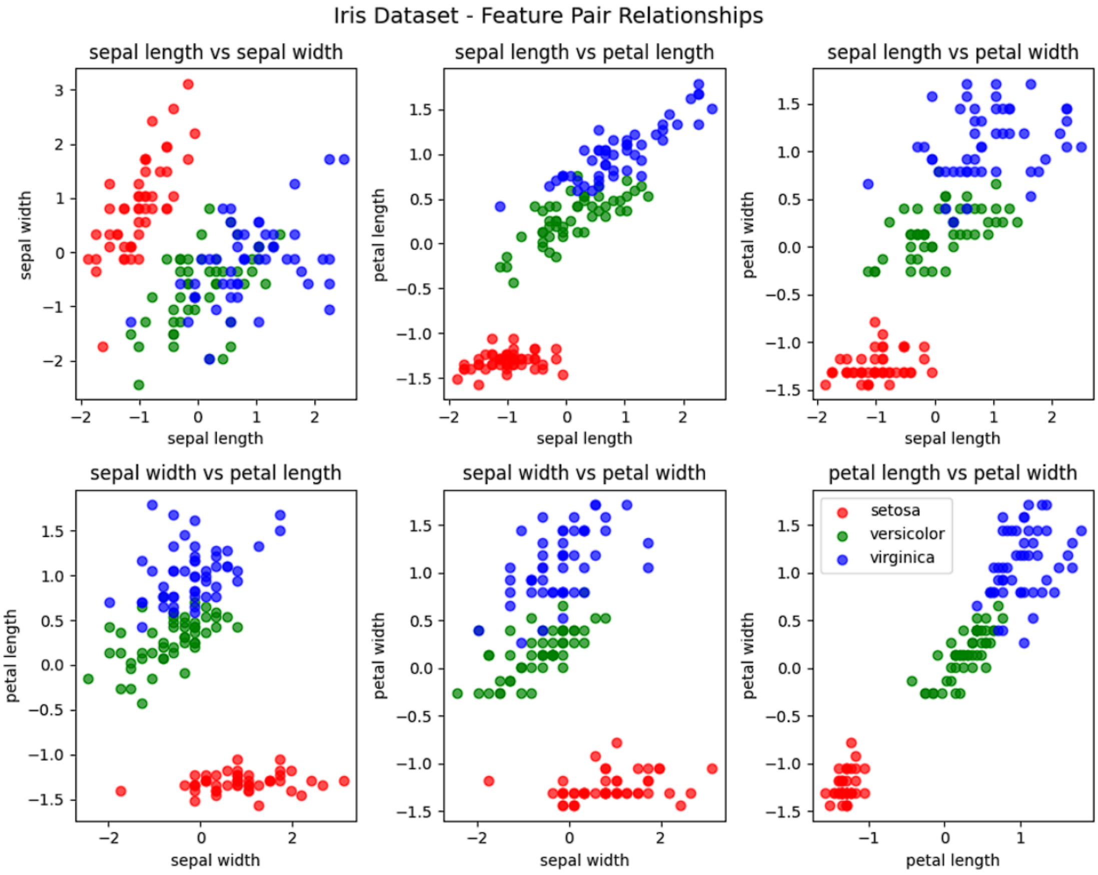
K-Nearest Neighbors (KNN) Implementation from Scratch

[**Student Name:** Anish Kumar **Student ID:** 23/CS/057 **Course:** BTech CSE]

1. Analysis of Exploratory Data Analysis (EDA)

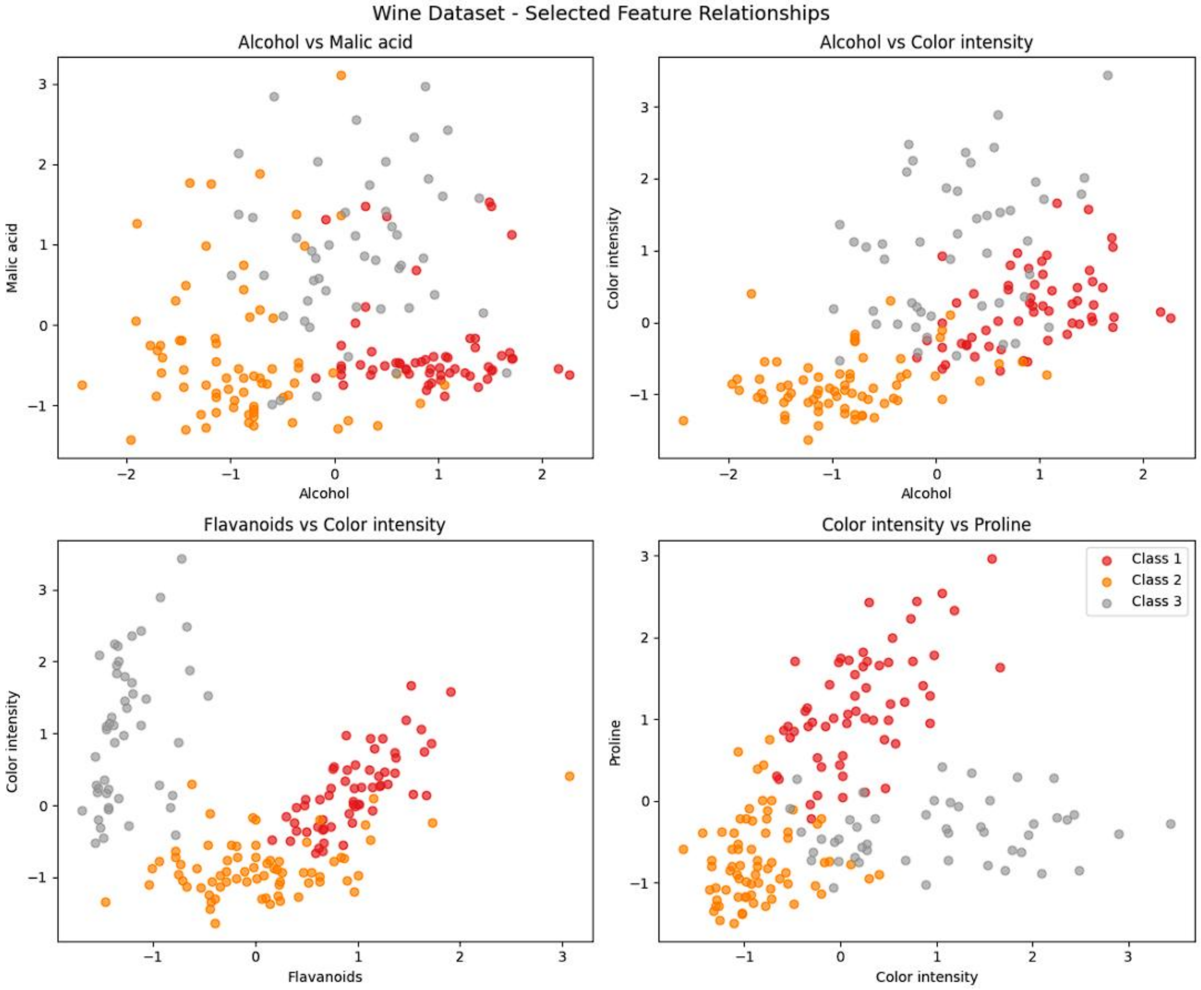
Iris Dataset EDA



Analysis: The Exploratory Data Analysis (EDA) for the Iris dataset involved plotting pairs of all four features (sepal length, sepal width, petal length, petal width) against each other, color-coded by their species.

- **Best Class Separation:** The feature pair that provides the best class separation is clearly **"petal length vs. petal width"**. In this plot, the 'setosa' class is perfectly separated from 'versicolor' and 'virginica' with a simple linear boundary.
- **Ease of Distinction:** The **'setosa'** class is inherently the easiest to distinguish. In all plots involving petal length or petal width, it forms a tight, isolated cluster that does not overlap with the other two classes. The 'versicolor' and 'virginica' classes show some overlap, particularly in the sepal-related plots, making them harder to distinguish from each other.

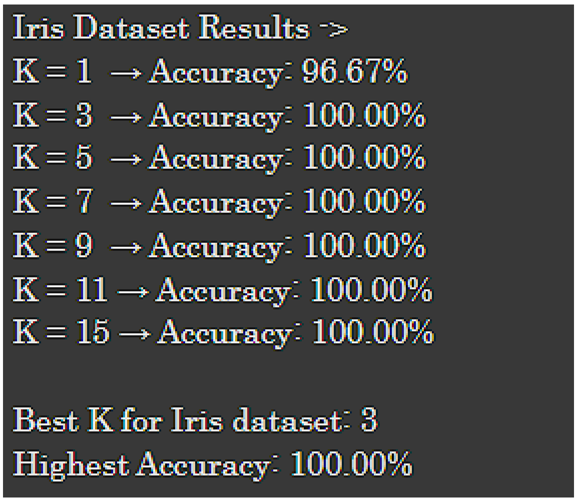
Wine Dataset EDA



Analysis: The EDA for the Wine dataset shows the relationships between selected feature pairs.

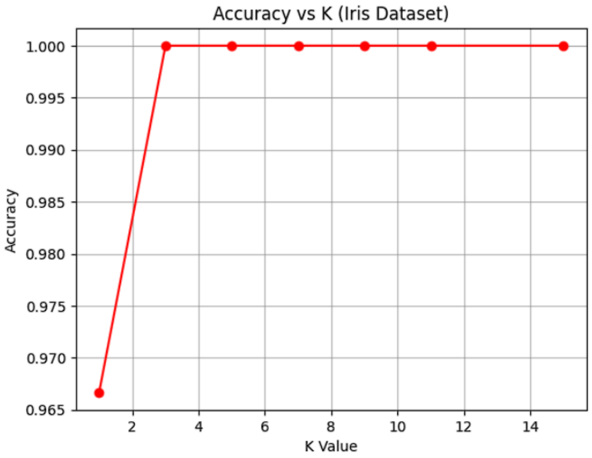
- **Best Class Separation:** Of the pairs plotted, **"Alcohol vs. Color intensity"** and **"Color intensity vs. Proline"** appear to provide the most significant class separation. In these plots, the three classes form relatively distinct clusters. For example, in the "Alcohol vs. Color intensity" plot, Class 2 (orange) is clustered in the bottom-left (low alcohol, low intensity), while Class 1 (red) is generally in the top-right (high alcohol, high intensity), with Class 3 (grey) occupying the middle-to-top-left region.
- **Ease of Distinction:** Unlike the Iris dataset, no single class is perfectly isolated. However, **Class 2** consistently forms a more distinct cluster that is separated from Classes 1 and 3. Classes 1 and 3 show a greater degree of overlap in these feature spaces, suggesting they are more similar and will be more difficult for the classifier to separate.

2. Iris Dataset: Classification Results



```
Iris Dataset Results ->
K = 1  → Accuracy: 96.67%
K = 3  → Accuracy: 100.00%
K = 5  → Accuracy: 100.00%
K = 7  → Accuracy: 100.00%
K = 9  → Accuracy: 100.00%
K = 11 → Accuracy: 100.00%
K = 15 → Accuracy: 100.00%

Best K for Iris dataset: 3
Highest Accuracy: 100.00%
```



Final Accuracy (k=3)

The final classification accuracy achieved on the Iris dataset using the implemented KNN algorithm with **k=3** was:

Accuracy: 100.00%

Analysis of "Accuracy vs. k-value" Plot

- **Best Performance:** The plot shows that the best performance (100.00% accuracy) was first achieved at **k=3** and was maintained for all tested k-values (k=5, 7, 9, 11, 15). Given this, **k=3** is the optimal choice as it represents the simplest model (lowest k) that achieves maximum accuracy.
- **Sub-optimal k-values:**
  - **Very Small k (e.g., k=1):** A very small k value, like k=1, makes the model highly sensitive to noise and outliers. The classification of a new point is determined by its *single* nearest neighbor. If that one neighbor happens to be a mislabeled data point or an anomaly, the prediction will be incorrect. This is known as **high variance** or **overfitting**. As seen in the results, k=1 yielded a lower accuracy (96.67%) than k=3, demonstrating this sub-optimality.
  - **Very Large k:** A very large k value makes the model overly general. The decision boundary becomes "too smooth," and the model loses its ability to capture local patterns in the data. It essentially begins to predict the majority class of the entire dataset.

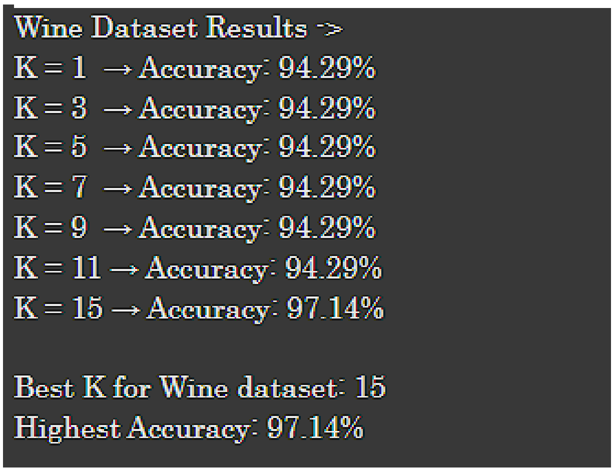
This is known as **high bias** or **underfitting**. While accuracy did not drop in this specific experiment for Iris (likely due to the dataset's high separability), a large k often performs poorly on more complex datasets.

3. Wine Dataset: Classification Results

Final Accuracy

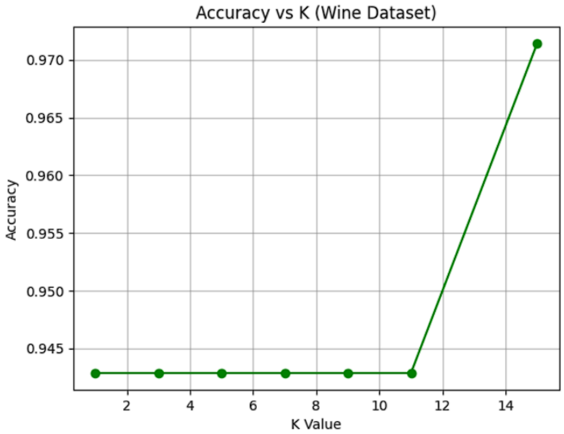
The final (and highest) classification accuracy achieved on the Wine dataset was:

Accuracy: 97.14% (at k=15)



```
Wine Dataset Results ->
K = 1  → Accuracy: 94.29%
K = 3  → Accuracy: 94.29%
K = 5  → Accuracy: 94.29%
K = 7  → Accuracy: 94.29%
K = 9  → Accuracy: 94.29%
K = 11 → Accuracy: 94.29%
K = 15 → Accuracy: 97.14%

Best K for Wine dataset: 15
Highest Accuracy: 97.14%
```

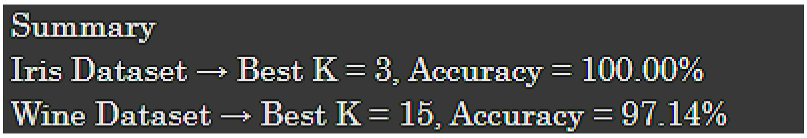


Analysis of "Accuracy vs. k-value" Plot

The results for the Wine dataset show a different pattern. The accuracy remained stable at ~94.29% for k values from 1 to 11. The accuracy then increased to **97.14%** at **k=15**.

This suggests that, for the Wine dataset, a "smoother" decision boundary (a larger k) is beneficial. This aligns with the EDA, which showed more overlap between classes. A larger k considers more neighbors, making the model more robust to the noise and intermingling of data points at the class boundaries.

4. Conclusion and Key Learnings



```
Summary
Iris Dataset → Best K = 3, Accuracy = 100.00%
Wine Dataset → Best K = 15, Accuracy = 97.14%
```

This lab involved the successful implementation of the K-Nearest Neighbors (KNN) algorithm from scratch. The process provided several key insights:

1. **EDA is Crucial:** The preliminary EDA was vital for understanding the data. It immediately highlighted which features would be most informative for classification (e.g., petal dimensions for Iris) and which classes would be easy ('setosa') or difficult ('versicolor' vs. 'virginica') to separate.
2. **KNN Implementation:** The core of the lab involved implementing the Euclidean distance calculation and a "voting" mechanism. The main challenge was efficiently calculating distances between all test points and all training points, sorting these distances, and identifying the labels of the k closest neighbors to determine the majority class.
3. **k is a Critical Hyperparameter:** The experiments clearly demonstrated that k is not a one-size-fits-all parameter. The optimal k is data-dependent.
    - The **Iris dataset**, being highly separable, performed best with a small k (k=3), which created a precise decision boundary.
    - The **Wine dataset**, with more class overlap, performed best with a larger k (k=15), which created a smoother, more generalized boundary that was less susceptible to noise.
4. **Bias-Variance Trade-off:** This lab was a practical demonstration of the bias-variance trade-off. A small k (like 1) has low bias but high variance (overfits), while a large k has high bias and low variance (underfits). The goal is to find a k that balances this trade-off, which our experiments successfully identified for both datasets.