Ensemble Learning Approaches For Predicting Occurance of Kidney Stones in NHANES 2017 - 2020

David Nguyen 300584723 (Individual)

Individial Github Repository

Group 2 Github Repository
Katie Chu-Fong 300601174; Thinh Nguyen 300611544;
Brook Thomson 300653729

2024-10-22

Executive Summary

Kidney stones affect approximately 10% of people in their lifetime, causing severe pain and potentially leading to serious complications. Our research aimed to develop a predictive model to identify individuals at high risk for kidney stone formation, using data from the National Health and Nutrition Examination Survey.

We analyzed various factors and found that several key elements are strongly associated with an increased risk of kidney stones: High blood pressure, Gender (being female), Saturated fatty acid intake (particularly dodecanoic acid), Abdominal pain history, Vitamin A intake, Age, Race (particularly non-Hispanic White and non-Hispanic Black), History of gallstones

Our predictive model demonstrated strong performance in identifying individuals at risk for kidney stones. Importantly, the model revealed that risk factors can vary significantly between individuals. For example, high blood pressure emerged as one of the most significant risk factor, while certain dietary elements like saturated fatty acid intake also played a crucial role.

These findings highlight the complex nature of kidney stone formation and underscore the importance of personalized risk assessment. By leveraging these insights, healthcare providers can implement targeted preventive measures, potentially reducing the incidence of kidney stones, improving patient outcomes, and decreasing healthcare costs associated with treatment and complications.

Additionally, we recommend implementing targeted screening programs and patient education focusing on modifiable risk factors. While our model is promising, further research is needed to fully understand kidney stone formation mechanisms and refine predictive models, particularly regarding the interplay between physiological conditions and dietary factors across diverse populations.

Contents

1. Background

Is kidney stone prevalence associated with factors such as diet, lifestyle, and other existing medical conditions? Kidney stones are solid deposits of minerals and salts that develop in the urinary tract, and can cause blockage and severe pain when urine is passed (National Institutes of Health, n.d.). It is a common condition that affects approximately 10% of people at least once in their lifetime, and in some cases may require significant and/or recurrent treatment (Abufaraj et al., 2020).

This report presents an exploratory data analysis, investigating variables previously shown to be associated with kidney stone occurance. It is based on the National Health and Nutrition Examination Survey (NHANES) from the National Center for Health Statistics, of the Centers for Disease Control and Prevention. NHANES is an ongoing program of surveys in the United States that assesses the health and nutritional status of adults and children. The surveys collect health-related data ranging over a number of topics, which are organised broadly into Demographics, Dietary, Examination, Laboratory, and Questionnaire. It is widely used to analyse or identify associative or causal factors of various conditions and/or diseases, such as diabetes, hypertension, and hearing loss. Thus, the comprehensive scope of NHANES makes it an ideal resource to investigate factors associated with kidney stones, especially given the diverse range of known causes.

In the following sections, we evaluate the completeness, quality, and distributions of kidney stone-relevant data in NHANES. Data from the most recent cycle is used, NHANES 2017 - March 2020.

2. Data Description

2.1 Data Structures and Types

Data from each NHANES cycle is released as many tables, each containing a collection of similar features. For the specific focus on kidney stone disease, only a subset of tables is used, and from these tables, only a subset of key features. The integrated dataset used in this project is composed of 9208 instances/rows, and 146 columns. The column `SEQN` contains a unique identifier for each instance, and the column `KIQ026` contains the target variable. Thus, there are 144 informative features.

The target variable belongs to the Questionnaire component of NHANES, and is phrased as "Ever had kidney stones?". Possible answers of this question are "Yes", "No", "Refused", and "Don't know". Only Yes/No are used as the binary classification label of this project.

Counts (and proportions) of the binary target variable classes are as follows:

- Yes, has had kidney stones: 866 instances (9.4%)
- No, has not had kidney stones: 8342 instances (90.6%)

The key features are broadly described in the following:

- **Demographic**: gender, age, race, education, marital status, and income.
- **Dietary**: vitamin, water, nutrient, and dietary supplement intake. Everyday foods in the NHANES dietary interviews are deconstructed and aggregated into their nutritional components.
- **Examination**: body mass index (BMI), blood pressure, and pulse readings.
- **Laboratory**: aspects of biochemistry profile, and urine-associated tests.
- **Questionnaire**: past medical history (conditions and medicines), dietary and alcohol habits, urinary tract function, physical activity, smoking, and sleep habits.

Feature type ranges from numerical continuous and discrete to categorical binary, nominal, and ordinal. Dietary, examination, and laboratory data are mainly numerical, while demographic and questionnaire data

are mainly categorical. To avoid difficult or complicated natural languange processing or text mining, free-text data was not selected.

Counts of feature types and brief examples are as follows:

- **97 numerical features**, e.g. energy in kilocalories (continuous); age in years (discrete)
- **49 categorical features**, e.g. gender (binary: male, female); race (nominal: Mexican American, other Hispanic, white, etc.)

2.2 Summary of the Data

27 features have no missing values (not including the unique identifier and target variable columns).

Features that do have missing data can be summarised as follows:

- 98 features have under 25% missing data;
- 5 features have 25 - 50% missing data;
- 7 features have 50 - 75% missing data;
- 6 features have 75% - 100% missing data.

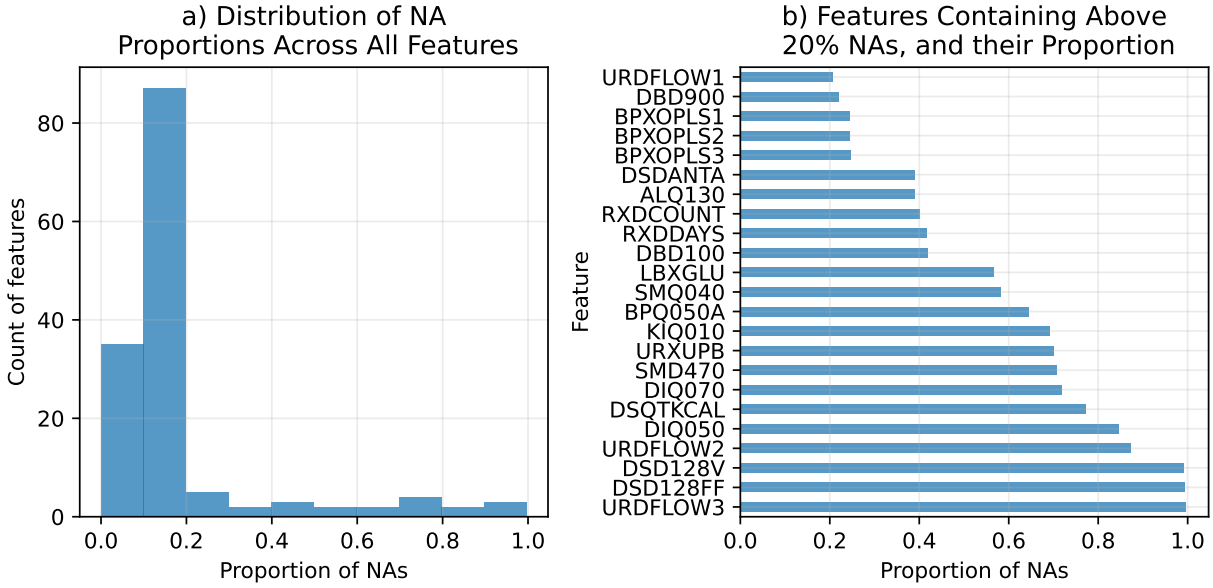

Figure 1: a) Count Of Features With Various Missing Data Proportions. b) Missing Data Proportions Of The Features Containing Above 20% Missing Data.

Overall, the majority of features do not contain a substantial proportion of missing data (Figure 1a).

Over half of features have less than 20% missing data (Figure 1a). Features from the table "Dietary Interview - Total Nutrient Intakes" (P_DR1TOT) are the main contributors to this particular proportion. A large number of features were selected from that table, and data collected within pertains to a consistent subset of people.

3. Ethics, Privacy, and Security

3.1 Ethical Considerations

- Epistemic – evidence related ethical issues. This includes bias within a dataset, opacity in decision making, and inaccurate decisions made from inconclusive evidence.
- Normative – learning and transforming processes of algorithms which may result in discriminatory outcomes and/or outcomes where the reasoning is unclear and therefore difficult for a patient to understand, or refute.
- Traceability – the ability to trace mistakes, and identify responsibility for mistakes. Additionally, that information about the algorithm be accessible and comprehensible (Mittelstadt et al., 2016)

Ethical issues in healthcare machine learning fall into three categories: epistemic (evidence-related), normative (fairness and transparency), and traceability (accountability and comprehensibility). These are interrelated; for instance, dataset bias (epistemic) can lead to discriminatory outcomes (normative). In kidney stone prediction, while not immediately life-threatening, accuracy remains crucial for informed consent and appropriate preventative measures. The algorithm should be traceable, with efforts made to reduce dataset bias and ensure clear reasoning behind risk assessments.

3.2 Privacy Concerns

Personal health data privacy is a major concern, governed by extensive legislation. As we're using NHANES data from a U.S. government agency, we must consider U.S., local, and international laws, as well as social norms. Our moral duty extends beyond data collection to its use. Four U.S. federal laws apply to NHANES data privacy:

1. The Privacy Act of 1974
2. The Confidential Information Protection and Statistical Efficiency Act
3. The Cybersecurity Enhancement Act
4. Section 308(d) of the Public Health Service Act

These laws ensure individual consent, data use restrictions, non-identifiability, and cybersecurity measures. NHANES complies with these acts, and our statistical use aligns with requirements. U.S. states have varying data privacy laws, which are evolving. While NHANES collects geographic information, its use is restricted. The public NHANES data contains no identifiable information, with special permissions required for accessing certain identifiers, ensuring patient confidentiality.

3.3 Security Measures

To comply with NCHS requirements, we've implemented security measures for the NHANES dataset. Data is stored in access-controlled repositories with separate directories for raw and processed data. Access is restricted to authorized team members based on specific tasks. We use encrypted data transfers, maintain audit trails, and employ version control. Two-factor authentication and regular access reviews are enforced. Only essential project-related data is stored and processed. Upon project completion, all data will be securely deleted from all storage locations, ensuring NCHS compliance throughout the project lifecycle.

4. Exploratory Data Analysis

4.1 Demographic Analysis

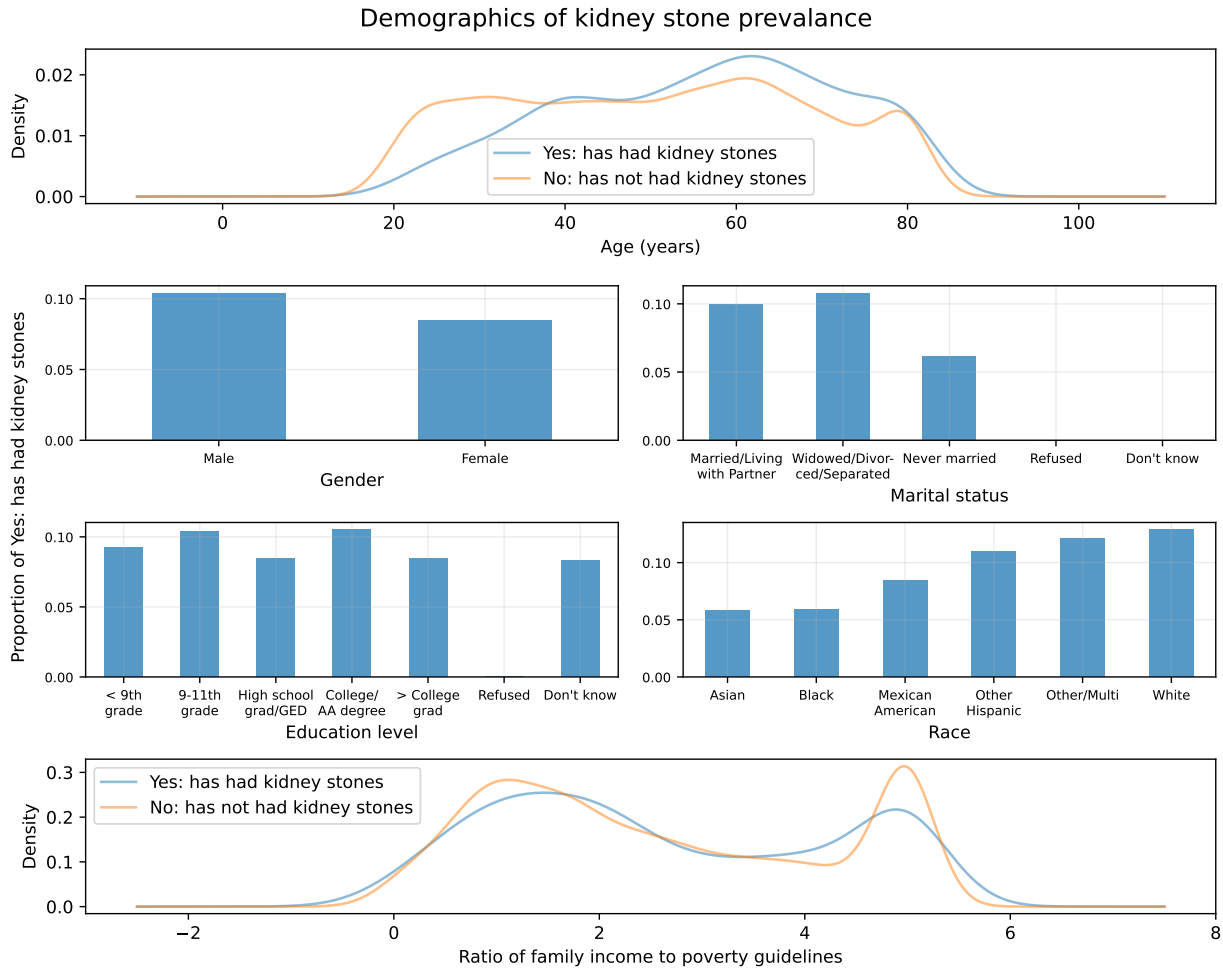

Figure 2: Demographic Factors and Kidney Stone Prevalence

Age demonstrates a significant correlation with kidney stone occurrence. The prevalence increases steadily from ages 20-40, plateaus briefly, then peaks around age 60. Notably, individuals over 50 show a higher likelihood of having experienced kidney stones compared to younger cohorts. Gender plays a significant role in kidney stone prevalence, with males showing a slightly higher rate (10%) compared to females (8%). We have also noticed Racial Disparities in kidney stone prevalence are evident. Asian and Black populations exhibit the lowest rates (just above 5%), while White and other/multiracial groups show the highest (close to 15%).

Education level appears to have minimal impact on kidney stone prevalence. The difference between the highest prevalence (college/AA degree at ~10%) and lowest (high school grad/GED at ~8%) is relatively small. The ratio of family income to poverty guidelines suggests an inverse relationship with kidney stone prevalence. Higher ratios (>4), indicating income well above poverty thresholds, correspond with lower kidney stone occurrence.

Age, gender , and income-to-poverty ratio all show associations with kidney stone prevalence. While education level appears less influential, it may still provide contextual information. These demographic factors could prove valuable in predictive models for kidney stone risk.

4.2 Health Conditions Analysis

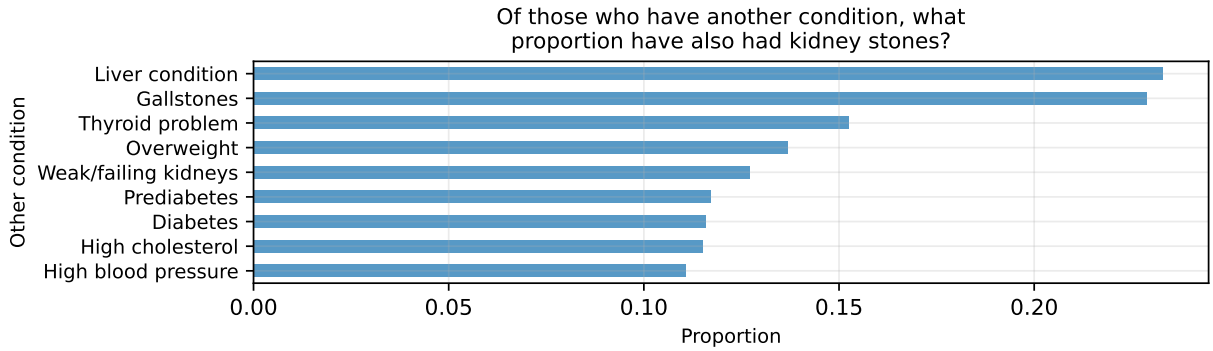

Figure 3: The Proportion Of People With Various Other Conditions Who Also Have Had Kidney Stones.

Gallstones and weak/failing kidneys show the strongest association with kidney stone occurrence, with about 25% of individuals having these conditions also experiencing kidney stones. Other health conditions exhibit a 11-15% co-occurrence rate with kidney stones, all exceeding the overall prevalence of 9.4%. These elevated rates suggest these features could be valuable for predictive modeling.

4.3 Dietary Analysis

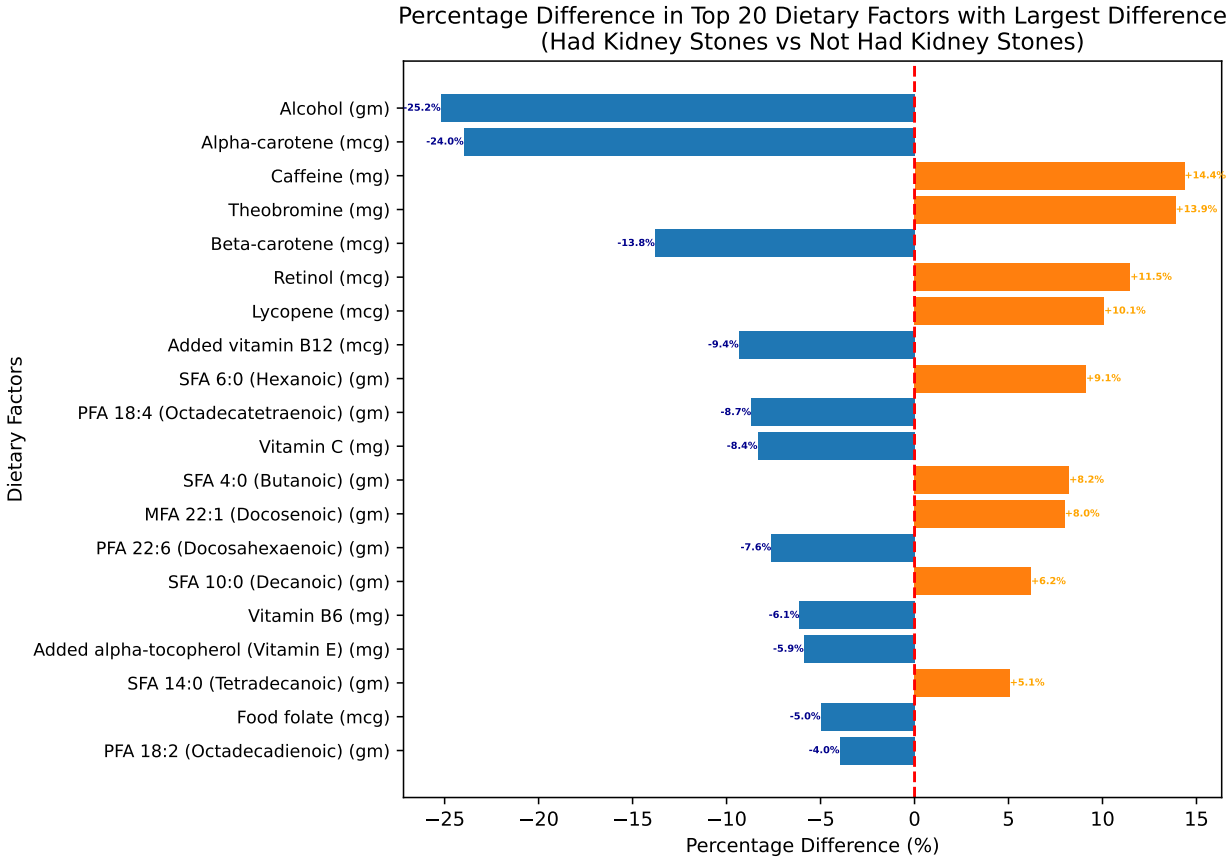

Figure 5: Dietary Differences.

Alpha-carotene and beta-carotene (both forms of vitamin A) display substantial negative differences (-24.0% and -13.8% respectively), with those who have had kidney stones consuming less. In contrast, other form of vitamin A, Retinol intake is 11.5% higher in the kidney stone group, contrasting with the carotenoid findings. Some research has indicated that excessive vitamin A intake may increase kidney stone risk (Tang et al., 2012), which could explain the lower carotenoid but higher retinol intake in those with a history of stones. Caffeine and theobromine consumption is notably higher in those with a history of kidney stones (14.4% and 13.9% more, respectively). While caffeine has been associated with increased risk of kidney stones in some studies (Ferraro et al., 2014), the higher intake in those with a history of stones could reflect changes in fluid consumption patterns post-diagnosis.

Vitamin C intake is 8.4% lower in individuals who have had kidney stones. This aligns with studies suggesting that high-dose vitamin C supplementation may increase kidney stone risk (Thomas et al., 2013), potentially leading to reduced intake in those with a history of stones. Among the top factors, we see a trend in vitamins and antioxidants, particularly forms of vitamin A, vitamin C, and vitamin E (alpha-tocopherol).

4.4 Laboratory Analysis

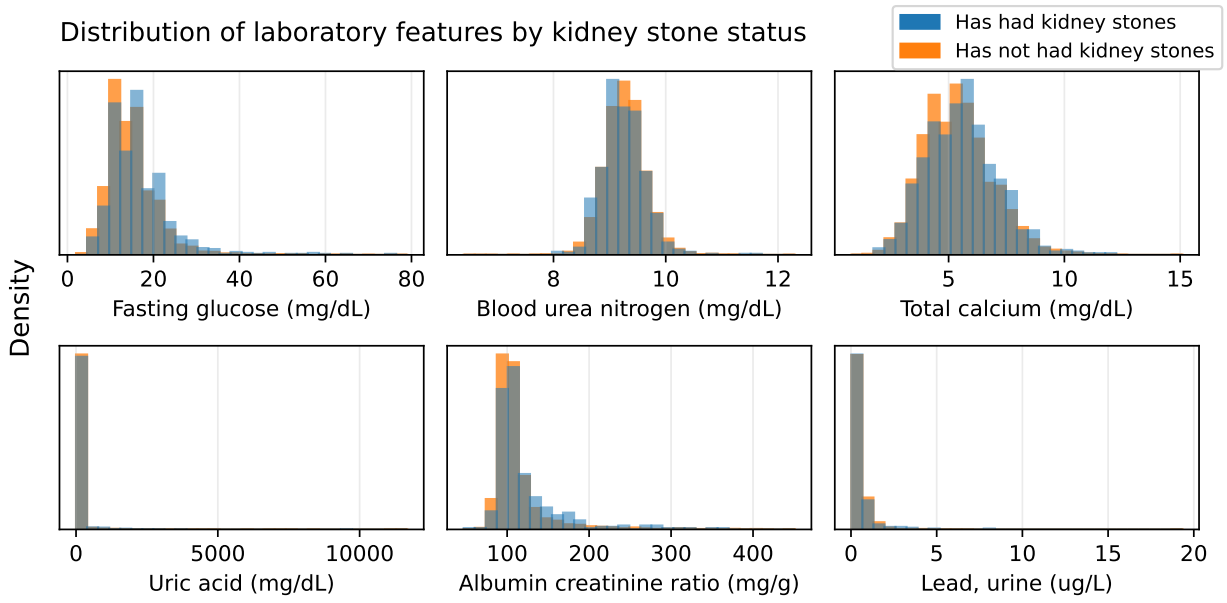

Figure 4: Density distributions of laboratory features, split by those who have had kidney stones and have not had kidney stones. a) Fasting glucose. b) Blood urea nitrogen. c) Total calcium. d) Uric acid. e) Albumin creatinine ratio. f) Lead in urine

Figure 4 shows that most laboratory features appear within expected ranges, with the exception of uric acid (Figure 4d). There appears to be outlier(s) skewing this feature with up to 10000 mg/dL uric acid, which is likely to be an error as ordinary uric acid levels should not exceed the single-digit mg/dL range.

Distribution shape of laboratory features remains relatively identical, regardless of kidney stone status. Distributions for fasting glucose (Figure 4a) and total calcium (Figure 4c) are shifted slightly right (towards higher values) for those who have had kidney stones. The peak bin for blood urea nitrogen (Figure 4b) is at a marginally lower value for those who have had kidney stones in comparison to those who have not. Albumin creatinine ratio appears to peak later, and remain slightly higher, at increasing mg/g for those who have had kidney stones (Figure 4e). Distributions for lead (Figure 4f) and uric acid are consistent for both kidney stone statuses - however, detail in the uric acid histogram may be obscured by the outlier(s). Therefore this indicates that uric acid and lead are not associated with kidney stone prevalence.

5. Detailed Analysis Results

5.1 Modelling Methodology Overview

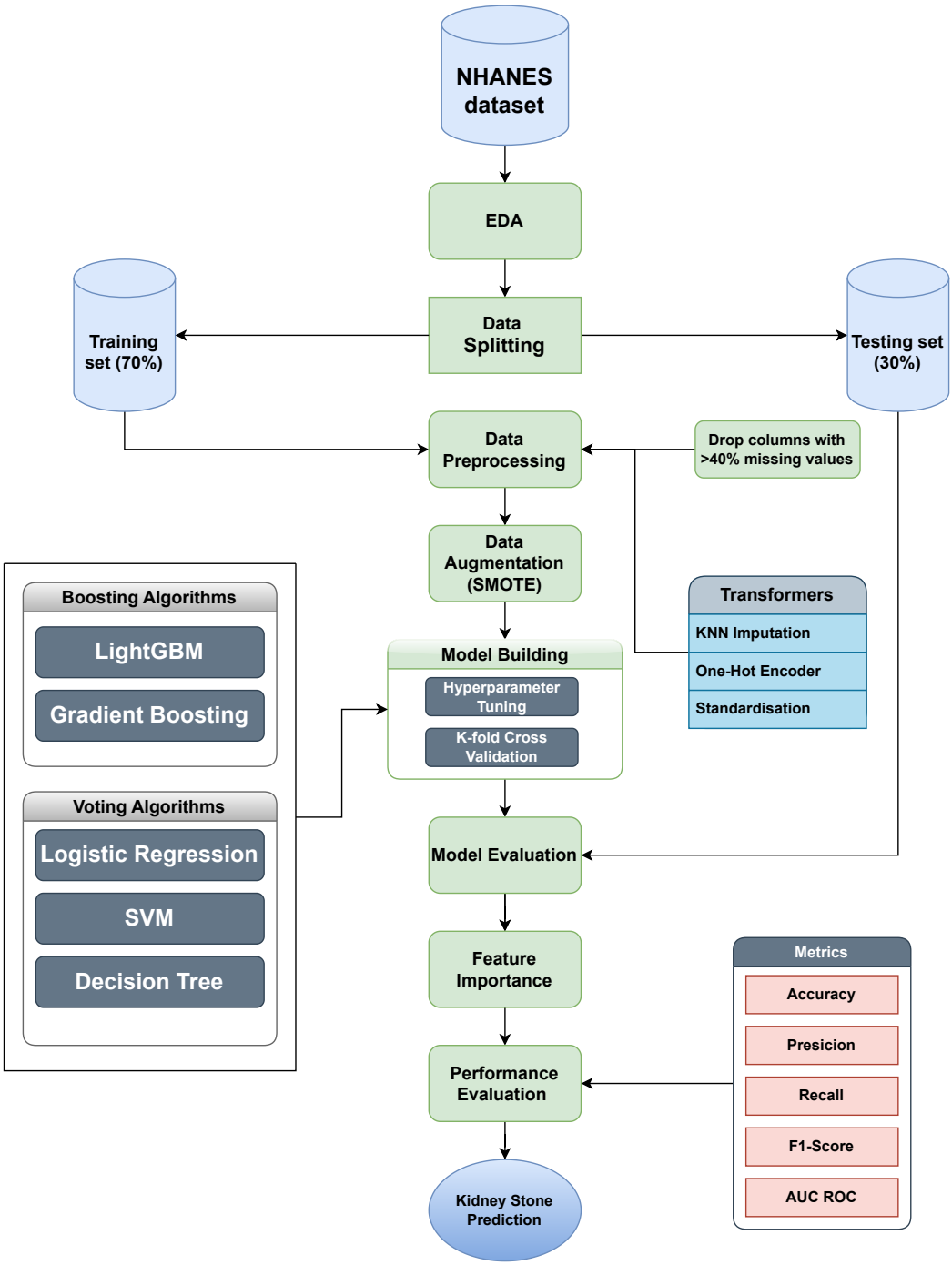

Figure 6: Overview Of The Kidney Stones Disease Prediction Methodology.

5.2 Data Preprocessing and Data Augmentation

5.2.1 Data Preprocessing Techniques

Our data preprocessing pipeline involved several key steps to prepare the dataset for model training. We began by splitting the dataset into training and testing sets using a 70-30 ratio, ensuring sufficient data for both model training and unbiased evaluation. Feature selection was performed by removing features with more than 40% missing values, reducing the number of features from 145 to 128. For the remaining features, we employed K-Nearest Neighbors (KNN) imputation with k=5 neighbors to handle missing values, potentially preserving important data relationships. Numeric features were then standardized using Standard Scaler to have zero mean and unit variance, while categorical variables were encoded using one-hot encoding with the 'ignore' option for handling unknown categories during model inference. These steps collectively focused our analysis on the most complete and potentially informative variables while preparing the data for effective model training.

5.2.2 Data Augmentation

In our initial training set, we observed a significant class imbalance, with 6,445 total records comprising 5,818 patients without kidney stones (90.3%) and only 627 patients with kidney stones (9.7%). This imbalance posed a potential challenge, as models could be biased towards the majority class, potentially leading to poor predictive performance for the minority class.

To mitigate this issue, we employed the **Synthetic Minority Over-sampling Technique (SMOTE)**. SMOTE works by creating synthetic examples of the minority class, effectively increasing its representation in the dataset without simply duplicating existing instances. Post-SMOTE, our training set achieved balance with 5,818 records in each class, resulting in a total of 11,636 records. With equal representation of both classes, our models can learn patterns associated with kidney stone presence more effectively.

For a visual representation of the data distribution before and after SMOTE balancing, please refer to Appendix C.

5.3 Model Development

5.3 Model Selection Evaluation

We chose these models for their complementary strengths:

Gradient Boosting is know for effective for handling complex interactions and non-linear relationships in medical data. LightGBM is chosen for its efficiency with large datasets and ability to handle categorical features directly. Voting Classifier, an ensemble model that combines Logistic Regression, Decision Tree, and Support Vector Machine (SVM) classifiers. This approach leverages the strengths of different algorithms to make predictions.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
GradientBoosting	0.9102	0.8523	0.9102	0.8726	0.6313
LightGBM	0.9121	0.8565	0.9121	0.8729	0.6282
VotingClassifier	0.8625	0.8494	0.8625	0.8558	0.5858

Table 1: Performance Comparison Between Classifiers.

The performance metrics of our three classifiers - Gradient Boosting, LightGBM, and Voting Classifier - are presented in Table 1, with detailed confusion matrices available in Appendix A. All models exhibit high accuracy (>0.86), suggesting strong predictive capability for kidney stone occurrence. However, given

the imbalanced nature of our dataset, accuracy alone is insufficient for comprehensive evaluation. The ROC AUC scores, ranging from 0.5858 to 0.6313, are lower than expected, indicating potential difficulties in distinguishing more nuanced cases despite good overall classification. LightGBM marginally outperforms the other models across most metrics, suggesting its effectiveness in capturing underlying patterns in our dataset. Gradient Boosting shows comparable performance, with only slight differences in ROC AUC. These results highlight the models' strengths in general classification while also revealing areas for potential improvement in handling more challenging instances.

5.4 Interpretation of Model Results

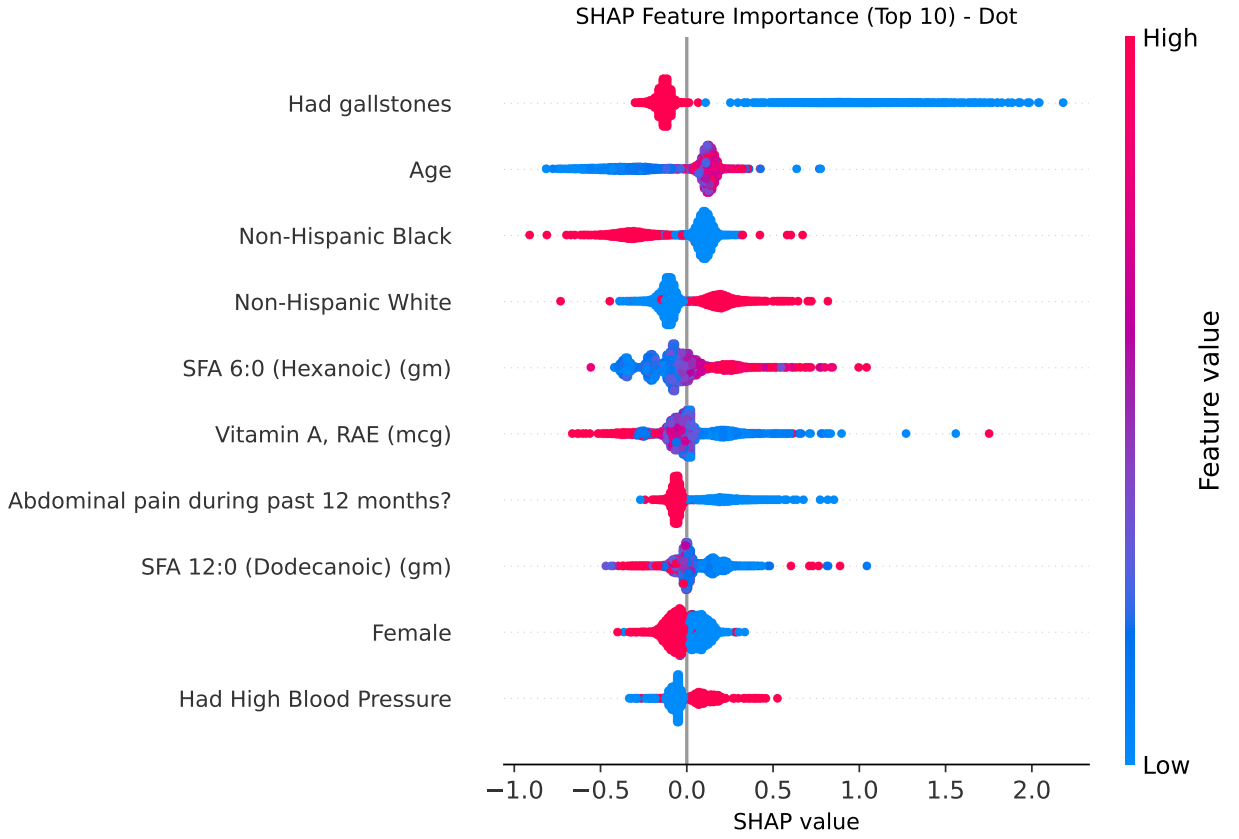

Figure 7: Top 10 Predictors of Kidney Stone Risk: Impact and Distribution (SHAP Summary Plot)

The beeswarm plot illustrates SHAP values for each feature, showing their impact on model predictions. Gallstones and age generally increase kidney stone likelihood, with age showing a positive correlation. Non-Hispanic Black ethnicity indicates reduced risk, while Non-Hispanic White suggests increased risk. Dietary factors show varied impacts: short-chain saturated fatty acids and vitamin A have wide-ranging effects, while very long-chain polyunsaturated fatty acids suggest a potential protective effect. Abdominal pain is a significant predictor, and the albumin creatinine ratio shows variable impact across patients.

To understand how these factors interact in individual cases, we'll examine a specific instance demonstrating the complex interplay between different risk factors.

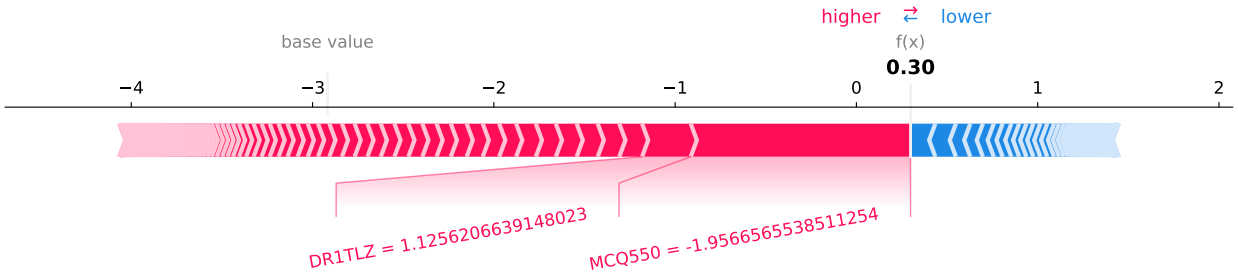

Figure 8: Individual Risk Factor Analysis for Kidney Stone Formation (SHAP Force Plot)

In this case, the model's final prediction f(x) is 0.30, which is greater than the base value. This positive difference indicates that for this specific patient, the combined effect of the variables increases the predicted risk above the average model output.

The plot highlights two key variables, MCQ550 (history of gallstones) shows a substantial negative impact of -1.9567, suggesting that for this individual, a history of gallstones significantly lowers the predicted risk of kidney stones. On the other hands, DR1TLZ (Lutein + Zeaxanthin intake in mcg) demonstrates a positive influence of 1.1256, partially offsetting the negative effect of gallstones.

The interplay between these factors results in a net positive effect, pushing the final prediction above the base value. This example illustrates how factors that may have a general trend in the population (such as gallstones typically increasing risk) can have different effects in individual cases, and how dietary factors like Lutein and Zeaxanthin intake can modulate the overall risk.

5.5 Estimates Of Risk Behind The Models

The NHANES dataset, while comprehensive, may not be fully representative of all demographics. Certain populations might be under-represented, potentially skewing our predictions (i.e Non-Hispanic in SHAP feature importance). In addition, some predictors, particularly dietary intake measures, rely on self-reporting, which can be subject to recall bias. The dataset is also cross-sectional, capturing a snapshot in time. This limits our ability to infer causal relationships or account for how risk factors might change over time.

While we've used ensemble methods to reduce potential bias, our models may still have inherent biases. For example, the high accuracy but lower ROC AUC scores suggest our models might be overfitting to certain patterns in the majority class. In addition, SMOTE helped address class imbalance, it generates synthetic examples but may not represent real-world data distributions.

6. Conclusions, Recommendation and Limitations

6.1 Conclusions

Our kidney stone prediction model successfully identified several key factors associated with increased risk of kidney stone formation. These include age, race (particularly Non-Hispanic White), presence of certain health conditions (high blood pressure, gallstones), body mass index, and specific biochemical markers (albumin-creatinine ratio).

Dietary factors, including intake of certain fatty acids (e.g., SFA 6:0 Hexanoic) and vitamins (particularly Vitamin A), play a complex role in kidney stone risk, with their effects not always being straightforward.

6.2 Recommendations

Patient at risks should be more educated about the importance of modifiable risk factors, particularly hydration and dietary choices. The protective effect of adequate fluid intake should be a key message. Healthcare systems should also consider implementing screening programs for kidney stone risk, especially for individuals with identified high-risk factors such as high blood pressure or gallstone.

6.3 Limitations

The model is based on cross-sectional data from the NHANES survey, which limits our ability to establish track risk factors over time. While our model incorporates a wide range of factors, it may not capture all possible influences on kidney stone formation, such as genetic predisposition or certain environmental factors. The model's performance may vary across different populations not well-represented in the NHANES dataset.

6. References

Abufaraj, M., Xu, T., Cao, C., Waldhoer, T., Seitz, C., D'andrea, D., Siyam, A., Tarawneh, R., Fajkovic, H., Schernhammer, E., Yang, L., & Shariat, S. F. (2020). Prevalence and Trends in Kidney Stone Among Adults in the USA: Analyses of National Health and Nutrition Examination Survey 2007–2018 Data. European Urology Focus, 7(6). https://doi.org/10.1016/j.euf.2020.08.011

Bayne, D. B., Usawachintachit, M., Armas-Phan, M., Tzou, D. T., Wiener, S., Brown, T. T., Stoller, M., & Chi, T. L. (2019). Influence of Socioeconomic Factors on Stone Burden at Presentation to Tertiary Referral Center: Data From the Registry for Stones of the Kidney and Ureter. Urology, 131, 57–63. https://doi.org/10.1016/j.urology.2019.05.009

Cappuccio, F. P., Strazzullo, P., & Mancini, M. (1990). Kidney stones and hypertension: population based study of an independent clinical association. BMJ, 300(6734), 1234–1236. https://doi.org/10.1136/bmj.300.6734.1234

Chen, Y., Lee, J., Shen, J.-T., Wu, Y., Tsao, Y.-H., Jhan, J., Wang, H.-S., Lee, Y., Huang, S.-P., Chen, S.-C., & Geng, J.-H. (2023). The impact of secondhand smoke on the development of kidney stone disease is not inferior to that of smoking: a longitudinal cohort study. BMC Public Health, 23(1). https://doi.org/10.1186/s12889-023-16116-6

Chewcharat, A., & Curhan, G. (2020). Trends in the prevalence of kidney stones in the United States from 2007 to 2016. Urolithiasis, 49, 27–39. https://doi.org/10.1007/s00240-020-01210-w

Chewcharat, A., Thongprayoon, C., Vaughan, L. E., Mehta, R. A., Schulte, P. J., O'Connor, H. M., Lieske, J. C., Taylor, E. N., & Rule, A. D. (2022). Dietary Risk Factors for Incident and Recurrent Symptomatic Kidney Stones. Mayo Clinic Proceedings, 97(8), 1437–1448. https://doi.org/10.1016/j.mayocp.2022.04.016

Coenen, P., Huysmans, M. A., Holtermann, A., Krause, N., van Mechelen, W., Straker, L. M., & van der Beek, A. J. (2018). Do highly physically active workers die early? A systematic review with meta-analysis of data from 193 696 participants. British Journal of Sports Medicine, 52(20), 1320–1326. https://doi.org/10.1136/bjsports-2017-098540

Curhan, G. C. (2007). Epidemiology of Stone Disease. Urologic Clinics of North America, 34(3), 287–293. https://doi.org/10.1016/j.ucl.2007.04.003

Federal Chief Information Officers. (2002). Confidential Information Protection and Statistical Efficiency Act. Federal Chief Information Officers; U.S. Federal Government. https://www.cio.gov/handbook/it-laws/cipsea/

Ferraro, P. M., Curhan, G. C., Sorensen, M. D., Gambaro, G., & Taylor, E. N. (2015). Physical Activity, Energy Intake and the Risk of Incident Kidney Stones. Journal of Urology, 193(3), 864–868. https://doi.org/10.1016/j.juro.2014.09.010

Ferraro, P. M., Taylor, E. N., Gambaro, G., & Curhan, G. C. (2013). Soda and Other Beverages and the Risk of Kidney Stones. Clinical Journal of the American Society of Nephrology, 8(8), 1389–1395. https://doi.org/10.2215/cjn.11661112

Ferraro, P. M., Taylor, E. N., Gambaro, G., & Curhan, G. C. (2014). Caffeine intake and the risk of kidney stones. The American Journal of Clinical Nutrition, 100(6), 1596–1603. https://doi.org/10.3945/ajcn.114.089987

Hellerstein, J. (2008). Quantitative Data Cleaning for Large Databases. https://dsf.berkeley.edu/jmh/papers/cleaning-unece.pdf

Jeong, I. G., Kang, T., Bang, J. K., Park, J., Kim, W., Hwang, S. S., Kim, H. K., & Park, H. K. (2011). Association Between Metabolic Syndrome and the Presence of Kidney Stones in a Screened Population. American Journal of Kidney Diseases, 58(3), 383–388. https://doi.org/10.1053/j.ajkd.2011.03.021

Li, C. - H., Sung, F. - C., Wang, Y. - C., Lin, D., & Kao, C. - H. (2014). Gallstones increase the risk of developing renal stones: a nationwide population-based retrospective cohort study. QJM: An International Journal of Medicine, 107(6), 451–457. https://doi.org/10.1093/qjmed/hcu017

Library of Congress. (2014, December 18). S.1353 - 113th Congress (2013-2014): Cybersecurity Enhancement Act of 2014. Library of Congress. https://www.congress.gov/bill/113th-congress/senate-bill/1353

Lieske, J. C. (2013). New Insights Regarding the Interrelationship of Obesity, Diet, Physical Activity, and Kidney Stones. Journal of the American Society of Nephrology, 25(2), 211–212. https://doi.org/10.1681/asn.2013111189

Lieske, J. C., Rule, A. D., Krambeck, A. E., Williams, J. C., Bergstralh, E. J., Mehta, R. A., & Moyer, T. P. (2014). Stone composition as a function of age and sex. Clinical Journal of the American Society of Nephrology: CJASN, 9(12), 2141–2146. https://doi.org/10.2215/CJN.05660614

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The Ethics of algorithms: Mapping the Debate. Big Data & Society, 3(2), 1–21.Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. Social Science & Medicine, 260(113172). https://doi.org/10.1016/j.socscimed.2020.113172

National Center for Health Statistics. (2013). Vital and Health Statistics Report Series 1, Number 56 August 2013. National Center for Health Statistics. https://www.cdc.gov/nchs/data/series/sr_01/sr01_056.pdf

National Center for Health Statistics. (2021, July 22). NHANES - Your Privacy Matters. National Center for Health Statistics; Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/nhanes/participant/participant-confidentiality.htm

National Institutes of Health. (n.d.). Kidney Stones. National Institute of Diabetes and Digestive and Kidney Diseases. https://www.niddk.nih.gov/health-information/urologic-diseases/kidney-stones?dkrd=hispt0421

NHS. (2018, October 3). Kidney stones - Causes. NHS. https://www.nhs.uk/conditions/kidney-stones/causes/

Office of the Legislative Counsel. (2020). Public Health Service Act. Office of the Legislative Counsel. https://www.govinfo.gov/content/pkg/COMPS-8773/pdf/COMPS-8773.pdf

Prezioso, D., Strazzullo, P., Lotti, T., Bianchi, G., Borghi, L., Caione, P., Carini, M., Caudarella, R., Gambaro, G., Gelosa, M., Guttilla, A., Illiano, E., Martino, M., Meschi, T., Messa, P., Miano, R., Napodano, G., Nouvenne, A., Rendina, D., & Rocco, F. (2015). Dietary treatment of urinary risk factors for renal stone formation. A review of CLU Working Group. Archivio Italiano Di Urologia E Andrologia, 87(2), 105. https://doi.org/10.4081/aiua.2015.2.105

Sorensen, M. D., Chi, T., Shara, N. M., Wang, H., Hsi, R. S., Orchard, T., Kahn, A. J., Jackson, R. D., Miller, J., Reiner, A. P., & Stoller, M. L. (2014). Activity, energy intake, obesity, and the risk of

incident kidney stones in postmenopausal women: a report from the Women's Health Initiative. Journal of the American Society of Nephrology: JASN, 25(2), 362–369. https://doi.org/10.1681/ASN.2013050548

Sorensen, M. D., Hsi, R. S., Chi, T., Shara, N., Wactawski-Wende, J., Kahn, A. J., Wang, H., Hou, L., & Stoller, M. L. (2014). Dietary Intake of Fiber, Fruit and Vegetables Decreases the Risk of Incident Kidney Stones in Women: A Women's Health Initiative Report. Journal of Urology, 192(6), 1694–1699. https://doi.org/10.1016/j.juro.2014.05.086

Soueidan, M., Bartlett, S. J., Noureldin, Y. A., Andersen, R. E., & Andonian, S. (2015). Leisure time physical activity, smoking and risk of recent symptomatic urolithiasis: Survey of stone clinic patients. Canadian Urological Association Journal, 9(7-8), 257. https://doi.org/10.5489/cuaj.2879

Spatola, L., Angelini, C., Badalamenti, S., Maringhini, S., & Gambaro, G. (2016). Kidney stones diseases and glycaemic statuses: focus on the latest clinical evidences. Urolithiasis, 45(5), 457–460. https://doi.org/10.1007/s00240-016-0956-8

Sun, Y., Zhou, Q., & Zheng, J. (2019). Nephrotoxic metals of cadmium, lead, mercury and arsenic and the odds of kidney stones in adults: An exposure-response analysis of NHANES 2007–2016. Environment International, 132, 105115. https://doi.org/10.1016/j.envint.2019.105115

Tang, J., McFann, K. K., & Chonchol, M. B. (2012). Association between serum 25-hydroxyvitamin D and nephrolithiasis: the National Health and Nutrition Examination Survey III, 1988-94. Nephrology Dialysis Transplantation, 27(12), 4385–4389. https://doi.org/10.1093/ndt/gfs297

Thomas, L. D. K., Elinder, C.-G., Tiselius, H.-G., Wolk, A., & Åkesson, A. (2013). Ascorbic Acid Supplements and Kidney Stone Incidence Among Men: A Prospective Study. JAMA Internal Medicine, 173(5), 386. https://doi.org/10.1001/jamainternmed.2013.2296

Trivedi, R. B., Ayotte, B., Edelman, D., & Bosworth, H. B. (2008). The association of emotional well-being and marital status with treatment adherence among patients with hypertension. Journal of Behavioral Medicine, 31(6), 489–497. https://doi.org/10.1007/s10865-008-9173-4

U.S. Department of Justice. (2014, June 16). Office of Privacy and Civil Liberties | Privacy Act of 1974. Office of Privacy and Civil Liberties; U.S Department of Justice. https://www.justice.gov/opcl/privacy-act-1974#:~:text=The%20Privacy%20Act%20prohibits%20the

University of Florida Health. (2019). Kidney stones. Department of Urology; University of Florida Health. https://ufhealth.org/conditions-and-treatments/kidney-stones

Wang, X., Sun, M., Wang, L., Li, J., Xie, Z., Guo, R., Wang, Y., & Li, B. (2023). The role of dietary inflammatory index and physical activity in depressive symptoms: Results from NHANES 2007–2016. Journal of Affective Disorders, 335, 332–339. https://doi.org/10.1016/j.jad.2023.05.012

Winitzki, D., Zacharias, H. U., Nadal, J., Baid-Agrawal, S., Schaeffner, E., Schmid, M., Busch, M., Bergmann, M. M., Schultheiss, U., Kotsis, F., Stockmann, H., Meiselbach, H., Wolf, G., Krane, V., Sommerer, C., Eckardt, K.-U., Schneider, M. P., Schlieper, G., Floege, J., & Saritas, T. (2022). Educational Attainment Is Associated With Kidney and Cardiovascular Outcomes in the German CKD (GCKD) Cohort. Kidney International Reports, 7(5). https://doi.org/10.1016/j.ekir.2022.02.001

Xue, W., Xue, Z., Liu, Y., Yin, P., Liu, L., Qu, S., Wu, S., & Yang, C. (2024). Is Kidney Stone Associated with Thyroid Disease? The United States National Health and Nutrition Examination Survey 2007–2018. Endocrine, Metabolic & Immune Disorders, 24(11). https://doi.org/10.2174/0118715303268738231129093935

Yencilek, E., Yencilek, F., Ozcan, C., Demirel, A., Coskun, S., & Basaran, M. (2010). The effect of lycopene on the kidney stones: a preliminary study. Urological Research, 38(4).

Singh, P., Harris, P. C., Sas, D. J., & Lieske, J. C. (2022). The genetics of kidney stone disease and nephrocalcinosis. Nature Reviews Nephrology, 18(4), 224-240.

Appendices

Appendix A: Detailed Model Performance Metrics

Table A1: Confusion Matrix comparison of Gradient Boosting, Light GBM and Voting Classifiers.

Model	True Positive	True Negative	False Positive	False Negative
Gradient Boosting	3	2512	12	236
LightGBM	2	2518	6	237
Voting Classifiers	26	2357	167	213

Appendix B: Performance Metrics Formulas

The evaluated model performance using a comprehensive set of metrics:

Accuracy: The proportion of correct predictions (both true positives and true negatives) among the total number of cases examined.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

TP : The number of samples correctly classified as had kidney stones

TN : The number of samples correctly classified as not had kidney stones

FP : The number of samples incorrectly classified as had kidney stones

FN : The number of samples incorrectly classified as not had kidney stones

Precision: The proportion of correct positive identifications.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (also known as Sensitivity): The proportion of actual positive cases that were correctly identified.

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN}$$

F1 Score: The harmonic mean of **Precision** and **Recall**, providing a single score that balances both measures.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

ROC AUC: A measure of the model's ability to distinguish between classes. It's calculated by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings.

$$\text{TPR} = \frac{TP}{TP + FN}$$

$$\text{FPR} = \frac{FP}{FP + TN}$$

The AUC is then calculated as the area under this ROC curve, with values ranging from 0 to 1. An AUC of 0.5 represents a model with no discriminative ability, while an AUC of 1.0 represents a perfect model.

Appendix C: Data Balancing Visualization