**📄 Justification and Explanation for Prediction Model**

**Project Title:** Chronic Kidney Disease (CKD) Prediction using Classification Algorithms  
**Client:** Hope Artificial Intelligence | Healthcare Sector  
**Prepared by:** Santhosh Jecob  
**Date:** 06/06/2025

**1. Project Objective**

The objective of this project was to develop a reliable and accurate machine learning model that can predict **Chronic Kidney Disease (CKD)** based on patient medical parameters. The goal is to assist the hospital management in early diagnosis and proactive treatment planning for patients at risk.

**2. Dataset Overview**

* **Source:** Hospital client-provided CSV dataset
* **Total Records:** 400
* **Target Variable:** Presence or absence of CKD
* **Features:** Age, blood pressure, albumin, blood glucose, blood urea, serum creatinine, sodium, potassium, red/white blood cells, and several yes/no categorical indicators.

**3. Preprocessing Summary**

* Removed missing and inconsistent values (e.g., '?', 'NaN')
* Converted all numerical columns appropriately
* Applied **Label Encoding** and **One-Hot Encoding** for categorical features
* Scaled features using StandardScaler where necessary
* Split the dataset: 80% training / 20% testing

**4. Modeling and Evaluation**

Six different classification models were built, tested, and compared:

| **Model** | **Accuracy** |
| --- | --- |
| Logistic Regression | 97.06% |
| K-Nearest Neighbors | 95.59% |
| Decision Tree | 94.12% |
| **Random Forest** ✅ | **98.24%** |
| Support Vector Machine | 97.06% |
| Naive Bayes | 91.18% |

**5. Justification for Final Model**

After careful evaluation of all models using **accuracy**, **precision**, **recall**, **F1-score**, and **confusion matrix**, the **Random Forest Classifier** was selected as the final model for deployment.

**Reasons for Selection:**

* Achieved the **highest accuracy** (98.24%) among all tested models
* Demonstrated excellent performance in **detecting CKD and non-CKD** cases (balanced precision/recall)
* Robust to **noise, outliers**, and **non-linear feature interactions**
* Performs well even when some features are missing or unimportant
* **Low risk of overfitting** due to ensemble method

**6. Sample Output - Classification Report (Random Forest)**

precision recall f1-score support

0 0.96 0.96 0.96 25

1 0.99 0.99 0.99 55

accuracy 0.98 80

macro avg 0.97 0.97 0.97 80

weighted avg 0.98 0.98 0.98 80

**7. Conclusion**

The implemented solution successfully meets the requirement of the hospital management by providing a **highly accurate CKD prediction model** using real-world patient data. This model can now be integrated into a user interface (e.g., Streamlit dashboard) or embedded into the hospital’s diagnostic systems for early intervention and care planning.