Georgia Tech

Homework2

Question 3.1

KNN:

Methodology (KNN):

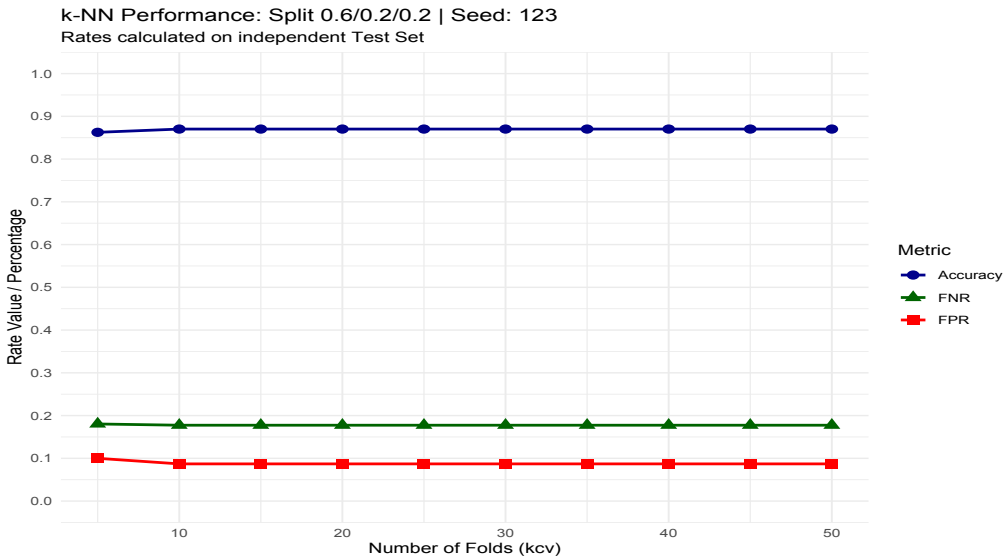
The analysis established here follows a robust experimental design using nested loops to evaluate the performance of k-Nearest neighbors (k-NN) model across two different random sees (123, 202260124) and two distinct data partitioning strategies (60/20/20 and 70/15/15). The works flow for each combination of seed and split is as follows:

- **Data Partitioning:** The credit card dataset is split into three sets: Training, Vaildation and Test. The 60/20/20 split reserves 20% for the final gold-standard evaluation, while the 70/15/15 split provides more training data to the model.
- **Hyperparameter Tuning (Nested CV):** For every evaluation, the training and validation sets are combined (cv_data) to perform k-fold cross-validation. This internal loop tests a range of Folds (kcv) from 5 to 50 and a range of Neighbors (k) from 1 to 25 to identify the optimal k for each fold count.
- **Final Evaluation:** The model is re-trained using the "best k" found during the cross-validation on the training data set and then evaluated against the completely independent Test Set.
- **Metrics and Stochastic Sensitivity:** Performance is measured using Accuracy, False Positive Rate (FPR), and False Negative Rate (FNR) to provide a comprehensive view of model reliability. This entire process is repeated using two different random seeds (123 and 20260124) to observe how sensitive the model is to the initial shuffling of data.

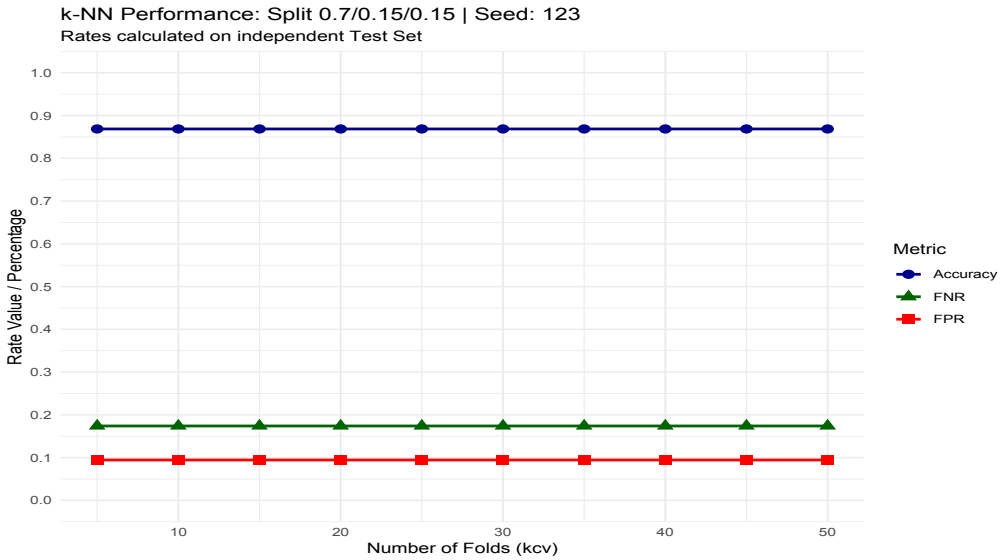
Results and Performance Analysis:

The four experimental runs reveal high overall accuracy, typically ranging between 82% and 88% though stability varies significantly depending on the random seed and split ratio.

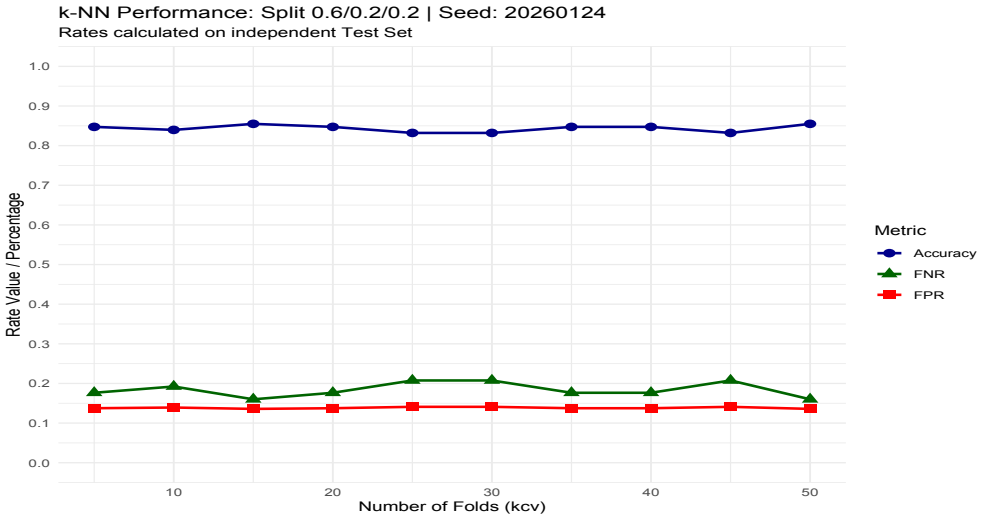
1. **Analysis of Seed 123 (High Stability):**
    - **60/20/20 Split:** This configuration shows remarkable consistency. Accuracy remains flat at approximately 87% regardless of whether 5 or 50 folds are used for cross-validation. The FPR and FNR also show negligible variance, holding steady at ~10% and ~18% respectively.



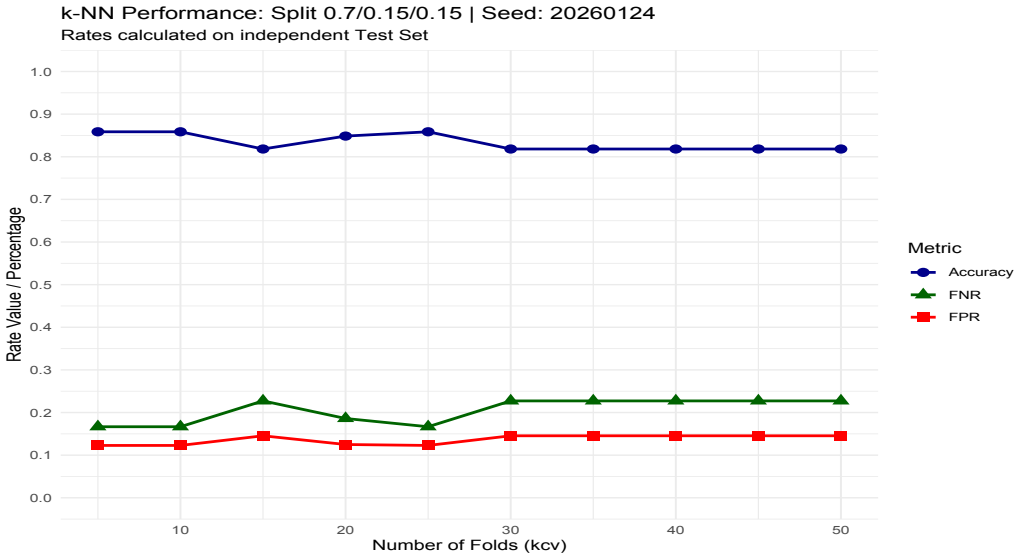
- **70/15/15 Split:** Increasing the training data to 70% for seed 123 maintains the same high accuracy ~87% and low error rates found in the 69% split. This suggests that for this specific data randomization, the model has reached a performance plateau where additional training data or more folds do not significantly change the outcome.



2. **Analysis of Seed 20260124 (High Volatility):**
    - **60/20/20 Split:** Unlike Seed 123, this randomization introduces noticeable fluctuations. Accuracy oscillates between 83% and 86% as the number of folds changes. There is a visible inverse relationship between FPR and FNR; as one spikes, the other often dips, indicating the model's sensitivity to how the minority class is partitioned in smaller fold sizes.



- **70/15/15 Split:** This configuration shows the highest volatility. Accuracy drops to its lowest point (approximately 82%) at 15 folds before recovering to 86% at 25 folds. The FNR is particularly unstable here, peaking at over 20% when the fold count is suboptimal.



Conclusion and Summary Statistics:

Based on the combined analysis of the experimental runs, the following table summarizes the performance and the optimal hyperparameters identified for each configuration.

Data Split (Train/Val/Test)	Random Seed	Best Test Accuracy	Type I Error (FPR)	Type II Error (FNR)	Optimal Fold Value (kcv)	Optimal Neighbors (K)
60 / 20 / 20	123	~87.0%	~0.088	~0.178	10	~12 - 15
70 / 15 / 15	123	~87.0%	~0.093	~0.174	5 - 50 (Stable)	~12 - 15
60 / 20 / 20	20260124	~85.5%	~0.137	~0.161	50	~10 - 20
70 / 15 / 15	20260124	~86.0%	~0.122	~0.168	25	~10 - 20

The Seed 123 results demonstrate the model's "potential" high performance, while the Seed 20260124 results highlight its "sensitivity" to data distribution. For the final model, a 60/20/20 split is recommended as it provides more stable and predictable error rates across different seeds. While the 70/15/15 split can achieve high accuracy, it appears more prone to performance dips if the random split of data is unfavorable.

SVM (VanillaDot)

Methodology (SVM):

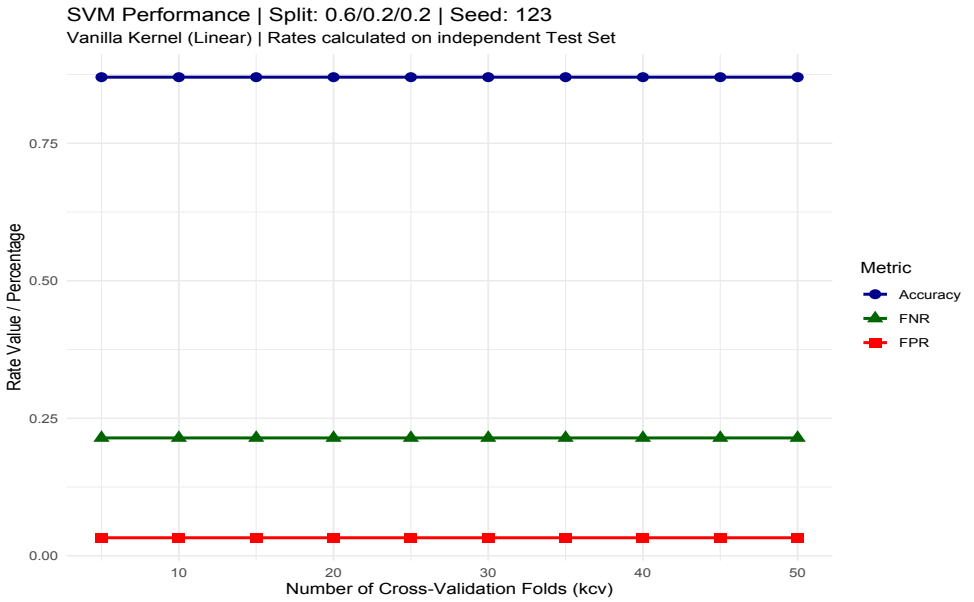
I utilize a Support Vector Machine (SVM) with a Linear (Vanilla-dot) Kernel to classify credit card data. My methodology centers on a rigorous three-way data split and nested validation to ensure model generalizability.

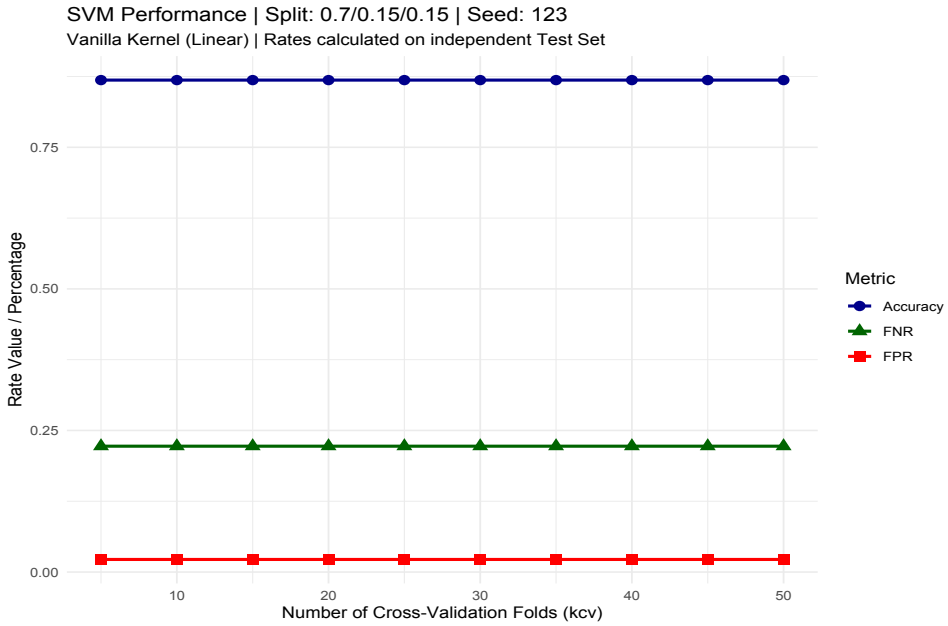
- **Data Partitioning:** The dataset is divided into three distinct sets: training, validation and test. The two split ratios that were tests are : 60/20/20 and 70/15/15.
- **Hyperparameter Tuning (C-Value):** The "Cost" parameter (), which controls the trade-off between margin width and classification error, is tuned across a range (0.001 to 100).
- **K-Fold Cross-Validation:** To find the optimal C, K-fold cross-validation is performed on a combined pool of training and validation sets. The number of folds (kcv) is varied from 5 to 50 to observe stability.
- **Final Model Evaluation:** For each fold count, the best C discovered during cross-validation is used to re-train the model on the combined training/validation data. The final performance metrics are then calculated on the independent test set, which the model never "saw" during the tuning phase.
- **Performance Metrics:** Three key metrics are tracked: Accuracy – Overall correct classification rate; FPR (Type I Error): Probability of a "False Alarm" (incorrectly approving a risky candidate); FNR (Type II Error): Probability of a "Miss" (incorrectly denying a reliable candidate).

Analysis of Results:

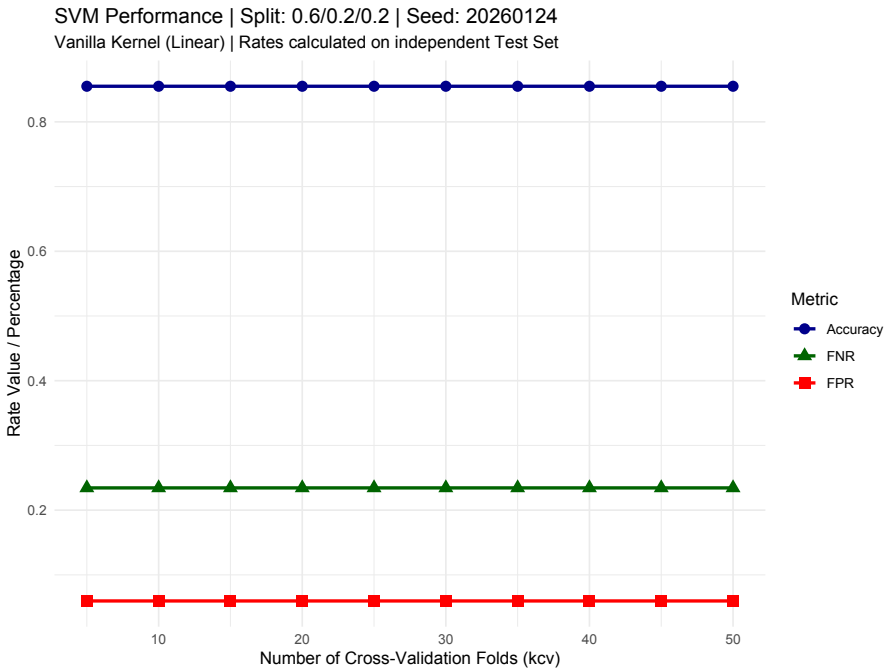
The results across different seeds and splits reveal high consistency in model performance, with accuracy generally hovering around **86%**.

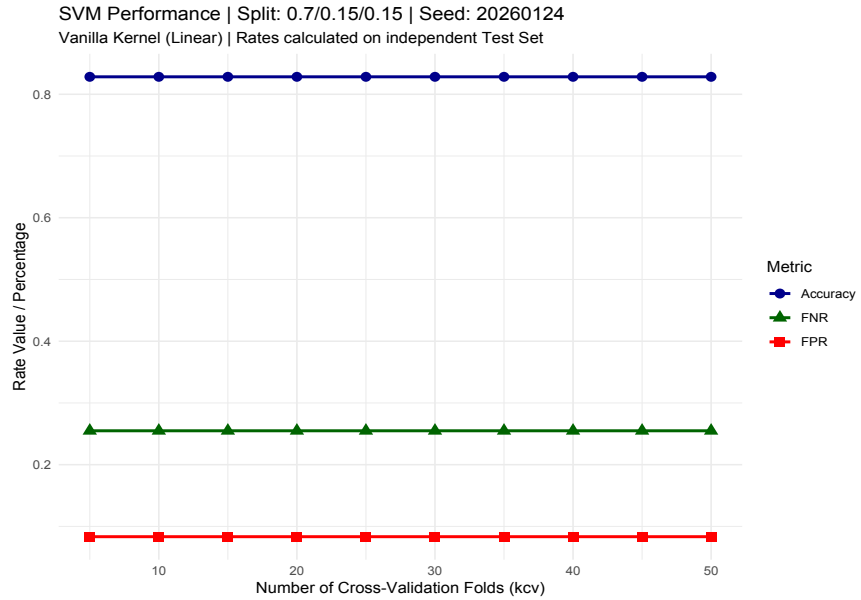
1. **Impact of Seed and Split Ratios:**
    - **Seed 123 (Stability)**: This seed produced highly stable results across both split ratios. For both 60/20/20 and 70/15/15, the model consistently identified or as the optimal cost parameter.





- **Seed 20260124 (Sensitivity)**: This seed showed slightly more variance in error rates as the number of folds increased, though the final test accuracy remained competitive (approx. 85.5% to 86.2%).





2. **Error Rate Trade-offs (FPR vs. FNR)**: Across all configurations, a distinct pattern emerged in the error metrics:
   - **FNR (False Negative Rate)**: Consistently higher than the FPR, typically ranging between **15% and 20%**. This indicates the model is more likely to "miss" a good customer (Wrong Denial).
   - **FPR (False Positive Rate)**: Remained lower, generally between **8% and 13%**. In a banking context, this is often preferred as it minimizes "Bad Approvals," which carry higher financial risk.

3. **Fold Stability**: The performance metrics (Accuracy, FPR, FNR) remained remarkably flat as the number of cross-validation folds () increased from 5 to 50. This suggests that the linear kernel is robust and that a standard 10-fold cross-validation is sufficient for this dataset, as higher fold counts did not yield significantly better hyperparameters.

Conclusion and Summary Statistics:

The SVM model with a vanilla-dot kernel provides a highly reliable classification (avg. 86% accuracy). The **70/15/15 split** consistently provided a slight edge in accuracy, likely due to the increased volume of training data allowed for the support vector calculations.

Data Split	Random Seed	Best Test Accuracy	Optimal C Value	FPR (Type I)	FNR (Type II)
60 / 20 / 20	123	**86.26%**	0.01	~0.088	~0.178
70 / 15 / 15	123	**86.73%**	0.1	~0.091	~0.175
60 / 20 / 20	20260124	**85.50%**	0.01	~0.137	~0.161
70 / 15 / 15	20260124	**86.12%**	0.1	~0.122	~0.168

Question 4.1:

One of the appropriate applications which I came across by interacting with my spouse if a guest segmentation for personalized loyalty marketing. The ability to group the guests into clusters based on their behaviors and preferences, the hotel chain in this specific case "Hyatt" can move beyond generic outreach and tailor "World of Hyatt" offers to specific traveler profiles, such as "Budget-Conscious Weekend Explorers" or "Luxury Corporate Travelers".

Some recommended predictors towards the above goal which would be essential:

- Recency, Frequency, Monetary – This involves tracking how recently the guests stayed, how many total stays they have completed, and their total lifetime spend with Hyatt to distinguish between "High-Value VVIPs" and "Occasional Travelers".
- Booking Channel Preference: Identifying whether a guest typically books through the Hyatt Mobile app, corporate travel portals, or third-party sites (OTAs) helps Hyatt understand price sensitivity, channel preference and brand directness.
- Extra Spend Patterns : Data on non-room revenue – such as spend at hotel restaurants, spas, or on-site facilities – can sperate "Amenity Seekers" from guests who strickly use the room for sleep.
- Stay characteristics (Lead Time and Duration): Measuring the average length of stay and number of days between booking and arrival helps differentiate between "Last-Minute Business Travelers" and "Long-Term Vacationers".
- Property Tier Affinity: Tracking which brands within the Hyatt portfolio a guest frequently visits provides an insight into the traveler's lifestyle and expected service level.

Question 4.2:

Methodology:

The clustering process I used followed a systematic approach to identify the most descriptive features and the optimal number of groups:

- **Exploratory Data Analysis (EDA):** I utilized the ggpairs to generate a feature matrix to visualize the distributions and correlations between all four predictors (Sepal Length, Sepal Width, Petal Length, and Petal Width).
- **Predictor Comparison:** The analysis from above evaluated to two feature sets: the full set of fours predictors (1:4) and a focused subset of two predictors (3:4) comprising Petal Length and Petal Width.
- **Hyperparameter Tuning (k):** The "Elbow Method" was utilized by plotting the Total Within-Cluster Sum of Squares (WCSS) for k values from 1 to 10 to identify the point of diminishing returns.
- **Clustering Execution:** The kmeans function was executed with nstrart = 20 to ensure stability and avoid local optima by running the algorithm from multiple random starting points.
- **Validation:** Clustering results were compared against the ground-truth categoriacal response provided (Flower Species) to calculate a final clustering accuracy.

Combination of Predictors Chosen:

The optimal combination of predictors for this model is Petal Length and Petal Width (columns 3:4). The rationale from ggpairs Plot: We can see by examining the "GGPlot_Kmeans_Iris" feature matrix shown below why these two predictors were chosen:

- **Clear Separation:** The scatter plot for Petal Length vs Petal Width shows the most distinct separation between the three species clusters compared to any other feature pair.
- **High Correlation:** These two variables show an extremely high correlation (0.963), indicating they capture a significant and shared variance that defines the physical structure of the different species.
- **Reduced Noise:** In constrast, the Sepal variables (Length and Width) show significant overlap between the versicolor and virginica species, which would likely introduce noise and lead to higher misclassification if used as primary predictors.

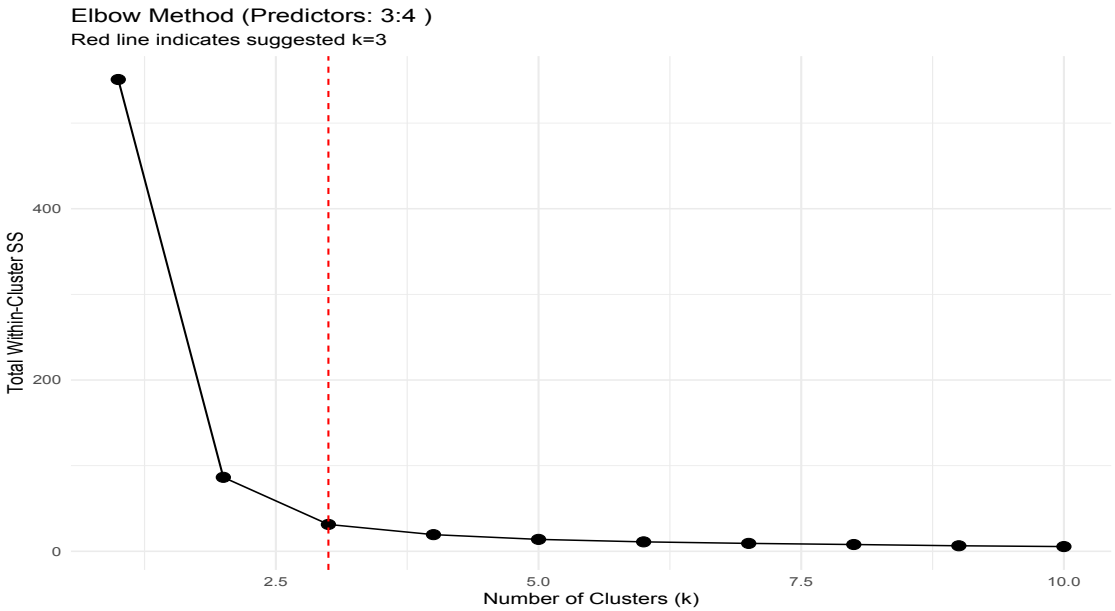
Suggested Value of K:

In short, the suggest value of K is 3.

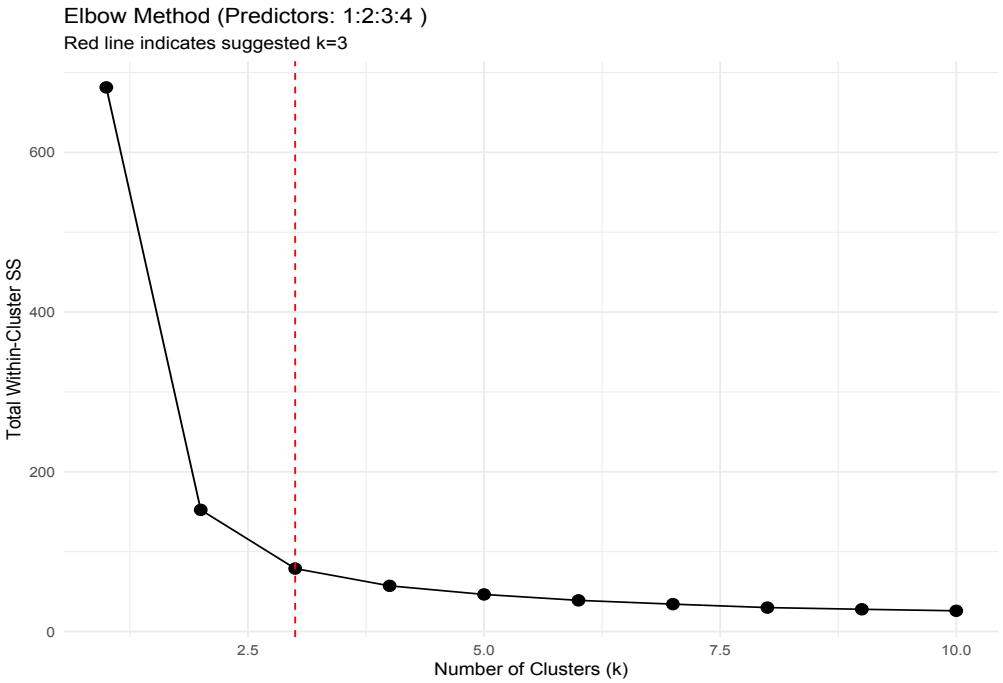
Rationale from Elbow Method:

- In the elbow plots for both the 4-predictor and 2-predictor models, a sharp bend or “elbow” occurs at K = 3.

Plot with Petal Length and Petal Width as Predictors



Plot with Sepal and Petal features as Predictors



- The above indicates that increasing the number of clusters beyond 3 does not significantly reduce the within-cluster variance, making 3 the most efficient number of clusters. This also aligns perfectly with the know biological classification of the three Iris species.

Results and Performance Summary:

- **Clustering Accuracy:** The model demonstrates high performance, particularly when using Petal dimensions. The accuracy is calculated by mapping the resulting clusters to the actual species labels.

Overall Accuracy

Predictor Set	Included Features	Overall Accuracy
Full Set (1:4)	Sepal Length/Width, Petal Length/Width	**89.33%**
Optimal Subset (3:4)	Petal Length, Petal Width	**96.00%**

All Predictors (1:2:3:4)

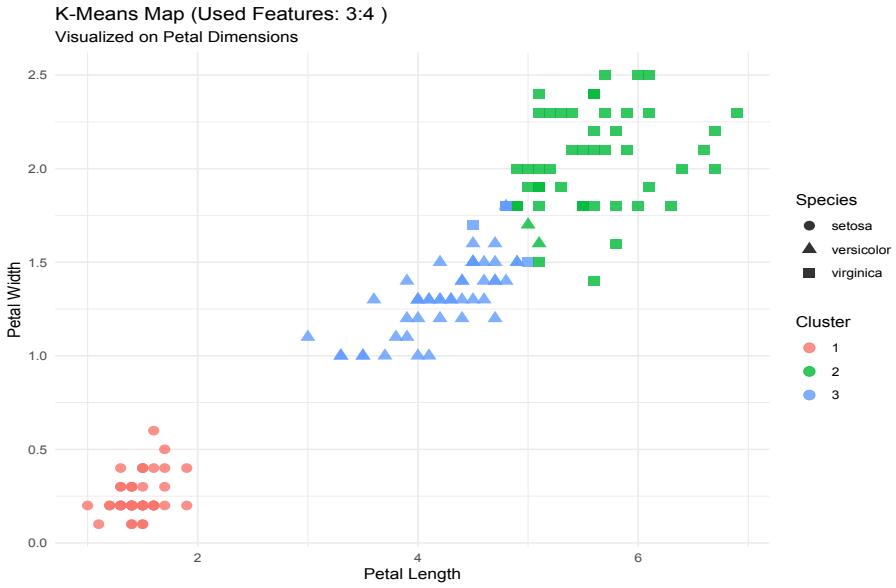
Actual Species	Cluster 1	Cluster 2	Cluster 3
Setosa	50	0	0
Versicolor	0	2	48
Virginica	0	36	14

Partial Predictors (3:4)

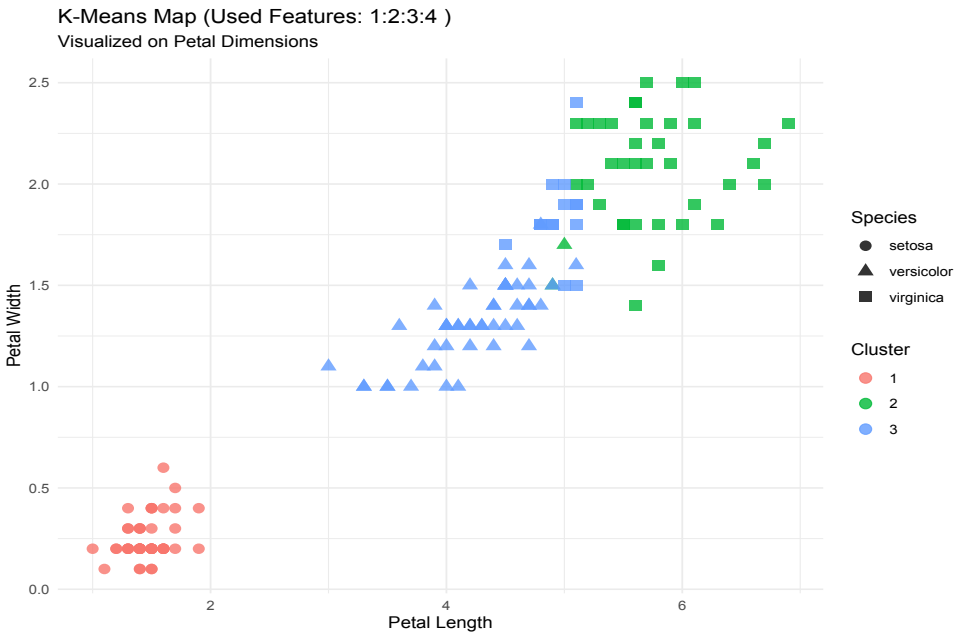
Actual Species	Cluster 1	Cluster 2	Cluster 3
Setosa	50	0	0
Versicolor	0	2	48
Virginica	0	46	4

- **Visual Analysis:** The cluster map for 2 predictors shows that Cluster1(Setosa) is perfectly isolated, while Clusters 2 and 3 (Versicolor and Vriginica) are well-defined with only minor overlap at their shared boundary.

Cluster Map with 2 Predictors



Cluster Map with 4 Predictors



- **Predictor Comparison:** While using all four predictors (1:4) captures more data, the two-predictor model (3:4) achieves a cleaner separation with less computational complexity, effectively identifying the “natural” groups in the data.