Proyecto Final

DOCENTE	CARRERA	CURSO
PhD(c) Vicente Machaca Arceda	Maestría en Ciencia de la Computación	Algoritmos y Estructura de Datos

1. Integrantes

- Grupo: 03
  - Edwin Fredy Chambi Mamani
  - Gludher Quispe Cotacallapa
  - Erwin Cruz Mamani
  - Omar Castillo Alarcón

2. Resumen

1. La estructura KD-Tree es una estructura multidimensional de k dimensiones.Esta permite implementar búsquedas por similitud como K Nearest Neighbor o Closest point. Adicionalmente, se puede usar esta estructura como un clasificador. Usted debe implementar este clasificador en el tema de su preferencia. A continuación detallamos lo realizado: Primeramente se hizo análisis del dataset iris y de diabetes del repositorio de datasets de la Universidad de Irvine. `https://archive.ics.uci.edu/ml/`.

   Sobre el clasificador KNN, se hizo dos pruebas , uno sobre datos de flor iris (dataset popular), y otro sobre el dataset diabetes, primeramente sin reducción de dimensiones, tomando las variables con mayor correlación entre ellas, esto se hizo con estadística. Seguidamente usando Análisis de componentes principales se hizo reducción de dimensiones para convertir los 4 parámetros del dataset de iris en solo 2, lo mismo se aplico para el caso del dataset de diabetes.

   a) Se tienen 4 resultados, los cuales son variantes del mismo clasificador:
      1) Un clasificador KNN para el dataset iris sin reducción de dimensiones, considerando solo las características de largo y ancho de petalos de la flor, así como las especies virginica y versicolor.
      2) Un clasificador KNN para el dataset diabetes usando las características que consideramos mas relevantes, la glucosa en ayunas y el índice de masa corporal.
      3) Un clasificador KNN para el dataset iris usando reducción dimensional por análisis de componentes principales para convertir los 5 parámetros mas a 2 mas etiqueta.
      4) Un clasificador KNN para el dataset diabetes usando reducción dimensional por análisis de componentes principales para convertir los 9 parámetros a 2 mas etiqueta.

   b) No se hizo ningún tipo de reducción de dimensiones, solo se tomaron dos parámetros mas correlacionados (en el caso de iris), donde para dataset iris se uso las variantes de iris-versicolor y la iris-virginica, que muestran en el scatter una separación compatible con el uso de KNN, no se uso la iris-setosa, esto para mantener simple el ejercicio. Para esta etapa según lo analizado los parámetros a usar en el el plano XY del KD Tree son *petal lenght* y *petal width*.

c) En el caso del dataset de diabetes. De la misma forma no se hizo reducción de dimensiones y de los parámetros del dataset, de un total de 8 mas el TAG de 1 y 0 de estado de diabetes, nos quedamos con las características IMC, Glucemia en ayunas y índice de masa corporal, *Glucose* y *BMI*.

d) Para la segunda parte del Para los dos datasets se hizo una normalización de los parámetros usando python, de esta manera los datos pasaron a tomar valores de 0 a 1 en punto flotante.

e) Esto se describe a continuación.

3. Materiales y Métodos

1. Análisis del dataset iris mediante *Orange3*. Se cargo el CSV del dataset iris a *Orange3* para el tratamiento estadístico y se ve que los pétalos son lo mas característico entre los tipos de iris.(Figura 1 y 2).

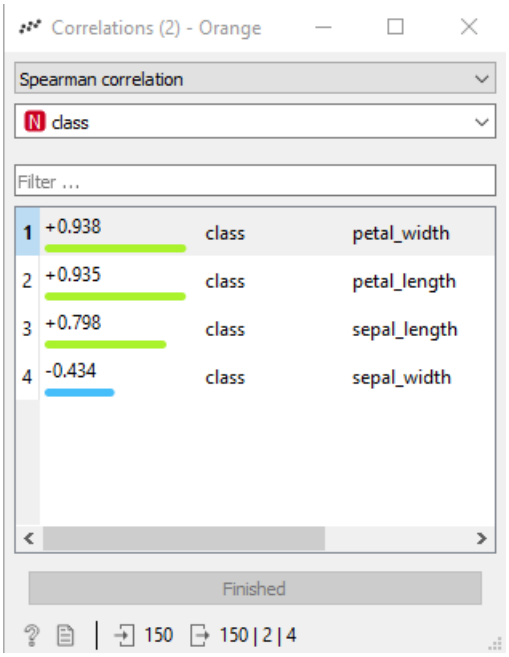

Figura 1: Observamos que los parámetros de ancho y alto de pétalo son los que aportan mas a la clasificación entre iris.

2. De la misma manera se analizó el dataset de diabetes, y del análisis se decidió usar los parámetros de BMI y Glucosa. Esto aun sin reduccion dimensional por PCA. (Figura 3).

3. El dataset resultante se normalizo y se uso en el KNN desarrollado en la practica anterior.

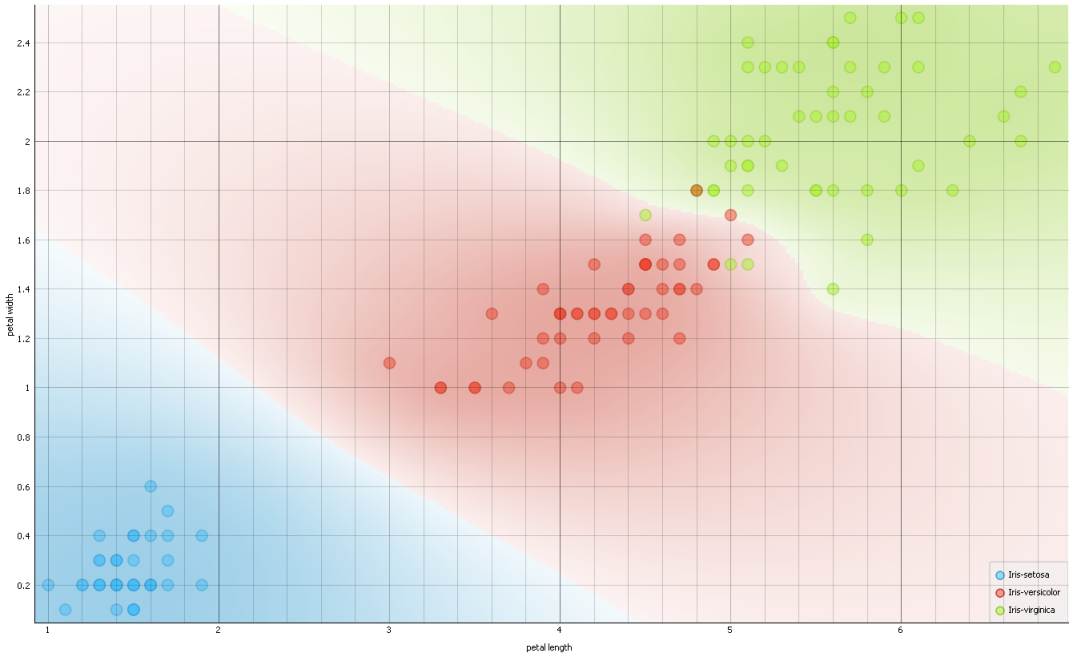

Figura 2: Observamos que los parámetros de ancho y alto de pétalo son los que aportan mas a la clasificación entre iris.

4. Para el tercer resultado, se uso Análisis de componente principales en el dataset iris, esto usando python y sklearn. De esta manera reducimos la dimensión del dataset de 4 características a dos componentes principales, y de la misma forma se eligió dos tipos de iris para mantener la clasificación como 0 y 1. Para el punto de consulta de la misma forma hay que convertir los datos a consultar a PCA y luego pasarlo al clasificador. Aparte un porcentaje de datos del dataset se reservo para pruebas.

5. Para el cuarto resultado también se uso PCA y se convirtió el dataset de 8 características a 2, y las clases de salida se mantuvieron como 0 y 1, es decir la presencia o ausencia de diabetes.

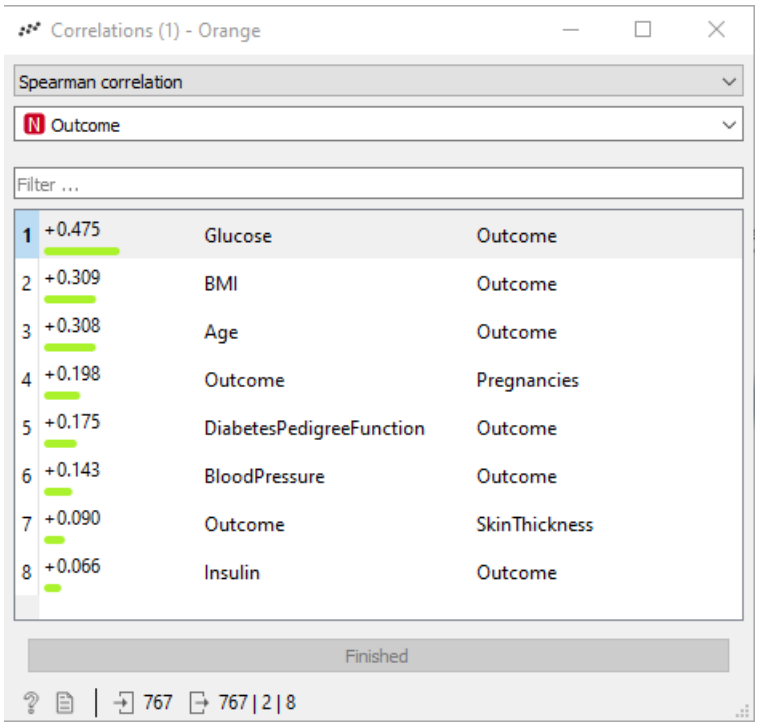

Figura 3: Observamos dos parámetros importantes, la glucosa y el Índice de masa corporal BMI (Body Mass Index).

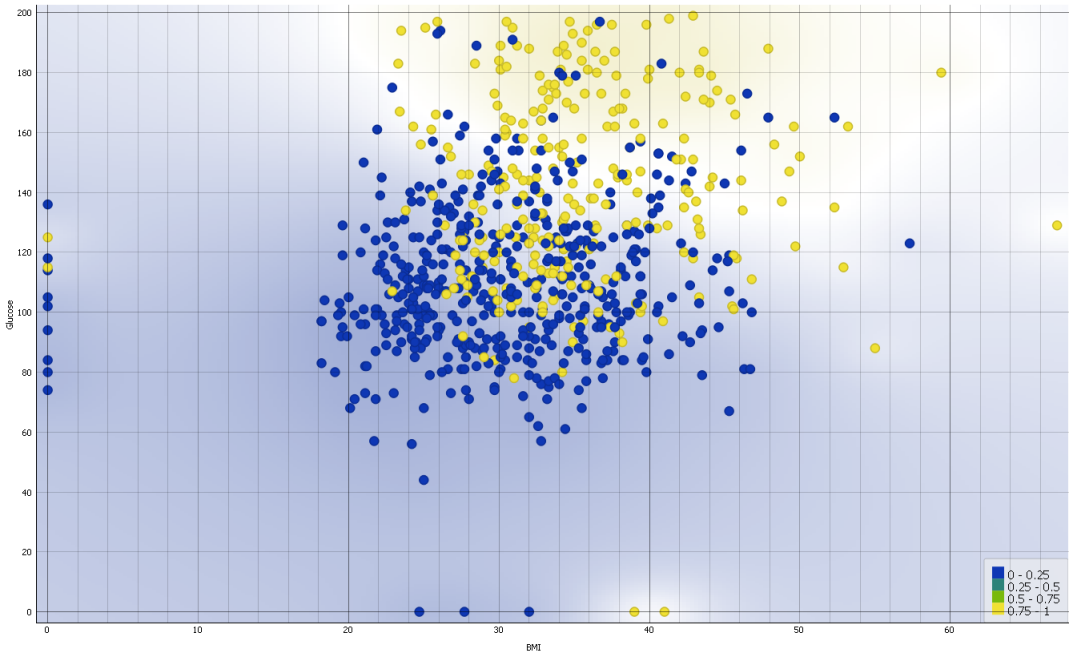

Figura 4: Observamos a los parámetros de Índice de masa corporal y de Glucosa en el plano.

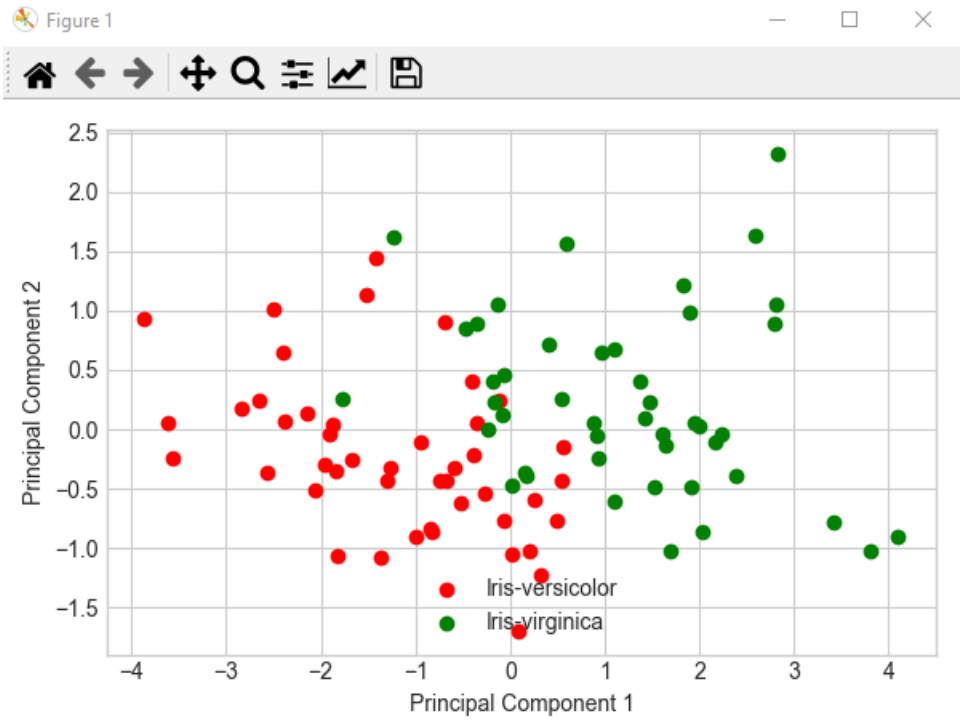

Figura 5: Gráfico Scatter del dataset iris con dos tipos de flor y con reducción dimensional a 2 con PCA, aun sin normalización.

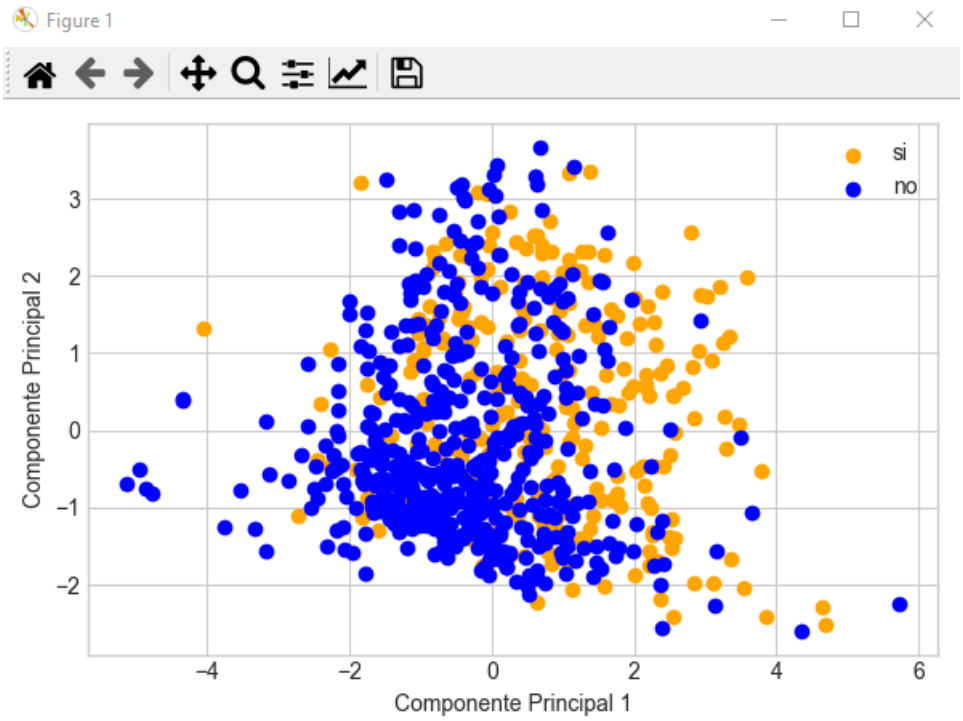

Figura 6: Gráfico Scatter del dataset diabetes con reducción dimensional a 2 mediante PCA, aun sin normalización para ser usado en el KNN.

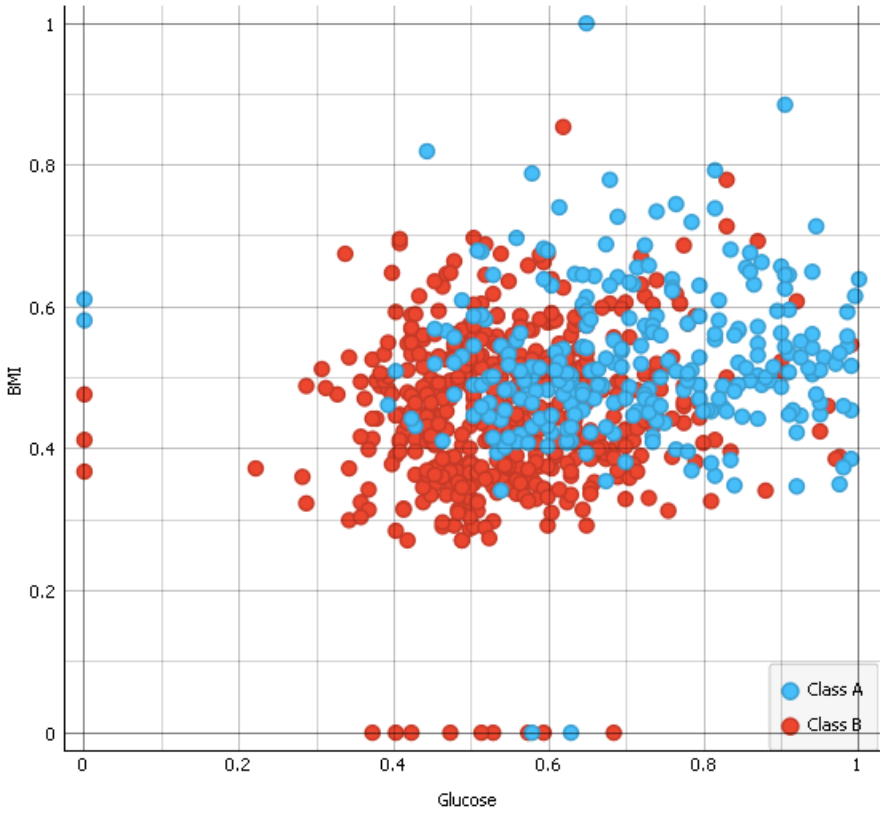

Figura 7: Gráfico Scatter del dataset diabetes con solo dos características del dataset, el IMC y la Glucosa (Clase A = Diabetes True). Datos Normalizados

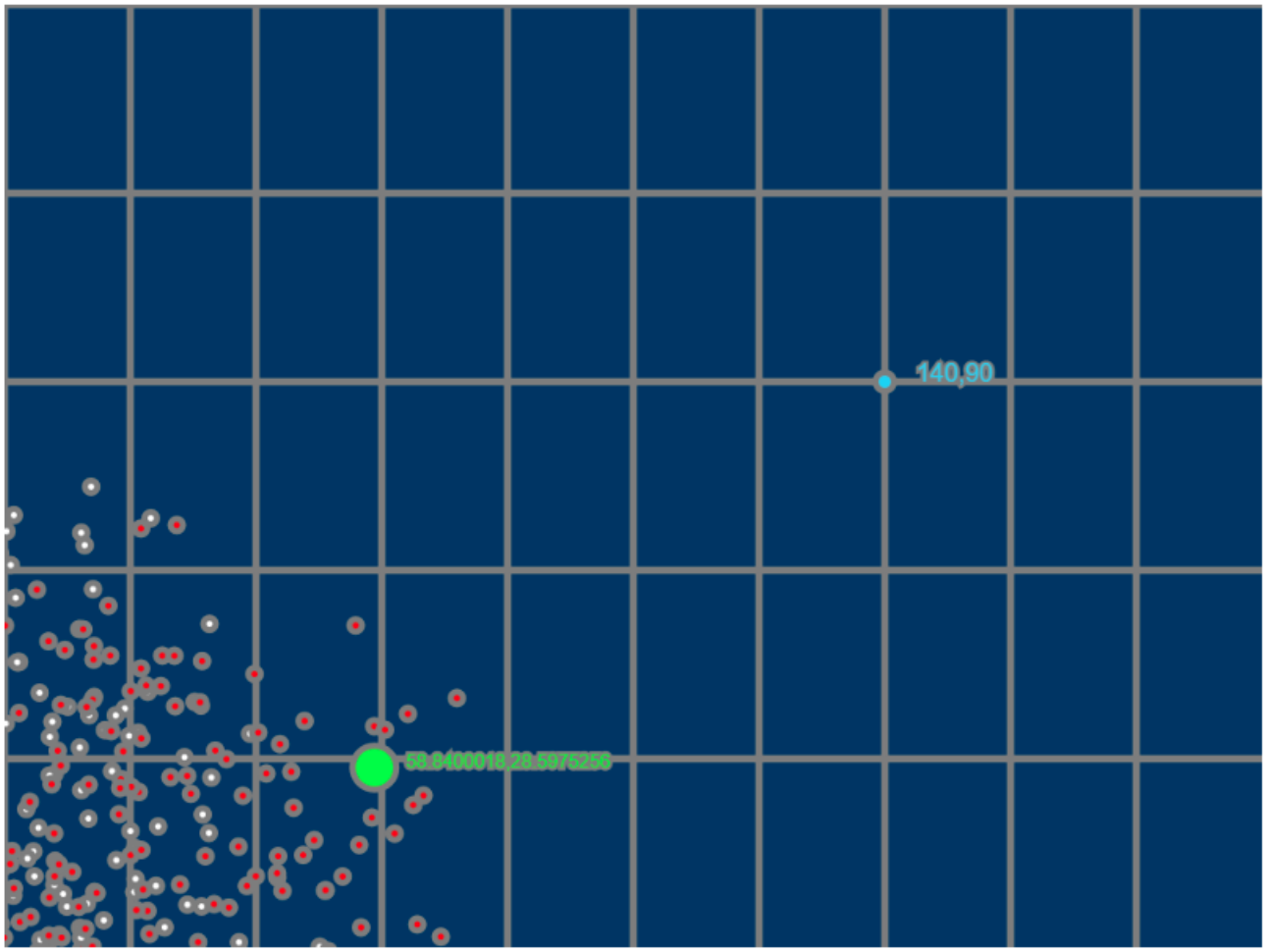

Figura 8: Salida del clasificador KNN usando datos con dimensión reducida por PCA y normalizados al intervalo 0-1, ponderados por 50 para mejor visualización