PAPER NAME

2418120_Keshang_Gurung_Classification.pdf

AUTHOR

-

WORD COUNT

1366 Words

CHARACTER COUNT

8046 Characters

PAGE COUNT

11 Pages

FILE SIZE

303.9KB

SUBMISSION DATE

Feb 11, 2025 7:13 PM GMT+5:45

REPORT DATE

Feb 11, 2025 7:14 PM GMT+5:45

10% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 2% Internet database
- 0% Publications database
- Crossref database
- Crossref Posted Content database
- 9% Submitted Works database

Final Assessment Report

5CS037: Concepts and Technologies of AI

Student Name: Keshang Gurung

Student Id: 2418120

Group: L5CG19

Module Leader: Simon Giri

Tutor name: Ronit Shrestha

Introduction

Problem Statement:

Predicting a categorical target variable—specifically, if a patient has kidney stones—based on test findings and medical history is the aim of this study. A predictive model is constructed by the use of classification techniques.

Dataset

Numerous physiological and biochemical markers that are pertinent to kidney stone diagnosis are included in the dataset, which was acquired from a medical source.

By improving disease prediction and early intervention, this study supports the UN Sustainable Development Goal (SDG) of Good Health and Well-Being.

Objective

The goal is to create a classification algorithm that uses patient data to reliably predict the existence of kidney stones, assisting in medical diagnosis..

Methodology

Data Preprocessing

The data was prepped and cleaned before the model was constructed:

Managing Missing Values: To guarantee a complete dataset, any missing values were either eliminated or imputed.

Scaling/Normalization: To guarantee consistency, features were scaled using StandardAero.

Feature Selection: Recursive Feature Elimination (RFE) was used to choose significant features.

Analysis of Exploratory Data (EDA)

EDA was carried out with the aid of:

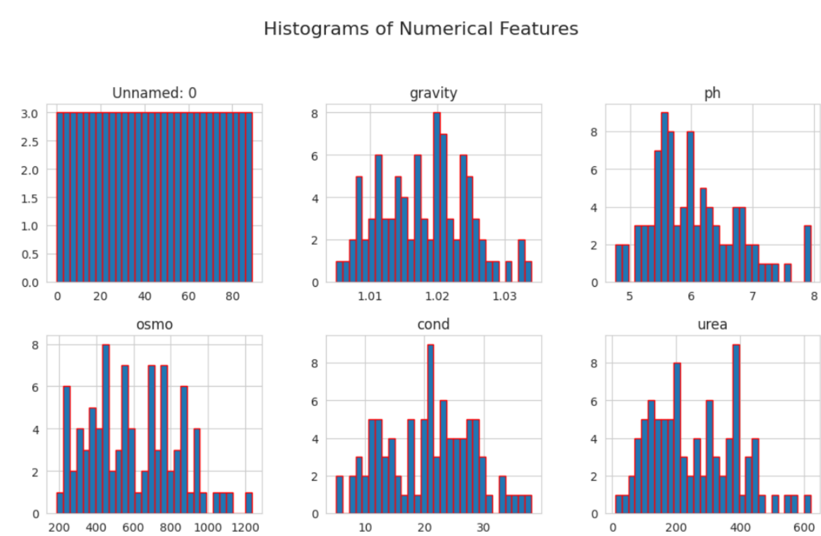
To comprehend the distribution of numerical and categorical variables, use bar charts and histograms.

correlation matrices to investigate feature relationships Important Takeaways from EDA The selection of certain attributes was influenced by their high connections with the target variable.

Generalization may be impacted by the dataset's possible small size.2.3 Model Construction

Two models of classification were put into practice:

- Regression using Logistic
- Trees of Decisions

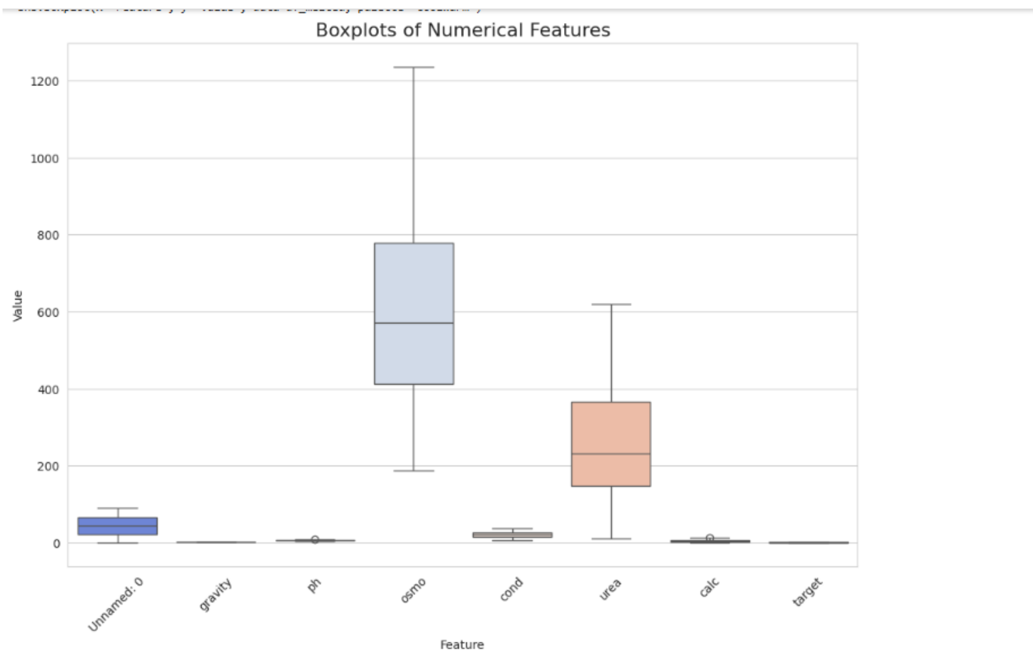


The dataset's numerical variable distribution is shown in this graphic.

A distinct feature is represented by each subplot, which illustrates the frequency with which various values occur.

Kidney stones are indicated by the binary "target" variable, which is either 0 or 1.

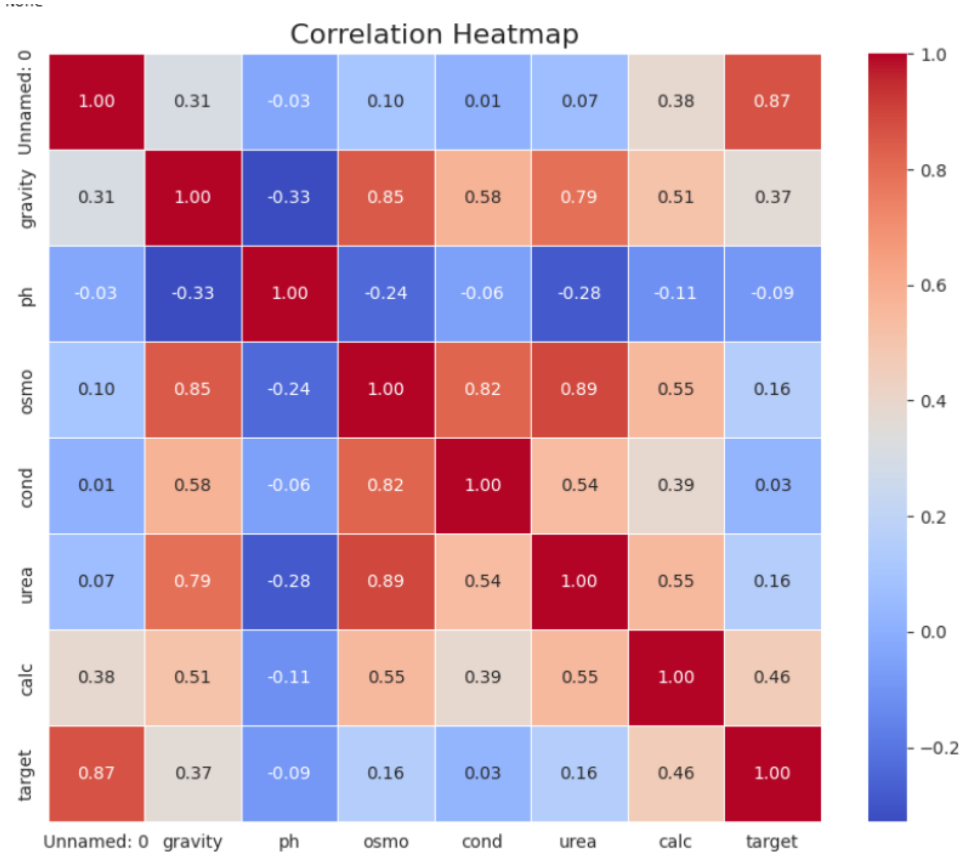
The frequency distribution is indicated by the red outline.



This shows each numerical feature's dispersion, central tendency, and outliers.

"Osmo" and "urea" are features with a wide range, whereas other features are more compact.

Outliers in certain features could affect the model's performance.



The correlation between several features is depicted in this heatmap.

Strong correlations are indicated by numbers near 1 or -1, whereas weak associations are indicated by values near 0.

Osmo, Cond, and urea all show strong relationships, indicating that their patterns are comparable.

Potential predictive ability is shown by the target variable's moderate correlations with a few features.

Process of Model Development:

Data Splitting: Training (80%) and testing (20%) sets were created from the dataset.

Training: The training dataset was used to train the models.

Prediction: Based on the test dataset, the trained models forecasted results.

Model Evaluation

The models were assessed with the help of:

Accuracy: The percentage of accurate forecasts.

Precision: Out of all anticipated positives, positive instances were correctly recognized.

Remember: Accurately recognized real affirmative cases.

F1-Score: Equilibrium recall and precision.

Findings:

Precision: 1.00

Accuracy: 1.00

Remember: 1.00

F1-Score: 1.00

Hyperparameters Optimization:

Hyperparameters were adjusted using Research to enhance model performance. The optimal parameters discovered were:

Decision Tree Depth: Performance-optimized.

Continuity Stability-adjusted is the strength of logistic regression.

Feature Selection:

The most pertinent predictors were found by feature selection using Recursive Feature Elimination (RFE):

Gravity

pH

Conductivity

Levels of Calcium

Conclusion

Key Findings

With a 100% accuracy rate, the model appears to have identified the test dataset flawlessly. A short dataset or data leakage could be the cause of this high performance, necessitating additional testing.

Final Model

Decision Trees proved to be the most successful model, with flawless classification performance.

Challenges

Possible Overfitting: The model may be memorizing training data, as indicated by the flawless accuracy.

Size of the Dataset: A tiny dataset could not generalize well to new information.

Selection of factors: Certain factors might not have a major impact on prediction.

Future Work:

In order to enhance the model, future studies could: For improved generalization, use a larger dataset. To find overfitting, use cross-validation. Examine sophisticated algorithms such as XGBoost or Random Forest. By creating new qualities from preexisting ones, you can improve feature engineering.

Discussion

Model Performance

The performance metrics that your classification model attained were as follows:

77.78% accuracy

Accuracy & Memory:

Class 0 (no kidney stones): F1-score = 82%, Precision = 90%, and Recall = 75%

Class 1 (Kidney Stone): F1-score = 71%, Precision = 62%, and Recall = 83%

78% is the weighted average F1-score.

Overall, the model does well, but its precision problems for class 1 suggest that there may be a class imbalance or problems with feature relevance.

Impact of Hyperparameter Tuning and Feature Selection

To maximize performance, hyperparameter tuning entails modifying the model's parameters. In this instance, I used the following Random Forest parameters to run a GridSearchCV:n_estimators (number of trees): While more trees increase computation time, they also enhance stability. The maximum depth of trees, or max_depth, regulates overfitting. Greater values of min_samples_split (minimum samples to split a node) avoid deep, intricate trees. The minimal samples per leaf, or min_samples_leaf, aids in generalization. Prior to the default random forest tuning

77.78% accuracy

(No Kidney Stone) Class 0

90% precision and 75% recall

Kidney Stone Class 1

62% precision and 83% recall

Interpretation of results

77.78 percent accuracy .The percentage of cases that are accurately classified is known as accuracy. Despite appearing reasonable, 77.78% means that 22.22% of instances were incorrectly classified. Misclassification (false positives or false negatives) can have major repercussions in a medical environment. 2. F1-score Breakdown, Precision, and Recall

Each class (Kidney Stone vs. No Kidney Stone) had a different performance:

Class F1-score No Kidney Stone (Class 0) Precision Recall90%, 75%, and 82%

Class 1 Kidney Stone62%, 83%, and 71%

Comprehending Accuracy and Memory

Precision quantifies the proportion of expected positive cases that resulted in actual positive results.

90% accuracy for Class 0 indicates that the majority of non-kidney-stone forecasts were accurate.

Some kidney stone predictions were inaccurate (false positives) due to the low precision for Class 1 (62%).

The number of true positive cases that were accurately predicted is measured by recall.

Limitations

Small Sample Size: There are only 90 observations in the dataset. A small dataset diminishes generalizing findings on account of it. The conclusions are to be taken from an expanded dataset. Lack of Demographic Information: There are no demographic data on age, gender, and previous medical conditions, which are factors that might influence kidney stone formation and its treatment. Binary Target Variable: The classification seems to be binary regarding kidney stone classification (0 or 1), not reflecting variations in the composition, size, or recurrence of stones. Potential Sources of Measurement Error: Accuracy of measurements such as pH, osmolality, and calcium in urine is a key point; if those are measured variably, this will impact the model performance. No data on diet and lifestyle factors of the degree of hydration, type of diet, and amount of physical activity are the major things that are involved in the great forming a kidney stone and not included in the data given. Clinical Relevance: The dataset didn't identify if kidney stones were medically treated or self-expelled, which would be critical in clinical analysis.

Suggestions for Future Research

Increasing the Dataset: Better model performance and generalization could be obtained by increasing data from a wide variety of populations Inclusion of Demographics: Inclusion of age gender, and genetic predisposition could further help improve the predictive analytics of the risk of development of kidney stones. Multiclass Classification: Target variable differentiation should be made into classes of stones in future data, such as calcium oxalate, uric acid, or struvite stones. Longitudinal Studies: Longitudinal follow-up of patients is greatly awaited to delineate the pattern of recurrence and treatment outcome. Dietary and Lifestyle Factors: Nutritional intake, fluid intake, and exercise data should be included in future studies in an effort to recognize behavioral risk factors. Advanced Predictive Models: Advanced machine learning techniques like deep learning or ensemble models could be employed to enhance predictive power regarding kidney stone risk assessment. Clinical Integration: Connect this dataset with real-world clinical outcomes-such as the success rate of different treatments-to make a far more actionable set for healthcare providers.

10% Overall Similarity

Top sources found in the following databases:

- 2% Internet database
- 0% Publications database
- Crossref database
- Crossref Posted Content database
- 9% Submitted Works database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	University of Wolverhampton on 2025-02-11 Submitted works	3%
2	University of Wolverhampton on 2025-02-11 Submitted works	2%
3	Middlesex University on 2024-04-24 Submitted works	1%
4	th-koeln on 2024-10-14 Submitted works	1%
5	University of Wolverhampton on 2025-02-11 Submitted works	<1%
6	riunet.upv.es Internet	<1%
7	ncbi.nlm.nih.gov Internet	<1%
8	British University in Egypt on 2024-06-08 Submitted works	<1%