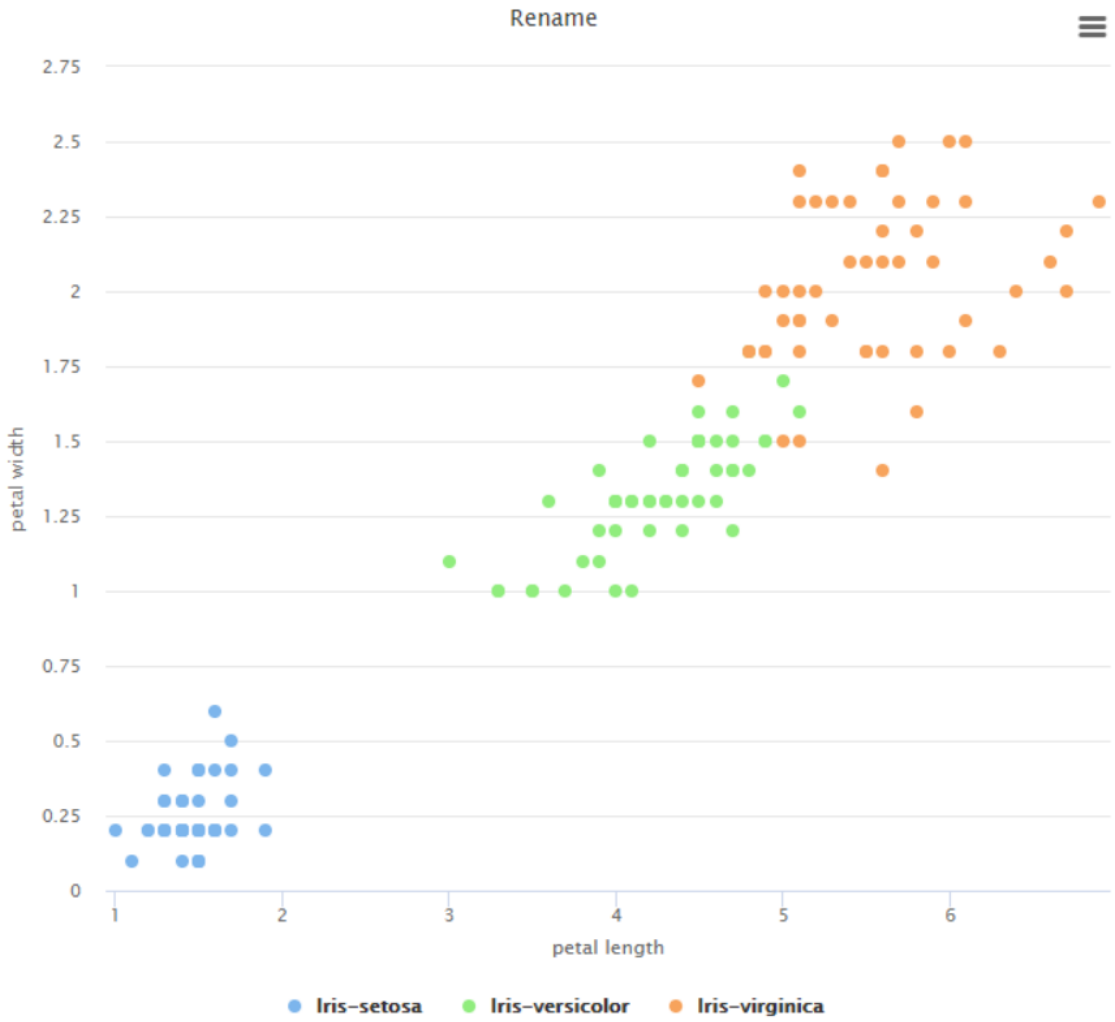
Ejercicio 2



Iris setosa se diferencia perfectamente de iris versicolor e iris virginica, estas últimas dos clases están bastante cerca.

Para preparar los datos voy a quitar los atributos que no son los usados en la gráfica, para que se refleje la gráfica hecha. Además, se estandarizarán los datos para evitar que las magnitudes de los datos incidan en los cálculos de distancia.

El parámetro Weighted Vote hace que la distancia entre el punto objetivo y sus vecinos sea considerada en la decisión de la clase, de esta forma los vecinos más cercanos tienen más importancia.

Tipos de medición:

El parámetro se usa para seleccionar el tipo de la medida usada para encontrar a los vecinos más cercanos. Las opciones son:

- **MixedMeasures**: Se usa para calcular las distancias en caso de atributos nominales y numéricas. Opciones de medida:

    - **Mixed Measure:** Se usa la distancia euclideana para atributos numéricos. Para valores nominales se toma una distancia 0 cuando las instancias tienen el mismo valor, de lo contrario se toma 1.
- **NominalMeasures:** Se usa en caso de tener solo atributos nominales.

  e: Número de atributos para el cual ambos ejemplos tienen valores iguales distintos de cero.

  u: Número de atributos para el cual las dos instancias tienen valores distintos.

  z: Número de atributos cuyo valor es cero.

    - **Nominal Distance:** La distancia de los valores es 0 si son iguales o 1 si son distintos.
    - **Dice Similarity:** 2*e/(2*e+u)
    - **JaccardSimilarity:** e/(e+u)
    - **KulczynskiSimilarity:** e/u
    - **RogersTanimotoSimilarity:** (e+z)/(e+2*u+z)
    - **RussellRaoSimilarity:** e/(e+u+z)
    - **SimpleMatchingSimilarity:** (e+z)/(e+u+z)
- **NumericalMeasures:** En caso de que se usen solo atributos numéricos.

Y(I,j) = Valor del j ésimo atributo the la i ésima fila.

    - **EuclideanDistance:** Dist = Sqrt (Sum(j=1) [y(1,j) – y(2,j)]^2). Es la suma de las diferencias cuadráticas entre todos los atributos.
    - **CamberraDistance:** La suma de todos los atributos. El sumando es el valor absoluto de la resta entre el valor, dividido entre la sumatoria de los valores absolutos. Dist = Sum_(j=1) |y(1,j)-y(2,j)| / (|y(1,j)|+|y(2,j)|)
    - **ChebychevDistance:** El máximo de las diferencias entre atributos. Dist = max_(j=1) (|y(1,j)-y(2,j)|)
    - **CorrelationSimilarity:** La similitud es calculada como la correlación entre los vectores de atributos de las dos instancias.
    - **CosineSimilarity:** Medida de similitud que mide el cose del ángulo entre los vectores atributo de las dos instancias.
    - **DiceSimilarity:** La similitud del dado para atributos numéricos se calcula como 2*Y1Y2/(Y1+Y2). Y1Y2 = sumatoria sobre el producto de los valores = Sum_(j=1) y(1,j)*y(2,j). Y1 = sumatoria sobre los valores del primer ejemplo = Sum_(j=1) y(1,j) Y2 = sumatoria sobre los valores de la segunda instancia = Sum_(j=1) y(2,j)
    - **DynamicTimeWarpingDistance:** Se usa en el análisis de series de tiempo para medir la distancia entre dos secuencias temporales.
    - **InnerProductSimilarity:** La similitud es calculada como la sumatoria del producto de los vectores atributo de las dos instancias: Dist = -Similitud = Sum (j=1) y(1,j)*y(2,j).
    - **JaccardSimilarity:** Se calcula como Y1Y2/(Y1+Y2-Y1Y2).
    - **KernelEuclideanDistance:** La distancia se calcula como la distancia euclideana de los dos ejemplos en un espacio transformado. La transformación es definida por el kernel elegido que puede ser: gamma, sigma1, sigma2, sigma3, shift, degree, a, b.

    - **ManhattanDistance:** La sumatoria de las distancias absolutas de los valores de los atributos. Dist = Sum_(j=1) |y(1,j)-y(2,j)|
    - **MaxProductSimilarity:** La similitud es el máximo de todos los productos de los valores de los atributos. Si el máximo es menor o igual a cero la similitud no está definida. Dist = - Similitud = -max (j=1) (y(1,j)*y(2,j))
    - **OverlapSimilarity:** La similitude es una variante de matcheo simple para atributos numéricos y se calcula como minY1Y2 / min(Y1,Y2).
- **BregmannDivergences:** Son tipos de medida genéricos de "lejanía" que no satisfacen la desigualdad de triangulo o simetría.
    - **GeneralizedIDivergence:** Se calcula como sum1 * sum2. Esta técnica no se puede usar con valores negativos. Sum1 = Sum_(j=1) y(1,j)*ln[y(1,j)/y(2,j)] Sum2 = Sum_(j=1) [y(1,j)-y(2,j)]
    - **ItakuraSaitoDistance:** Puede ser calculada solo con datsets con 1 atributo y valores positivos mayores a cero. Dist = y(1,1)/y(2,1)-ln[y(1,1)/y(2,1)]-1.
    - **KLDivergence:** La divergencia de Kullback-Leibler es una medida de como una distribución de probabilidad diverge de una segunda distribución de probabilidad. Dist = Sum_(j=1) [y(1,j)*log_2(y(1,j)/y(2,j))]
    - **LogarithmicLoss:** Se puede calcular en datasets con 1 atributo cuyos valores son todos positivos. Dist = y(1,1)*ln[y(1,1)/y(2,1)]-(y(1,1)-y(2,1))
    - **LogisticLoss:** Se puede calcular en datasets con 1 atributo cuyos valores son todos positivos. . Dist = y(1,1)*ln[y(1,1)/y(2,1)]+(1-y(1,1))*ln[(1-y(1,1))/(1-y(2,1))]
    - **MahalanobisDistance:** Mide la distancia entre dos instancias bajo la asunción de que son vectores randomicos de la misma distribución. Se calcula la matriz de covarianza S para todo el dataset y la distancia se calcula como: Dist = Sqrt [(vecY1-vecY2) S (vecY1-vecY2)] vecY1 = vector de atributo de instancia 1 vecY2 = vector atributo de instancia 2.
    - **SquaredEuclideanDistance:** La sumatoria de las diferencias cuadráticas entre todos los atributos. Dist = Sum_(j=1) [y(1,j)-y(2,j)]^2
    - **SquaredLoss:** Puede ser calculada solo con datsets con 1 atributo y valores positivos mayores a cero. Dist = [y(1,1)-y(2,1)]^2.

Resultados

Tecnica de medida	K	Performance
Mixed Euclidean Distance	3	96%
Mixed Euclidean Distance	5	96%
Cosine Similarity	3	80%
Cosine Similarity	5	77.3%

La medida de distancia euclideana se ve favorecida ante la disposición de las instancias del dataset, si se calcula la distancia con la similitud del coseno del ángulo entre instancias el algoritmo pierde precisión, los datos tienen cierta disposición "circular" entonces en los puntos borde se hace más difícil clasificar.