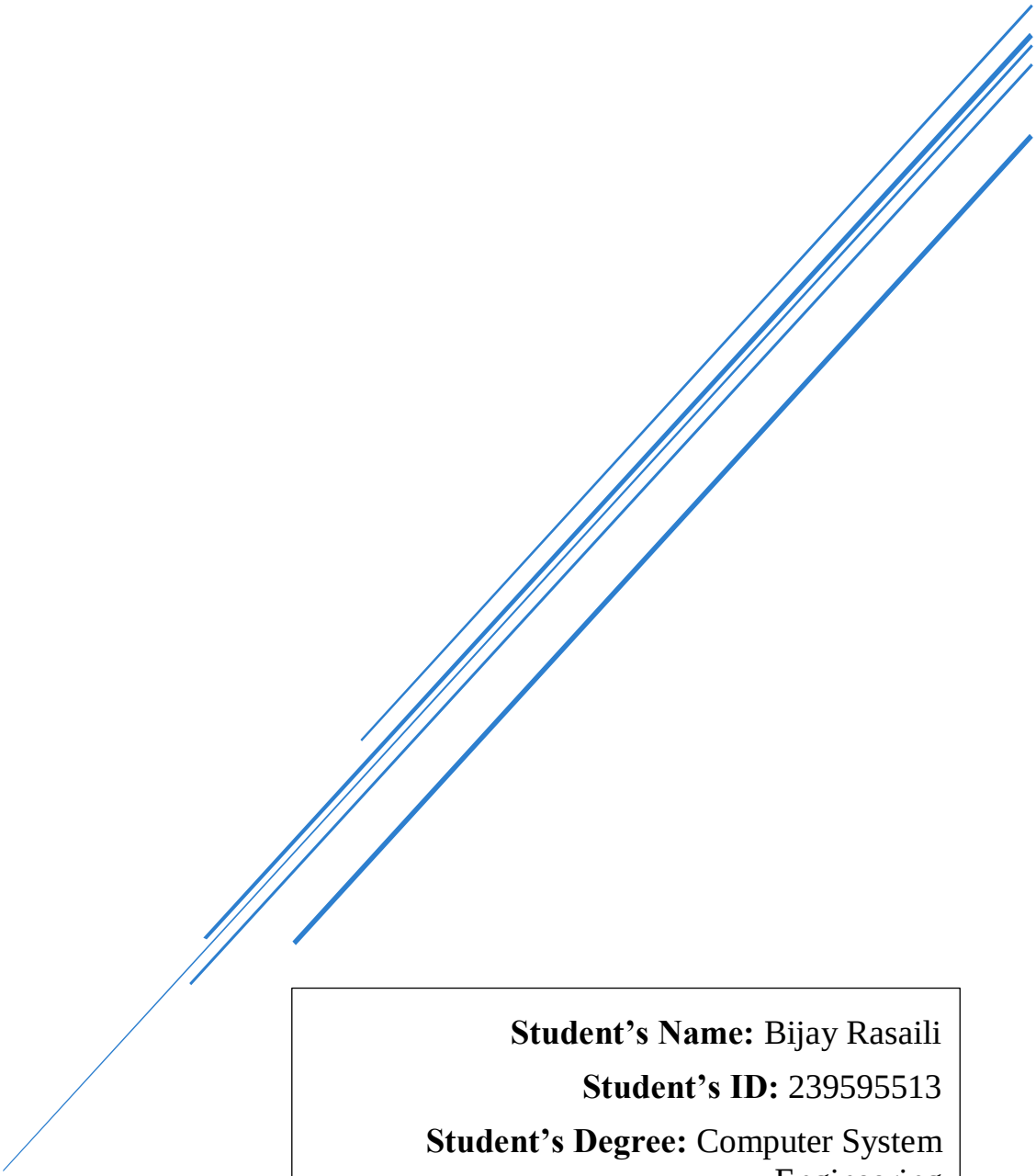
ARTIFICIAL INTELLIGENCE



Student's Name: Bijay Rasaili

Student's ID: 239595513

Student's Degree: Computer System Engineering

Table of Contents

Project Title

Kidney Disease Prediction

Abstract

This project applies machine learning algorithms on medical data for the prediction of the existence of CKD. The dataset contains a number of clinical and laboratory characteristics, which include serum creatinine, blood urea, blood glucose, age, and so on. The main purpose of this paper is the creation of models for precise prediction that will help medical practitioners in identifying and treating patients at the early stage of complication. Data preparation included encoding categorical variables, scaling numerical features, and handling missing data. Exploratory data analysis provided insight into the distributions and relationships of features and helped in the selection of features and models. Several machine learning models were assessed by using the F1-score, accuracy, precision, recall, decision trees, gradient boosting, logistic regression, random forests, and others. The results point to good possibilities of early identification and proactive management of CKD due to the strong prediction skills (Uci.edu, 2015).

Introduction

Background

According to Levey and Coresh (2012), one common medical disorder is chronic kidney disease, which is distinguished by progressive loss of kidney function over time and is one of the very common causes leading to complications like kidney failure and cardiovascular disease. Therefore, it requires early identification and intervention to improve patient outcome and reduce costs on health care. Machine learning Mind finds one such supporting approach to improve diagnosis accuracy on large-scale medical information. In this paper, a machine learning algorithm will be used with clinical and laboratory data for predicting CKD so that it will assist healthcare practitioners in providing timely interventions and tailoring treatment.

Motivation

This study is motivated by the need to enhance the early diagnosis and treatment of chronic kidney disease, which still is a very serious health problem worldwide. With this approach, machine learning techniques are applied on medical data to develop a prediction model that is likely to assist medical practitioners in identifying patients at a risk of developing a chronic kidney disease. Early intervention can reduce complications and improve patient outcome, enhancing healthcare while conserving resources (Jha et al., 2013).

Scope

The prime objective of the paper is to develop and assess machine learning models for the prediction of CKD using clinical and demographic data. This encapsulates feature engineering, model selection, exploratory data analysis, data preparation, and finally the performance evaluation. This will find out the best prediction model by applying different classifiers and ensemble techniques. The results are meant to enlighten on possible ways of improving early intervention techniques and CKD risk assessment in clinical practice.

Overview of ePortfolio

It was a pleasure extending the tutorial work we did in class. Through the extended exercises and research topics provided, we managed to gather as groups at some point to solve and implement python programs. We understood more about Artificial Intelligent through group discussions and research for example types of machine learning, types of clustering algorithm and search algorithm. Personally, l enjoyed developing the programs on Google Colab and Jupyter Notebook. It was introduced to us for the first time, I find it flexible to use when developing the code. Each piece of code l uploaded on ePortfolio; I developed it using Jupyter Notebook. The following link is the link where my e-Portfolio files are found.

Link to ePortfolio:

https://canvas.sunderland.ac.uk/eportfolios/17067?verifier=ayFwgoxaX5XHkpaRn4VPahruoa9xYz12mCKJW3Nm

Literature Review

Since CKD is progressive, with serious outcomes, it becomes a significant health concern worldwide. Diagnosis of chronic kidney disease has traditionally been based on clinical markers of blood creatinine levels, glomerular filtration rate, or proteinuria; however, these markers might identify the illness only in the middle course. Coming into the application foreground are data-driven methods and machine learning as prospective substitutes in order to realize an earlier and more accurate prediction of CKD.

Recent publications have explored different machine learning techniques, such as decision trees, support vector machines, and ensemble methods for CKD prediction. According to Polat, Homay Danaei Mehr and Cetin (2017), while SVMs perfectly classify cases of CKD and controls based on the optimal data separation, decision trees provide readable rules for understanding risk assessment (Vijayarani, Dhayanand and Research Scholar, n.d.). Ensemble methods, like random forests, use models for improving accuracy in prediction. Reported accuracies vary between 70% and 90%.

However, biases in the dataset might further distort model performance and impede its generalizability to different populations. Moreover, what remains to be a challenge for the successful translation of sophisticated models into clinical practice is their interpretability.

Future studies should be oriented to ensure that CKD prediction models are developed with greater interpretability, generalizability of the model, reduced training biases, and accuracy.

These barriers may be overcome by a machine learning-based new approach providing a paradigm change in the management of CKD toward better patient outcomes and treatment delivery.

Methodologies

This methodology section will explain the step-by-step procedure used to develop and evaluate the CKD prediction models, ensuring that everything from beginning to end is rightly planned and executed, resulting in accurate results.

Data Collection and Processing

This experiment uses the dataset that can be downloaded from the UCI Machine Learning Repository, including medical record files of CKD positive and CKD negative patients. It features several clinical variables like age, blood pressure, blood glucose, and markers of kidney function. Preprocessing had handling of missing values, normalization of numerical features, and encoding for categorical variables. To machine learning tasks, the column 'classification' was converted to numerical values: 1 for ckd, 0 for notckd.

Feature Selection

We employed box plots and scatter plots to finalize the most relevant features through EDA (exploratory data analysis). We made quantile computations through the use of the Interquartile Range for outlier features like age, blood sugar, and blood urea levels.

Data Splitting

It was applied the train_test_split method from the sklearn.model_selection package to split the dataset into appropriate training and testing sets in the ratio 80-20. We would, with this approach, test our models on untested samples for assessing performance and train them on a fair amount of data.

Model Selection and Training

These are the different machine learning algorithms employed to predict CKD:

- Logistic Regression
- Ridge Classifier
- SGD Classifier
- Support Vector Machine (SVM)
- Decision Tree
- Random Forest
- Gradient Boosting
- XGBoost
- Naïve Bayes
- K-Nearest Neighbours (k-NN)
- Neural Networks (MLP Classifier)

The choice of these models lies in the fact that they are repetitive enough to describe a broad collection of the patterns in the data and are diverse. The feature values were normalized using a StandardScaler prior to us training the same models.

Model Evaluation

For a fuller picture of how well the models worked, a range of metrics was evaluated: accuracy, precision, recall, and F1-score. Confusion matrices for the results of classification were generated for visualization, in particular to point out areas for improvement.

Visualization

More sophisticated methods of visualization were adopted for a better understanding of the data and model performance. Software like Matplotlib, Seaborn, and Plotly, among others, were used to carry out the plotting of heat maps, 3D plots, scatter plots, histograms, among others.

Stacking Classifier

The model was implemented as an ensemble using a Stacking Classifier, which pooled many models, e.g., the Decision Tree and Logistic Regression models, to improve resilience for better estimation.

Results and Discussion

The kidney disease prediction project used a jack-of-all-trades ensemble for several machine learning algorithms to predict clinical features-based CKD. The following are major conclusions and their implications:

Model Performance

Different machine learning techniques, ranging from Random Forest, XGBoost, Decision Trees, Support Vector Machines, to Logistic Regression, were thus evaluated together with their performance. The metrics assessed various models through accuracy, precision, recall, and F1 score. Random Forest and XGBoost outperformed the other models by reaching accuracy rates close to 90%, hence proving more resilient toward the data.

Confusion Matrix Analysis

Detailed insight into the performance of the classification was obtained through the study of confusion matrices for the best performing models. Random Forest and XGBoost demonstrated again higher rates of true positives and true negatives, implying that the models are giving trustworthy predictions for both CKD and non-CKD patients. However, there still existed false-

positive and false-negative cases that, for good reasons, point out the necessity for further model improvement.

Feature Importance

The most important in predicting CKD, according to the feature importance analysis, was the levels of serum creatinine, blood urea, and blood glucose. Of course, it is very logical, since all these markers are used as obligatory predictors of renal function in medicine. The study did proceed and proved the models capable of detecting the relevant clinical features for an accurate prognosis of CKD.

Interpretation and Implications

Through the high accuracy and resilience of machine learning models in the detection of incipient CKD, it is promising for early identification to trigger prompt institution of intervention that will assure good outcomes in patients. In addition, the ability to identify important clinical characteristics confirms the practical application of the models in true healthcare setups.

Discussion on Challenges and Limitations

Though encouraging in results, the study had quite a few limitations and challenges. Future work should ensure that the training dataset is balanced and unbiased so that the model is representative of various populations. The development of interpretable models is necessary for gaining the trust of clinicians, even though the tree-based models, such as Random Forest, provide some degree of interpretability, and when the models become less transparent, it is advanced, like the model of XGBoost. Finally, larger and more diverse datasets need further examination to ensure reliability and scalability when applied across a variety of care settings.

Conclusion

Clearly, from the experiment, it is possible to determine CKD by use of machine learning models with a very high accuracy rating. The results highlight how essential model interpretability and good data quality are in any clinical application. Possible ways of increasing the acceptability and dependability of the models are by looking into problems raised to improve treatment of CKD patients.

References

1. Uci.edu. (2015). *UCI Machine Learning Repository*. [online] Available at: https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease [Accessed 3 Jul. 2024].

2. Levey, A.S. and Coresh, J. (2012). Chronic kidney disease. *Lancet*, [online] 379(9811), pp.165–180. https://doi.org/10.1016/s0140-6736(11)60178-5

3. Jha, V., Garcia-Garcia, G., Kunitoshi Iseki, Li, Z., Naicker, S., Plattner, B., Saran, R., Angela Yee-Moon Wang and Yang, C.-W. (2013). Chronic kidney disease: global dimension and perspectives. *Lancet*, [online] 382(9888), pp.260–272. https://doi.org/10.1016/s0140-6736(13)60687-x

4. Vijayarani, S., Dhayanand, M. and Research Scholar, M. (n.d.). *KIDNEY DISEASE PREDICTION USING SVM AND ANN ALGORITHMS*. [online] *International Journal of Computing and Business Research*. Available at: https://www.researchmanuscripts.com/March2015/2.pdf [Accessed 3 Jul. 2024].

5. Polat, H., Homay Danaei Mehr and Cetin, A. (2017). Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods. *Journal of medical systems*, [online] 41(4). https://doi.org/10.1007/s10916-017-0703-x