Chronic Kidney Disease Prediction: Model Summary and Rationale

Machine Learning: CS379

Antoine Gaton

October 20, 2024

**Dataset**

The data set I used was the "Chronic Kidney Disease" dataset from the UCI Machine Learning Repository (ID: 336). This dataset was created in 2015 and contains medical attributes collected over a period of about 2 months from a hospital. The reason I chose this dataset was due to my personal history and curiosity about CKD.

*Key dataset characteristics*:

* Number of instances: 400
* Number of features: 24
* Target variable: 'class' (ckd or notckd)
* Contains missing values

**Feature Overview**

The dataset includes a mix of numerical and categorical features, such as:

* *Numerical*: age, blood pressure, blood glucose, blood urea, serum creatinine, sodium, potassium, hemoglobin, etc.
* *Categorical*: specific gravity, albumin, sugar, red blood cells, pus cell, bacteria, hypertension, diabetes mellitus, etc.

**Algorithm Selection: Logistic Regression**

I chose logistic regression for this binary classification task for several reasons:

1. Simplicity and interpretability
2. Efficiency with linearly separable classes
3. Probabilistic output, useful in medical contexts
4. Feature importance insights

**Results**

* *Model Performance*
  + Accuracy: 0.99
  + Precision: 0.98 (ckd), 1.00 (notckd)
  + Recall: 1.00 (ckd), 0.96 (notckd)
  + F1-score: 0.99 (ckd), 0.98 (notckd)
  + These metrics indicate that the model performs exceptionally well in predicting chronic kidney disease, with near-perfect accuracy and balanced performance across both classes.
* A blue squares with white text

  Description automatically generated*Confusion Matrix*
  + The confusion matrix shows that our model correctly identified:
    - 52 true positives (ckd)
    - 27 true negatives (notckd)
    - With only 1 false negative and no false positives.
* A graph showing a number of blue lines

  Description automatically generated with medium confidence*Feature Importance*
  + The feature importance graph reveals the most influential factors in predicting chronic kidney disease according to the model.

**Conclusion and Discussion**

1. *Model Performance*: The logistic regression model demonstrates outstanding performance in predicting chronic kidney disease, with 99% accuracy. This high accuracy suggests that the selected features are highly indicative of the presence or absence of CKD.
2. *Balanced Performance*: The model shows excellent balance between precision and recall for both classes (ckd and notckd), indicating it's equally effective at identifying both positive and negative cases.
3. *Data Quality*: The dataset contains missing values, which were handled using mean imputation. This approach, while simple, has proven effective given the model's high performance.
4. *Potential Limitations*:
   1. The extremely high accuracy (99%) could potentially indicate overfitting. It's important to validate this performance on external datasets.
   2. The dataset is relatively small (400 instances), which might limit its generalizability to larger, more diverse populations.
5. *Future Work*:
   1. Implement cross-validation to ensure the model generalizes well to unseen data.
   2. Explore more complex models (e.g., random forests, gradient boosting) to compare performance.
   3. Investigate the impact of different imputation methods for handling missing values.
   4. Collect more diverse data to further validate the model's performance.
   5. I would like to consult, if possible, with medical professionals to ensure the identified important features align with clinical knowledge.

This model serves as a strong starting point for predicting chronic kidney disease and could be a valuable tool in early detection and prevention efforts. However, further validation and refinement are recommended before any clinical application.