# Chronic Kidney Disease prediction Using Bayesian Logistic Regression

**Sree Paada Reddy**
Information Science
University of Arizona
pallikila@arizona.edu

## Abstract

CKD is one of the serious health conditions that affects millions of people in the world. The early diagnosis of CKD is very essential for ensuring effective management and improving patient chances. This paper classifies CKD patients using clinical data with Bayesian logistic regression along with MAP estimation. Bayesian methods provide a robust framework for integrating prior knowledge and addressing uncertainty in parameter estimation, which a traditional logistic regression lacks. Through iterative weight optimization involving gradients and Hessians, the model demonstrates great predictive capabilities. A range of regularization hyperparameters were explored to evaluate model performance across metrics .Using a publicly available CKD dataset, the model achieves impressive performance metrics Results demonstrate that hyperparameter =1 achieves optimal performance, with an accuracy of 97.75%, a precision of 98.80%, a recall of 97.58%, and an F1 score of 98.17%.These results evidence the capabilities of Bayesian logistic regression on healthcare decision-making, pointing toward its relevance on such essential tasks as disease classification. Of course, being so promising, this also underlines a number of other challenges one has to address in extending the approach further, e.g., computational cost and transparency of features.

## 1 Introduction

CKD is an irreversible, progressive clinical condition that is characterized by a gradual loss of kidney function, leading to severe complications such as kidney failure, cardiovascular disease, and premature death. The timely diagnosis and intervention in the disease are crucial to reduce these poor outcomes. Recent advances in machine learning have revolutionized healthcare diagnostics by offering new ways to classify and predict diseases with high accuracy. Logistic regression is one of the most popular classification algorithms, but it has some drawbacks in handling uncertainty and incorporating prior domain knowledge. Bayesian logistic regression extends this by combining data likelihood with prior distributions, yielding models that are more robust to noise and better equipped to manage uncertainty.

This paper explores the application of Bayesian logistic regression with MAP estimation for the classification of CKD patients. This Bayesian framework not only increases the flexibility of a model but also enhances its interpretability with an estimation of uncertainty in the predictions. The study used the publicly available CKD dataset, after preprocessing by imputing missing values and encoding categorical features. It yields remarkable metrics by employing a Gaussian prior leveraged to optimize weights iteratively and taking multiple $\sigma 2$

values. This paper discusses methodology, results, and implications for the use of Bayesian logistic regression in the classification of CKD, providing valuable lessons for its applicability in a real-world healthcare setting.

## 2 Methods

The dataset used for the study includes clinical and diagnostic features regarding CKD. Missing values were cleaned by a combination of mean imputation for numeric variables and forward filling for categorical variables. The target variable, "classification," was binary encoded, where CKD was set as 1 and non-CKD as 0. Non-numeric columns are encoded with Label Encoding. After that, the data is divided into an 80:20 split for training and testing, respectively.

Bayesian logistic regression was performed by placing a Gaussian prior with variance $\sigma 2$ over the weights (w) MAP estimation is made to run in iterative method variations in weights through using gradient and Hessian-based methods for optimizations. The gradient brings together the prior and likelihood of the data, while in turn, the Hessian provides information of the curvature of the posterior distribution. Optimizer iterates for 10 steps with weight initialization at zero value. Probabilities would get predicted with the usage of logistic function, using the threshold 0.5 for the classification of outcomes. Hyperparameter tuning involved the testing of four values of $\sigma 2$ , namely 1, 10, 100, and 1000, to see the effect caused by the strength of regularization. Evaluation metrics such as accuracy, precision, recall, and F1 score were computed on the test set to evaluate the model performance.

## 3 Results

The results show that the performance of the models was highly dependent on the value of the regularization hyperparameter $\sigma 2$ . In particular, for $\sigma 2$ =1, the model attained the best performance on all metrics, with an accuracy of 97.75%, precision of 98.80%, recall of 97.58%, and an F1-score of 98.17%. For $\sigma 2$ =10 and $\sigma 2$ =100, metrics started to degrade gradually, indicating poor generalization. For $\sigma 2$ =1000, the metrics showed a sharp decline: the accuracy was 83.00%, precision-87.03%, recall-83.72%, and the F1-score equaled 85.33%, reflecting over-regularization. For small values $\sigma 2$ , the performance remains relatively stable, while at higher values, it deteriorates substantially. These results support the conclusion that regularization is critical to the balance between model complexity and generalization: $\sigma 2$ =1 emerging as the optimal choice

## 4 Conclusion

This study pinpoints Bayesian logistic regression with a Gaussian prior and Maximum A Posteriori estimation as a robust tool for classifying CKD. This model achieved the best trade-off between the quality of fitting and the quality of generalization because of regularization introduced by the hyperparameter $\sigma 2$ . The optimal value of this hyperparameter was $\sigma 2$ =1, which gave accuracy of 97.75%, precision of 98.80%, recall of 97.58%, and an F1-score of 98.17%. The study underlines the practical utility of Bayesian methods in healthcare, where usually noisy or incomplete data demand robust models that can tackle uncertainty. Besides its high predictive performance, the Bayesian framework offers interpretability, an essential ingredient in clinical applications. By integrating prior knowledge and generating probabilistic outputs, the approach could support clinicians in understanding which factors lead to the classification of CKD and therefore allow informed decision-making. The high recall also ensures that the model is well-suited for healthcare contexts where the cost of a false negative is extremely high, such as diagnosis of chronic diseases. Beyond CKD, the methodology proves to be flexible, indicating its possible use in other medical classification tasks. In summary, this study illustrates the potential of Bayesian logistic regression as a trustworthy and interpretable diagnostic tool for CKD, laying the ground for further research in Bayesian methods within healthcare analytics. Balancing technical performance, interpretability, and robustness, this article lays the foundation for using

Bayesian approaches in clinical decision-making, with promising applications across many domains of disease prediction and diagnosis. .

## 5 Limitations

While the results of this study are promising, several limitations must be acknowledged. Firstly, the dataset on which this test case analysis was done is little and may restrict the performance capability of the model for bigger and more diverse populations. Most medical datasets also manifest variability across demographics, clinical practices, and measurement techniques, which a limited dataset cannot capture. Second, the adopted preprocessing steps themselves-missing value imputation with mean, forward-filling for categorical features-may lead to some biases that further call into question the reliability of the results. These common methods are simplifications of the data that might fail to take into account underlying patterns of missingness that could influence model performance. While superior in interpretability to standard logistic regression, the Bayesian framework poses its own challenges when it comes to communicating results to clinical stakeholders.

## References

[1] Neha Sonone and A. Daniel. Early Prediction and Progrssion of Chronic Kidney Disease Using Machine Lerning Techniques. In: *2024 2nd International Conference on Networking and Communications (ICNWC)*. 2024, pp. 1–6.

[2] *Chronic Kidney Disease Dataset*. `https://www.kaggle.com/datasets/mansoordaku/ckdisease/data`. 2016.

[1] [2]