CP468 Project Report

Andre Grandon

190831360

Johnson Huynh

190170320

Tanisha Mehta

201918330

CP468 - Artificial Intelligence

Dr. Sumeet Kaur Sehra

Wilfrid Laurier

August 9, 2023

Table of Contents

Abstract:

Chronic Kidney Disease (CKD) is a widespread medical condition that necessitates early detection and intervention for effective treatment. In this project, we delve into the realm of artificial intelligence and machine learning to construct a predictive model capable of identifying the presence of CKD based on a range of medical attributes. Through a systematic approach, encompassing exploratory data analysis, data preprocessing, and the implementation of a Logistic Regression model, we aim to contribute to the advancement of medical diagnostics.

Introduction:

Chronic Kidney Disease poses a significant healthcare challenge, impacting millions of lives worldwide. Timely diagnosis can significantly improve patient outcomes and quality of life. As technology continues to permeate healthcare, machine learning offers a promising avenue for enhancing medical diagnostics. Our project addresses this challenge by developing and evaluating a CKD prediction model that harnesses the power of machine learning algorithms.

Project Description:

Our project is structured around the following key stages:

Exploratory Data Analysis (EDA):

> The initial phase involves delving into the dataset to uncover valuable insights. We visualize the distribution of CKD cases, identify missing values, explore correlations between variables, and gain an understanding of the dataset's characteristics.

Data Preprocessing:

> To facilitate model training, we employ data preprocessing techniques. This includes standardization to ensure all variables are on a common scale, and the transformation of categorical features through one-hot encoding. The dataset is then split into training and testing sets to enable robust model evaluation.

Logistic Regression:

Logistic Regression, a widely used classification algorithm, forms the core of our predictive model. Leveraging the LogisticRegression class from the scikit-learn library, we train the model on the preprocessed data to predict the likelihood of CKD presence.

Model Evaluation:

Comprehensive model evaluation is critical to assess its performance. We present classification reports detailing precision, recall, F1-score, and support for both the training and testing sets. Additionally, confusion matrices provide a visual representation of the model's predictions.

Methodology:

Exploratory Data Analysis (EDA):

- CKD Cases Count: Visualizing the distribution of CKD cases provides a fundamental understanding of the dataset's class balance, setting the stage for subsequent analysis.
- Missing Values: An exploration of missing values column-wise offers insights into data quality and potential areas for data imputation or cleaning.
- Category Graph Observations: We delve into category graphs and histograms to uncover patterns within categorical and continuous variables, highlighting potential relationships.
- Correlation Matrix: The correlation matrix provides a comprehensive view of the relationships between variables, guiding feature selection and engineering.

Data Preprocessing:

- Standardization: By standardizing selected features using StandardScaler, we ensure that each feature contributes equally to model training without being influenced by different scales.
- Categorical Transformation: Transforming categorical variables using one-hot encoding enables the model to comprehend categorical data while preserving the integrity of the original features.

Logistic Regression:

- Train-Test Split: To avoid overfitting, we split the dataset into training and testing subsets, ensuring the model's ability to generalize to unseen data.
- Logistic Regression Model: We employ the LogisticRegression class to construct a logistic regression model, enabling us to model the probability of CKD presence based on input attributes.

Model Evaluation:

- Classification Report: The classification report furnishes a comprehensive overview of the model's performance, encompassing metrics such as precision, recall, F1-score, and support for each class.
- Confusion Matrix: Visual representations of confusion matrices provide a clear depiction of the model's predictive accuracy and error rates, aiding in the interpretation of results.

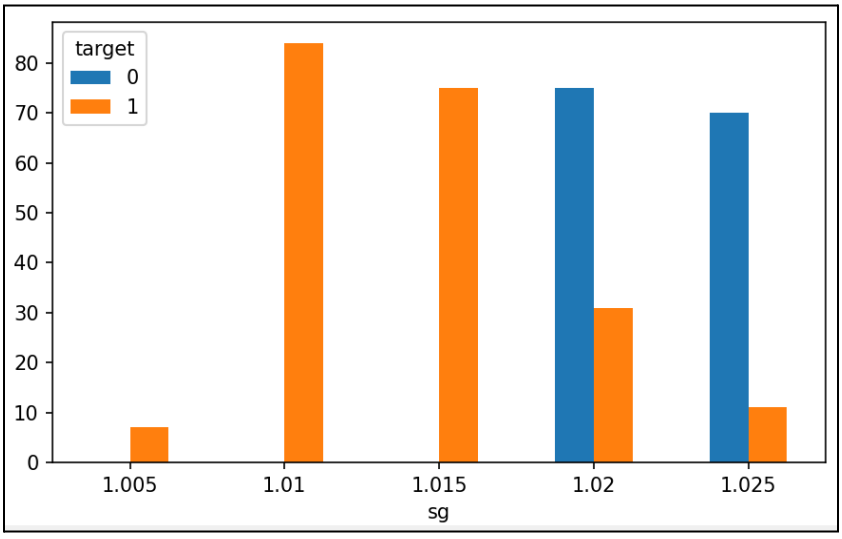
Results:

The outcomes of our project are twofold:

Exploratory Data Analysis:

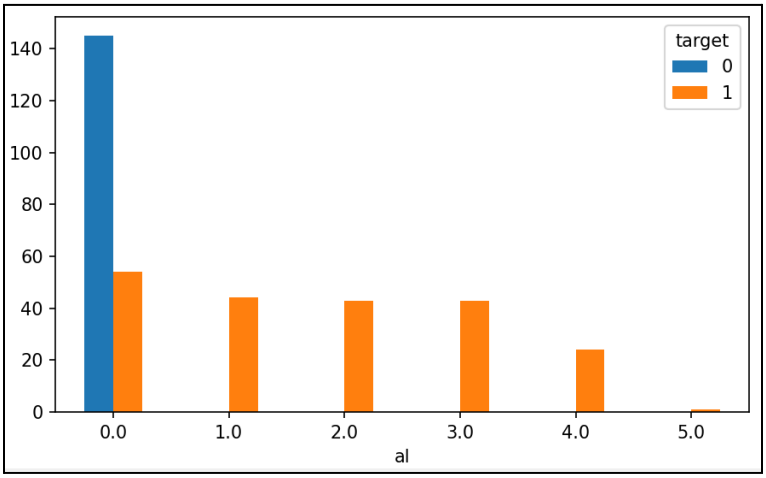
- CKD Cases Distribution: The visualization of CKD cases highlights the dataset's class distribution, revealing insights into the prevalence of the disease.
- Missing Values Analysis: Our analysis of missing values sheds light on potential data quality issues and informs imputation strategies.
- Category Graph Observations: Graphical observations reveal trends and patterns within categorical variables, providing a foundation for feature engineering.
- Correlation Insights: The correlation matrix unveils relationships between variables, guiding feature selection and offering potential insights into predictive factors.

Category Graph Observations

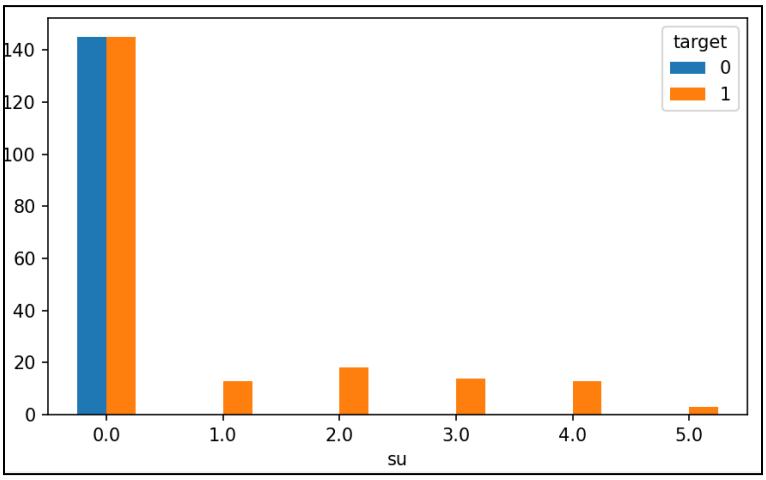
- 

(Specific Gravity)

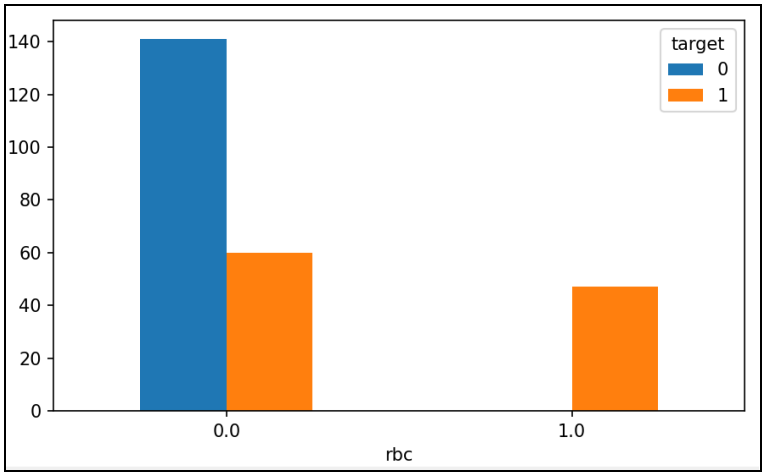
- Those with a sg (specific gravity) of 1.015 or under are more likely to have CKD judging from the graph. According to a 2017 article from Schott et al. under *Chapter 14: Disorders of the Urinary System*, “with chronic renal insufficiency, the ability to produce concentrated (specific gravity greater than 1.025) or dilute (specific gravity less than 1.008) urine is lost”.

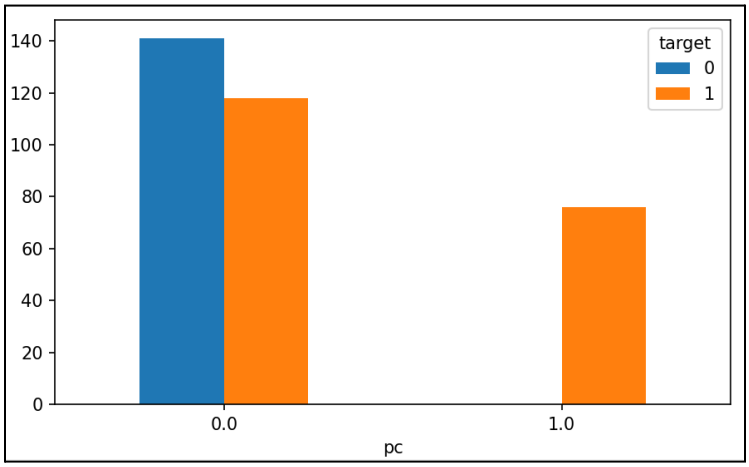
- 

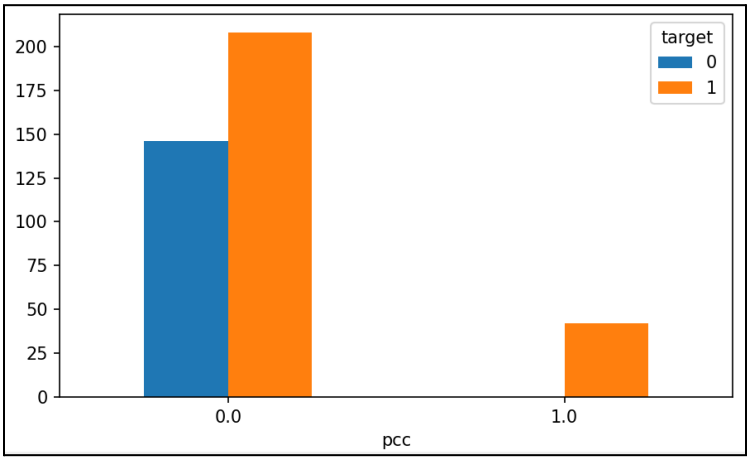
(Albumin)
    - From the graph and fact checking on Albuminuria, it is just better to not have much Albumin in normal urine. According to the National Institute of Diabetes and Digestive and Kidney Diseases, the term Albuminuria is about having “too much albumin in your urine” and also “a sign of kidney disease” (2016). Note that Albumin is a protein commonly found in blood.

- 

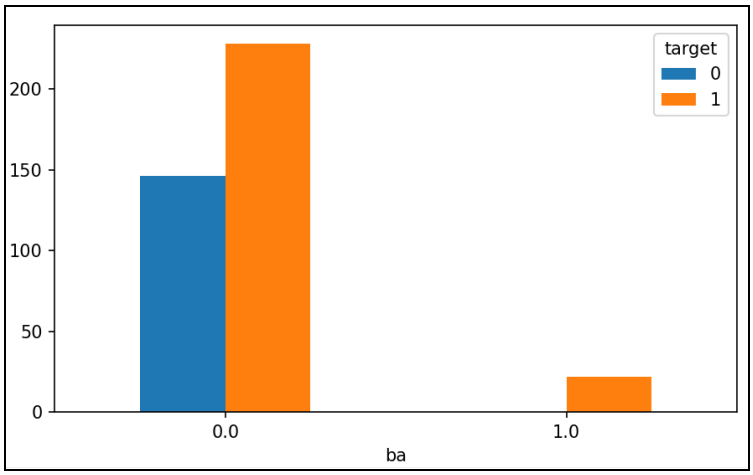
(Sugar)
    - There is another description for Sugar under the Blood Glucose graph.

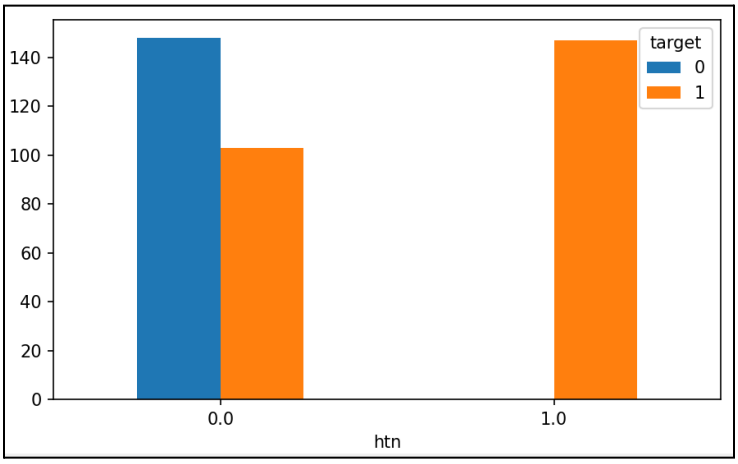
- 

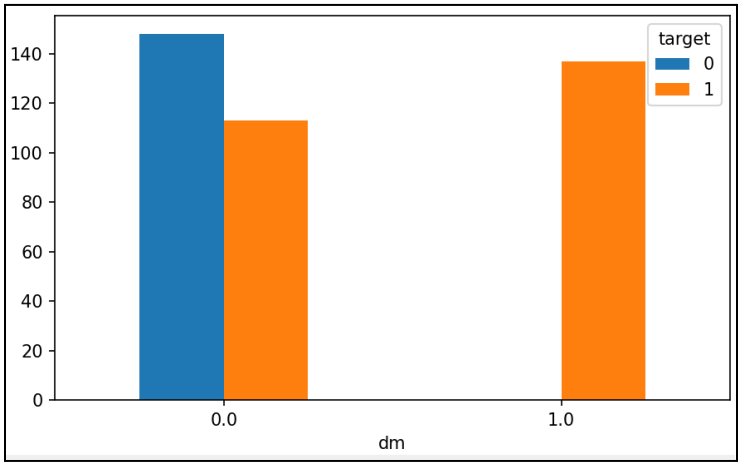
(Red Blood Cells, 0 is normal, 1 is abnormal)
    - For the next few similar graphs, it just shows that being young and healthy can reduce the risk of having chronic kidney disease in the first place.
- 

(Pus Cell, 0 is normal, 1 is abnormal)
- 

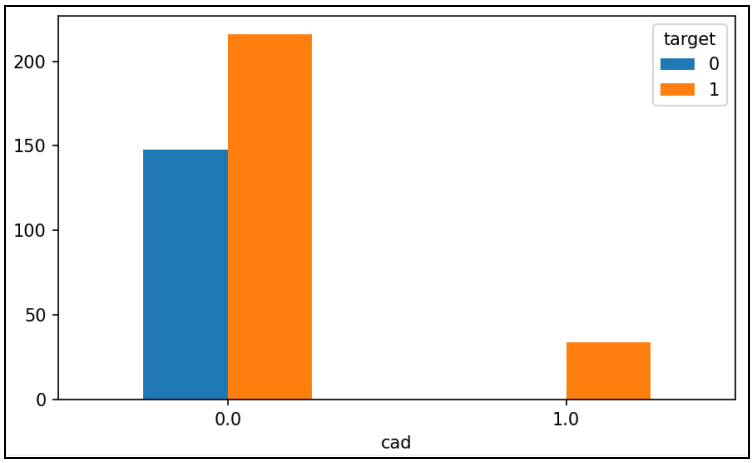
(Pus Cell clumps, 1 is present, 0 is not present)

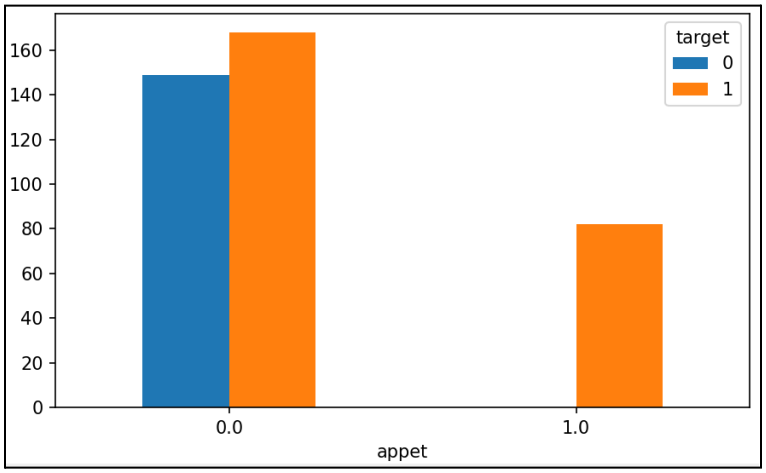
- 

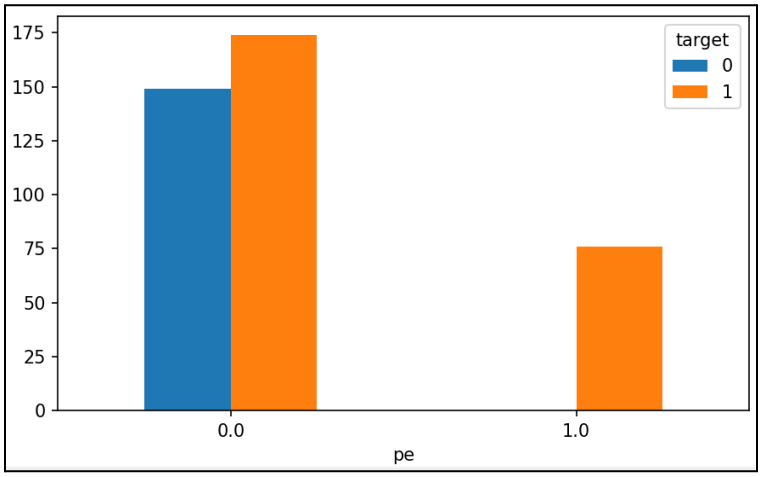
(Bacteria, 1 is present, 0 is not present)
- 

(Hypertension)
- 

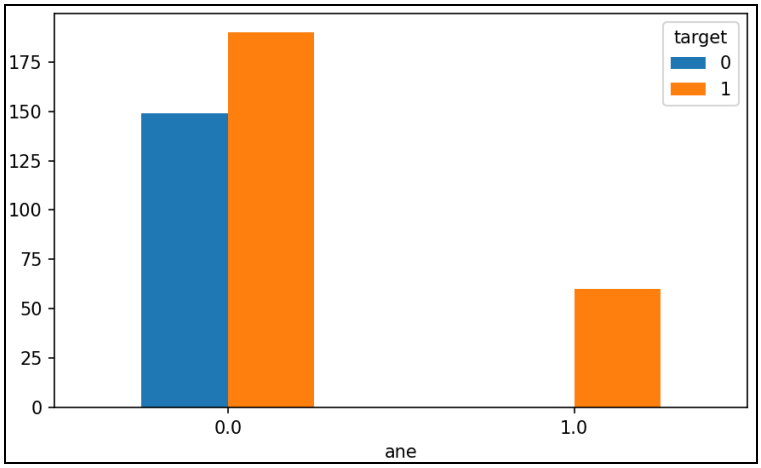
(Diabetes Mellitus)
    - Linked to Sugar and Blood Glucose graphs, so look under the Blood Glucose graph.

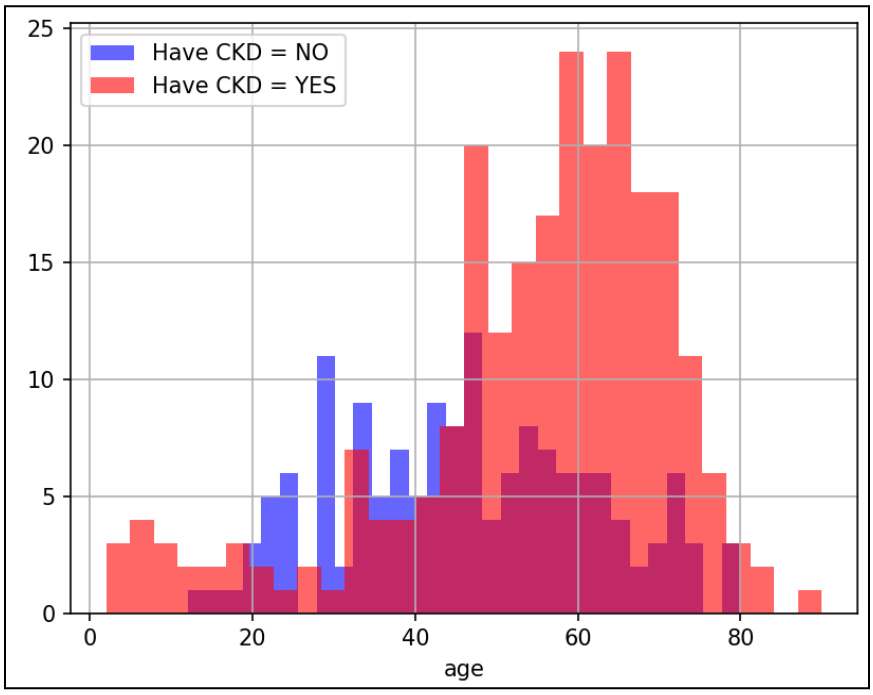
- 

(Coronary Artery Disease)
- 

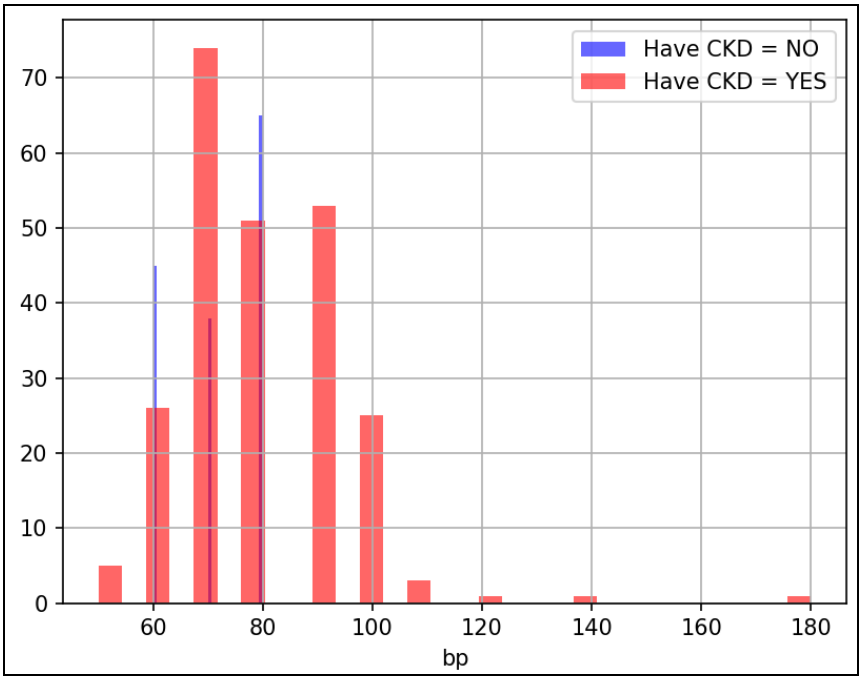
(Appetite, 0 is good, 1 is poor)
- 

(Pedal Edema)

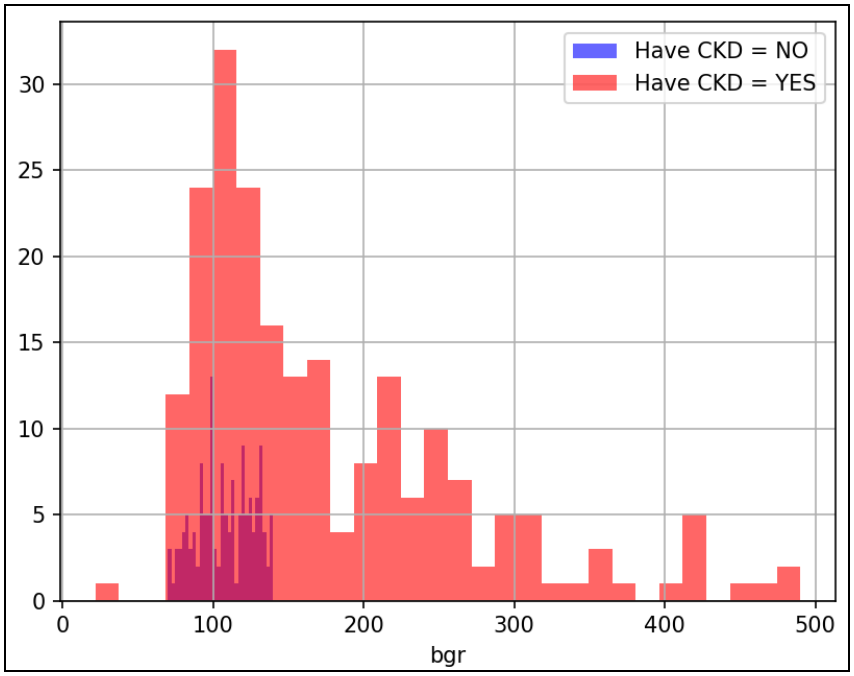
- 

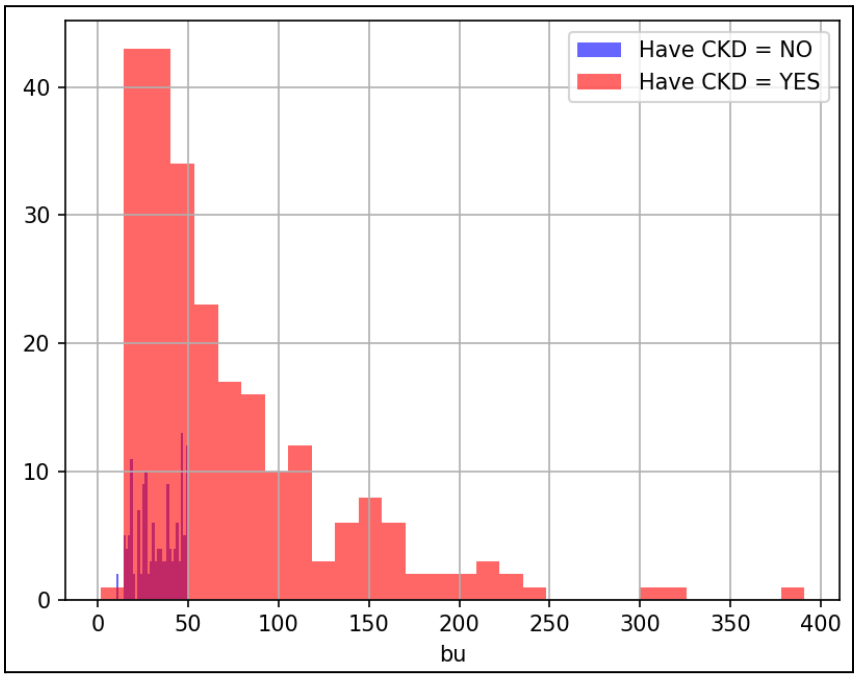
(Anemia)
    - Those suffering from anemia has a higher likelihood of having CKD then not. For more check the “Packed Cell Volume” graph description.
- 

    - The likelihood of having Chronic Kidney Disease(CKD) increases after around 50 years old

- 

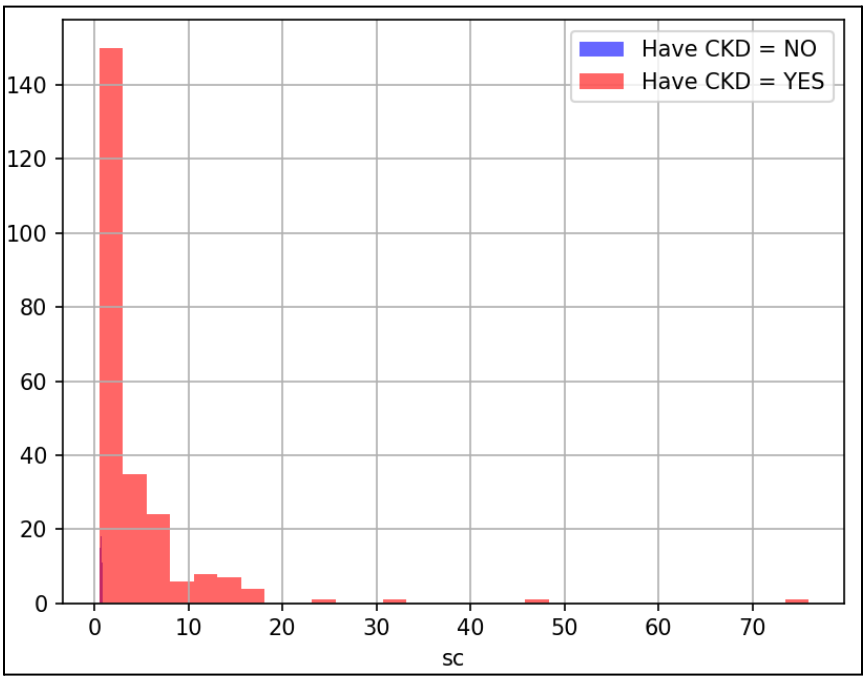
(Blood Pressure in mm/Hg)
    - The Blood Pressure between 70 to 90 mm/Hg is concerning. According to the National Health Service from UK, the “ideal blood pressure is considered to be between 90/60mmHg and 120/80mmHg” (2022). It is stated on the topic of blood pressure that people with around 90/60 or less blood pressure are considered to have low blood pressure (the two values are active and resting blood pressure).

- 

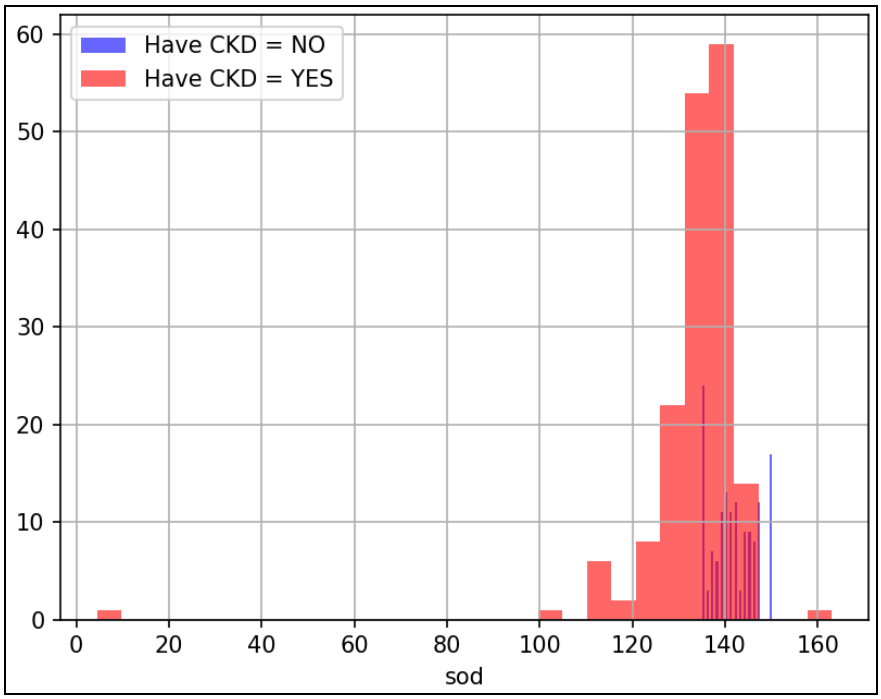
(Blood Glucose Random in mgs/dl)
    - Blood Glucose here is referring to sugar levels in the blood, there is another graph already showing that having high amounts of sugar increases the chances of having a Kidney disease. According to the National Institute of Diabetes and Digestive and Kidney Diseases, “Diabetes is the leading cause of kidney disease”.
- 

(Blood Urea in mgs/dl)

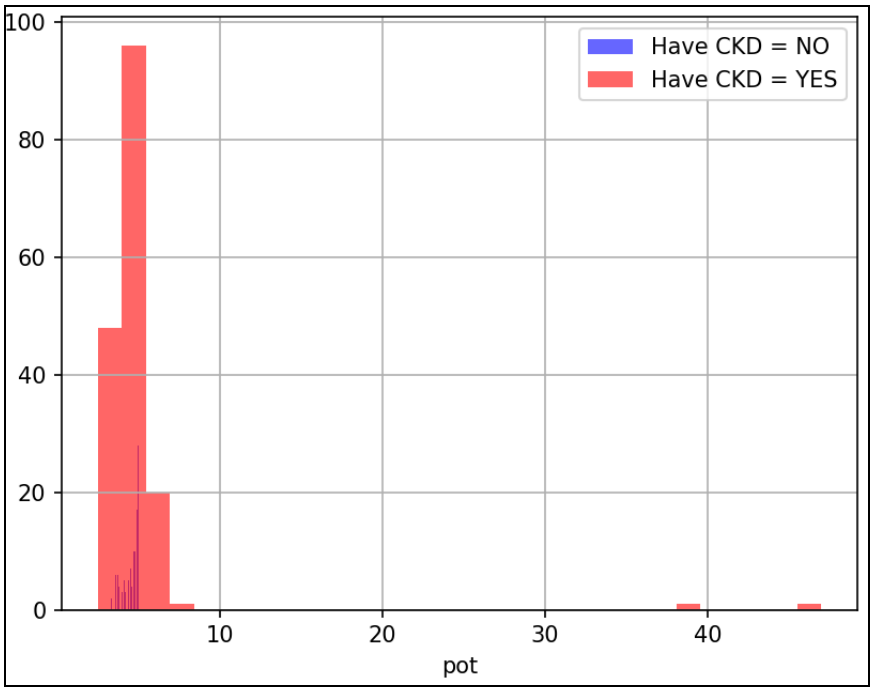
- For general information, according to the U.S. National Library of Medicine “Urea nitrogen is a waste product that your kidneys remove from your blood” (n.d.). Therefore, having higher then normal Urea nitrogen in blood is a sign that the kidney is not functioning well.

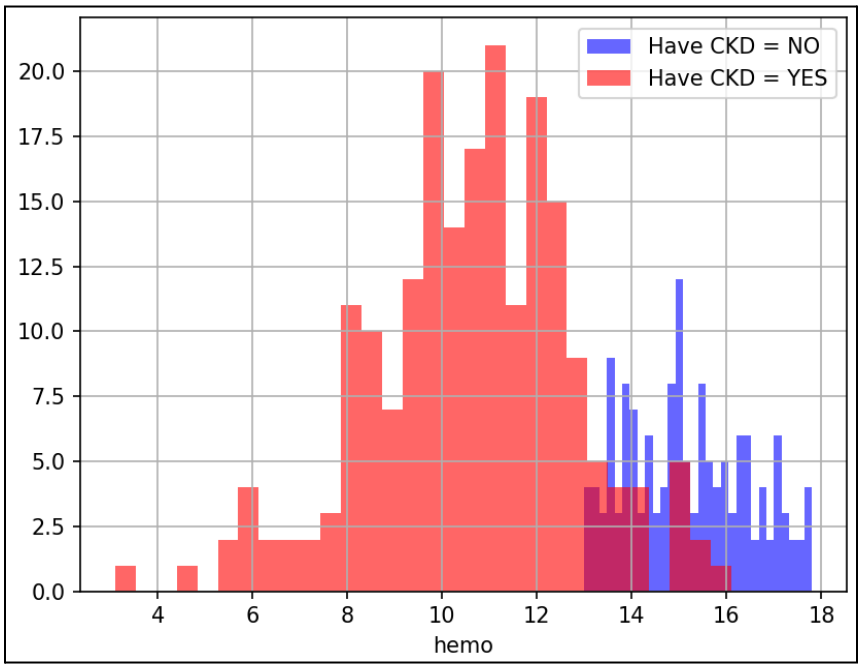
- 

(Serum Creatinine in mgs/dl)
    - It maybe that most of the missing values from the dataset are for Serum Creatinine from healthy patients. According to the National Kidney Foundation in 2023, “Creatinine is a waste product that comes from the digestion of protein in your food and the normal breakdown of muscle tissue”. So even if the Serum Creatinine is high, it is possible that there is a problem in kidney health or the patient has eaten too much cooked meat, had recently done high intensity exercise or has high muscle mass. Otherwise, this depends on a case by case basis on what would be abnormal levels.

- 

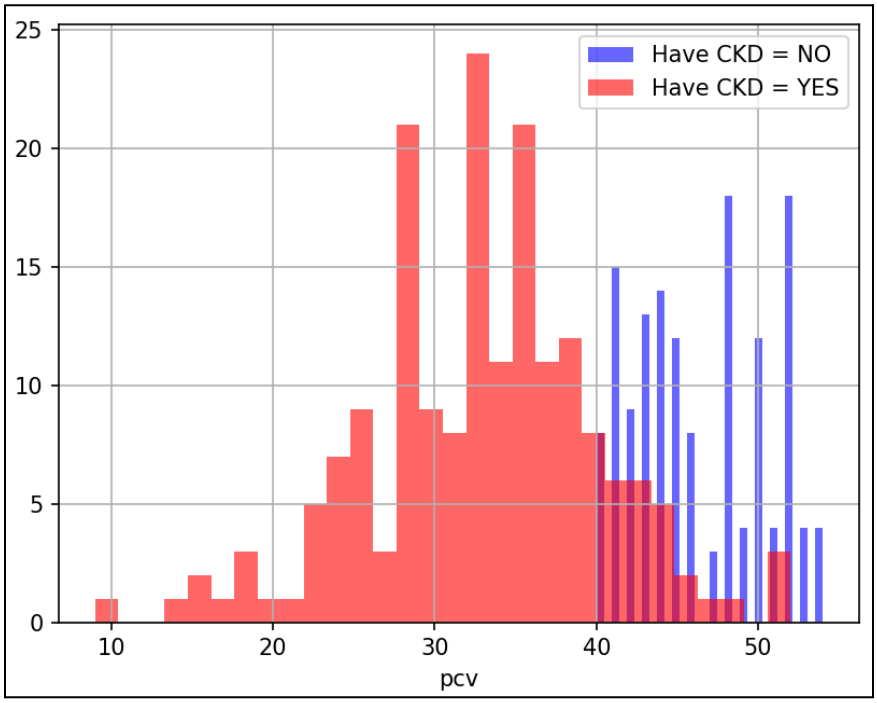
(Sodium in mEq/L)
    - Having Sodium above 130 mEq/L are more likely to have chronic kidney disease (CKD). When searching about the relationship between sodium and kidney disease, according to the Ontario Renal Network in 2012 while the kidney is not functioning correctly having “too much sodium (salt) can cause high blood pressure, which can cause further damage to your kidneys”.

- 

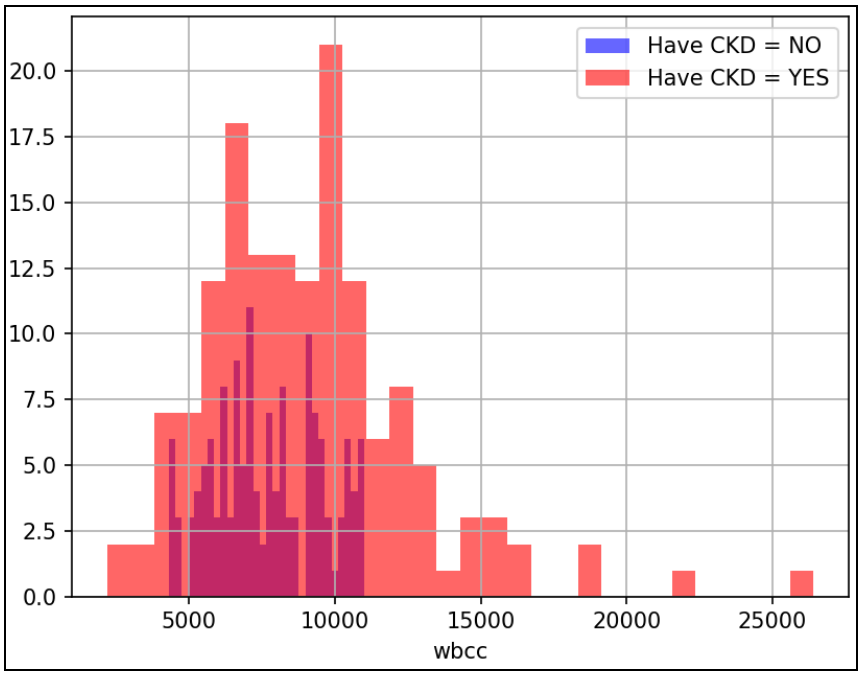
(Potassium in mEq/L)
    - According to the graph, people with potassium levels in the blood around 5 are more likely to have CKD. But according to the National Kidney Foundation, people with kidney diseases has “trouble removing extra potassium from the blood” and are “at risk for low potassium, especially during earlier stages of kidney disease” (2023).
- 

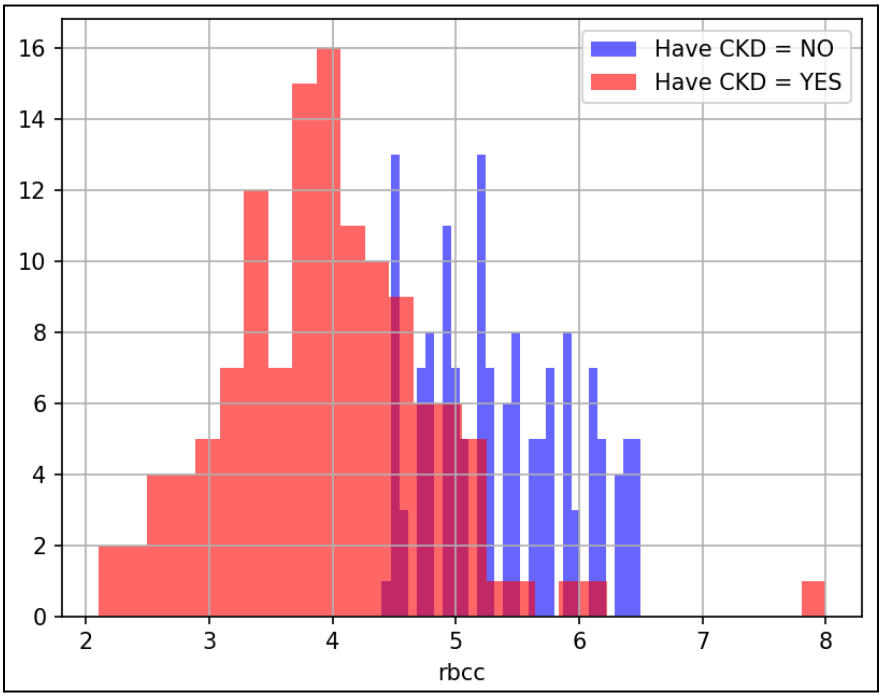
(Hemoglobin in gms)

- Those with a hemoglobin amount of 9 to 13 gms is more likely to have CKD. From Henny Billett's book on clinical methods, "The normal Hb level for males is 14 to 18 g/dl; that for females is 12 to 16 g/dl" (1990). So the range of 9 to 13 gms is considered low.



- (Packed Cell Volume)
    - Those with a packed cell volume of 38 or lower has a greater chance of having CKD. According to Yashoda Hospital's website page in 2023 on PCV tests, the normal range of PCV is 35.5 to 44.9% in females and 38.3% to 48.6% in males. This appears to be where most of the normal results exist in the graph. Note that from the same source, "a low PCV implies that the patient has a low number of red blood cells and is suffering from anemia" and that in both of these graphs there is an increased chance of having CKD (PCV Test, 2023).

- 

(White Blood Cell Count in cells/cumm)
  - Having a White Blood Cell Count above 5000 seems to be a concern as it appears in the graph. According to the 2017 CKD study by Arai et al. “elevated white blood cell (WBC) count is a well-known predictor of chronic kidney disease (CKD) progression”. So the findings from the case graph matches the relation results found by Arai et al.
- 

(Red Blood Cell Count in millions/cmm)

    - From the graph, those with 4.5 millon/cmm or lower in red blood cell count are at a higher risk of CKD. According to the National Kidney Foundation, when a patient has a kidney disease their kidneys cannot make enough erythropoietin hormones (EPO) which can cause their “red blood cell count to drop and anemia to develop” (2020).

Logistic Regression Model:

- Model Training: Our Logistic Regression model is trained on preprocessed data, learning to predict the likelihood of CKD presence based on medical attributes.
- Model Evaluation: Classification reports and confusion matrices elucidate the model's performance on both training and testing datasets, showcasing its predictive capabilities.

Conclusion:

In conclusion, this project epitomizes the symbiotic relationship between medical science and machine learning. By harnessing the power of artificial intelligence, we have developed a predictive model with the potential to revolutionize CKD detection. Our journey, encompassing data exploration, preprocessing, model construction, and evaluation, underscores the significance of data-driven methodologies in addressing real-world challenges. As we embrace the digital age, this project underscores the transformative potential of AI in reshaping the landscape of healthcare diagnostics.

Reference

Arai, Y., Kanda, E., Iimori, S., Naito, S., Noda, Y., Sasaki, S., Sohara, E., Okado, T., Rai, T., & Uchida, S. (2017, July 11). *Low white blood cell count is independently associated with chronic kidney disease progression in the elderly: The CKD-route study*. Clinical and experimental nephrology. https://pubmed.ncbi.nlm.nih.gov/28699033/

Billett, H. H. (1990). *Hemoglobin and hematocrit - clinical methods*. NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK259/

Kharwal, A. (2020, November 10). *Heart disease prediction using machine learning*. The Clever Programmer. https://thecleverprogrammer.com/2020/11/10/heart-disease-prediction-using-machine-learning/

National Health Service (NHS). (2022, December 12). *What is blood pressure?*. NHS choices. https://www.nhs.uk/common-health-questions/lifestyle/what-is-blood-pressure/

National Health Service (NHS). (2023, April 14). *Low blood pressure (hypotension)*. NHS inform. https://www.nhsinform.scot/illnesses-and-conditions/heart-and-blood-vessels/conditions/low-blood-pressure-hypotension

National Kidney Foundation. (2020, October 30). *Anemia and chronic kidney disease*. National Kidney Foundation. https://www.kidney.org/atoz/content/what_anemia_ckd

National Kidney Foundation. (2023a, May 26). *Potassium*. National Kidney Foundation. https://www.kidney.org/atoz/content/about-potassium

National Kidney Foundation. (2023b, July 25). *Serum (blood) creatinine*. National Kidney Foundation. https://www.kidney.org/atoz/content/serum-blood-creatinine

Ontario Renal Network. (2012). *Chronic Kidney Disease Nutrition Fact Sheet*. Ontario Renal Network. https://www.ontariorenalnetwork.ca/sites/renalnetwork/files/assets/fnimfactsheet-sodium-metis-english.pdf

PCV Test. Yashoda Hospitals. (2023, February 17). https://www.yashodahospitals.com/diagnostics/pcv-test/

Rubini, L. J., Soundarapandian, P., & Eswaran, P. (2015, July 2). *Chronic_Kidney_Disease*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease

Schott, H. C., Waldridge, B., & Bayly, W. M. (2017, November 17). *Disorders of the urinary system*. Equine Internal Medicine (Fourth Edition). https://www.sciencedirect.com/science/article/abs/pii/B9780323443296000140

U.S. Department of Health and Human Services. (2016, October). *Albuminuria: Albumin in the urine*. National Institute of Diabetes and Digestive and Kidney Diseases.

https://www.niddk.nih.gov/health-information/kidney-disease/chronic-kidney-disease-ckd/tests-diagnosis/albuminuria-albumin-urine

U.S. Department of Health and Human Services. (2017, February). *Diabetic Kidney Disease*. National Institute of Diabetes and Digestive and Kidney Diseases. https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/diabetic-kidney-disease

U.S. National Library of Medicine. (n.d.). *BUN (Blood Urea Nitrogen): Medlineplus medical test*. MedlinePlus. https://medlineplus.gov/lab-tests/bun-blood-urea-nitrogen/