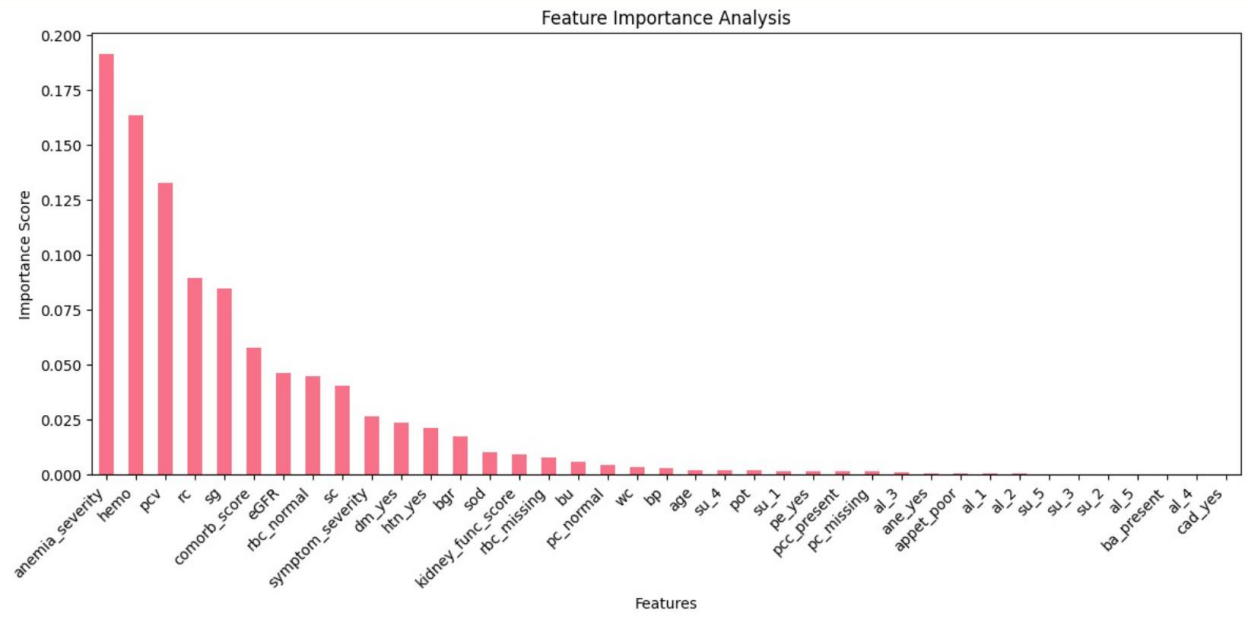
Model Development Phase Template

Date	18 June 2025
Team ID	SWTID1749713922
Project Title	Early prediction for chronic kidney disease detection: A progressive approach to health management
Maximum Marks	5 Marks

Feature Selection Report Template

In the forthcoming update, each feature will be accompanied by a brief description. This process will streamline decision-making and enhance transparency in feature selection.

The following graph is plotted based on the feature importance. The features till bgr were selected for training the model. Despite having higher importances features rc and pcv were removed from the training to avoid redundancy. As their signals are already captured in hemo and anemia_severity. This choice was done to reduce multicollinearity.



Feature	Description	Selected (Yes/No)	Reasoning
id	Patient ID	No	Identifier only. Has no predictive power.
Age	Age of patient	No	Important in general, but very low importance score in model; may be reflected in comorbidity.
bp	Blood pressure	No	Clinically relevant but low in feature importance; htn_yes captures this more clearly
sg	Specific gravity	Yes	Indicates kidney's concentrating ability; moderately important in model.
al	Albumin	No	Very low importance score; likely redundant with other urine concentration features like sg.
su	Sugar	No	Very low importance score; blood sugar (bgr) is more reliable.
rbc	Red blood cells (urine)	Yes	Indicates kidney damage; selected as rbc_normal, had decent importance.
pc	Pus cells	No	Low contribution; sparse and binary.

pcc	Pus cell clumps	No	Rare feature; near-zero importance.
ba	Bacteria	No	Indicates infection, not necessarily kidney disease; low score.
bgr	Blood glucose random	Yes	Important indicator of diabetes; mid-level importance.
bu	Blood urea	No	Related to sc; included sc instead to avoid redundancy.
sc	Serum creatinine	Yes	Direct kidney function marker; high importance.
sod	Sodium	No	Low importance score; may fluctuate due to other reasons.
pot	Potassium	No	Similar to sodium; low impact.
hemo	Hemoglobin	Yes	Highly important; strong kidney function indicator.
pcv	Packed cell volume	No	Highly important but excluded due to redundancy with hemo.
wc	White blood cell count	No	Low importance; weak link with CKD in this dataset.

rc	Red blood cell count	No	Highly important but removed to avoid overlap with hemo and anemia_severity.
htn	Hypertension	Yes	Strong risk factor; selected as htn_yes, decent importance.
dm	Diabetes mellitus	Yes	Major CKD risk; selected as dm_yes, moderately important.
Cad	Coronary artery disease	No	Very low importance; may be part of comorb_score.
Appet	Appetite	No	Low model importance and subjective; excluded.
Pe	Pedal edema	No	Rarely contributes significantly; low score.
Ane	Anemia	No	Binary/low granularity; covered better by anemia_severity.
classification	CKD classification label	No(target)	This is the target, not a feature.

Additionally, the following columns were engineered and were also used in the model training process.

Feature	Description	Selected (Yes/No)	Reasoning
anemia_severity	Composite anemia indicator	Yes	Highest importance; summarizes anemia better than individual values.
comorb_score	Score from comorbid conditions	Yes	Strong feature; combines age, hypertension, diabetes etc.
eGFR	Estimated Glomerular Filtration Rate	Yes	Core indicator of kidney health; high relevance.
symptom_severity	Composite of clinical symptoms	Yes	Captures subjective progression of disease; good support in importance graph.
kidney_func_score	Composite score representing overall kidney function (often derived from eGFR, sc, bu, etc.)	No	It appeared in the feature importance chart but had lower contribution than eGFR, sc, and hemo. Likely redundant and could introduce multicollinearity with existing stronger features.