 81 variables, only 5 were found to be relevant to the analysis: ActivityLocation/LongitudeMeasure (sample longitude), ActivityLocation/LatitudeMeasure (sample latitude), CharacteristicName (contaminant that was tested for), ResultMeasureValue (test result), and ResultMeasure/MeasureUnitCode (unit of measurement of the test result). The first two variables started as float data types while the remaining 3 variables started as object data types.

Any row containing null value(s) was dropped because if an entry was missing, there was no feasible way to fill in the missing information. Therefore, the information in the row was not useful for the analysis. Any non-numeric characters of the ResultMeasureValue column were removed, and any results that contained the less-than symbol or were reported as not detectable were replaced with 0. The ResultMeasureValue column was then changed from object to float data type. The resulting dataset contained 5 variables with 247,876 entries for each.

The initial cleaning of the Water Quality dataset was performed in GoogleColab using Python.

[4] Iqbal, Mansoor. "Chronic Kidney Disease Dataset." Kaggle, April 13, 2017. https://www.kaggle.com/datasets/mansoordaku/ckdisease.

[5] Water Quality Data Home. Accessed May 1, 2023. https://www.waterqualitydata.us/#mimeType=csv&providers=NWIS&providers=STEWARDS&providers=STORET.

Air Quality Dataset

The air quality dataset was downloaded from the US Environmental Protection Agency (EPA) air data downloadable files and contains air quality data collected from air monitors along with geographic location data of the air monitors[6]. The dataset consists of test result data from air monitored in the US during 2022. The original dataset consisted of 55 variables and up to 51,887 entries per variable. Of the original 55 variables, only 6 were found to be relevant to the analysis: Latitude (sample latitude), Longitude (sample longitude), Parameter Name (pollutant that was tested for), Sample Duration (length of sampling time), Units of Measure (unit of test result), and 99th Percentile (test result that all test results were less than or equal to). The first two variables and the last variable started as float data types while the remaining 3 variables started as object data types.

Of the selected data, there were no rows with null entries, so no rows were dropped. No additional changes were needed for the data or data types. The resulting dataset contained 6 variables with 51,887 entries for each.

The initial cleaning of the Air Quality dataset was performed in GoogleColab using Python.

Food Access Dataset

The food access data included two data sets. The first data set was the food atlas dataset that was downloaded from the US Department of Agriculture website[7]. It contained information about supermarket availability as well as population demographics in the US. The data is from 2019 for supermarket information and the 2010 census data for population and demographic information. The main columns of this data set included population distances from supermarkets. This data set included 147 columns and 72,530 rows. No nulls were included in the data set and no rows were dropped.

The other data set used in this analysis was downloaded from Kaggle[8] and showed the location of each fast food establishment in the United States. There are 15 columns and 10,000 rows. The columns included items such as establishments ID, the longitude and latitude, the city, and state of the establishment. No rows were removed as the null values were in columns that were not used for the analysis.

[6] "AirData Website File Download Page." EPA. Environmental Protection Agency. Accessed May 1, 2023. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual.

[7] "Download the Data." USDA ERS - Download the Data, June 14, 2021. https://www.ers.usda.gov/data-products/food-access-research-atlas/download-the-data/.

[8] Datafiniti. "Fast Food Restaurants Across America." Kaggle, May 30, 2019. https://www.kaggle.com/datasets/datafiniti/fast-food-restaurants.

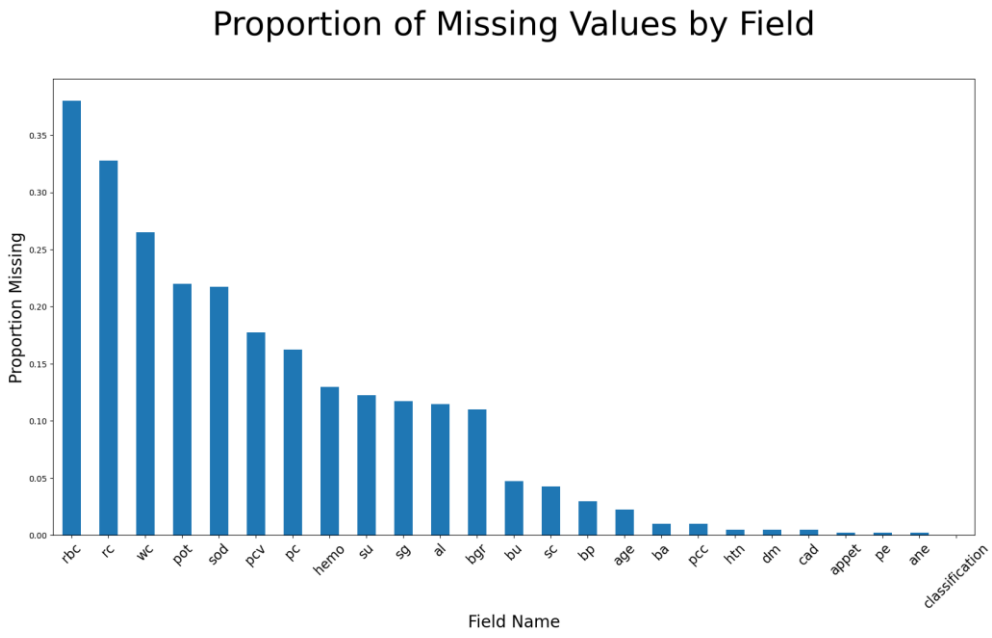
Description of Transformation of Data

CKD Dataset

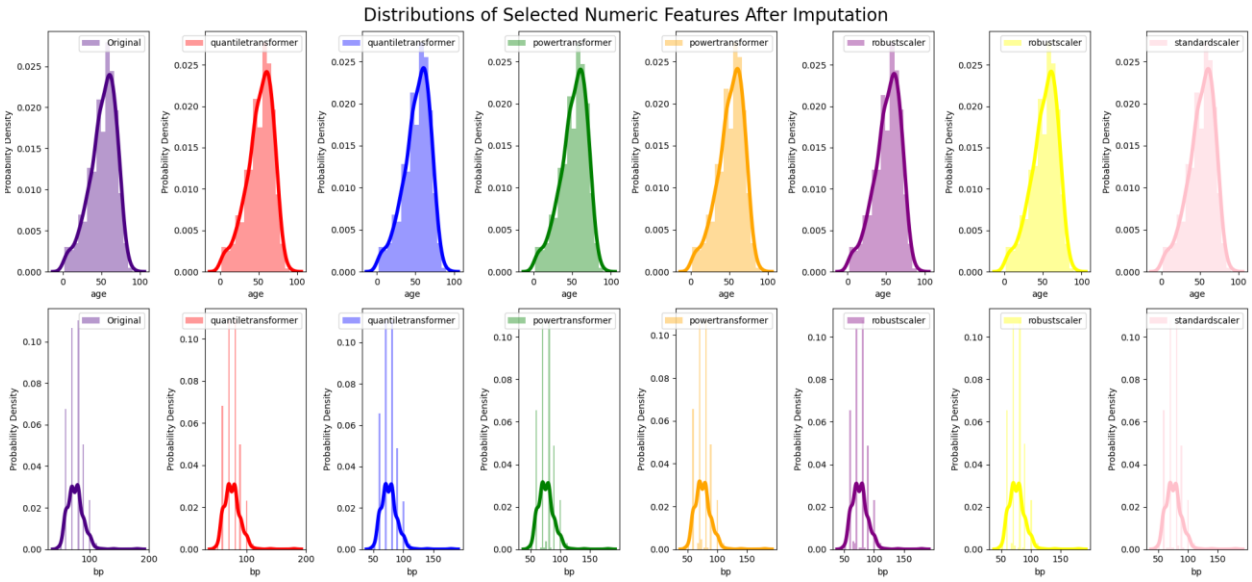
Investigation of the CKD dataset yielded insights into several issues with the data that needed to be corrected before EDA and the model development/selection process could begin. When investigating the values that were present in the independent variables, a number of values were identified that were either erroneous or became malformed during the import process. For example, several instances of tab values ('\t') were identified and would be inappropriate for inclusion in a clean dataset. As all work for this effort has been conducted in GoogleColab, any erroneous or malformed values were corrected using Python.

There were also several independent variables that were initially imported with the wrong data type, specifically Packed Cell Volume (pcv), White Blood Cell Count (wc), and Red Blood Cell Count (rc), which were imported as strings rather than numeric values. The data types for these columns were corrected using Python to ensure later EDA and model development activities were conducted appropriately.

In addition to the aforementioned issues, there were several rows that contained null values in certain columns. After investigation, it was determined that every column that represented an independent variable had at least one value missing, with several columns missing a significant portion of the values. Given the relatively small size of the dataset, it was decided that imputing these values would be more appropriate than dropping the rows that contained missing values in certain columns.



To populate these values, the TechStat team evaluated several different transformers and scalers from the sklearn.preprocessing module in Python as a means of imputation, with the success criteria being the least amount of impact on the distribution of the original feature. After analysis, the optimal transformer, for both numeric and categorical data types, was found to be the Quantile Transformer, and this transformation was used to impute all missing values in the data set that is used in later stages.



The final cleaned dataset contains 400 rows, 24 variable (independent) columns, and 1 target (classification) column.

Water Quality Data

Specific water contaminants, arsenic, lead, cadmium, and mercury were identified as being highly linked to CKD diagnosis and prevalence[9]. So, to study each contaminant individually, four data frames were created from the initial cleaned Water Quality dataset, one for each of the identified high-risk contaminants. It was determined that the test results for each contaminant were not all reported using the same units of measurement. The standard reporting unit for water contaminants is micrograms per liter (ug/L), but many different units were reported such as milligrams per liter (mg/L), micrograms per gram (ug/g), and parts per million (ppm), in addition to others. All test results were converted to ug/L using appropriate calculations, and all units of measurement were changed to ug/L to match the results.

[9] "Environmental Pollution and Kidney Disease." National Kidney Foundation, May 20, 2018. https://www.kidney.org/newsletter/environmental-pollution-and-kidney-disease.

For the arsenic data, two entries were "None" for the test result unit of measurement, so these rows were dropped since the unit of measurement was unknown. After the separation into four data frames, each with 5 variables, and conversion of results to ug/L, the arsenic data frame contained 17,085 entries, the lead data frame contained 14,793 entries, the cadmium data frame contained 10,219 entries, and the mercury data frame contained 3,168 entries.

The transformation of the Water Quality dataset was performed in GoogleColab using Python.

Air Quality Dataset

Specific air pollutants, PM 2.5 (particulate matter that are 2.5 microns in diameter or smaller), Sulfur Dioxide (SO2), and Carbon Monoxide (CO), were identified as being highly linked to kidney damage and CKD[10,11]. So, to study each contaminant individually, three data frames were created from the initial cleaned Air Quality dataset, one for each of the identified high-risk pollutants.

After the separation into three data frames, each with 6 variables, the PM 2.5 data frame contained 9,417 entries, the SO2 data frame contained 1,899 entries, and the CO data frame contained 493 entries.

The transformation of the Air Quality dataset was performed in GoogleColab using Python.

Food Access Dataset

For the food access atlas dataset, the columns used were population, county, state, and population that lived either outside of a 1-mile radius from a supermarket in urban areas and a 20-mile radius from a supermarket in rural areas. The rows were summed by the state to get the total population of the state and the total population living in a 'food desert'. A new variable was created that was the percentage of the population living in a food desert.

The fast-food dataset was used to analyze the number of fast-food establishments by state and population. The number of establishments was summed by state and then concatenated with the food atlas data frame. The total population was divided by 100,000 to get the number of 100,000 people in each state. A new column was created to get the number of fast-food establishments per 100,000 people per state.

The transformation of the Food Access dataset was completed in Jupyter Notebook using Python.

[10] Bonavitacola, Julia. "Long-Term Exposure to Air Pollution Associated with CKD." AJMC. AJMC, July 20, 2022. https://www.ajmc.com/view/long-term-exposure-to-air-pollution-associated-with-ckd.

[11] Fidler, Jessie. "Why Polluted Air May Be a Threat to Your Kidneys." Michigan Medicine University of Michigan, August 24, 2018. https://www.michiganmedicine.org/health-lab/why-polluted-air-may-be-threat-your-kidneys.

Analysis of Data

CKD Dataset Predictive Models

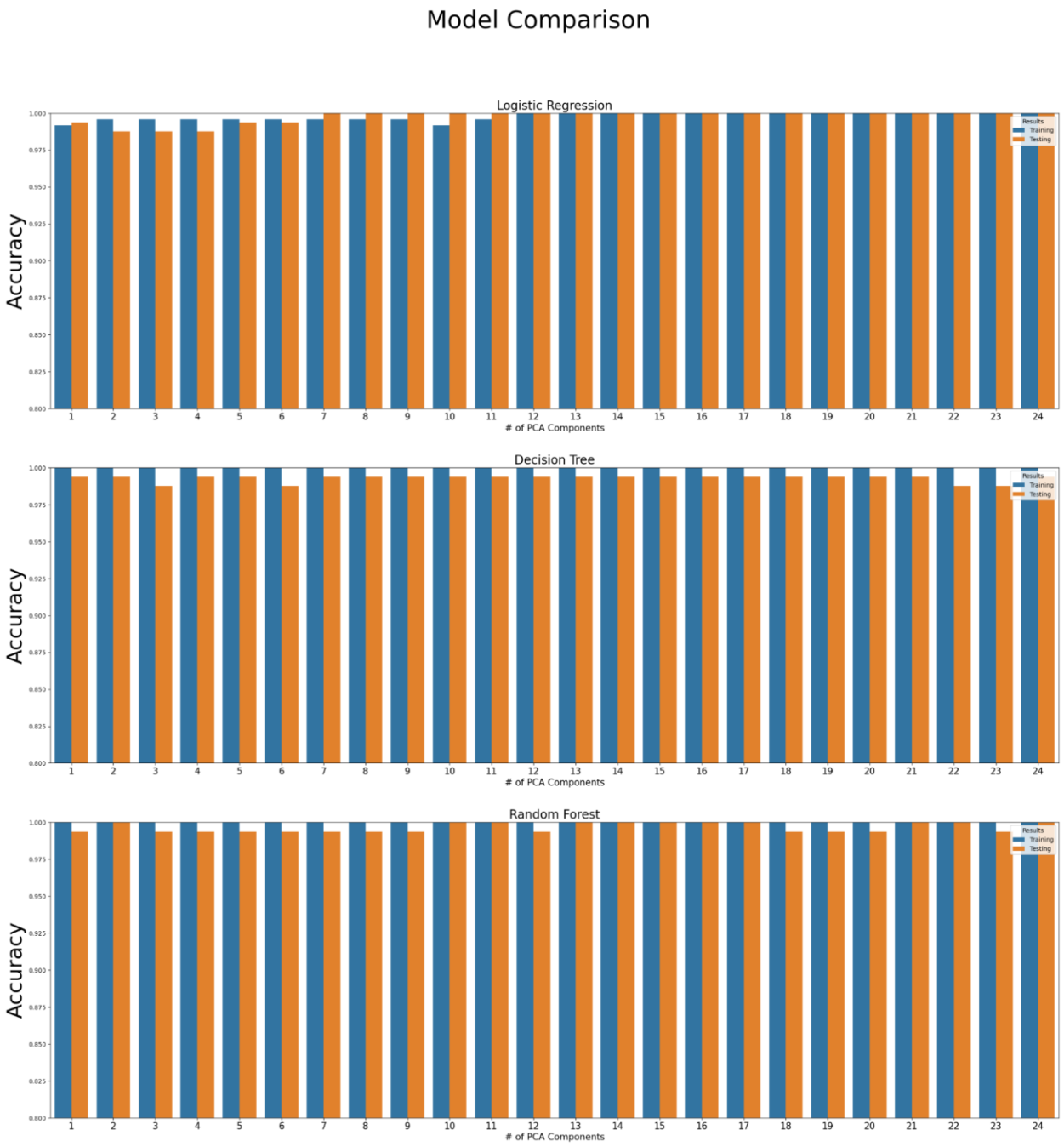
To support the goal of incorporating a predictive model into the project for early detection of CKD, the project team utilized Python to develop several different machine learning models in order to compare the performance and eventually select the optimal model for the predictive component of the application. These models leverage the CKD dataset, with missing values imputed via K-Nearest Neighbor imputation, as the input to train the models, and the following types were evaluated: Random Forest, Decision Tree, Logistic Regression, Gaussian Naive-Bayes, K-Nearest Neighbor (one model for 3 nearest, one model for 8 nearest, and one model for 15 nearest), Support Vector Machine with Radial Basis Function as the kernel, Support Vector Machine with 2nd Degree Polynomial kernel, and Support Vector Machine with 3rd Degree Polynomial Kernel. It should be noted that Principal Component Analysis (PCA) was incorporated with each of these models during training, as well as when evaluating the model on the test split, with the number of PCA components ranging from 1 to 24 (the full range of independent variables).

The models all performed well, with the Logistic Regression, Decision Tree, and Random Forest models providing the most consistent results. The following table shows a summary and comparison of the train and test accuracies for each preliminary model.

Preliminary Model Summary: Scaled Data with 2 PCA Components

Model Type	Train Accuracy	Test Accuracy
Logistic Regression	99%	99%
Decision Tree	100%	99%
Random Forest	100%	100%
Naive Bayes	100%	99%
KNN - Weighted 3 Nearest Neighbors	100%	100%
KNN - Weighted 8 Nearest Neighbors	100%	100%
KNN - Weighted 15 Nearest Neighbors	100%	100%
SVM - Radial Basis Function	99%	100%
SVM - 2nd Degree Polynomial	88%	86%
SVM - 3rd Degree Polynomial	99%	100%

The following figure shows a comparison of the Logistic Regression, Decision Tree, and Random Forest models' train and test accuracy with varying numbers of principal components.



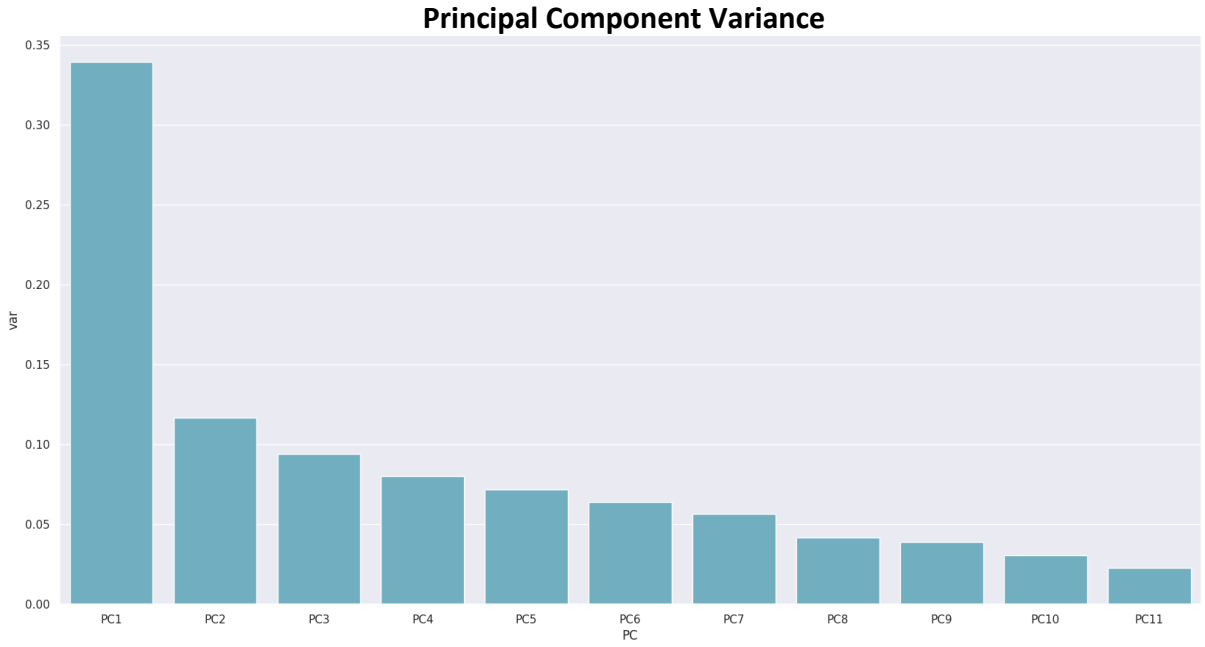
Interestingly, the inclusion of more PCA components did not make a significant difference in the accuracy of the models, except for the Support Vector Machine with 2nd Degree Polynomial kernel; however, this model performed relatively poorly when viewed in the context of the other models, particularly when evaluated against testing data. Additional charts that detail the

accuracy of the training and testing data, for each of the models developed, can be viewed in Appendix A2.

Further effort was dedicated to these models to examine the potential impact of outliers and to reduce the dimensionality of the data during a more robust PCA and t-SNE analyses. Feature selection was also performed to decrease the number of variables needed for predictive modeling. Discussion on PCA, t-SNE, and feature selection are presented in the next few sections.

CKD Dataset PCA

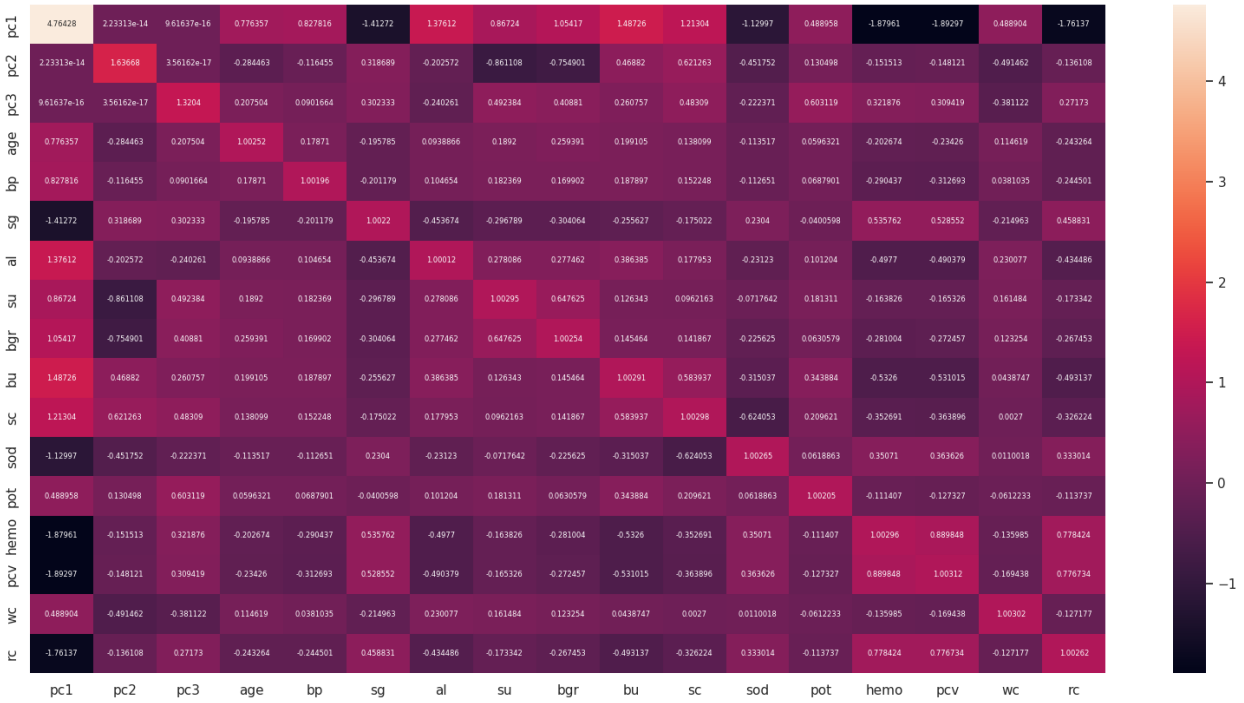
Principal component analysis (PCA) was performed using the scikit-learn package in Python. This utilized 95% of the information from the dataset and reduced the dimensions of the dataset from twenty-four to eleven. The Principal Component Variance figure displays a bar chart of the variance of each principal component.



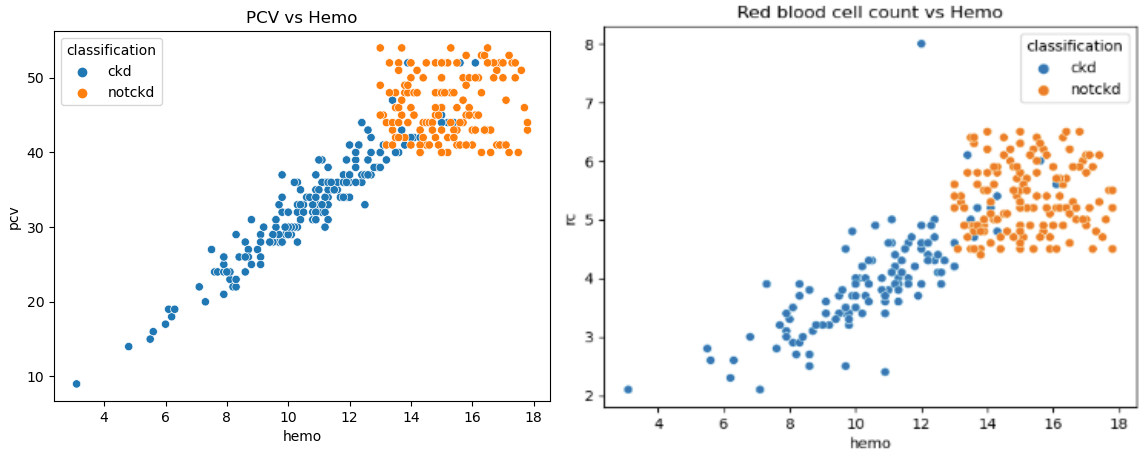
Roughly 55% (33.95%, 11.66%, 9.41% for the first three components, respectively) of the dataset's information is encapsulated by the first three principal components.

These three components were evaluated using the subsequent heatmap. Certain features of the data are highly positively and negatively correlated with the top three principal components.

Heatmap of Dataset Features and Principal Components 1, 2, and 3



Based on the correlation heatmap, red blood cell count (rc), packed cell volume (pcv), hemoglobin (hemo), sodium (sod), and specific gravity (sg) have a high negative correlation with the first principal component. Conversely, serum creatinine (sc), blood urea (bu), blood glucose random (bgr), and albumin (al) are all positively correlated to the first principal component. High rc, pcv, hemo, and sg levels are indicative that the patient is healthy. As shown in the scatterplots of pcv versus hemo and rc versus hemo, patients without ckd are clustered with the higher levels of pcv, rc, and hemo.



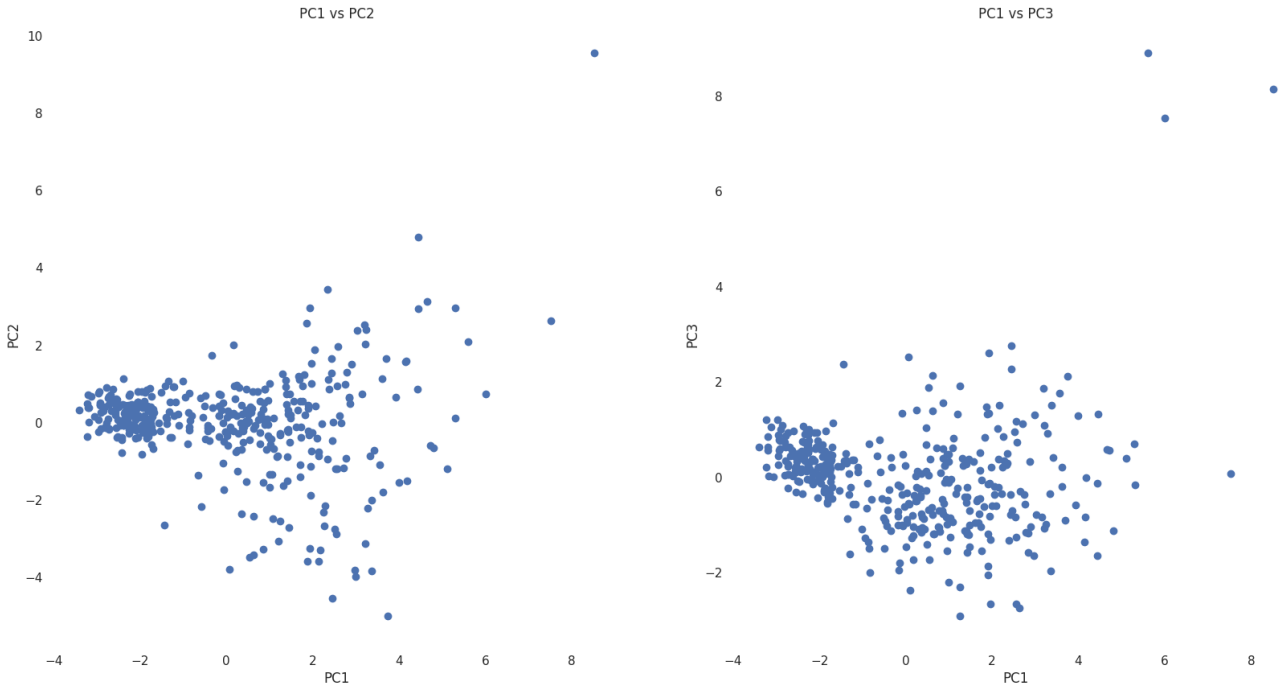
Sg refers to the specific gravity of the patient's urine. If the sg is high, this indicates that the kidneys are filtering waste from the blood properly. This implies that the first principal

component is correlated to patients with ckd. This finding agrees with the positive correlation between PC1 and sc, bu, bgr, and al. Serum creatinine level in the blood is an indication of how well kidneys are filtering waste products from the blood. Creatine is a waste product of normal muscle function that is filtered out of the blood through the kidneys. A higher level of creatine (measured as serum creatinine) in the blood, indicates that the kidneys are not functioning optimally[12]. Similarly, blood urea is also a waste product that gets filtered by the kidneys. Higher levels of urea in the blood could indicate that a patient's kidneys are not functioning properly. Since sc and bu are positively correlated with PC1, this is further evidence that PC1 is likely correlated with patients with ckd.

Principal components two and three have an assortment of correlations; however, not one feature has a strong negative or positive correlation to these principal components. For this reason, it is difficult to draw any conclusions from these.

Focusing on the serum creatinine feature, it is evident that the first principal component is highly correlated with the ckd classification. Therefore, it is likely that any observations that are highly correlated with PC1 are patients with ckd.

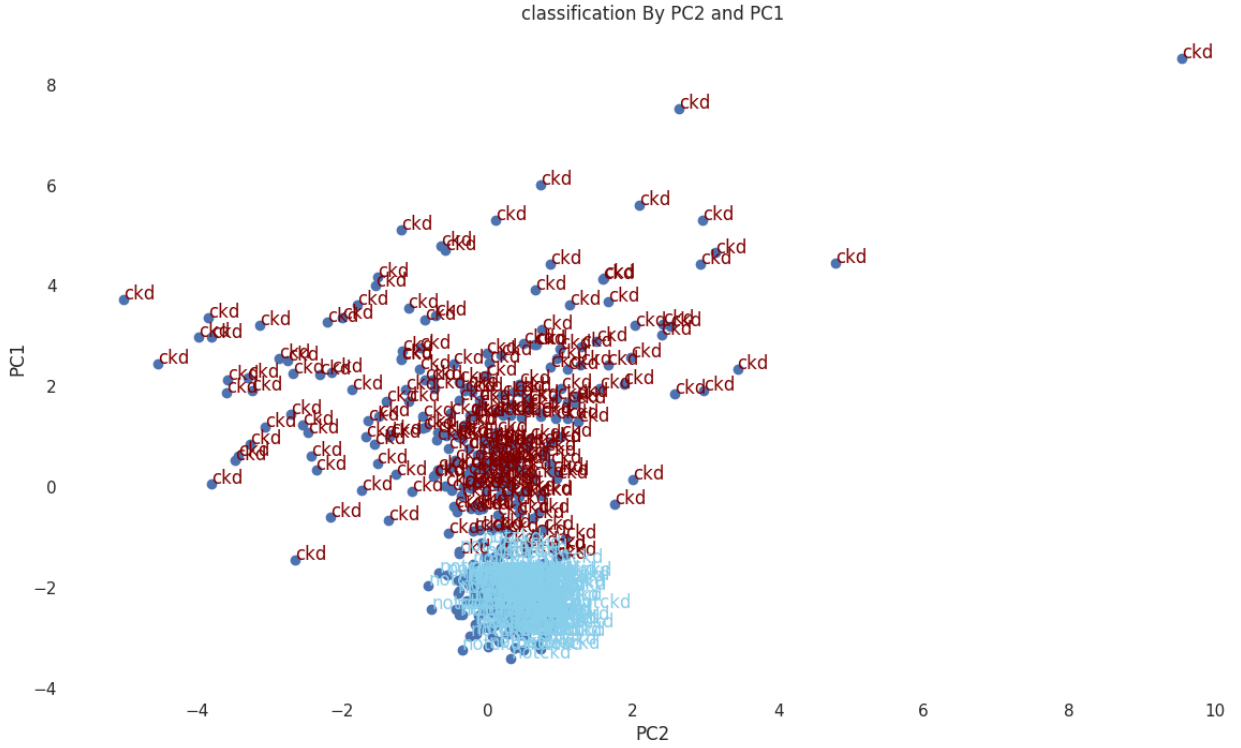
Two scatter plots were created for PC2 versus PC1 and PC3 versus PC1.



[12]"What Is Creatinine?" DaVita. Accessed May 3, 2023. https://www.davita.com/education/kidney-disease/symptoms/what-is-creatinine.

In the first scatter plot, the records are widely dispersed as the axis for PC1 increases. Furthermore, the bulk of the data is fairly level with no correlation to PC2. Accordingly, it is difficult to draw any conclusions on the relationship between PC1 and PC2. Conversely, the second scatter plot illustrates a slight inverse relationship between PC1 and PC3. In other words, as the records are less correlated to PC1, the relationship with PC3 strengthens. This indicates that PC3 is somewhat correlated with the records for non-ckd patients.

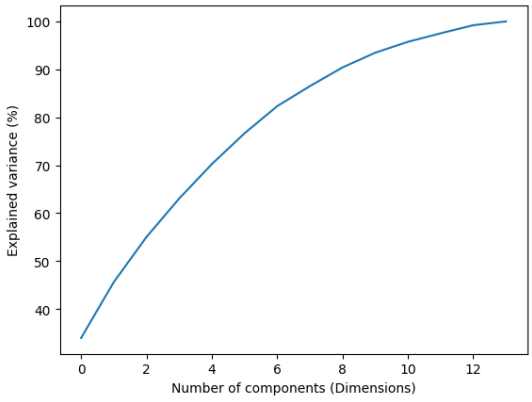
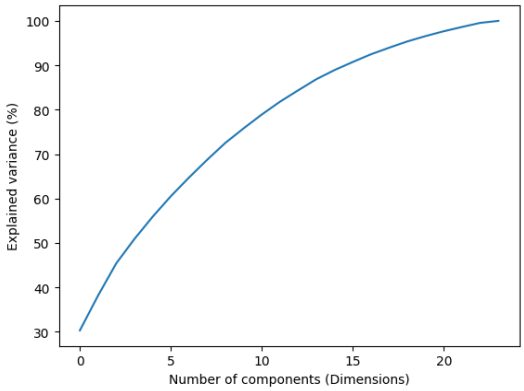
A scatter plot of the classification variable by PC1 and PC2 was created.



Patients with ckd have a higher correlation with the first principal component, but the correlation is weak, as most occurrences happen between zero and two. However, it is significant that there is a clear distinction between ckd and non-ckd records. Non-ckd records are almost purely neutral when compared to PC2. When compared to PC1, all non-ckd records are at least slightly inversely related. This shows that PC1 and ckd class records are related in some capacity.

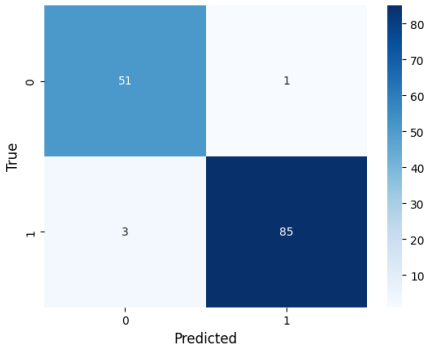
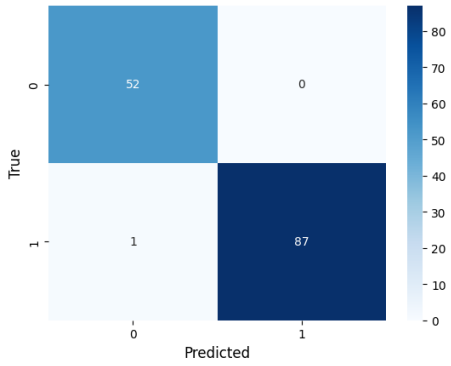
Our goal was to successfully integrate PCA into a predictive model. PCA will cut out some of the noise within the ckd dataset. Ultimately, this will allow the model to be trained quicker without compromising any crucial information. We compared the results of a logistic regression model with and without PCA while experimenting with the use of categorical data. We reduced dimensionality and still captured 90% of the information in the dataset.

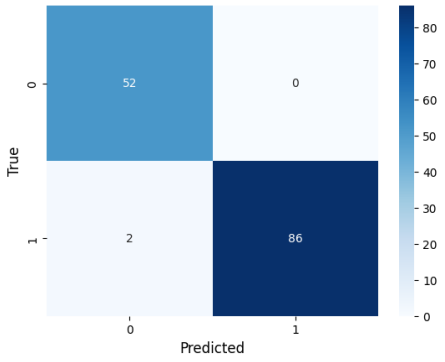
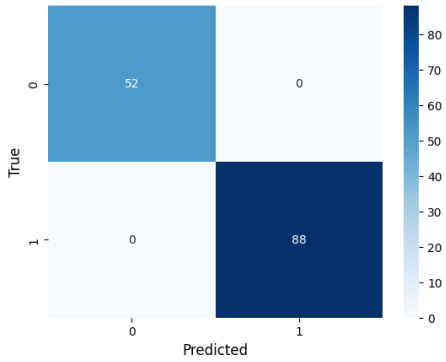
Explained Variance: With and Without Categorical Data

Explained Variance with Respect to Number of Principal Components	
Without Categorical Data	With Categorical Data
	

90% of the dataset's explained variance is captured with 9 principal components when it did not include categorical data. When it did include categorical data, 90% of the information was captured with 15 principal components. These are the number of components used for the respective logistic regression models. The results can be observed in the following confusion matrices.

Logistic Regression Model Confusion Matrices

	Without Categorical Data	With Categorical Data
Without PCA		

	Without Categorical Data	With Categorical Data
With PCA		

Further, the accuracies from these outcomes are illustrated in the table below.

Model Summary: PCA Versus Non-PCA Logistic Regression Model Performance

PCA Logistic Regression Model Comparisons				
Model Description	**Number of Dimensions**	**Test Size**	**Training Accuracy**	**Test Accuracy**
Without PCA, without categorical data	**14**	**35%**	**98%**	**97%**
With PCA, without categorical data	**9**	**35%**	**100%**	**99%**
Without PCA, with categorical data	**24**	**35%**	**99%**	**99%**
With PCA, with categorical data	**9**	**35%**	**100%**	**100%**

In both instances, with and without categorical data, PCA's dimensionality reduction helped the models perform better than without PCA.

CKD Dataset t-SNE

Another method for dimensionality reduction was performed for the CKD dataset: t-distributed stochastic neighbor embedding (t-SNE). Both PCA and t-SNE are unsupervised techniques for visualization and dimensionality reduction of high dimensional data; PCA is a linear method that preserves the global structure of the data, while t-SNE is a non-linear method that preserves the local structure of the data. T-SNE is useful for data that cannot be separated by a straight line.

Using the cleaned and scaled data output from the CKD dataset EDA, the t-SNE algorithm was performed with 2 components and default hyperparameters (perplexity = 30). A t-SNE projection was created in Python Seaborn for the data in a two-dimensional space, as shown in the following figure.

In the t-SNE projection, the red points show those associated with the notckd classification, while the blue points show those associated with the ckd classification. There are two well-defined clusters, one for each classification, with barely any overlap of points. The separation of the two classes in the projection indicates that there is sufficient difference between the two classes for supervised machine learning models to be successful. It also indicates that non-linear classification models will likely perform well with this data.

Several Machine Learning models were performed using the t-SNE dimensionality reduced dataset. It is important to note that there are several challenges when feeding t-SNE reduced

features into an ML model. T-SNE is non-deterministic, meaning that each time it runs, the results can be different even with the same hyperparameters. In addition, t-SNE is more difficult to use when new data is added to the dataset. Once additional data is added to the dataset, the t-SNE embeddings will be recalculated.

The following table summarizes the train and test accuracy for 6 different models using cleaned and scaled data, PCA (2-component) of the scaled data, and t-SNE of the scaled data. All models used a 70/30 split of training/testing data.

Model Summary: Scaled Data, PCA Data, t-SNE Data

Model Type	Scaled Data		PCA		t-SNE	
	Train Accuracy	Test Accuracy	Train Accuracy	Test Accuracy	Train Accuracy	Test Accuracy
Logistic Regression	100%	100%	99%	98%	100%	100%
Decision Tree	100%	100%	100%	99%	100%	100%
Random Forest	100%	99%	100%	99%	100%	100%
SVC	100%	100%	99%	99%	99%	99%
KNN	97%	97%	100%	99%	100%	99%
Naive Bayes	97%	96%	100%	100%	98%	98%

Overall, the models performed best using the PCA of the scaled data, and the decision tree model performed the best on all three datasets. Train accuracy and test accuracy did reach 100% in some cases. This is usually unrealistic and impractical in ML models, but since our dataset is relatively small, with 400 rows, it is not uncommon. The training and test accuracies will likely change once additional patient data is collected and added to the dataset in the future. Adding more patient data in the future will be required to develop a more robust model.

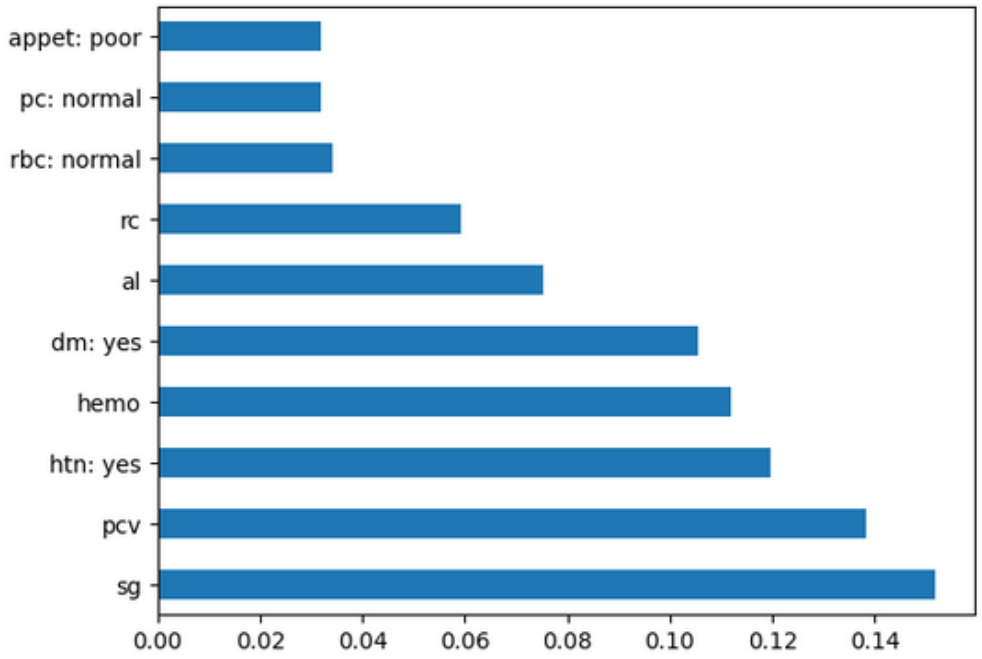
The t-SNE algorithm and machine learning modeling was performed using Python sklearn in Google Colab.

CKD Dataset Feature Selection

The original dataset has 24 variables but not all variables are necessarily common health data tested for at yearly physical appointments, etc. In addition, it can be cumbersome and time consuming for patients and healthcare professionals to enter 24 data points into the predictive model application. So, in the interest of saving time and computing resources, feature selection was used to determine the most important variables for classification of patients with and without CKD.

The Extra Trees Classifier from sklearn in Python was used for feature selection. A bar chart of the top 10 features and their feature importances was created, as shown in the following figure.

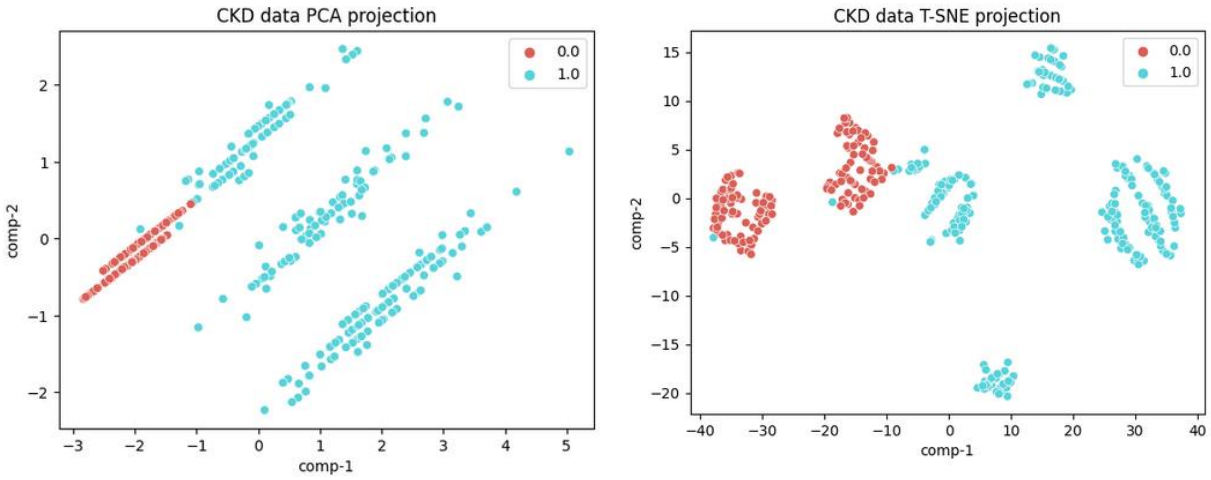
Feature Selection Bar Chart



Based on the output, the top 5 important features of the CKD dataset are sg (specific gravity), pcv (packed cell volume), htn (hypertension), hemo (hemoglobin), and dm (diabetes). Using the cleaned dataset, a new dataset, containing just these top 5 variables, was created to use for further modeling.

PCA and t-SNE dimensionality reduction techniques were performed with the abridged dataset only containing the top 5 variables. Each technique reduced the dataset to two components. The PCA and t-SNE projections are shown in the following figures.

PCA and t-SNE with Top 5 Selected Features



Based on the projections, there are still distinct clusters for the notckd (red) and ckd (blue) classifications. For each projection, the ckd class is shown in 3 or 4 distinct clusters. In the t-SNE projection, the notckd class has two distinct clusters. With the selected features, there is a small amount of overlap between the ckd and notckd classes. Only a few points from the ckd class are within the mostly notckd clusters.

The 6 different models were performed again using the abridged dataset with the top 5 selected features for the scaled data, PCA data (2-component), and t-SNE data. Train and Test accuracy for each model are summarized in the following table. All models used a 70/30 split of training/testing data.

Model Summary: Scaled Data, PCA Data, t-SNE Data with Top 5 Selected Features

Model Type	Scaled Data		PCA		t-SNE	
	Train Accuracy	Test Accuracy	Train Accuracy	Test Accuracy	Train Accuracy	Test Accuracy
Logistic Regression	99%	99%	99%	98%	100%	99%
Decision Tree	100%	100%	100%	99%	100%	99%
Random Forest	100%	99%	100%	99%	100%	99%
SVC	99%	99%	100%	98%	99%	99%

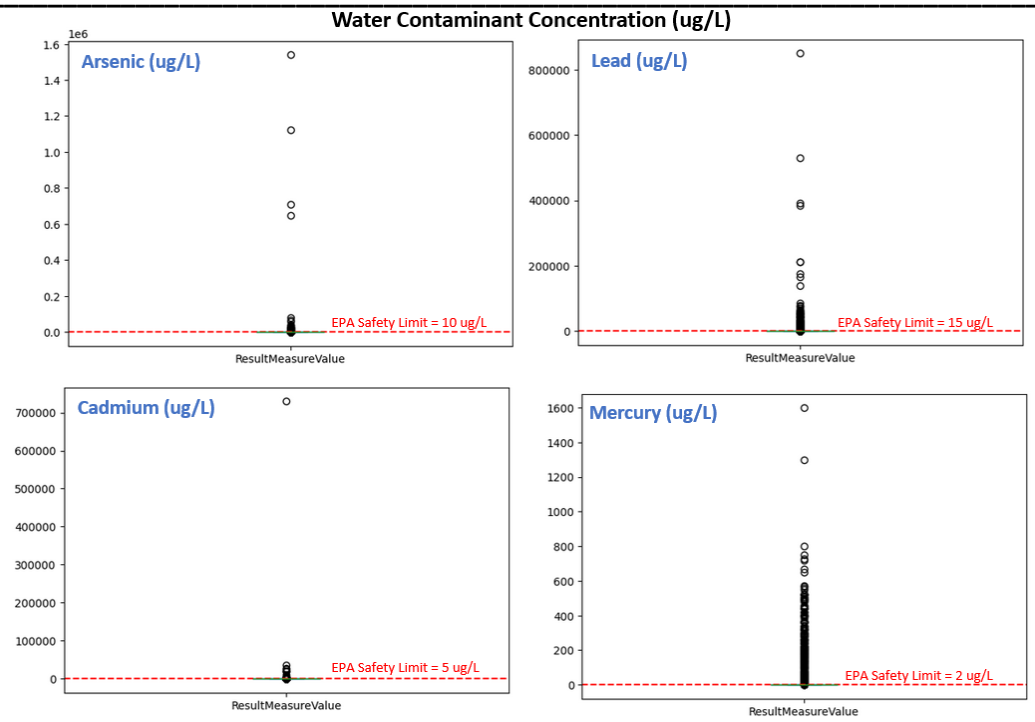
Model Type	Scaled Data		PCA		t-SNE	
	Train Accuracy	Test Accuracy	Train Accuracy	Test Accuracy	Train Accuracy	Test Accuracy
KNN	99%	99%	100%	99%	99%	99%
Naive Bayes	85%	83%	99%	99%	98%	98%

Overall, the models performed best using the PCA of the scaled data, and the decision tree and random forest models performed the best on all three datasets. The Naive Bayes model had the lowest train and test accuracy, with results in the 80% range. Comparing this model to the data with PCA and t-SNE applied shows a large improvement in the Naive Bayes model train and test accuracy, indicating that dimensionality reduction is working as intended. Train accuracy and test accuracy did reach 100% in some cases. This is usually unrealistic and impractical in ML models, but since our dataset is relatively small, with 400 rows, it is not uncommon. The training and test accuracies will likely change once additional patient data is collected and added to the dataset in the future. Adding more patient data in the future will be required to develop a more robust model. The PCA, t-SNE, and machine learning modeling was performed using Python sklearn in Google Colab.

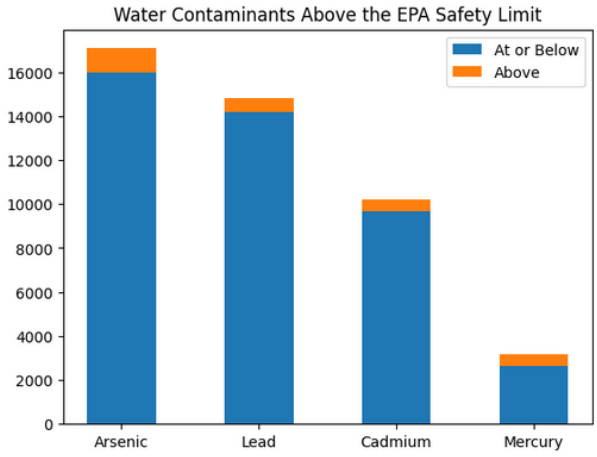
Water Quality Data

The results of each water contaminant were plotted in boxplots to display the spread of the data for each. A dotted reference line in red was plotted onto each visualization to show the maximum concentration limit of each contaminant for safe drinking water per the Environmental Protection Agency (EPA)[13], as shown in the figure below.

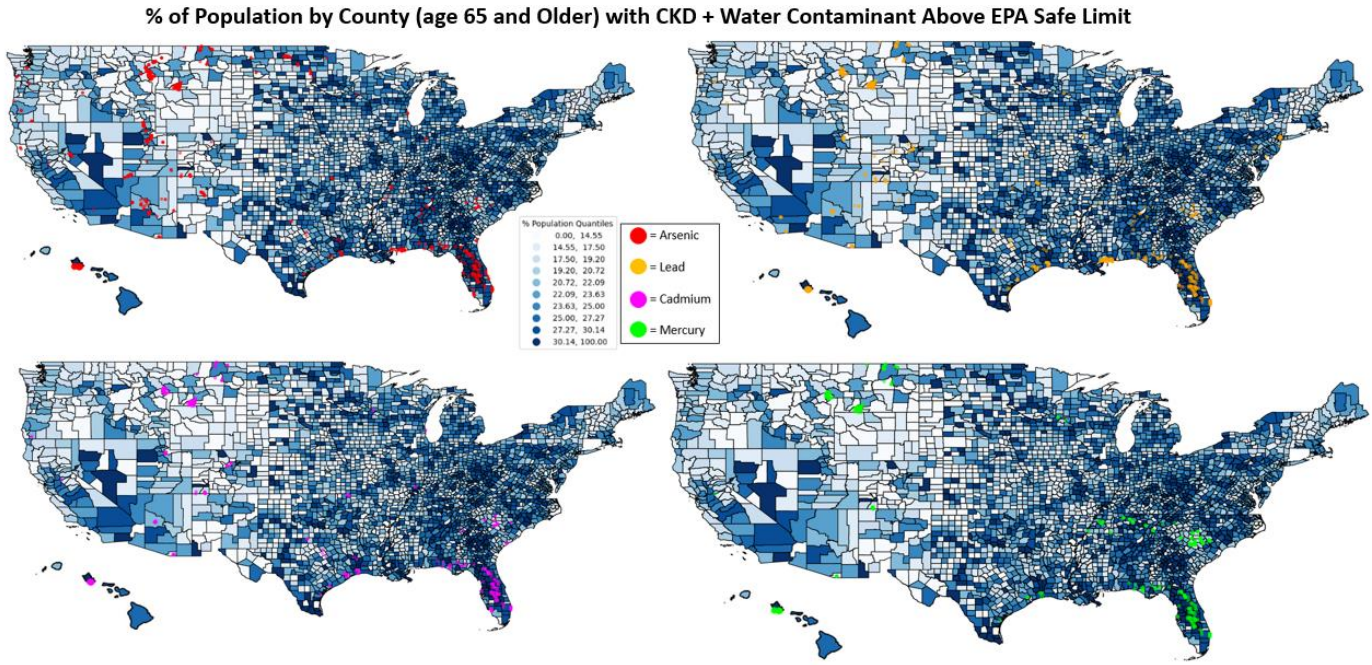
[13] "National Primary Drinking Water Regulations." EPA. Environmental Protection Agency, January 9, 2023. https://www.epa.gov/ground-water-and-drinking-water/national-primary-drinking-water-regulations.



Most of the test results are equal to or below the EPA maximum concentration safety limit, however based on the boxplots, there are many results above the safety limit. For this analysis, the goal is to identify locations of unsafe water, i.e., water with contaminant concentrations above the EPA safety limit. Therefore, all test results for each contaminant that were at or below the EPA safety limits were dropped from the data frames, leaving only results above the EPA limit. After dropping results meeting the EPA limit, the arsenic data frame contained 1,113 entries, the lead data frame contained 586 entries, the cadmium data frame contained 546 entries, and the mercury data frame contained 513 entries. The following bar chart shows the quantities of each contaminant that are at or below the EPA limit or above the EPA limit.



A map was created using data of the US population of medicare beneficiaries age 65 years and older with diagnosed CKD by US county[14]. The water quality data for each contaminant was displayed on top of this map to see if any clusters of contaminated water overlapped with higher incidence of CKD.



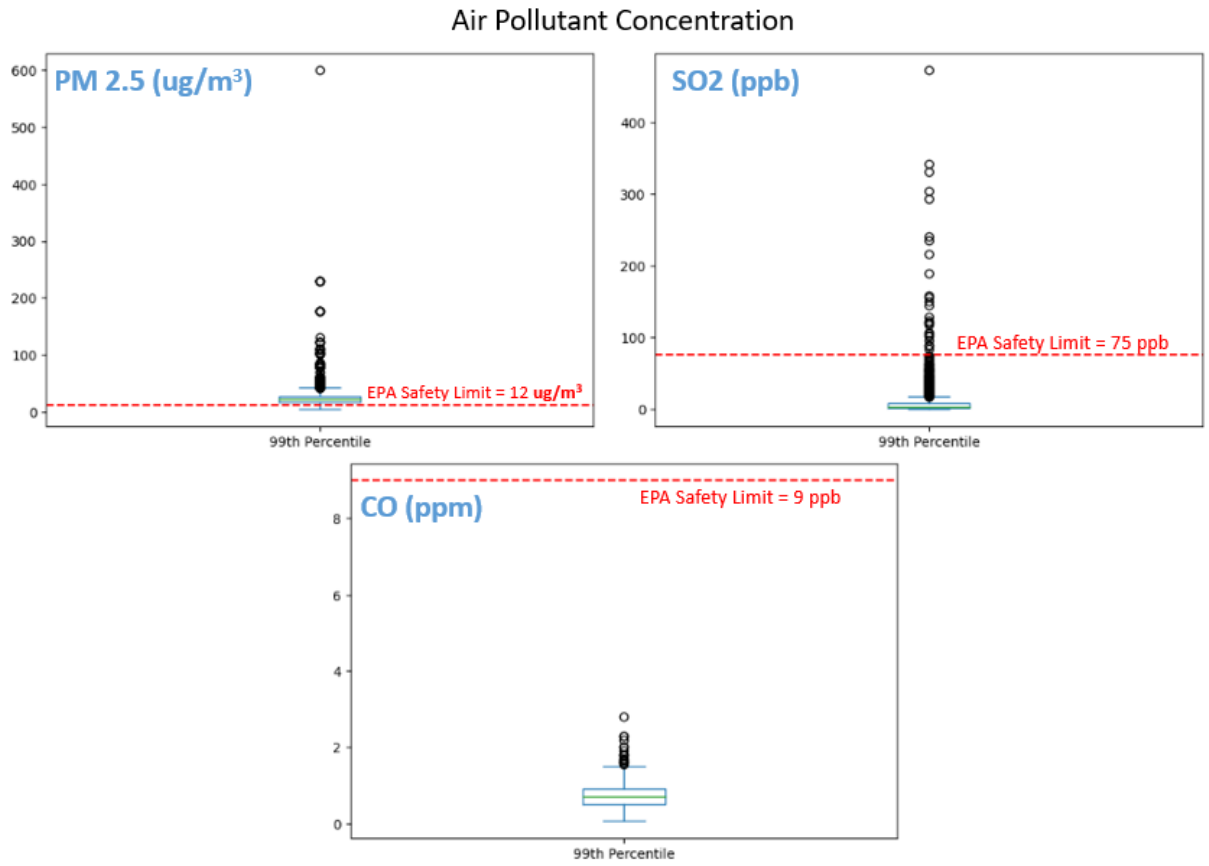
Focusing on the prevalence of diagnosed CKD, the counties with the highest population percentage with CKD are located in Hawaii, Southern California and Nevada, Florida, Appalachia, and the Midwest. (Note that Alaska is not shown because it does not have the highest prevalence of CKD or any water contaminants of the four analyzed). Based on the water contaminant data, Florida has a high prevalence of CKD with many counties above 30% and has a large cluster of geographical locations with water containing arsenic, lead, cadmium, and mercury above the EPA safety limit. Hawaii also has water contamination above the EPA limit in Oahu, which is also the island with the highest prevalence of CKD. There are also clusters of water contaminants in Montana and South Carolina, but these areas do not have the highest prevalence of CKD. Though Southern California and Nevada have a high prevalence of CKD, no elevated levels of the analyzed water contaminants were recorded there.

Air Quality Dataset

The results of each air pollutant were plotted in boxplots to display the spread of the data for each. A dotted reference line in red was plotted onto each visualization to show the primary

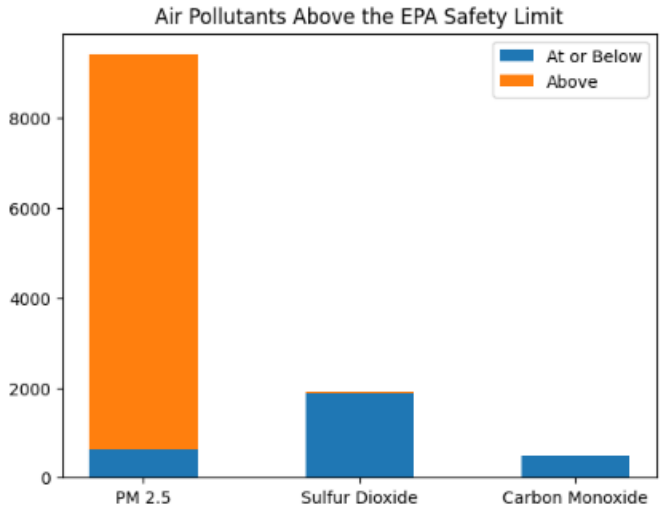
[14] "CDC Surveillance System: Diagnosed CKD among Medicare Beneficiaries ..." Kidney Disease Surveillance System. Accessed April 28, 2023. https://nccd.cdc.gov/ckd/detail.aspx?QNum=Q705.

concentration limit (providing public health protection) of each pollutant for safe air per the Environmental Protection Agency (EPA)[15], as shown in the figure below.

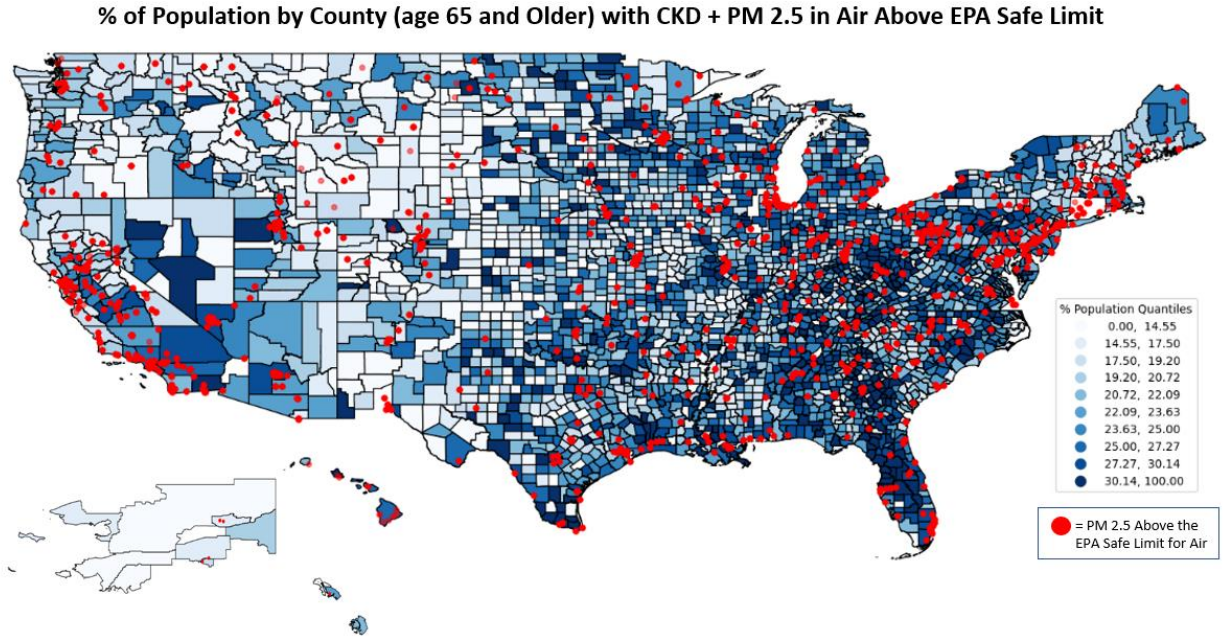


For this analysis, the 99th Percentile results were chosen to show the maximum level of each pollutant that occurred over the monitoring period. For SO2, most of the test results are equal to or below the EPA maximum concentration safety limit, and for CO the EPA safety limit was never exceeded. For PM 2.5, the EPA safety limit has been exceeded more times than not. For this analysis, the goal is to identify locations of unsafe air, i.e., air where pollutant concentrations have been recorded above the EPA safety limit. Therefore, all test results for each contaminant that were at or below the EPA safety limits were dropped from the data frames, leaving only results above the EPA limit. After dropping results meeting the EPA limit, the PM 2.5 data frame contained 8,781 entries, the SO2 data frame contained 26 entries, and there were no CO entries that were above the EPA safety limit. The following bar chart shows the quantities of each pollutant that were recorded at or below the EPA limit or above the EPA limit.

[15] "NAAQS Table." EPA. Environmental Protection Agency, March 15, 2023. https://www.epa.gov/criteria-air-pollutants/naaqs-table.



A map was created using data of the US population of medicare beneficiaries age 65 years and older with diagnosed CKD by US county[16]. The air quality data for PM 2.5 above the EPA limit was displayed on top of this map to see if any clusters of polluted air overlapped with higher incidence of CKD. Maps for SO2 and CO are not shown because there are few or no locations within the US where those pollutants exceeded the EPA safety limit.



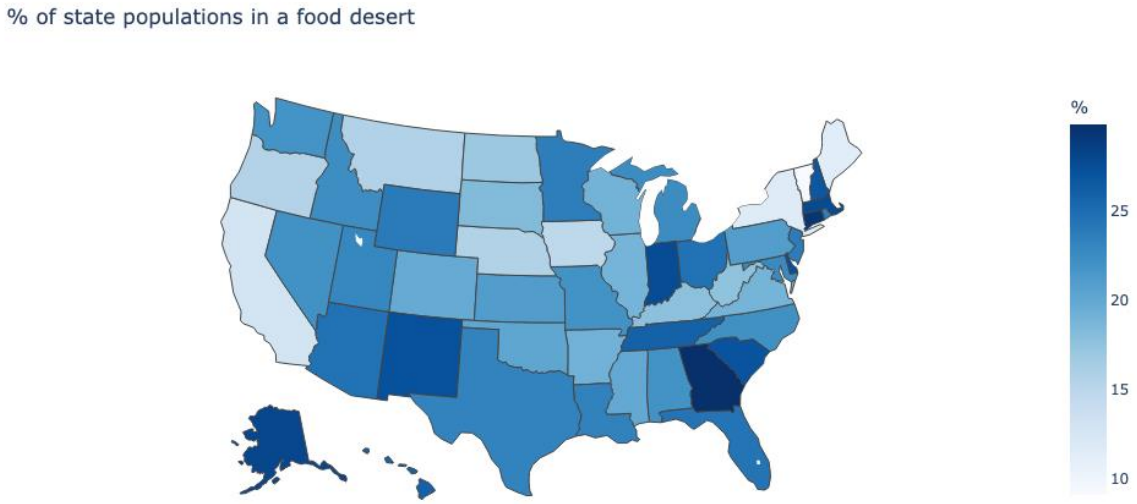
The counties with the highest population percentage with CKD are in Hawaii, Southern California and Nevada, Florida, Appalachia, and the Midwest. Based on the PM 2.5 concentration data, many of the counties that have higher prevalence of CKD are also areas where air with PM 2.5

[16] "CDC Surveillance System: Diagnosed CKD among Medicare Beneficiaries ..." Kidney Disease Surveillance System. Accessed April 28, 2023. https://nccd.cdc.gov/ckd/detail.aspx?QNum=Q705.

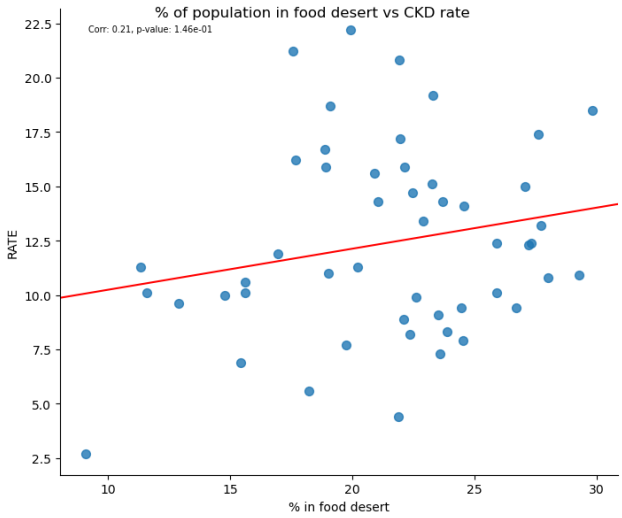
above the safe limit has been detected. In the Rocky Mountain region of the US and some clusters in the Midwest and South where there is the lowest prevalence of CKD, there are also fewer occurrences of PM 2.5 measured above the EPA safe limit. The measurements of PM 2.5 from this region above the EPA safe limit tend to be relatively lower.

Food Access Dataset

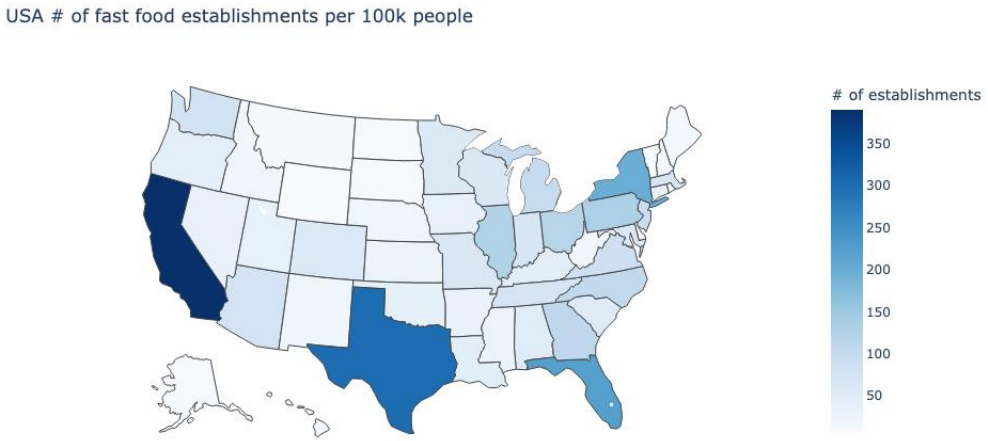
A map was created, using data of populations by state considered to be living in 'Food deserts'. in this case, a person living in a 'food desert' is defined as anyone living outside of a 1 mile radius from a supermarket in urban areas and outside a 20 mile radius of a supermarket in rural areas. For this analysis, the sum of the population living in a 'food desert' was divided by the sum of the population, which was provided by the U.S. census bureau.



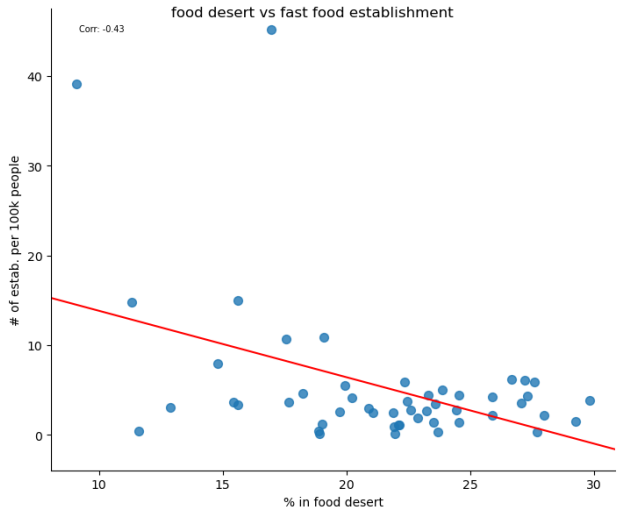
The goal of this analysis was to determine whether 'food deserts' have any impact on the rate of CKD. A correlation analysis was conducted and showed that there is a positive correlation of 0.21 between the population living in 'food deserts' and CKD rates.



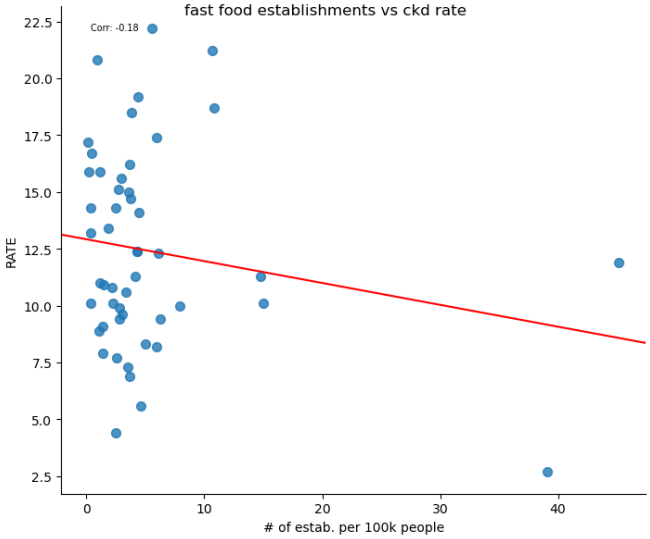
Initially, it was thought that food deserts might have a higher prevalence of fast-food establishments per 100k leading to a higher rate of CKD. So, an analysis was conducted to look at how many fast-food establishments there are by state per 100k people. Data with the location of each fast-food establishment in the United States was used. The number of fast-food establishments was summed by state and then divided by the state population divided by 100k.



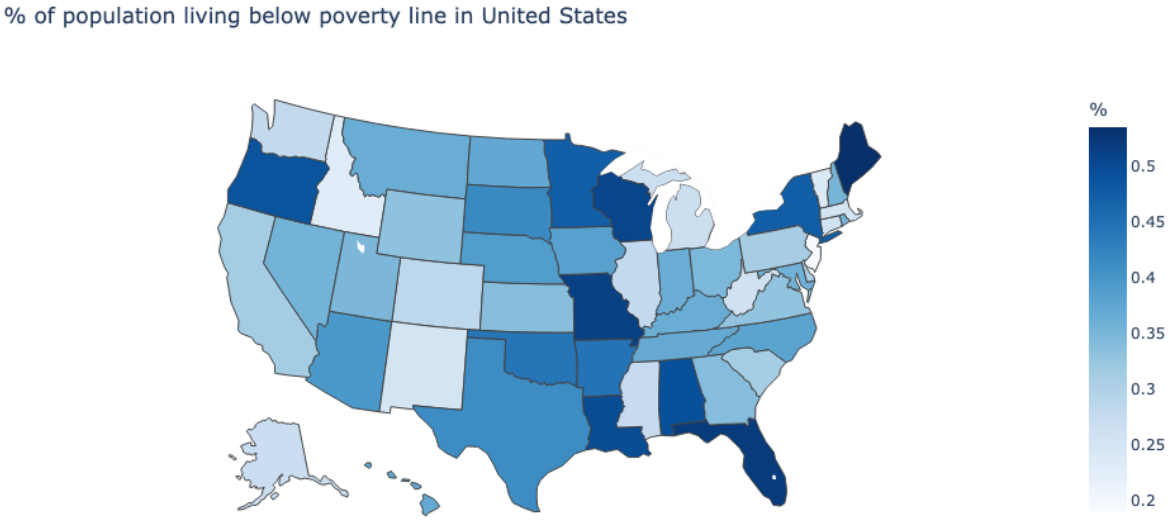
Then an analysis was conducted to measure the correlation between both the percentage of population residing in a food desert and the number of fast-food establishments per 100k people. There was a negative correlation between the two, as shown in the following figure.



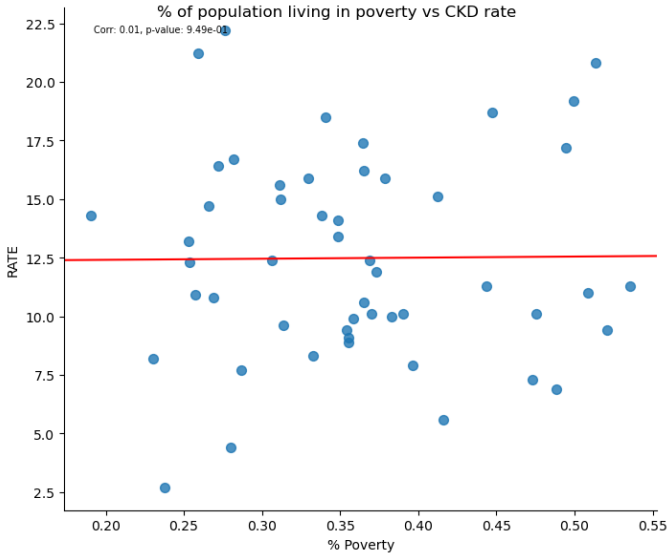
For the analysis of correlation between the number of fast-food establishments per 100k people and the rate of CKD, there was a weak, negative correlation, as shown in the following figure.



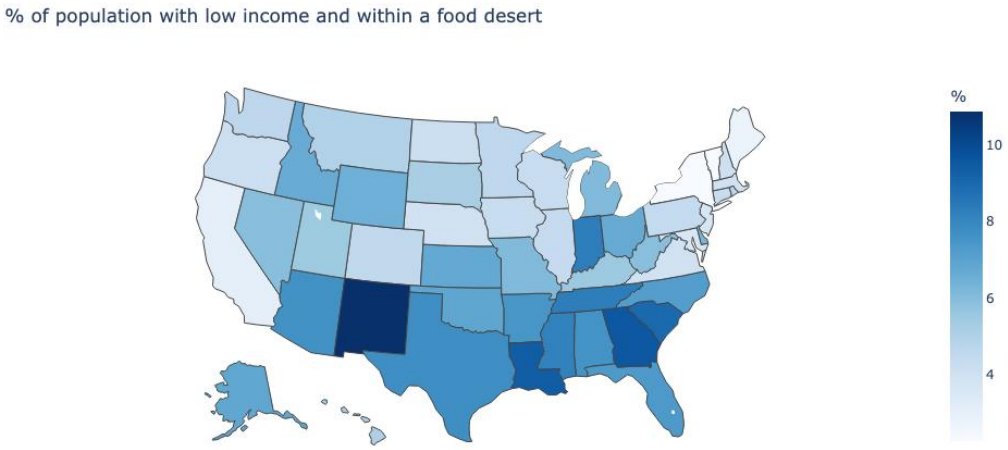
Income levels were thought to be a contributing factor to CKD rates, so an analysis was on the poverty rate in the U.S and compared to CKD rates. First, a map was generated to show the percentage of population in each state that was living at or below the poverty line. The analysis was completed by dividing the population living at or below the poverty line by the total population of the state and multiplying by 100 to get a percentage point. The rates of poverty by state are shown in the image below.



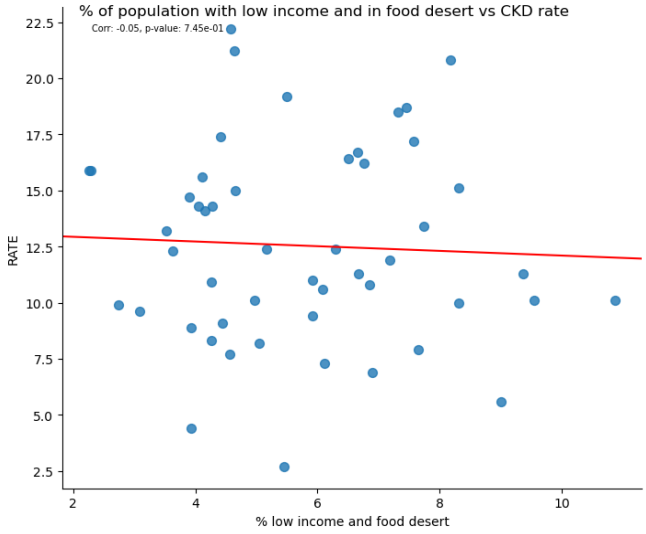
Next, an analysis was completed to show correlation between the population living at or below the poverty line and CKD rates in the U.S. The correlation coefficient given is 0.01 showing there is almost no correlation between poverty rate and CKD rates in the U.S. based on this analysis.



In addition, it was hypothesized that low income or poverty plus other factors combined might contribute to CKD rates. So, an analysis of the low income population was conducted and a map was generated to show the percentage of the population by state that was both in the low income category and lived in a food desert. The map was generated by dividing the population count that was considered low income and living in a 'food desert' by the total population of the state. The results are shown in the image below.



This was followed by an analysis to show the correlation between the percentage of the population that was low income and in a food desert with CKD rates by state. The idea behind this analysis was that low-income families in food deserts may have a higher risk of eating less healthy foods due to limited access and ability of purchase. There was a slight negative correlation between these two factors indicating that living with low income in a food desert does not correlate with CKD rates.



Web/Mobile Applications

Web Application

To support our client's needs, a web-based application was developed that includes a predictive component, as well as a dashboard for data analysis and retrieval. End users can leverage the predictive endpoint to input certain health-related data and receive output that indicates their level of risk regarding developing chronic kidney disease. The predictive component uses a simple form-based approach, making it very easy and quick for the users to input their information and receive a prediction within moments. Behind the scenes, the predictive component is using Principal Component Analysis to reduce the dimensionality of the dataset and then passing these values to the Decision Tree Classifier model that was developed for this project. The dashboard portion of the application provides a range of functionality that allows healthcare providers with an appropriate level of access to view healthcare related information for patients that are under their care; these providers can filter, sort, add/remove/edit data, and export data to an excel file format for their own use. Data visualizations have been included that allow for insights to be gleaned from a glance, and charts can be added or removed by the development team as needed to support our client.

The web application was developed using the Dash Package from Python; this package affords not only the ability to stand up a web-based interface, but also the ability to develop a highly customized and interactive application that can be easily scaled and extended as requirements change and grow. The modular nature of the dashboard design, and the supporting code, makes it very easy to update the content or add additional content at any time, and the application can be re-deployed at a moment's notice with no downtime.

Given the sensitive nature of the data that will be displayed in certain sections of the application, natively supported security and access management was built into the application. Users will be required to enter their login credentials when accessing the portions of the application that contain healthcare information; without this level of access, data that is sensitive in nature is inaccessible, securing patient information for only approved users with a real need to access this information.

The application has been designed so that it can be served in multiple ways, so that we can best accommodate our client's needs. The application can be run on a local computer or web server in an on-premises deployment, as a docker container on a local computer or web server in an on-premises deployment or can be integrated into a cloud environment and deployed by any major cloud provider. We have included the code, assets, and folder structure necessary for the locally hosted and dockerized versions of the application as part of our deliverables, and we have also

deployed the application to the Google Cloud Platform to serve as an example of how the application would look if hosted in this manner.

The application hosted by Google Cloud Platform can be accessed over the internet with a browser at the following link: https://techstat-app-u63grmon7a-uc.a.run.app/.

To log into the the 'Provider Overview' portion of the application, use the following username and password:

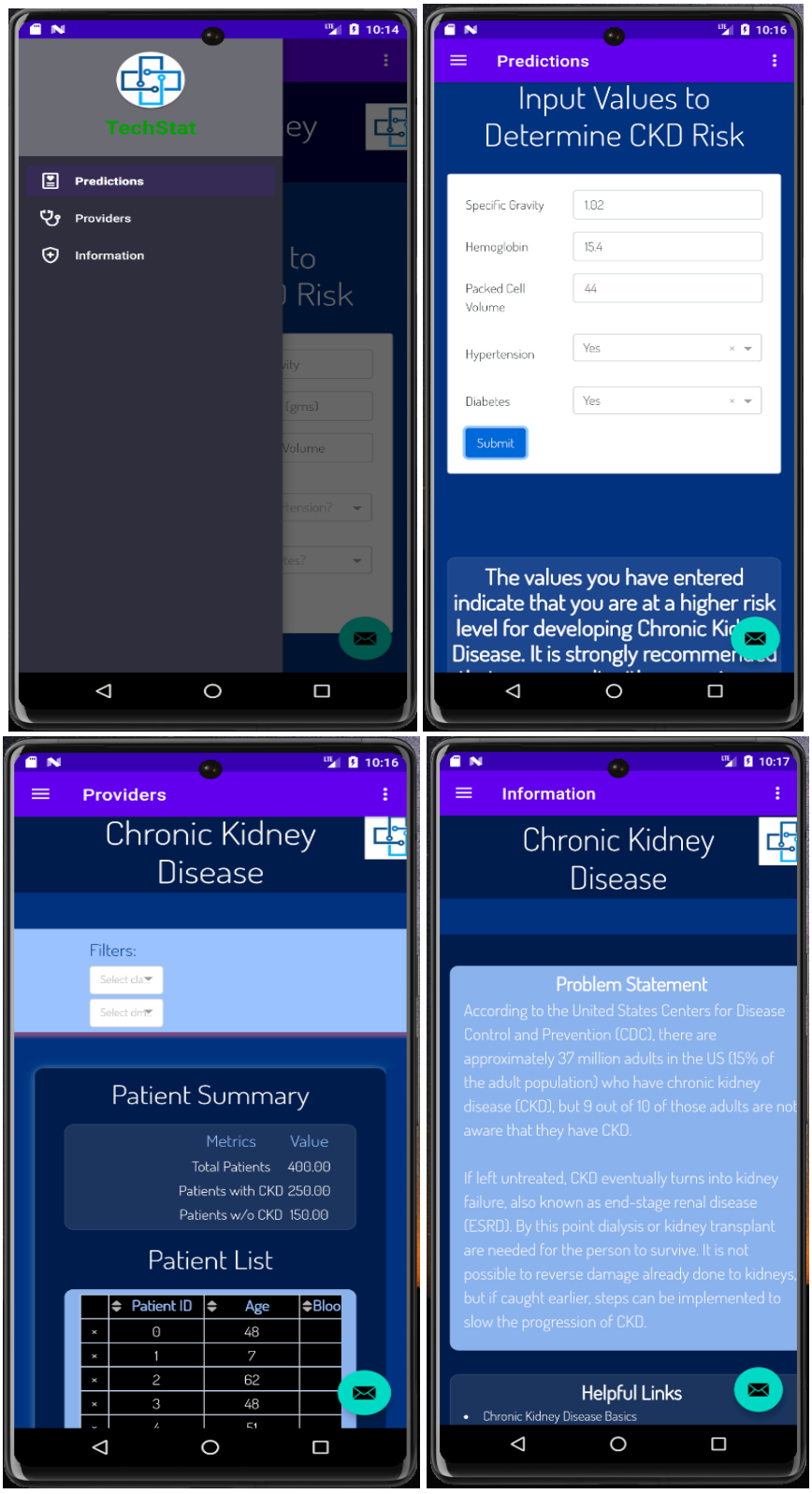
user: dr_jones
password: techstat_dashboard

Mobile Application

In addition to the web-based application, the current trend in our industry is to have a mobile application that also supports the needs of our clients; to that end, an Android-based mobile application was also developed that provides a means for our clients to offer an 'on-the-go' experience, with the ability to access the features and functionality of the web-based application anywhere from their mobile device. An iOS version of the application will also be developed in the next phase after funds are released, to ensure the accessibility of the mobile application by as wide a user base as possible.

The mobile application was developed using Android Studio, a freely downloadable software provided by Google. To ensure a similar experience between the web-based and mobile applications, a webView approach was taken, wherein the mobile application leverages the content of the web-application for its own content, sending requests to the web-application to serve its content and optimizing the display of the information for a smaller display size typically seen on mobile devices. The same user and password combination from above can be used to log into the 'Provider Overview' section of the mobile application. See the following images for mobile application screen captures.

Mobile Application Screen Captures



Conclusions

After the CKD dataset was cleaned and transformed, an in-depth PCA was performed to reduce the dimensionality of the dataset. Upon analysis of the first three principal components, it was determined that PC1 was correlated with the ckd classification while PC2 was correlated with the non-ckd classification. PC3 did not have a clear correlation with either classification but was slightly correlated with non-ckd more than ckd. Preliminary modeling was performed with 2 components (approximately 46% of the dataset information) and with 9 components (approximately 90% of the dataset information). Several of the models performed well, with 99% test accuracy, and it was determined that 2 component PCA would be sufficient for the predictive model.

Of the predictive models performed with the CKD dataset, most of the models performed well, with the Logistic Regression, Decision Tree, and Random Forest models providing the most consistent results. Other classification models that were evaluated in depth were SVC (support vector classifier), KNN (k-nearest neighbors, and Naive Bayes.

In addition to PCA, 2-component t-SNE analysis was also performed on the full cleaned dataset as an alternate technique to reduce dimensionality in the data. Based on the t-SNE projection, there are two well-defined clusters, one for each classification, with barely any overlap of points. The separation of the two classes in the projection indicates that there is sufficient difference between the two classes for supervised machine learning models to be successful. It also indicates that non-linear classification models will likely perform well with this data.

Several different models, Logistic Regression, Decision Tree, Random Forest, SVC, KNN, and Naive Bayes, were performed using the full cleaned dataset, 2 component PCA dataset, and t-SNE dataset and accuracies were compared. Overall, the models performed best using the PCA of the data, and the decision tree and random forest models performed the best on all three datasets.

Feature selection was also performed to reduce the number of variables needed for predictive modeling. Based on the feature importances ranking, the top 5 important features of the CKD dataset are sg (specific gravity), pcv (packed cell volume), htn (hypertension), hemo (hemoglobin), and dm (diabetes). Using the cleaned dataset, a new abridged dataset containing just these top 5 variables was created to use for further modeling.

Two-component PCA and t-SNE datasets were created using the abridged dataset to see the predictive modeling outcome with variable and dimensionality reduction in the data. Several different models, Logistic Regression, Decision Tree, Random Forest, SVC, KNN, and Naive Bayes, were performed using the abridged dataset, 2 component PCA abridged dataset, and t-SNE abridged dataset.

Overall, the models performed best using the PCA of the abridged data, and the decision tree and random forest models performed the best on all three datasets. The Naive Bayes model had the lowest train and test accuracy, with results in the 80% range. Comparing this model to the data with PCA and t-SNE applied shows a large improvement in the Naive Bayes model train and test accuracy, indicating that dimensionality reduction is working as intended. Train accuracy and test accuracy did reach 100% in some models. This is usually unrealistic and impractical in ML models, but since our dataset is relatively small, with 400 rows, it is not uncommon. The training and test accuracies will likely change once additional patient data is collected and added to the dataset in the future. Adding more patient data in the future will be required to develop a more robust model.

Based on the Water Quality dataset, there is no clear relationship with the identified water contaminants that are linked to kidney disease and prevalence of kidney disease in US counties. There are, however, some areas of contaminated water that need to be improved for public health and safety purposes, in Florida, South Carolina, and Montana.

Based on the Air Quality dataset, PM 2.5 is highly linked to counties in the US with increased prevalence of diagnosed CKD. PM 2.5 is particles in the air that are 2.5 microns in diameter or smaller. When people breathe in PM 2.5, these particles can cross the blood gas barrier and enter the kidneys where they accumulate. With enough accumulation of PM 2.5, the kidneys become damaged, causing renal injury, and eventually leading to CKD. The main current sources of PM 2.5 are due to industry and include cooking smoke, welding smoke, cigarette smoke, smoke from heating emissions, and traffic exhaust[17]. Governments should investigate ways to reduce PM 2.5 to ultimately reduce the harmful effects on public health, especially as it relates to CKD.

Lastly, the relationships between food access and fast-food restaurants with CKD rates were studied. Based on the data, there is a significant percentage of people living in the US that have low access to fresh food at grocery stores (i.e., food deserts), many states with 20% of the population or more. For fast food establishments per 100,000 people, only a few states have notably more restaurants (about 200 or more per 100,000 people): California, Texas, Florida, and New York.

Correlations between both food access and fast-food restaurants versus CKD rate were performed. There is a positive correlation between the percentage of a state's population in a food desert and the state's CKD rates. This makes sense because food deserts restrict people's access to healthy food. There is no clear correlation between the number of fast-food

[17] Xu, Wenqi, Shaopeng Wang, Liping Jiang, Xiance Sun, Ningning Wang, Xiaofang Liu, Xiaofeng Yao, et al. "The Influence of PM2.5 Exposure on Kidney Diseases." *Human & Experimental Toxicology* 41 (February 17, 2022). https://doi.org/10.1177/09603271211069982.

establishments per state and CKD rate per state. This also makes sense because most states have about the same number of fast-food restaurants per 100,000 people. Lastly, an analysis on income and its correlation to CKD was performed, but there was no correlation between lower income and rates of CKD in the US.

A web application and mobile application were developed to support our client's needs to detect and prevent CKD more easily. Utilizing TechStat's predictive model, patients can input their health data for the 5 identified features into the app. The app will feed the patient data into the predictive model and output a recommendation on whether to consult with their healthcare provider regarding getting further testing for CKD. The application also has a provider view that allows authorized healthcare providers to login and view and manipulate patient data in a spreadsheet. Lastly, the application has an information tab to provide visualizations for air and water contamination in the US and some useful links for app users to explore.

Recommendations

For the CKD predictive model, the TechStat team recommends the decision tree model using the two-component PCA dataset with only the top 5 features selected. This model performed very well with the CKD dataset and is also optimized to use fewer computing resources. While the two-component t-SNE dataset also performed well with this model, there are several challenges when feeding t-SNE reduced features into an ML model. T-SNE is non-deterministic, meaning that each time it runs, the results can be different even with the same hyperparameters. In addition, t-SNE is more difficult to use when new data is added to the dataset. Once additional data is added to the dataset, the t-SNE embeddings will be recalculated. Because of this, PCA was chosen since new data can be added to the current projection.

Environmental factors have potential to influence the development of CKD. The effects that air pollution can have on the body can be reduced with the use of air purifiers or filters in common indoor areas[18]. It is recommended to inform citizens that N95 Masks be worn when filters are not available and while spending time outdoors while exposure to the PM2.5 are at a high level[19]. While individuals take precautionary measures, it is recommended that regulations for people and businesses that contribute to higher levels of PM 2.5 are enforced by some financial and/or

[18] 1. "Extremely High Levels of PM2.5: Steps to Reduce Your Exposure," Extremely High Levels of PM2.5: Steps to Reduce Your Exposure | AirNow.gov, accessed May 18, 2023, https://www.airnow.gov/aqi/aqi-basics/extremely-high-levels-of-pm25/.

[19] 1. James Parsons, "What Is a PM2.5 Mask, and Is It Different than an N95?," Ellessco, July 4, 2022, https://ellessco.com/blog/2022/07/pm25-mask-different-n95.

legal penalty[20]. Despite the evidence that water contaminants likely have low impact on kidney health in the US, it is still recommended to inform citizens to filter drinking water within areas that have tested for levels of mercury and lead beyond the EPA safety limit. These have been noted to have potential cause for kidney damage over time from the EPA[21].

TechStat recommends deploying mobile and web applications to as many healthcare providers and patients as possible. This will allow for quicker detection and treatment of CKD in patients and will help to inform patients of environmental and lifestyle factors that affect their risk of developing CKD.

The client should create a database to house information input to the mobile and web applications for the purpose of improving the robustness of the CKD predictive model in the future. The current dataset is somewhat small, so additional data will improve the accuracy of the model for the future.

[20] 1. Yevgen Nazarenko, Devendra Pal, and Parisa A Ariya, "Air Quality Standards for the Concentration of Particulate Matter 2.5, Global Descriptive Analysis," *Bulletin of the World Health Organization* 99, no. 2 (2020), https://doi.org/10.2471/blt.19.245704.

[21] 1. "National Primary Drinking Water Regulations," EPA, accessed May 19, 2023, https://www.epa.gov/ground-water-and-drinking-water/national-primary-drinking-water-regulations.

Appendix

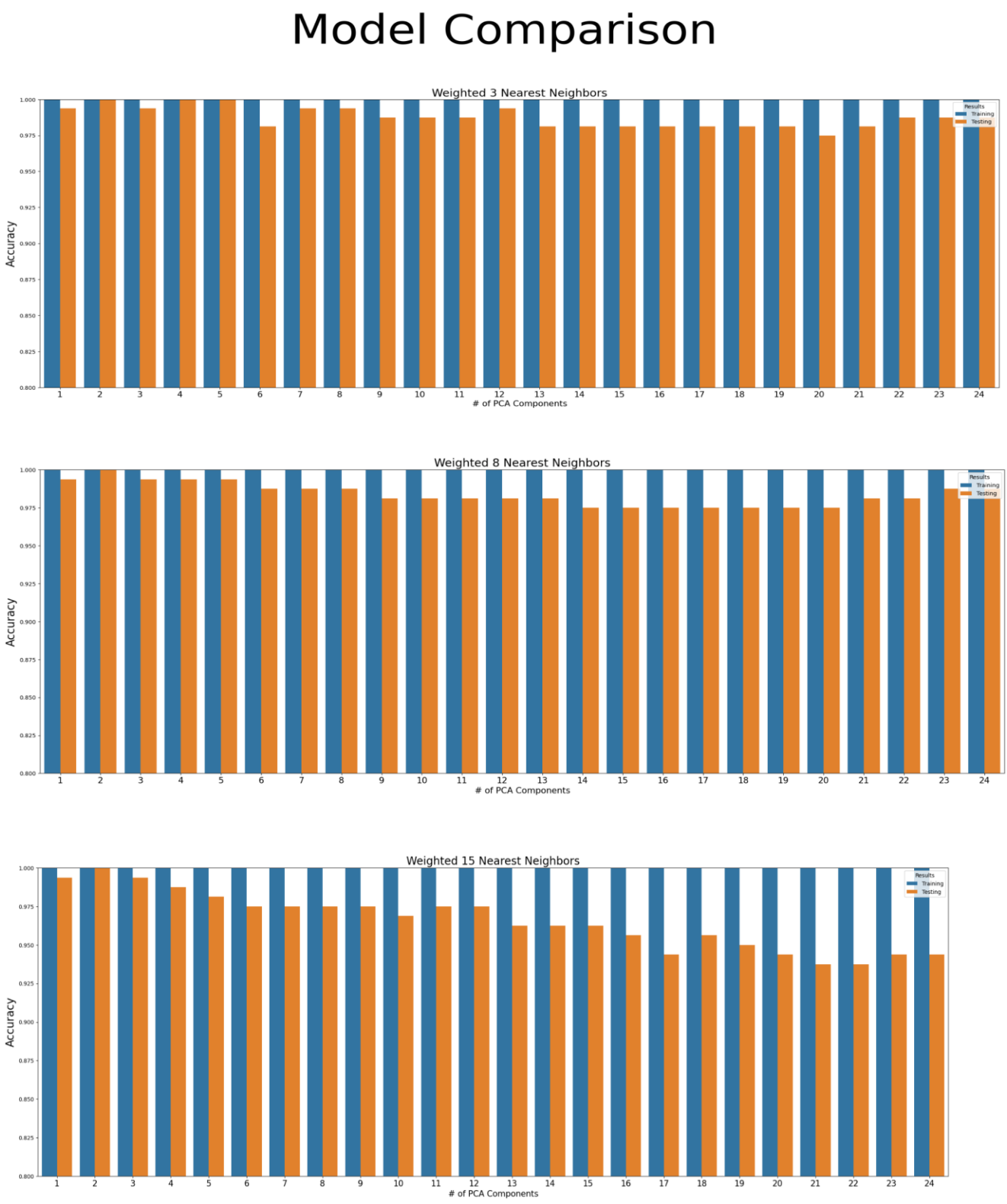
A1. CKD Dataset Dictionary

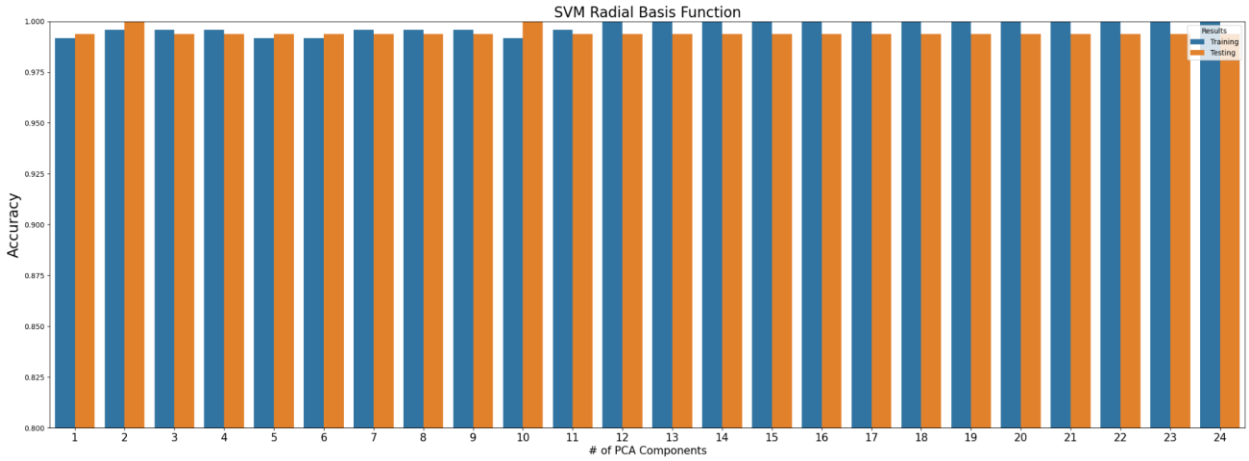
CKD Dataset Data Dictionary

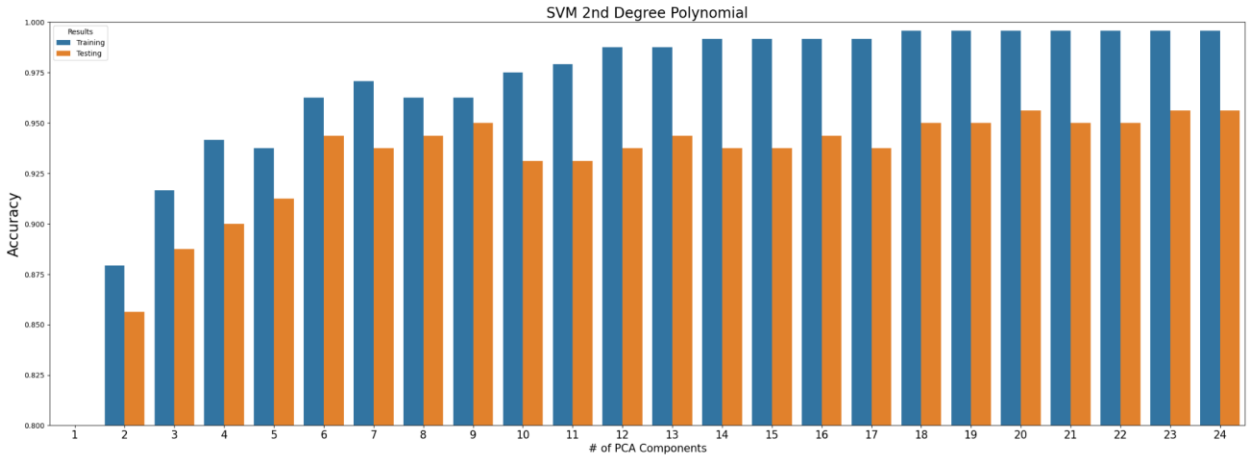
Field Name	Definition	Description
age	age	Age of the patient in years.
bp	Blood pressure	Systolic blood pressure of the patient in mm/Hg
sg	Specific gravity	Specific gravity of the patient's urine
al	Albumin	Albumin level in patient's blood
su	Sugar	Sugar level in patient's blood
rbc	Red blood cells	Describes if patient has normal or abnormal red blood cells
pc	Pus cell	Describes if patient has normal or abnormal pus cells
pcc	Pus cell clumps	Describes if pus cell clumps are present or not present
ba	Bacteria	Describes if bacteria is present or not present
bgr	Blood glucose random	Patient's blood glucose level in mg/dL
bu	Blood urea	Patient's blood urea level in mg/dL
sc	Serum creatinine	Patient's blood serum creatinine level in mg/dL
sod	Sodium	Sodium level in patient's blood in mEq/L
pot	Potassium	Potassium level in patient's blood in mEq/L
hemo	Hemoglobin	Hemoglobin level in patient's blood in g
pcv	Packed cell volume	Packed cell volume in patient's blood
wc	White blood cell count	White blood cell count in patient's blood in cells/mm^3
rc	Red blood cell count	Red blood cell count in patient's blood in million/mm^3

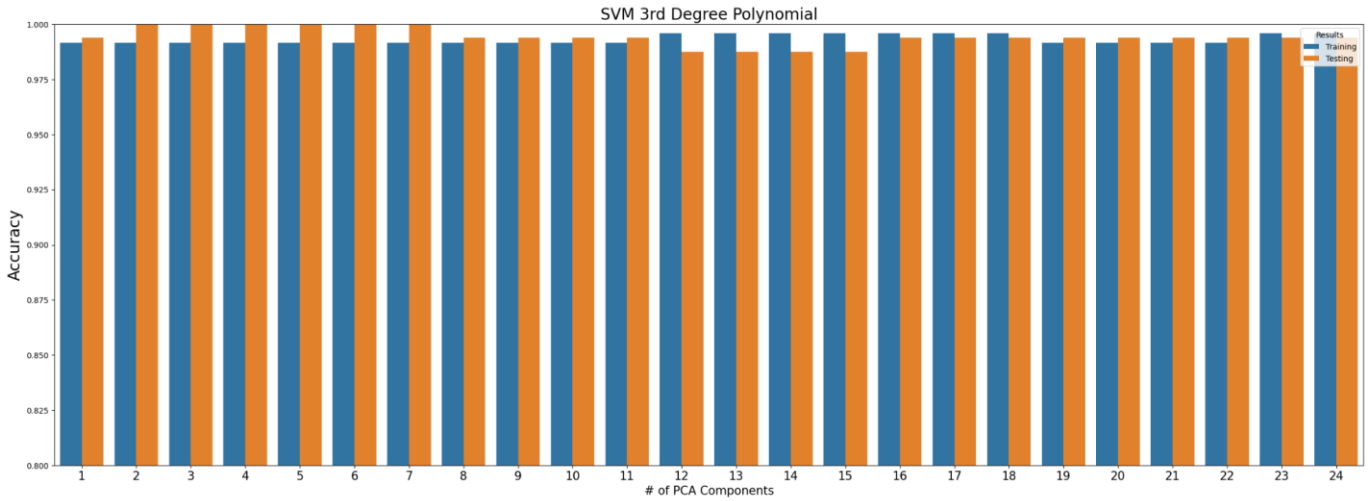
Field Name	Definition	Description
htn	Hypertension	Describes if patient has hypertension: yes or no
dm	Diabetes Mellitus	Describes if patient has diabetes: yes or no
cad	Coronary artery disease	Describes if patient has coronary artery disease: yes or no
appet	Appetite	Describes if patient's appetite is good or poor
pe	Pedal edema	Describes if patient has pedal edema: yes or no
ane	Anemia	Describes if patient has anemia: yes or no
class	Class	Target classification: ckd (patient has chronic kidney disease); notckd (patient does not have chronic kidney disease)

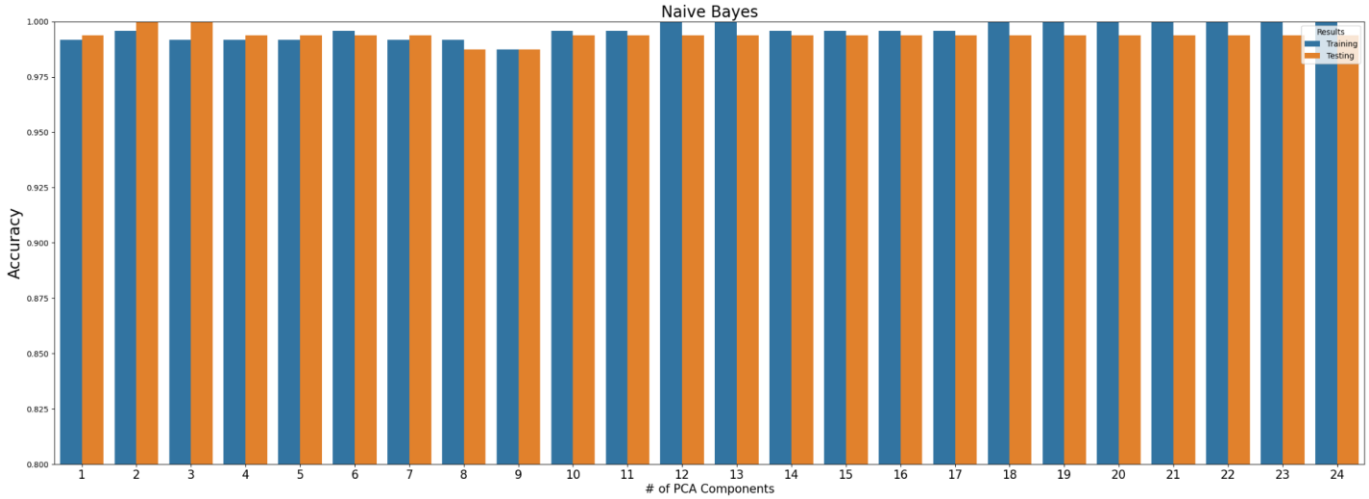
A2. Additional Predictive Model Visualizations

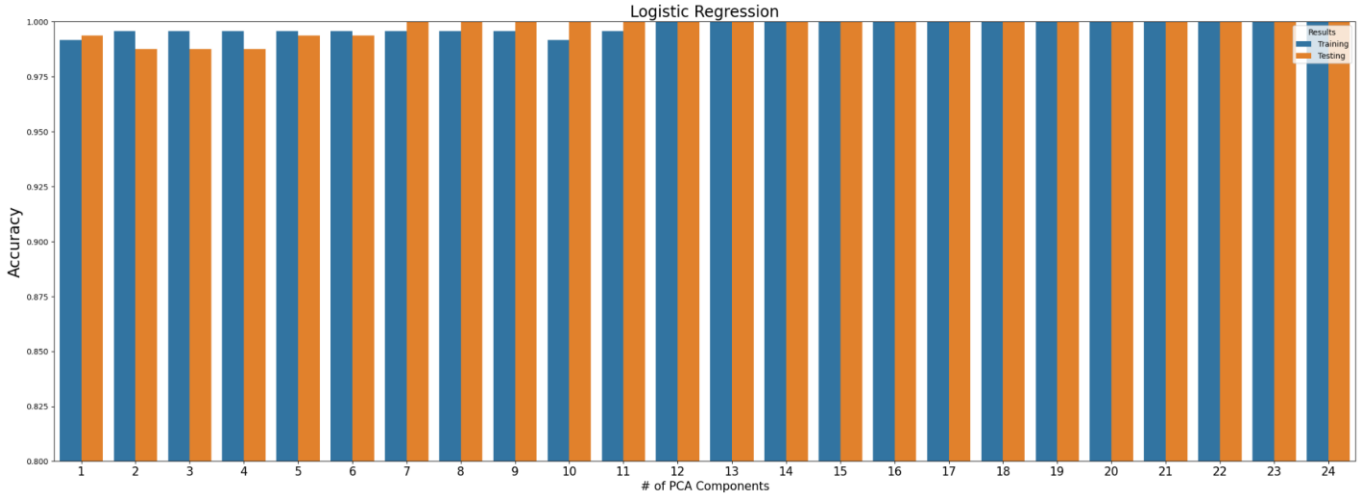


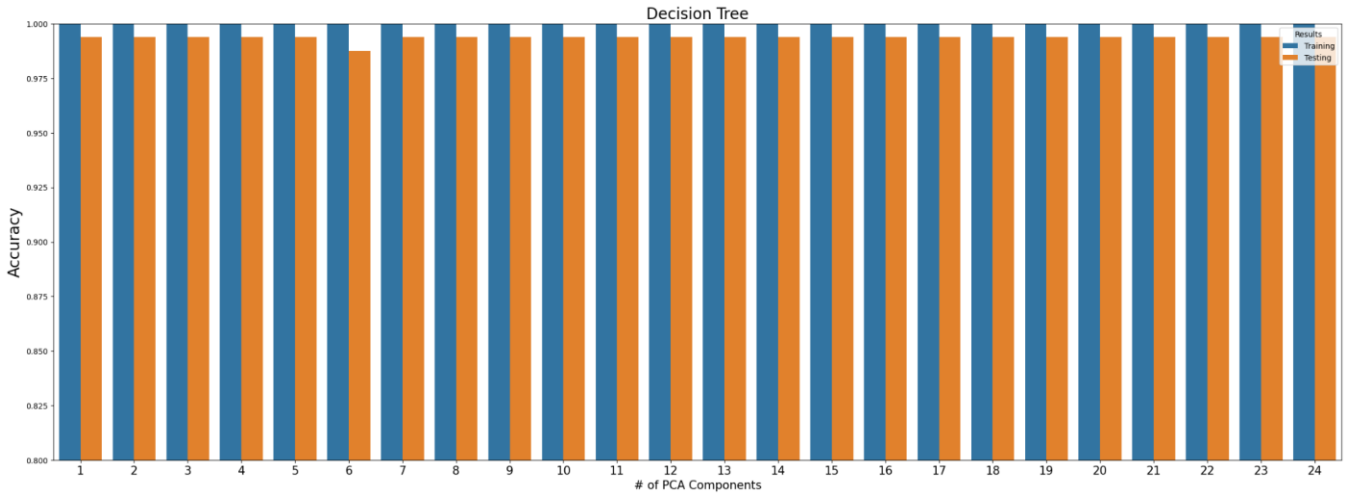

SVM Radial Basis Function
Accuracy
# of PCA Components
Results
Training
Testing


SVM 2nd Degree Polynomial
Accuracy
# of PCA Components
Results
Training
Testing


SVM 3rd Degree Polynomial
Accuracy
# of PCA Components
Results
Training
Testing


Naive Bayes
Accuracy
# of PCA Components
Results
Training
Testing


Logistic Regression
Accuracy
# of PCA Components
Results
Training
Testing


Decision Tree
Accuracy
# of PCA Components
Results
Training
Testing
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.800
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24