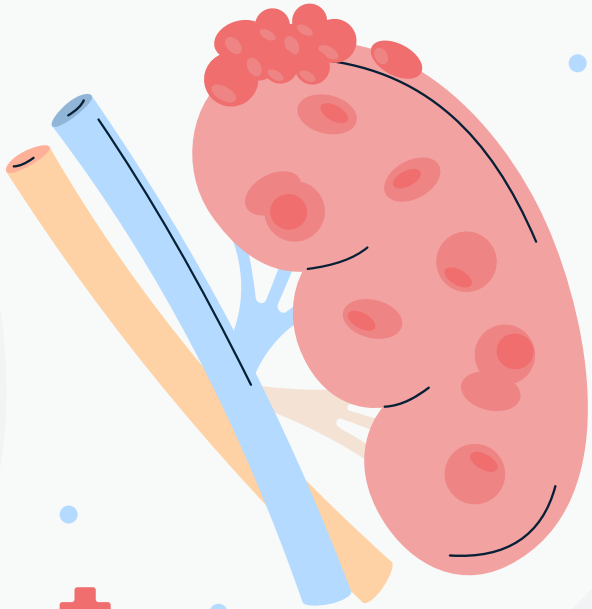
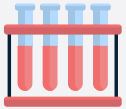
Chronic Kidney Disease Risk

Team 5
Ramon Gonzalez
Alanis Perez



Introduction



Chronic Kidney Disease

Progressive disease of the kidneys
Inhibits the ability to perform essential functions
Progression can result in more complications, e.g., heart disease

Objective

Classify patients as being “high risk” for CKD
Allow for intervention practices
Enable healthcare professionals to advance care

Feature Set and Target

Age

Numeric
Age in years

Creatinine Level

Numeric
Measured in mg/dL

Blood Urea Nitrogen (BUN)

Numeric
Measured in mg/dL

Diabetes

Categorical
1 - patient has diabetes
0 - no diabetes

Hypertension

Categorical
1 - patient has high blood pressure
0 - no high blood pressure

Urine Output

Numeric
Measured in ml/day

CKD Status

1 - patient has CKD
0 - no CKD

CKD Risk

“High Risk”
“Low/Moderate Risk”

Models Used

Logistic Regression

Baseline model
Estimate probability of CKD
Reflect magnitude and direction of each feature's influence

Random Forest

Feature importance
Ranking contribution of each feature to CKD risk

XGBoost

Depending on primary model outcomes, will serve as an attempt to evaluate more complex patterns and improve accuracy

Evaluation Metrics

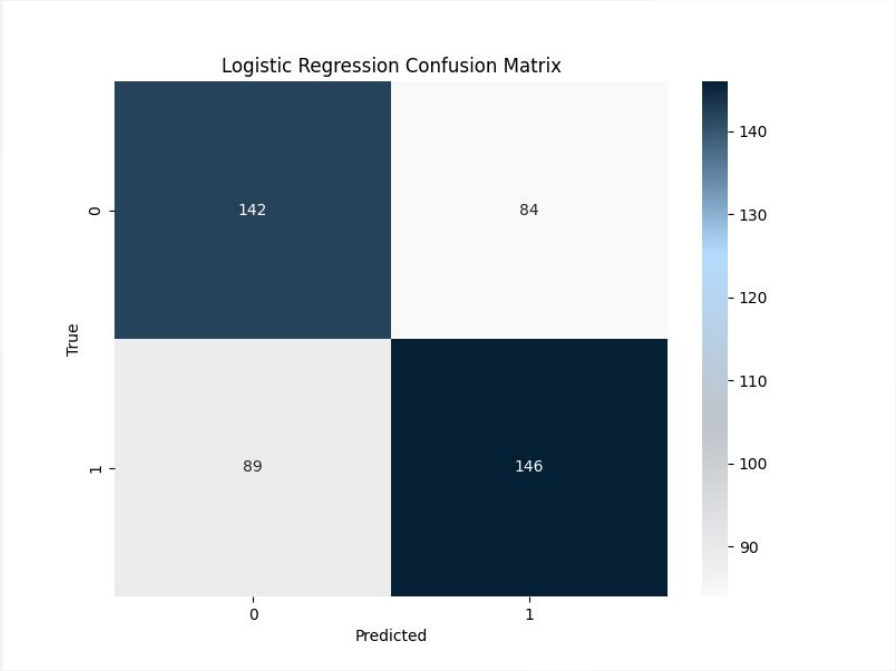
Logistic Regression:

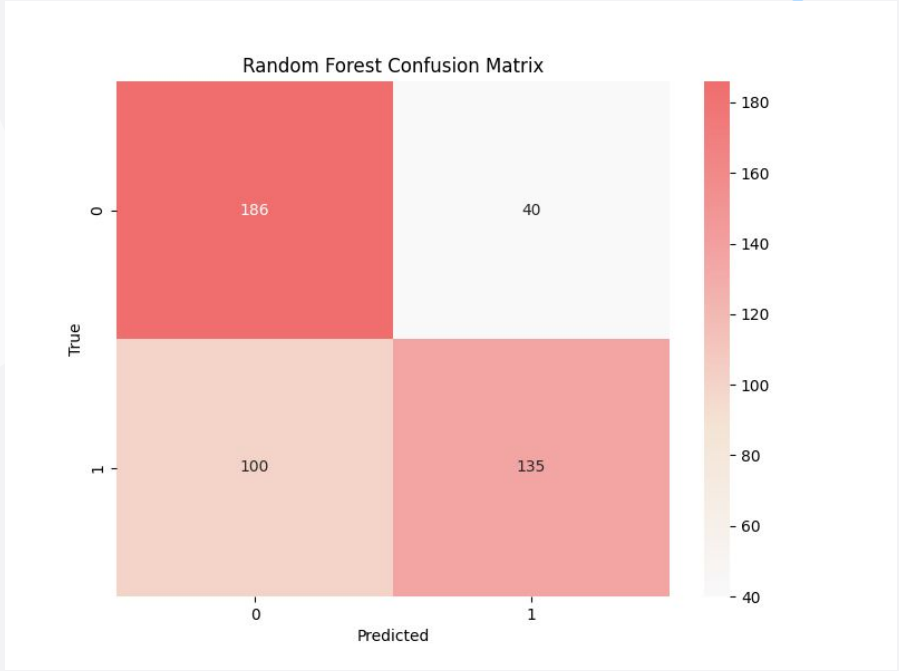
	precision	recall	f1-score	support
0	0.61	0.63	0.62	226
1	0.63	0.62	0.63	235
accuracy			0.62	461
macro avg	0.62	0.62	0.62	461
weighted avg	0.62	0.62	0.62	461

Random Forest:

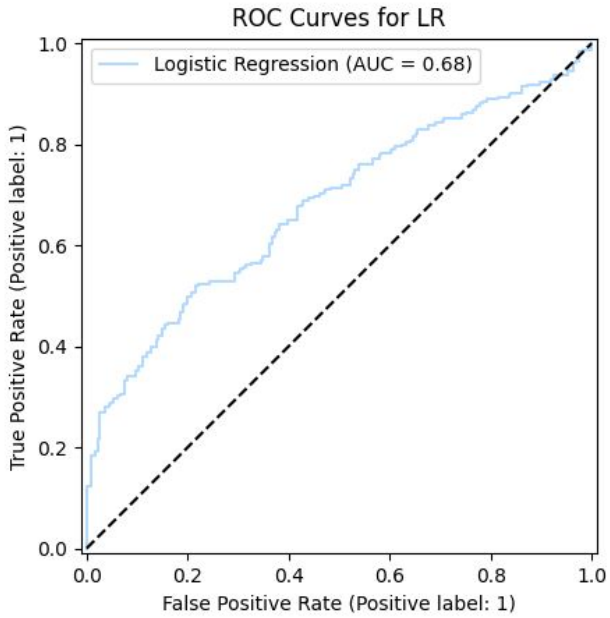
	precision	recall	f1-score	support
0	0.65	0.82	0.73	226
1	0.77	0.57	0.66	235
accuracy			0.70	461
macro avg	0.71	0.70	0.69	461
weighted avg	0.71	0.70	0.69	461

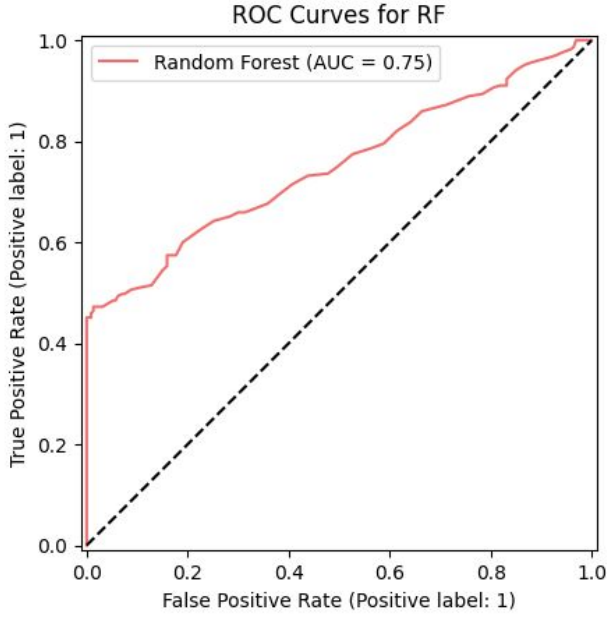
Confusion Matrices



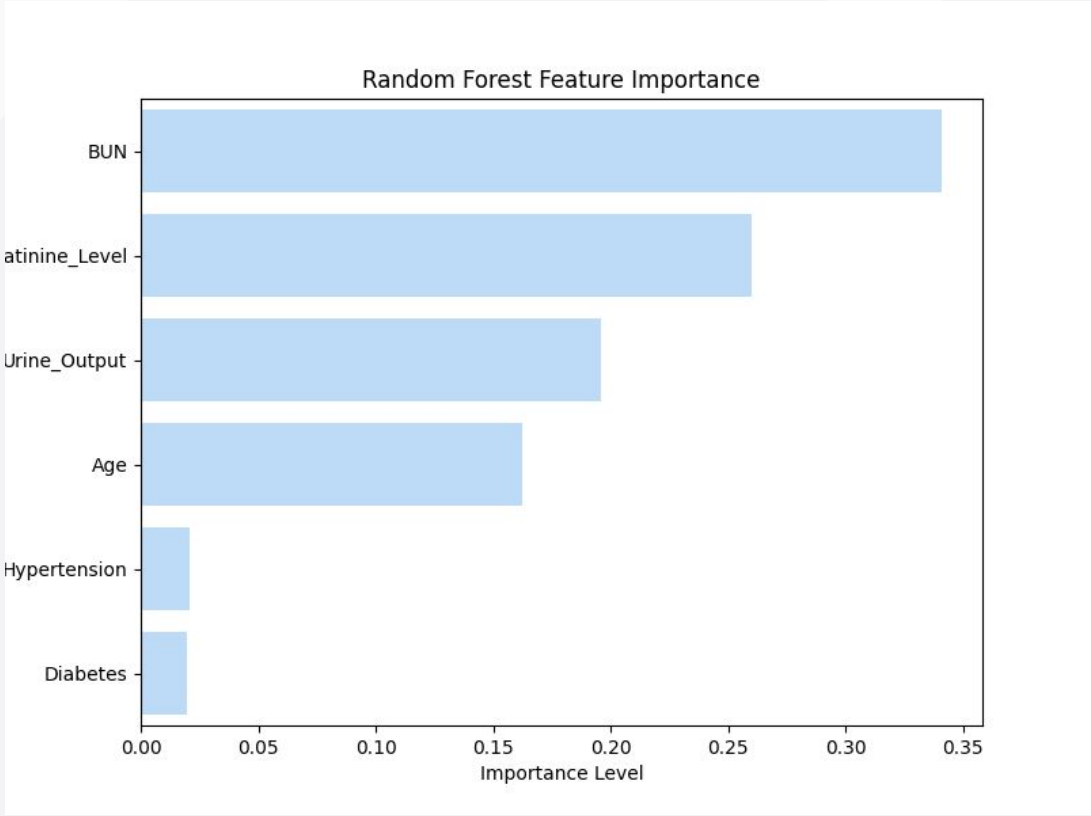


ROC Curves

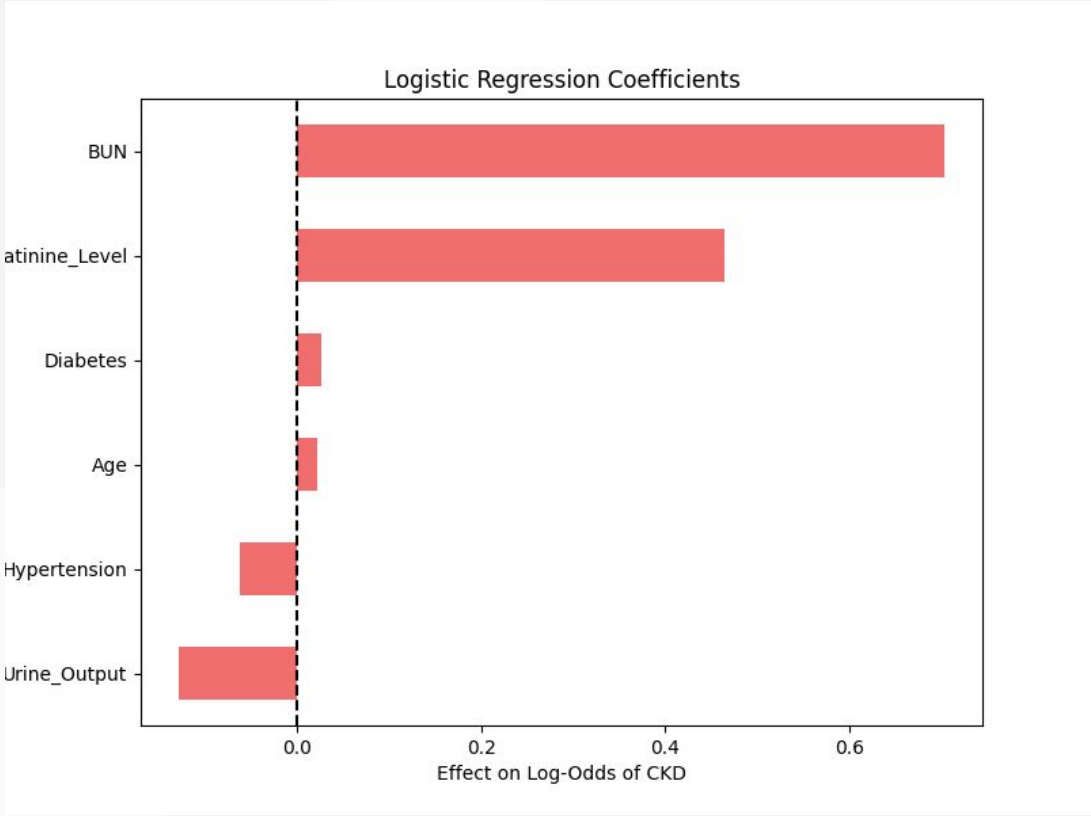

ROC Curves for LR
Logistic Regression (AUC = 0.68)
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curves for RF
Random Forest (AUC = 0.75)
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0

Feature Importance



Coefficients



Predictions and Label Creation

Predictions were ran on both LR and RF for the positive class (CKD = 1) on the test set - proceeded with RF due to higher accuracy

Test set: 461 patients (taken from the original 80/20 split)
Threshold: 0.7
Inclusion of a ‘CKD_Probability’ column after determining probability per patient
Inclusion of a ‘Risk_Label’ determined by threshold on the probability

Test set results:
112 patients at high risk of CKD
349 patients at low/moderate risk of CKD

Discussion

Conclusion

LR and RF were robust models
-Feature influences
-Magnitude
-Direction

Limitations

Synthetic dataset
Exclusion of GFR
Small sample size

Next Steps

Include GFR and other clinical measures
More complex models
Time-series: evaluation CKD progression
Apply models to real data
Post-diagnosis: ‘dialysis_needed’