Kidney Disease Prediction

P Ganga Bhavani, Assistant Professor, ECE , IARE

Ch . Venkata Sai Manoj Student ECE , IARE

Ch . Vinay Student ECE, IARE

---***---

Abstract:
Chronic Kidney Disease (CKD) is a growing public health concern, and early detection is critical for effective intervention. This study compares various machine learning models—Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), and Neural Networks—to predict CKD. Results show that ensemble methods, like Random Forests, outperform individual models in terms of accuracy and robustness due to their ability to reduce overfitting. However, simpler models like Logistic Regression and Decision Trees offer greater interpretability, which is essential in clinical settings. The study underscores the trade-off between model complexity and transparency, suggesting future research should focus on integrating predictive models with Electronic Health Records (EHRs) to facilitate real-time monitoring and improve patient care.

KeyWords: Chronic Kidney Disease, Machine Learning, Random Forest, Logistic Regression, K-Nearest Neighbors, Early Detection.

INTRODUCTION:

Chronic Kidney Disease (CKD) is a growing global health concern, affecting 1 in 10 individuals, with cases increasing by 17% worldwide. Early detection is critical for better management and treatment. This project aims to develop a simple yet effective predictive model for CKD by analyzing factors like age, blood pressure, diabetes status, and lifestyle habits.

We will use machine learning techniques such as logistic regression, decision tree classifiers, and k-nearest neighbors. Logistic regression offers clear probabilistic interpretations, decision trees provide understandable rules, and k-nearest Neighbors are effective for pattern recognition. By leveraging these methods, we aim to identify key CKD risk factors, providing a requirements and supports sustainable aquatic comprehensive toolkit for healthcare providers.

This project seeks to enhance early CKD detection and contribute to healthcare analytics by developing an accessible and interpretable predictive model. Early intervention can significantly improve patient outcomes, emphasizing the importance of data analytics in healthcare.

While focused on CKD, these methodologies can also be applied to other chronic conditions, further advancing healthcare delivery. The success of this model can inspire similar approaches for predicting other diseases, benefiting overall health outcomes.

Methodology:

The dataset for this study comprises 400 patient records with 26 attributes, including demographic, clinical, and laboratory variables. Attributes such as age, blood pressure, blood glucose levels, hemoglobin, and kidney-specific metrics were utilized. Data was sourced from publicly available datasets.

Data Preprocessing:

To ensure data quality, the following preprocessing steps were performed:

- Handling missing values through imputation or exclusion.
- Encoding categorical variables into numerical formats.
- Normalizing numerical features for consistent scaling.
- Splitting the data into training (75%) and testing (25%) sets.

Model Development:

Machine learning models implemented include:
1. Logistic Regression: Provides a probabilistic interpretation of CKD likelihood.
2. Decision Trees: Offers intuitive rule-based predictions.
3. Random Forest: Enhances accuracy through ensemble learning.
4. K-Nearest Neighbors: Classifies data based on proximity in feature space.4. K-Nearest Neighbors: Classifies data based on proximity in feature space.
Scikit-learn Documentation. Available at: https://scikit learn.org/stable/documentation.html
5. NumPy Documentation. Available at: https://numpy.org/doc/
6. Pandas Documentation. Available at: https://pandas.pydata.org/docs/
7. Dataset taken from https://www.kaggle.com/datasets/mansoordaku/ckdisease

Components:

- CKD Chronic Kidney Disease
- DT Decision Trees
- KNN K-Nearest Neighbors (KNN)
- IBM : Instance Based Models

Performance Metrics:

Models were evaluated using:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC

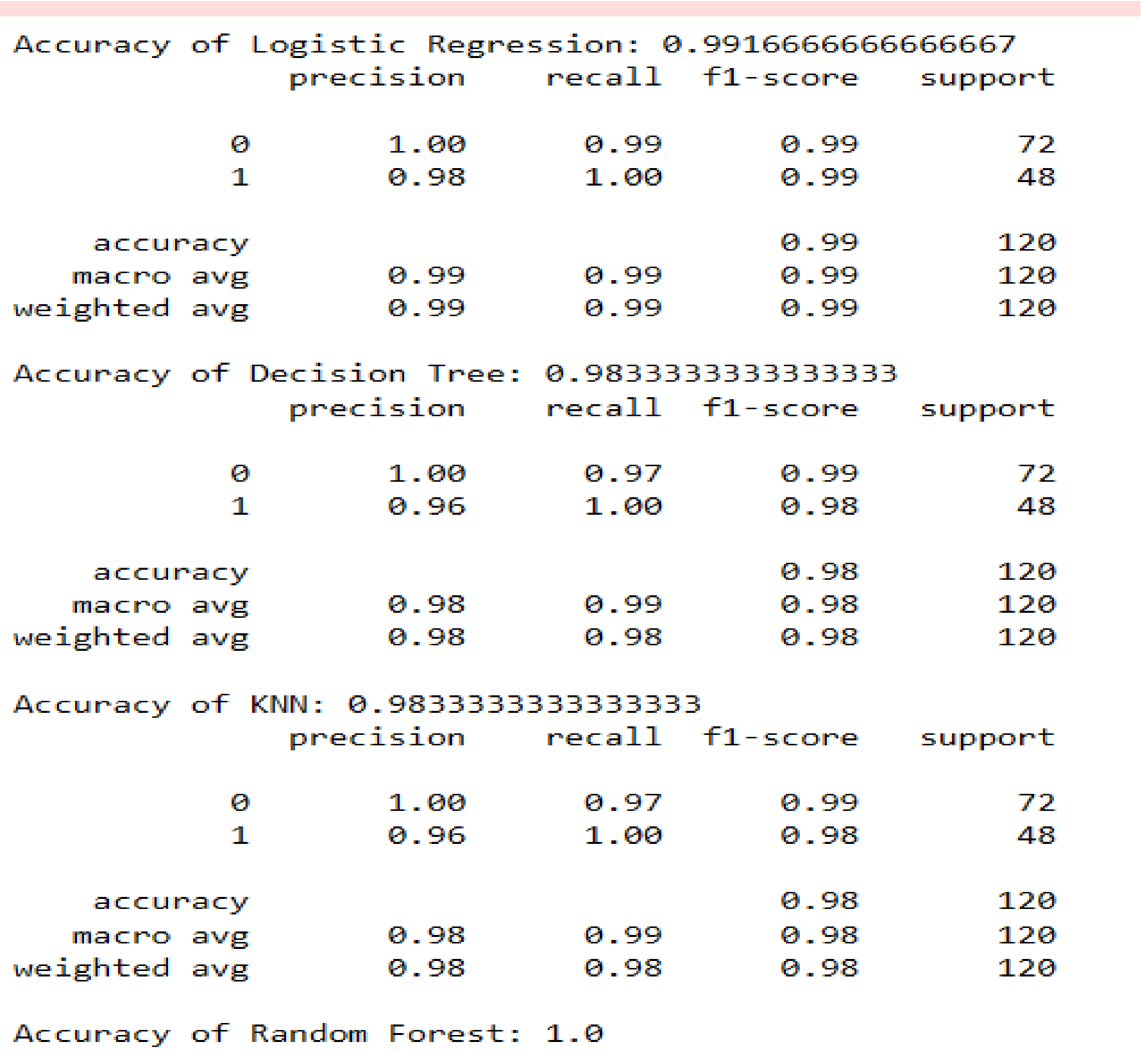
Implementation:

Python libraries such as Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn were used for data manipulation, visualization, and model implementation. The dataset was visualized to identify key patterns and correlations, aiding feature selection.

Visualization Highlights:

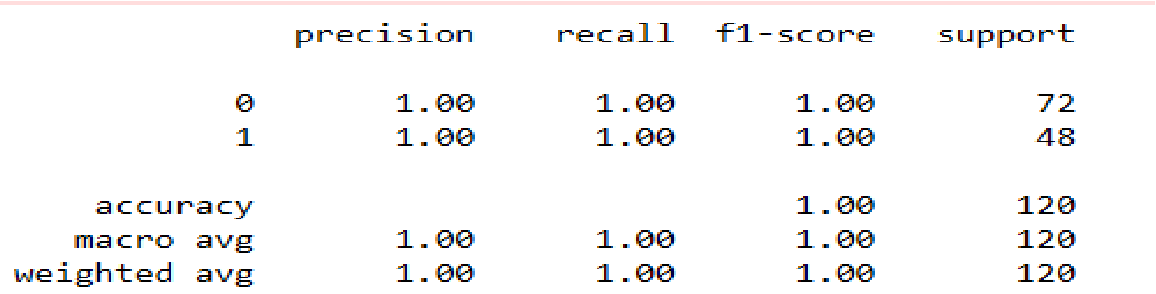
- Age distribution and CKD incidence.
- Correlation between hemoglobin levels and cell volume.
- Heatmaps illustrating attribute relationships

Results and Discussion Results and



```
Accuracy of Logistic Regression: 0.9916666666666667
              precision    recall  f1-score   support

           0       1.00      0.99      0.99        72
           1       0.98      1.00      0.99        48

    accuracy                           0.99       120
   macro avg       0.99      0.99      0.99       120
weighted avg       0.99      0.99      0.99       120

Accuracy of Decision Tree: 0.9833333333333333
              precision    recall  f1-score   support

           0       1.00      0.97      0.99        72
           1       0.96      1.00      0.98        48

    accuracy                           0.98       120
   macro avg       0.98      0.99      0.98       120
weighted avg       0.98      0.98      0.98       120

Accuracy of KNN: 0.9833333333333333
              precision    recall  f1-score   support

           0       1.00      0.97      0.99        72
           1       0.96      1.00      0.98        48

    accuracy                           0.98       120
   macro avg       0.98      0.99      0.98       120
weighted avg       0.98      0.98      0.98       120

Accuracy of Random Forest: 1.0
```

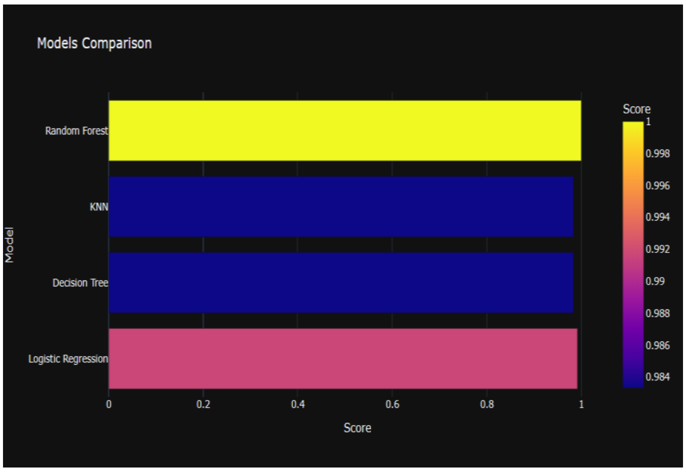
Discussion:



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        72
           1       1.00      1.00      1.00        48

    accuracy                           1.00       120
   macro avg       1.00      1.00      1.00       120
weighted avg       1.00      1.00      1.00       120
```

Random Forest emerged as the top-performing model, balancing accuracy and robustness. Simpler models like Logistic Regression offer greater interpretability, beneficial in clinical settings.

Model comparison:



Accuracies:
Logistic Regression=99%
Decision Tree Classifier=98%
KNN=99%
Random Forest=99

Discussion:
While complex models deliver high accuracy, their lack of transparency can impede clinical adoption. Hybrid approaches combining interpretability with performance may bridge this gap. Integrating predictive models with EHRs could facilitate real-time monitoring, enabling proactive intervention.

Conclusion:
This study underscores the potential of machine learning in CKD prediction, highlighting the strengths and limitations of various models. Future work should focus on enhancing model interpretability, integrating real-time systems, and expanding datasets for broader applicability. By advancing predictive analytics, we can improve CKD outcomes and contribute to sustainable healthcare solutions.

REFERENCES:

1. KNeighborsClassifier Documentation: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
2. LogisticRegression Documentation: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
3. DecisionTreeClassifier Documentation: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
4. Dataset Source: https://www.kaggle.com/datasets/mansoordaku/ckdisease
5. NumPy Documentation: https://numpy.org/doc/
6. Pandas Documentation: https://pandas.pydata.org/docs/