Predicting Diabetes Status Using Biomedical and Demographic Data

KELVIN MUTUA

Project Background

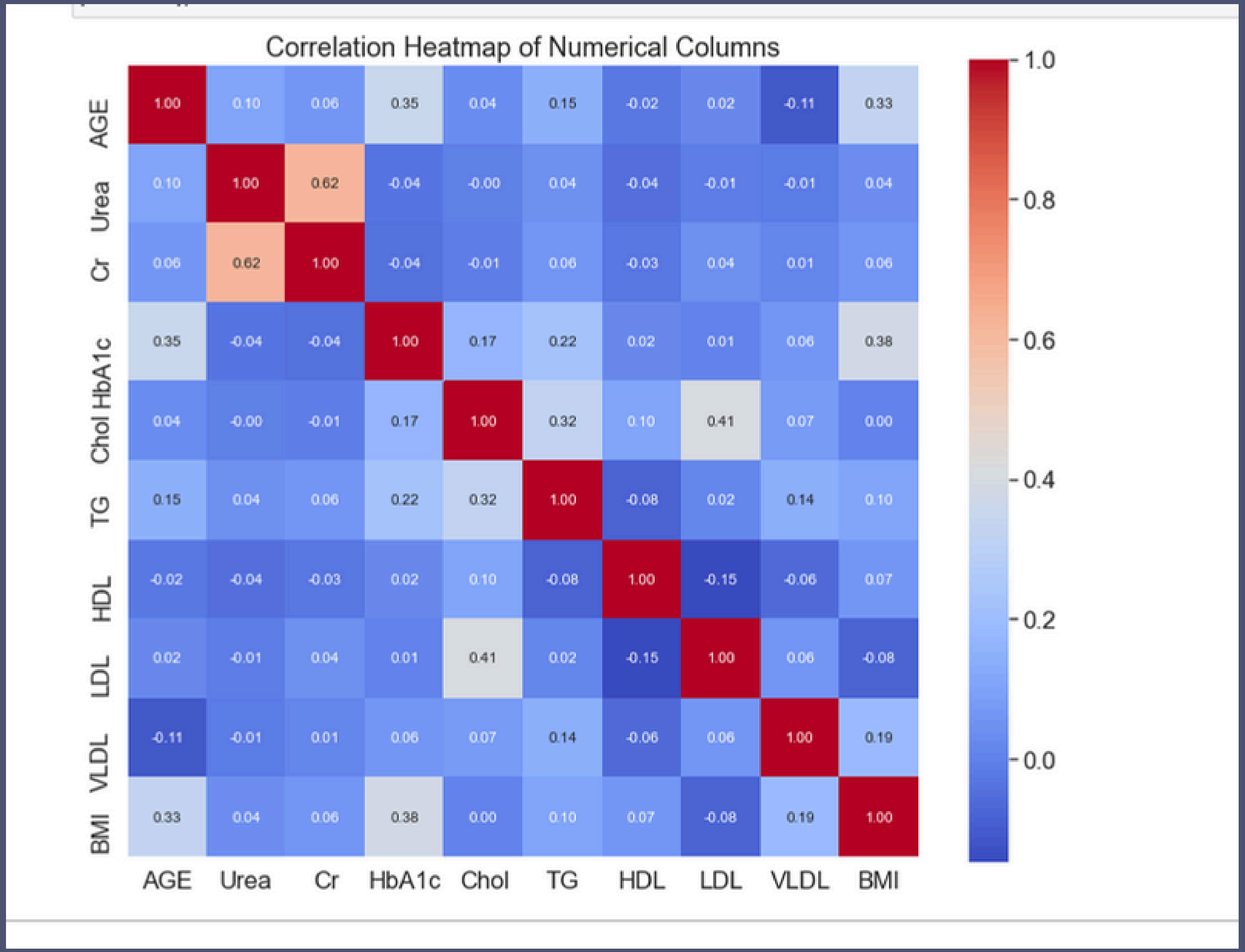
The Challenge

Diabetes is a rising global health concern. Early detection is crucial but limited by accessibility and cost of traditional diagnostic methods.

Our Solution

Use existing patient data (e.g., cholesterol, HbA1c, age, BMI) to automatically predict diabetes status using machine learning.

Exploratory Insights



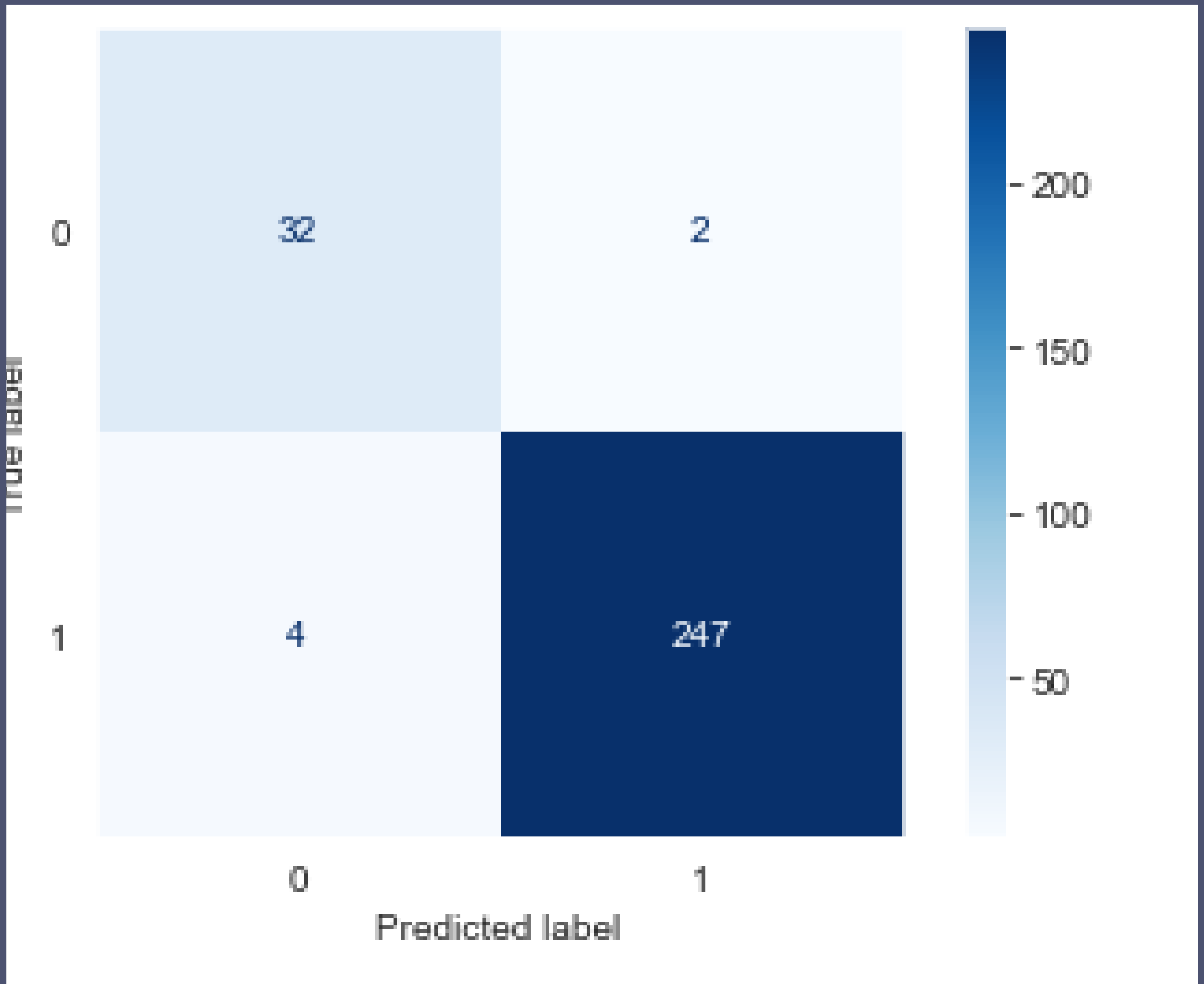
Urea and Creatinine (Cr) → 0.62

Both are indicators of kidney function and often rise together in diabetic nephropathy.

HbA1c and BMI → 0.38

Suggests individuals with higher BMI tend to have higher long-term blood sugar levels.

confusion matrix



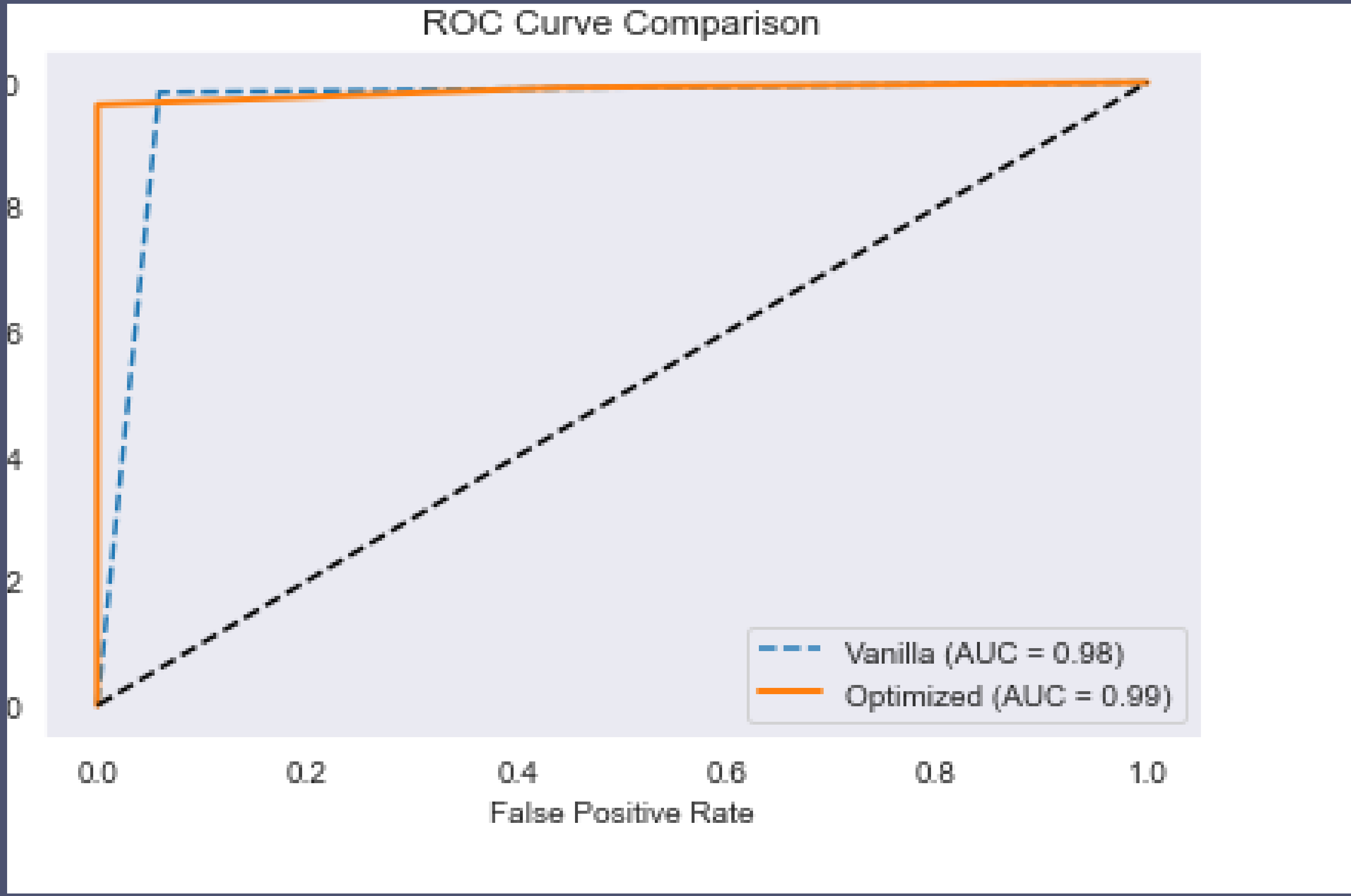
>The model is highly effective at identifying diabetic patients (only 4 missed cases).
>It also avoids over-diagnosing (only 2 false positives).
>This makes it very suitable for clinical decision support where missing a case is costly.

Decision Tree Classifier

The model achieved an accuracy of 97.2%,
correctly identifying most patients.
It showed very high precision (99.2%), meaning
almost no false diabetes predictions.
Recall was also strong at 98.4%, capturing nearly
all actual diabetic cases.
The F1-score of 98.8% confirms excellent balance
between precision and recall


HbA1c <= -1.06
gini = 0.5
samples = 1186
value = [593, 593]
class = 0
BMI <= -1.036
gini = 0.116
samples = 632
value = [593, 39]
class = 0
gini = 0.0
samples = 554
value = [0, 554]
class = 1
Chol <= 0.073
gini = 0.06
samples = 612
value = [593, 19]
class = 0
gini = 0.0
samples = 20
value = [0, 20]
class = 1
gini = 0.0
samples = 535
value = [535, 0]
class = 0
TG <= -0.29
gini = 0.372
samples = 77
value = [58, 19]
class = 0
VLDL <= -0.165
gini = 0.094
samples = 61
value = [58, 3]
class = 0
gini = 0.0
samples = 16
value = [0, 16]
class = 1
HbA1c <= -1.681
gini = 0.033
samples = 59
value = [58, 1]
class = 0
gini = 0.0
samples = 2
value = [0, 2]
class = 1
gini = 0.0
samples = 1
value = [0, 1]
class = 1
gini = 0.0
samples = 58
value = [58, 0]
class = 0

```
n: 6
 leaves: 7
```



ROC Curve & AUC

AUC Score: was at 0.99 -model is extremely good at distinguishing diabetic vs. non-diabetic.

Key Takeaways

> High potential to detect diabetes using readily available data

> Decision Tree performed well and is easy to interpret

> Logistic Regression offers simplicity and robustness

Recommendations

> Deploy model as a diagnostic support tool in clinics

> Train health workers to use it during patient intake

> Conduct follow-up studies to expand dataset and improve accuracy