Implementing Artificial Intelligence for Bioinformatics and Rapid Analysis of Cancer, CKD, and Heavy Metals in NHANES Datasets by Novice Programmers

The Oncology Explorer

New Mexico Supercomputing Challenge

Final Report

10 April 2024

School Name: Early College Academy/Justice Code

Team Member(s): Aileen Ukwuoma

Sponsor: Rebecca Campbell

Project Mentor(s): Robert Taylor, Ph.D. and Justin Baca, MD, Ph.D.

Implementing Artificial Intelligence for Bioinformatics and Rapid Analysis of Cancer, CKD, and Heavy Metals in NHANES Datasets by Novice Programmers

Table of Contents

Implementing Artificial Intelligence for Bioinformatics and Rapid Analysis of Cancer, CKD, and Heavy Metals in NHANES Datasets by Novice Programmers

Implementing Artificial Intelligence for Bioinformatics and Rapid Analysis of Cancer, CKD, and Heavy Metals in NHANES Datasets by Novice Programmers

Executive Summary

Chronic kidney disease (CKD) is a condition in which the kidneys progressively lose their ability to eliminate waste from the human body. Such inability can result in comorbidity, or the acquisition of two or more disorders simultaneously. This experiment examined the statistical correlation between cancer, chronic kidney disease, and heavy metals (HM) exposure. We compiled data related to cancer incidence, heavy metal burdens, and kidney dysfunction from the 2017-2018 National Health and Nutrition Examination Survey (NHANES) to create visuals and draw conclusions regarding our experimental questions. We analyzed various NHANES datasets, isolated features of interest related to our research question, determined correlations, and created machine learning models using Python programming and ChatGPT 3.5. Ultimately, we created two computational models predicting cancer incidence and kidney failure. In the future, we can replicate this experiment with a more comprehensive dataset and evaluate the use of our models in medical settings.

Introduction

Problem Statement

Chronic kidney disease (CKD) gradually inhibits the kidney's filtration process, causing an accumulation of fluid and waste in the body. In recent years, scientists have speculated the comorbidity of CKD and ailments such as cancer and heavy metal exposure. The purpose of this experiment is to explore the statistical correlation between cancer, chronic kidney disease, and heavy metals exposure. Another objective is to use machine learning to visualize data and enable predictions about the development of cancer and kidney dysfunction. Concurrently, we evaluate the efficacy of artificial intelligence (version 3.5 of ChatGPT) as a didactic tool in teaching computer science.

We hypothesize that CKD and heavy metals exposure will be correlated with a higher cancer incidence in the 2017-2018 National Health and Nutrition Examination Survey (NHANES) dataset. We also believe that ChatGPT will enable rapid analysis of NHANES datasets in conjunction with Python programming and machine learning.

This project provides novel insights into the link between HM exposure and the development of chronic ailments like cancer and CKD; such a project is especially pertinent to medically underserved regions. The insights obtained from this project will enrich our understanding of comorbidity and offer, through machine learning, a means of predicting one's susceptibility to the diseases in question.

Background Research

The New Mexico Epidemiology Office found that, in 2014, there were 28,473 (14.6%) hospitalizations for chronic kidney disease in New Mexico, a 6.0% increase from the number of CKD hospitalizations in 2013. [1] Furthermore, 15% of individuals across the United States struggle with chronic kidney disease, which may increase the risk of malignancy, or cancer. [2] Studies have found that CKD is a risk factor for mortality in cancer patients. [3] Also, exposure to heavy metals (such as cadmium, lead, and arsenic) can potentially result in CKD and, concurrently, cancer. [4] There is a gap in the scientific literature concerning the statistical correlation between CKD, cancer, and HM exposure and about methodologies to predict these ailments using machine learning. Our project aims to fill this gap.

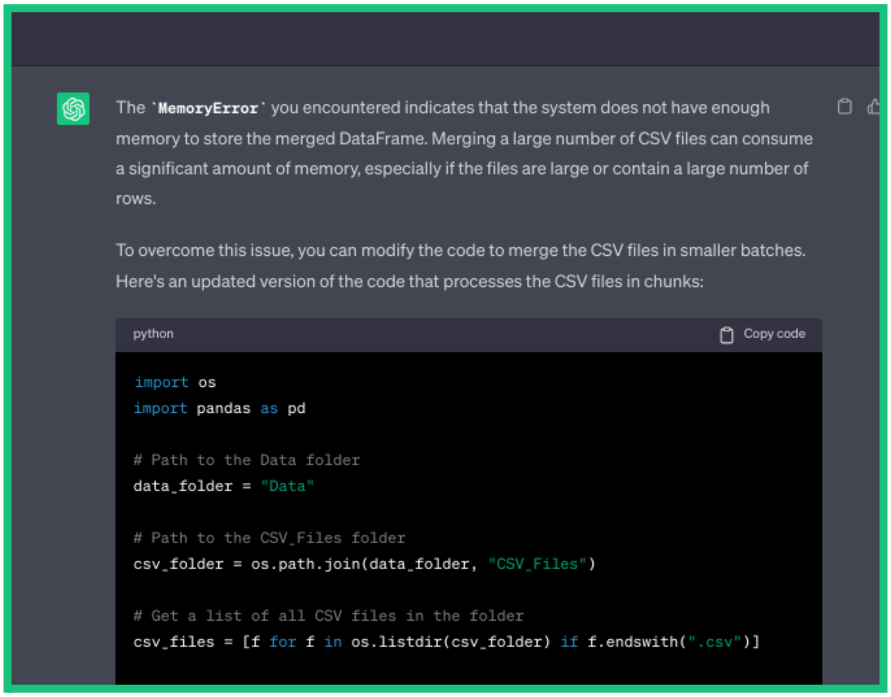
Computational Model

Selection

We selected Python as our primary programming language. Python is useful for machine learning and biostatistical projects such as ours. It allowed us to visualize cancer-related patient data, calculate statistics, determine correlations, and create predictive models using Scikit-Learn. We used Jupyter Notebooks to keep track of our work.

Modifications

While following the general steps of machine learning, we used version 3.5 of ChatGPT as a troubleshooting tool. For instance, while trying to merge crucial NHANES datasets, we encountered "memory" errors due to the large quantity of files. When we consulted ChatGPT 3.5 concerning the issue, the chatbot explained the error and provided alternative means of achieving our goal. This constant feedback allowed us to modify and ameliorate our code.



The `MemoryError` you encountered indicates that the system does not have enough memory to store the merged DataFrame. Merging a large number of CSV files can consume a significant amount of memory, especially if the files are large or contain a large number of rows.

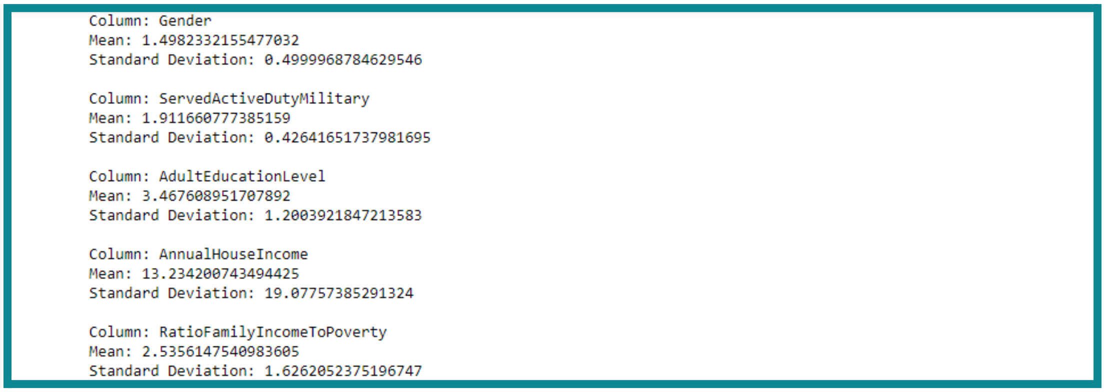
To overcome this issue, you can modify the code to merge the CSV files in smaller batches. Here's an updated version of the code that processes the CSV files in chunks:

```python
import os
import pandas as pd

# Path to the Data folder
data_folder = "Data"

# Path to the CSV_Files folder
csv_folder = os.path.join(data_folder, "CSV_Files")

# Get a list of all CSV files in the folder
csv_files = [f for f in os.listdir(csv_folder) if f.endswith(".csv")]
```

Figure 1. Image Capture of ChatGPT 3.5 Troubleshooting Advice

Visualization

Python allowed us to calculate statistics, determine correlations, and visualize data about cancer incidence, kidney failure, heavy metals exposure, and biochemical markers of the aforementioned conditions. We calculated the means and standard deviations of different columns in our combined NHANES dataset. Python calculated this data for both categorical and quantitative columns, so we had to use discretion in interpreting the platform's statistical results.



```
Column: Gender
Mean: 1.4982332155477032
Standard Deviation: 0.4999968784629546

Column: ServedActiveDutyMilitary
Mean: 1.911660777385159
Standard Deviation: 0.42641651737981695

Column: AdultEducationLevel
Mean: 3.467608951707892
Standard Deviation: 1.2003921847213583

Column: AnnualHouseIncome
Mean: 13.234200743494425
Standard Deviation: 19.07757385291324

Column: RatioFamilyIncomeToPoverty
Mean: 2.5356147540983605
Standard Deviation: 1.6262052375196747
```

Figure 2. Image Capture of Means and Standard Deviations Generated Using Python

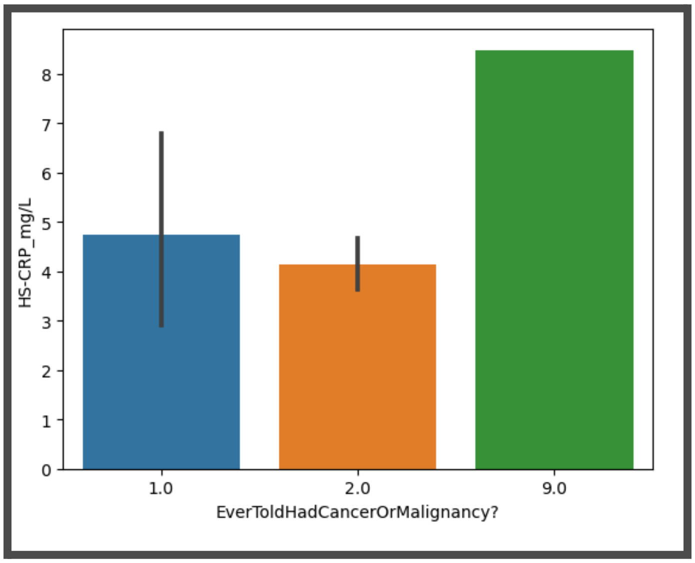


Figure 3. Image Capture of Python-Generated Barplot Comparing High-Sensitivity C-reactive Protein and Cancer Incidence

Limitations

The primary limitation we encountered while completing this project was the high levels of nullity in the NHANES datasets (due to survey respondents' incompletion of certain questions). Often, this forced us to disregard certain columns of data or impute null values with the mean of a column.

Problem Solving Method

Procedure

Given the format of our project, we followed the general methodology of data science and machine learning. The first step in our procedure was data conversion. We analyzed the NHANES 2017-2018 questionnaires and biochemical data, seeking information about cancer, CKD, and heavy metals. The information in these datasets is based on a "SEQN" number, which identifies an individual survey respondent.

We isolated the datasets with features related to our research question, converted them from ".xpt" (SAS, or Statistical Analysis Software, files) to ".csv" (text) files for easier access, and merged these datasets based on the subject identification (SEQN) number. To accomplish this, we consulted ChatGPT for written code that would enable us to combine large datasets.

The second step was exploratory data analysis. Since the NHANES datasets include questionnaires, many survey respondents failed to answer different questions. This resulted in nullity. We examined the structure of the combined datasets, removed duplicate columns, and ameliorated nullity (for instance, by writing code that filled null values or empty slots with the mean of a particular column). Afterward, we visualized our data, creating charts and graphs depicting information for patients with and without cancer.

Variable	**Diagnosed With Cancer (n=85)**	**Not Diagnosed With Cancer (n=763)**
Gender (2 levels)	41 (male), 44 (female)	385 (male), 378 (female)

Ever Told Had High Blood Pressure (4 levels)	50 (1), 34 (2), 1 (9)	287 (1), 475 (2), 1 (9)
Ever Told Had High Cholesterol (4 levels)	30 (1), 50 (2), 5 (9)	259 (1), 500 (2), 4 (9)
Dialysis in Past 12 Months? (4 levels)	9 (2)	4 (1), 15 (2)
Ever Told Had Weak or Failing Kidneys? (4 levels)	9 (1), 75 (2), 1(9)	19 (1), 742 (2), 2(9)
Ever Told Had Congestive Heart Failure? (4 levels)	5 (1), 80 (2)	23 (1), 738 (2), 2 (9)
Ever Told Had Coronary Heart Disease? (4 levels)	4 (1), 81 (2)	29 (1), 731 (2), 3 (9)
Ever Told Had Heart Attack? (4 levels)	3 (1), 81 (2), 1 (9)	40 (1), 720 (2), 3 (9)
Do You Smoke Cigarettes? (5 levels)	12 (1), 4 (2), 30 (3)	120 (1), 22 (2), 181 (3)

Table 1. Baseline comparisons of demographic characteristics between respondents with cancer and respondents without cancer.

Thirdly, we explored correlations in the data related to our target variable (cancer incidence). In this step, we encountered difficulty identifying variables specifically pertaining to CKD and HM exposure. The NHANES datasets do not explicitly stipulate whether respondents were diagnosed with CKD. However, these datasets have information about "weak or failing kidneys." For the purposes of this project, we considered this analogous to CKD. Regarding HM exposure, we consulted ChatGPT for assistance in combining columns in our dataset that contained information about the systemic presence of diverse metals into one column entitled "Total Urine Metal Burden (ug/L)." Through Python libraries like Pandas, NumPy, and Matplotlib, we imported the necessary dataframe (as a CSV file) and determined correlations correlations between cancer (recorded in NHANES as "EverToldHadCancerOrMalignancy?"), kidney failure ("EverToldHadWeakOrFailingKidneys?"), urine HM burdens ("TotalUrineMetalBurden_ug/L"), and different biological markers. These correlations were readily displayed in Spearman correlation heatmaps.

The next step in our process was building machine learning models to predict cancer and kidney failure. A crucial component of this step was feature selection. Given the large quantity of information in the NHANES datasets, we needed to narrow down the most important features to improve the predictive capacity of our machine learning (ML) models. For instance, while preparing to create a model predicting cancer, we needed to determine the features most related to cancer incidence. As such, our target variable, or "y," was cancer incidence. Through Python, we used three feature selection methods (Recursive Feature Elimination, SelectKBest, and Feature Importance from a Random Forest Classifier) to determine the features most important to the output variable (cancer incidence). Each feature selection method generated features (which corresponded with a specific NHANES questionnaire question or laboratory test category) that it considered most important when determining cancer incidence.

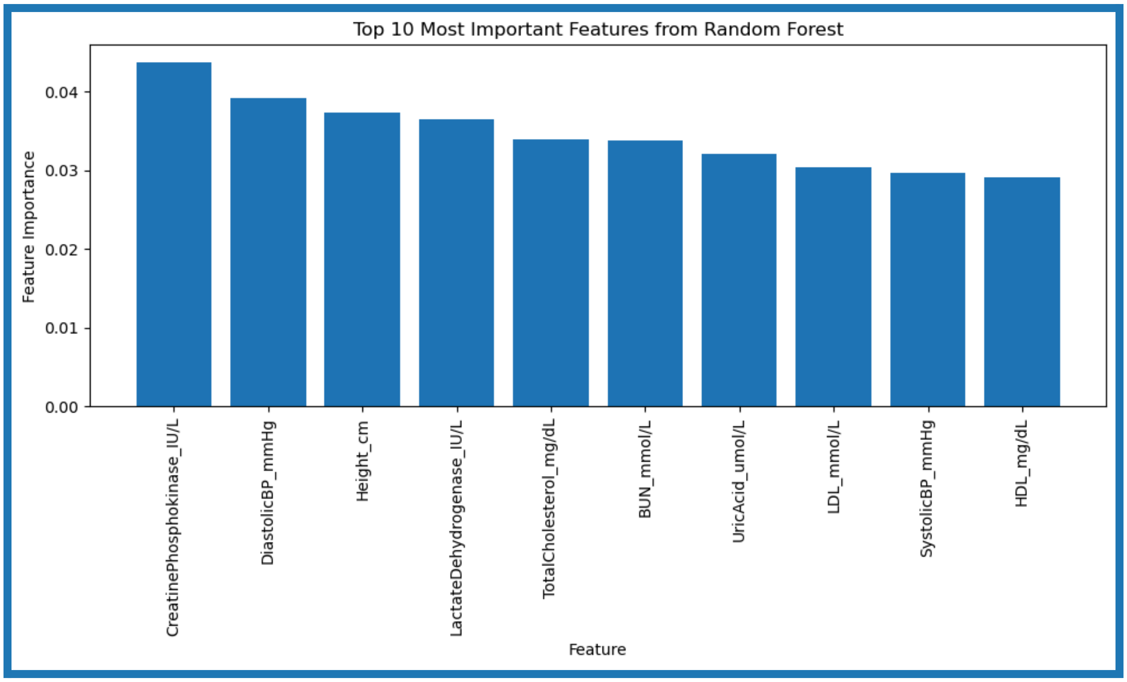
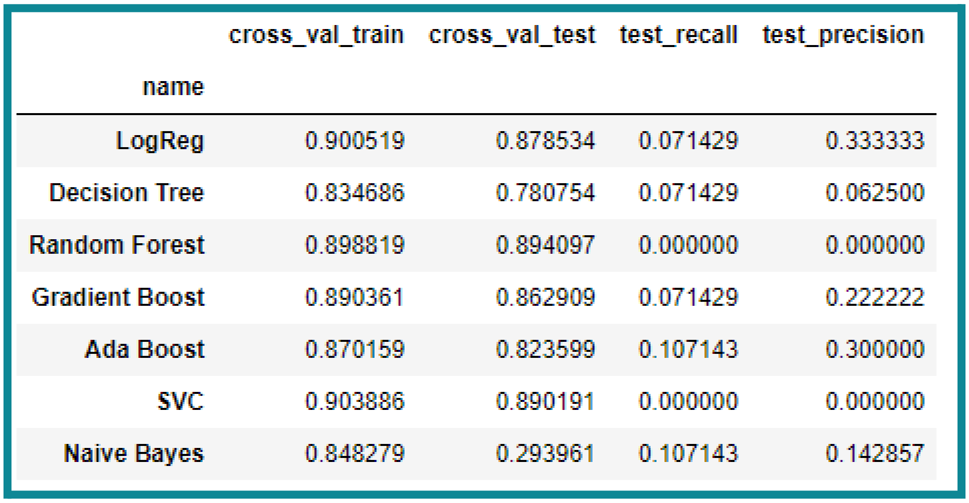


Figure 4. Image Capture of Ten Most Important Features Related to Cancer Incidence, According to the Random Forest Classifier Feature Selection Method

After cleaning the dataset and reducing it to incorporate only features most related to cancer incidence, we used Python to import different classification algorithms (such as logistic regression, decision trees, random forest, and Naive Bayes). Python ranked these algorithms based on their cross-validation, recall, and precision scores. We selected the best-performing algorithm based on these conditions and conducted hyperparameter tuning. In the case of the cancer predictive model, logistic regression had the highest cross-validation, recall, and precision results. We proceeded to perform hyperparameter tuning on this algorithm.



	cross_val_train	cross_val_test	test_recall	test_precision
name				
LogReg	0.900519	0.878534	0.071429	0.333333
Decision Tree	0.834686	0.780754	0.071429	0.062500
Random Forest	0.898819	0.894097	0.000000	0.000000
Gradient Boost	0.890361	0.862909	0.071429	0.222222
Ada Boost	0.870159	0.823599	0.107143	0.300000
SVC	0.903886	0.890191	0.000000	0.000000
Naive Bayes	0.848279	0.293961	0.107143	0.142857

Figure 5. Results of Classification Algorithm Tests for Cancer Incidence Predictive Model

Essentially, we used statistical metrics to compare the performance of models developed using the aforementioned classification algorithms. For the cancer prediction model, we determined the logistic regression algorithm to be the most reliable. For the kidney failure prediction model, we utilized the Adaptive Boosting algorithm. For both models, we created confusion matrices depicting the accuracy of our models in generating “false positives” or “true positives.”

Conclusion

Results

According to the Spearman correlation heatmaps we developed based on the NHANES dataset (utilizing the Spearman correlation coefficient values), cancer incidence is positively correlated with total protein (0.1), globulin (0.09), alanine aminotransferase (0.09), and urine thallium (0.07). The development of weak or failing kidneys is positively correlated with angina (0.27), high cholesterol (0.13), high-sensitivity C-reactive protein (0.10), and urine cadmium (0.08). The total heavy metal burden is positively correlated with urine creatinine (0.65) and urine albumin (0.44).

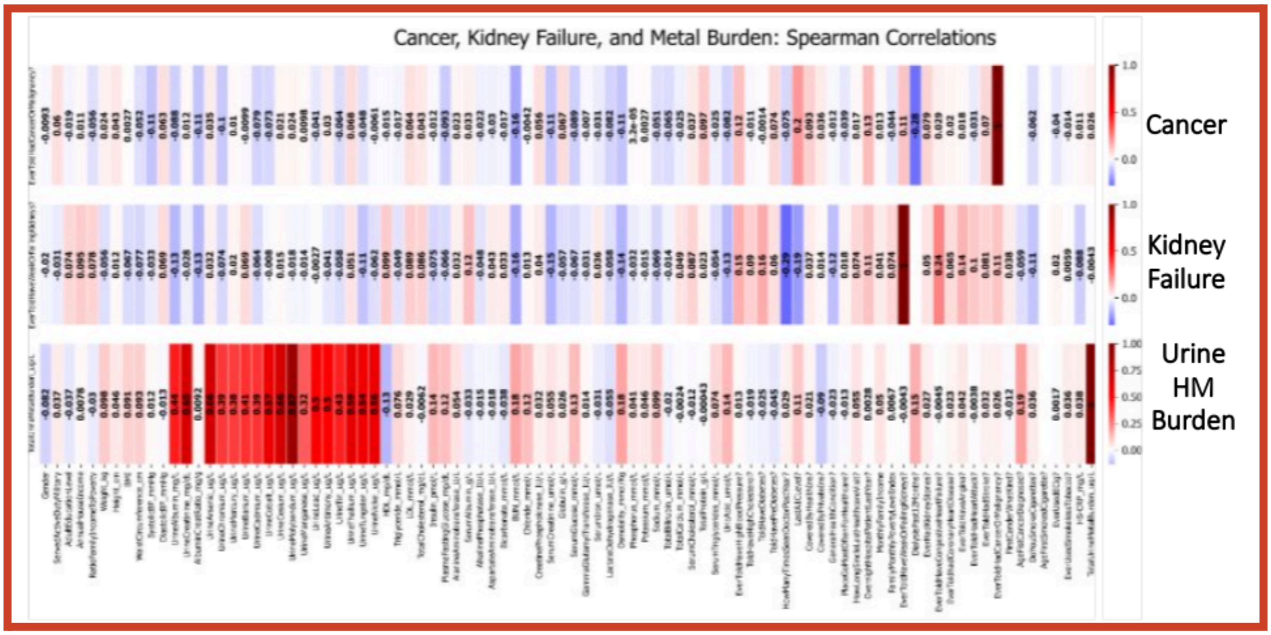


Figure 6. Correlation Heatmaps for Total Metal Burden, Kidney Failure, and Cancer Incidence

As previously mentioned, we developed machine learning models to predict cancer and kidney failure. The logistic regression model we created to predict cancer had better negative predictive power, meaning that it performed better in predicting the absence of cancer than it did in predicting the presence of cancer.

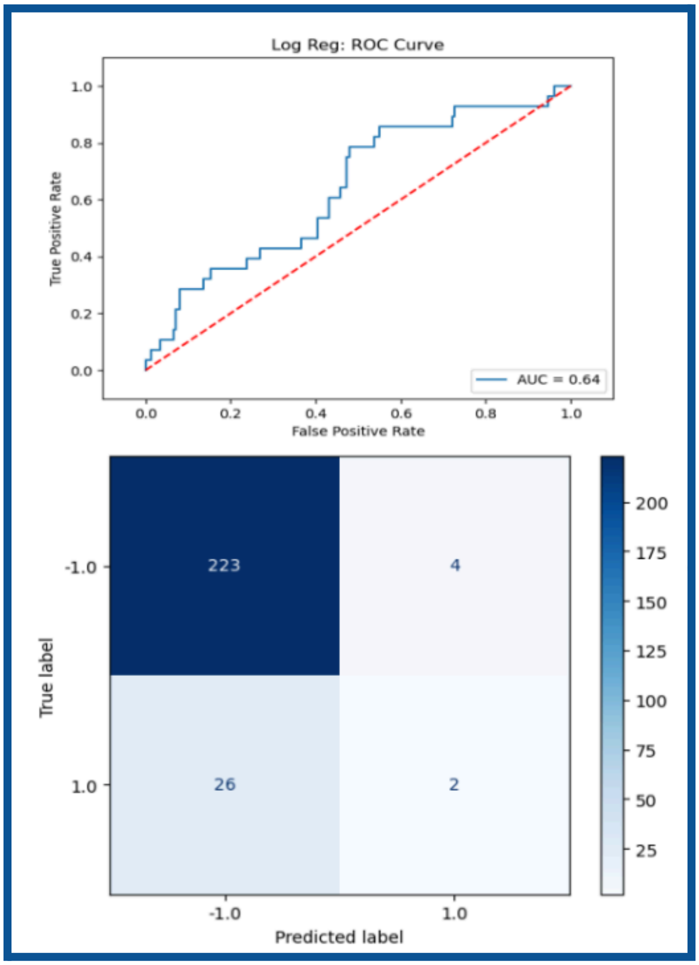


Figure 7. Logistic Regression Receiver Operator Characteristic (ROC) Curve and Confusion Matrix for Cancer Prediction Model

The adaptive boosting model we developed to predict kidney failure appears to function at 100% accuracy. This is highly improbable and an indicator that the model requires optimization.

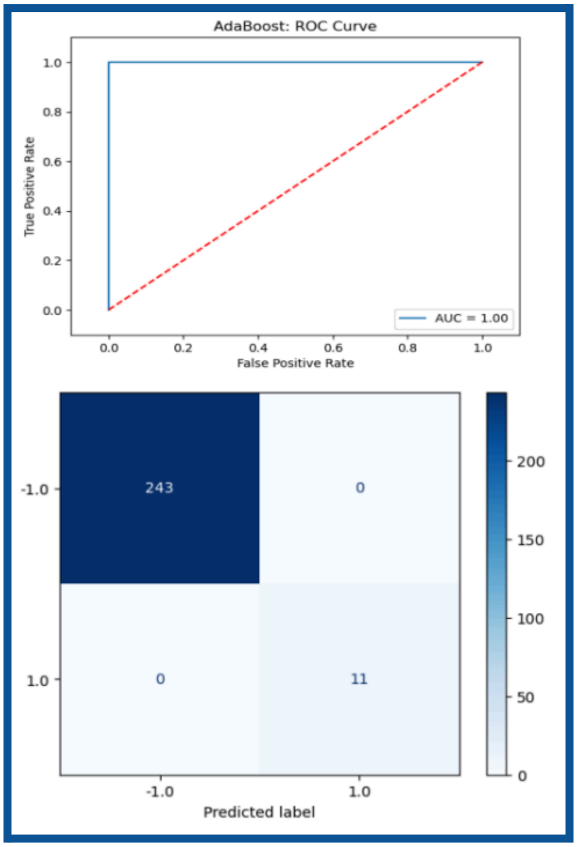


Figure 8. Adaptive Boosting Regression Receiver Operator Characteristic (ROC) Curve and Confusion Matrix for Kidney Failure Prediction Model

Validation and Verification

As discussed previously, our cancer and kidney failure machine learning models underwent validation when we, using Python, split our data into training and test sets. We adjusted the parameters on the training set, assessed performance on the validation set, and evaluated the performance of different classification algorithms to select the most suitable ones. Still, our cancer and kidney failure machine-learning models require further optimization and verification. We will work on this in the near future.

Discussion

Our Spearman correlation heatmap depicted a high correlation between the total metal burden, urine creatinine, and urine albumin levels. The latter two are indicators of kidney failure. As such, this result suggests a correlation between heavy metals exposure and kidney failure. The fact that cancer incidence is positively correlated with alanine aminotransferase (a kidney-and-liver-bound enzyme, the increased presence of which can result in CKD) and urine thallium (a metal) also suggests that cancer is connected to CKD and HM exposure. However, further research into the scientific implications of our statistical results is necessary to ensure the validity of these claims. Our most significant accomplishment in this project was the creation of two machine-learning models predicting cancer and kidney failure.

Future Work

High levels of nullity in the NHANES datasets (due to irresponsive subjects) may have skewed the results of our study. In the future, we intend to refine and optimize our predictive models for cancer and kidney failure while conducting literature searches that corroborate or allow us to verify our research results. We might also replicate our experiment on another, more complete dataset or on other years of the National Health and Nutrition Examination Survey.

Implementing Artificial Intelligence for Bioinformatics and Rapid Analysis of Cancer, CKD, and Heavy Metals in NHANES Datasets by Novice Programmers

Acknowledgments

Our involvement in this year's Supercomputing Challenge was made possible with help from many institutions and individuals.

We acknowledge the facilities and resources provided by the National Institute of Health, National Cancer Institute, University of New Mexico Comprehensive Cancer Center (UNMCCC), and the UNM Emergency Medicine Department. These resources were essential for completing this research project. We especially acknowledge the Continuing Umbrella of Research Experiences program, heralded by the UNMCCC.

We appreciate our mentors, Dr. Baca and Dr. Taylor, whose expertise in programming and biological sciences proved invaluable to our project. Dr. Baca and Dr. Taylor receive infrastructure support from the National Center for Advancing Translational Sciences, National Institutes of Health, under grant number UL1TR001449.

Furthermore, we are particularly thankful for the support provided by Ms. Rebecca Campbell from Justice Code and by the Supercomputing Challenge team, including Ms. Patty Meyer, Ms. Celia Einhorn, Ms. Mary Sagartz, and Mr. David Kratzer.

Finally, we extend our gratitude to the judges and scientists associated with the Supercomputing Challenge for their support throughout this endeavor.

Implementing Artificial Intelligence for Bioinformatics and Rapid Analysis of Cancer, CKD, and Heavy Metals in NHANES Datasets by Novice Programmers

References

1. Dirymer V. Chronic Kidney Disease in Persons with Multiple Chronic Diseases, NM, 2014. NMHealth. July 15, 2016. Accessed April 9, 2024. https://www.nmhealth.org/data/view/report/1925.
2. Tendulkar KK, Cope B, Dong J, Plumb TJ, Campbell WS, Ganti AK. Risk of malignancy in patients with chronic kidney disease. PLOS ONE. August 17, 2022. Accessed April 9, 2024. https://doi.org/10.1371/journal.pone.0272910.
3. Guo K, Wang Z, Luo R, Cheng Y, Ge S, Xu G. Association between chronic kidney disease and cancer including the mortality of cancer patients: national health and nutrition examination survey 1999-2014. Am J Transl Res. 2022;14(4):2356-2366. April 15, 2022. Accessed April 9, 2024
4. Jalili C, Kazemi M, Cheng H, et al. Associations between exposure to heavy metals and the risk of chronic kidney disease: A systematic review and meta-analysis. *Critical Reviews in Toxicology*. Published online May 7, 2021:1-30. doi:10.1080/10408444.2021.1891196
5. Nhanes questionnaires, datasets, and related documentation. Centers for Disease Control and Prevention. Accessed April 9, 2024. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017.