a) What is fraud detection, and why is it important?

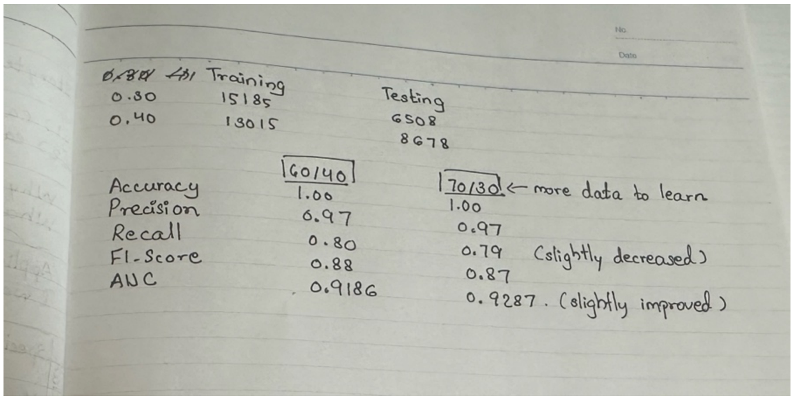
Fraud detection is the process of identifying suspicious or illegal activities, such as fake transactions, identity theft, or financial scams, by analyzing patterns in the data. It helps organizations detect and prevent fraud before it causes serious damage.

Fraud detection is important because it protects people and companies from financial loss, keeps systems safe, and maintains trust in services like banking, e-commerce, insurance, and even government programs. Without fraud detection, many systems could be easily misused or exploited, leading to huge economic and social impacts.

b) If you change the training and testing split to 70% training and 30% testing, how does the model's performance change?

When I changed the split from 60% training / 40% testing to 70% training / 30% testing, the model's performance stayed almost the same. The accuracy and precision remained very high, and the AUC score slightly increased from 0.9186 to 0.9287. This shows that giving the model more data to learn from helped it slightly improve in distinguishing between fraud and non-fraud.

Overall, the change in performance was small, which means the model is stable and generalizes well even when the training size changes.



~~[illegible]~~	Training	Testing
0.30	15185	6508
0.40	13015	8678

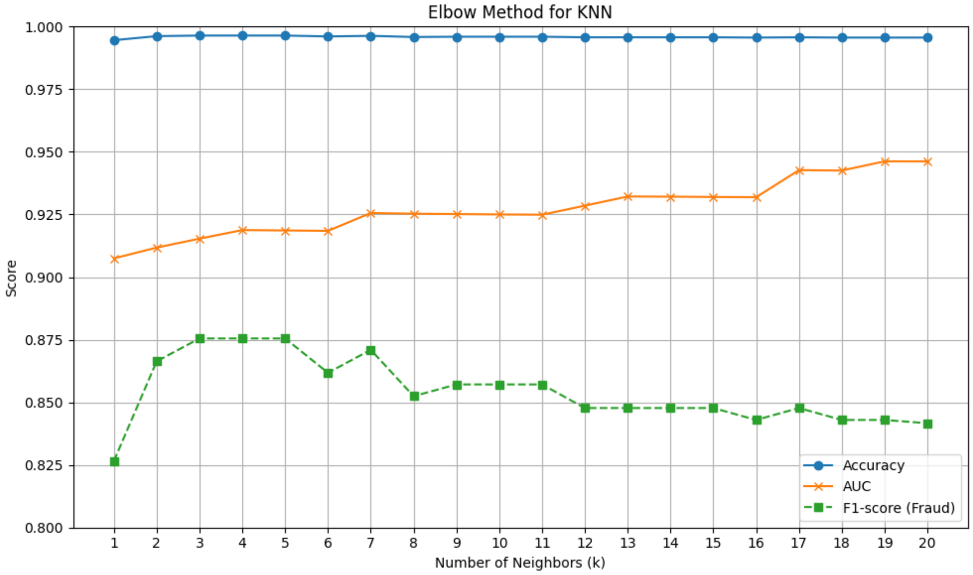
	60/40	70/30 ← more data to learn
Accuracy	1.00	1.00
Precision	6.97	0.97
Recall	0.80	0.79 (slightly decreased)
F1-Score	0.88	0.87
AUC	0.9186	0.9287. (slightly improved)

c) Keeping the test size fixed at 40%, try changing the number of neighbors (in KNN). How does the model's performance vary with different K values? Which value gives the best result, and how do you define what makes it the "best"?

I tested different values of k (from 1 to 20) to see how the number of neighbors affects the KNN model's performance. I found that accuracy stayed high for all values, but the F1-score for the fraud class was highest around k = 3 to 5, which means the model was best at detecting fraud in that range. On the other hand, the AUC score, which measures how well the model separates the two classes, was highest at k = 19 and 20.

So, the best value of k depends on what we care about more. If we want to catch more fraud cases accurately, then k = 5 is a good choice because it has a high F1-score and good AUC. But if we want the overall best separation between fraud and non-fraud, k = 19 or 20 gives the best AUC.

In this case, I would choose k = 5 as the best value because it balances both F1-score and AUC, and helps detect fraud more effectively.


Elbow Method for KNN
Score
Number of Neighbors (k)
Accuracy
AUC
F1-score (Fraud)