INFORME DE LA APLICACIÓN

INTELIGENCIA ARTIFICIAL

TRABAJO 02

Logotipo

Descripción generada automáticamente

Nombre de la aplicación: Water Classification

**Presentado por:**

Estudiante 1: Luis Alejandro Ordóñez Díaz

Estudiante 2: Fabrizio Ricardo Castellano Castillo

Contenido

[Descripción y fundamentación del problema a resolver 3](#_Toc74336084)

[Recolección de datos 4](#_Toc74336085)

[Descripción de la arquitectura de Red Neuronal 5](#_Toc74336086)

[Detalles del código fuente de la aplicación 6](#_Toc74336087)

[Pruebas de uso, ejecución y descripción de las funcionalidades 7](#_Toc74336088)

[Comparación funciones de activación: 7](#_Toc74336089)

[Comparación k-vecinos más cercanos: 8](#_Toc74336090)

[Comparación de algoritmos - menor MLP: 8](#_Toc74336091)

[Comparación de algoritmos - mayor MLP: 9](#_Toc74336092)

[Referencias bibliográficas 10](#_Toc74336093)

# Descripción y fundamentación del problema a resolver

En este proyecto haremos uso de ciertas propiedades del agua que, con ayuda de los estándares establecidos por el estado peruano y la organización mundial de la salud, serán de utilidad para la creación de modelos capaces de predecir la potabilidad del agua y por consecuencia determinar si es apta para el consumo humano o no.

Nuestro principal objetivo es presentar tres algoritmos provenientes de machine learning y pertenecientes al grupo de algoritmos de aprendizaje supervisado para determinar cuál de estos consigue la mayor precisión para dar solución a nuestro problema.

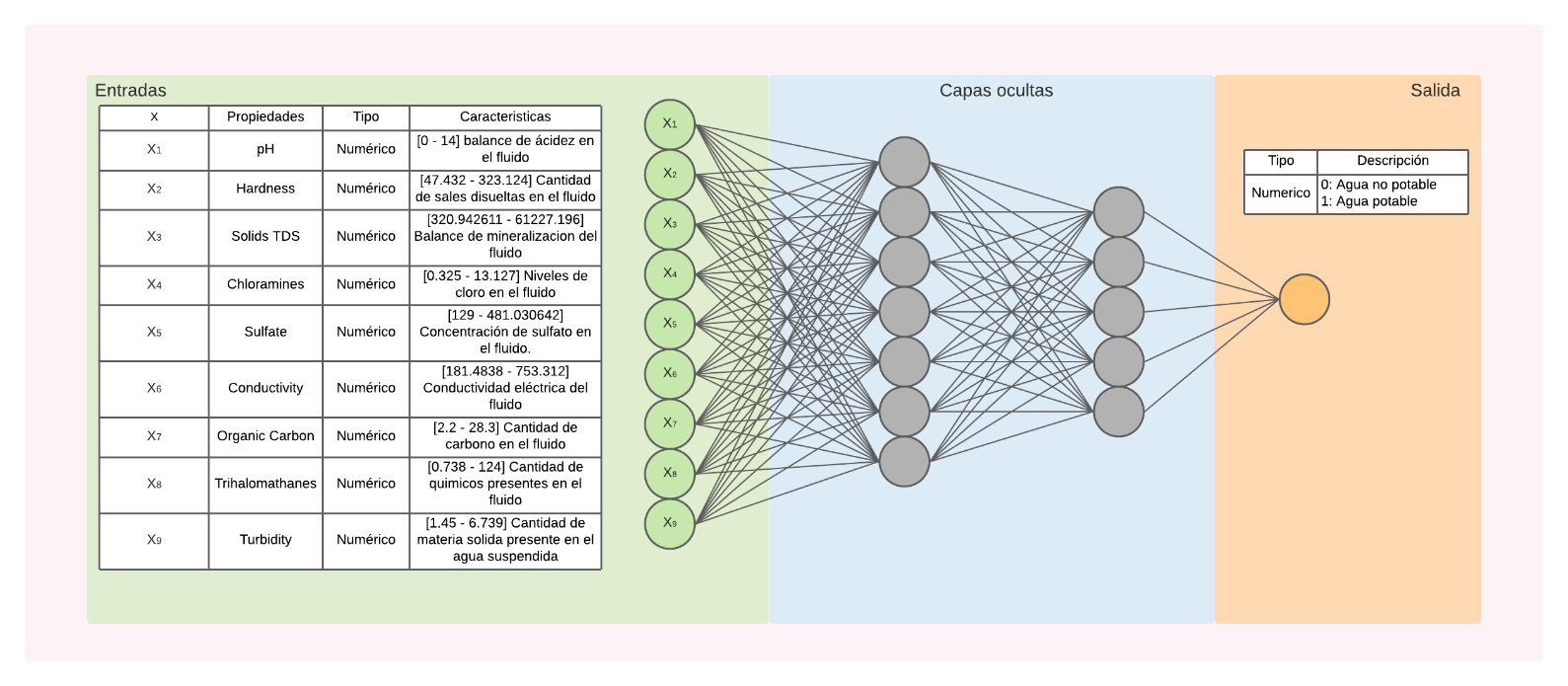
Para comenzar, elegiremos el modelo de redes neuronales de Perceptrón Multicapa (MLP) con mayor porcentaje de precisión según las 4 funciones de activación que se utilizan mayormente con este modelo (identidad, sigmoide, tangente hiperbólica y unidad lineal rectificada). Luego, se procederá a comparar el modelo escogido con los algoritmos K-Nerest Neighbor (KNN), Naive Bayes y el de Support Vector Classifier (SVM) y finalmente, para concluir mostraremos los resultados de estos modelos y/o algoritmos según su porcentaje de precisión.

# Recolección de datos

Para la recolección de nuestra data utilizamos el data set “Water Quality” que se encuentra en Kaggel, una comunidad en línea de científicos de datos y practicantes del machine learning, el enlace a data set se encuentra en las referencias bibliográficas. A continuación, presentaremos las propiedades del agua que nos brinda el data set en una tabla de datos.

|  |  |
| --- | --- |
| 1. pH value: [ 0 - 14] | 6. Conductivity: [ 181.483754 - 753.34162] |
| El pH es un parámetro importante para evaluar el balance de la acidez en el agua. Los valores de pH recomendados por La Organización Mundial de la Salud para el consumo humano del agua van desde 6.5 - 8.5. | El agua no es un buen conductor de electricidad, de hecho, es un buen aislante. Sin embargo, si aumentamos la concentración de iones mejora la conductividad eléctrica del agua. Generalmente la solución de solidos determina la conductividad eléctrica. De acuerdo con los estándares de la Organización Mundial de la Salud el valor de conductividad eléctrica no debe superar los 400 μS/cm. |
| 2. Hardness: [47.432 – 323.124] | 7. Organic\_carbon: [2.2 - 28.3] |
| La dureza del agua es causada por las sales de calcio y magnesio. Estas sales son disueltas en el agua durante el viaje de esta por los depósitos geológicos. La dureza del agua se definió originalmente como la capacidad del agua para precipitar el jabón (causado por el calcio y magnesio) | El carbono orgánico total (TOC) en las fuentes de agua proviene de la materia orgánica natural en descomposición (NOM), así como de fuentes sintéticas. TOC es una medida de la cantidad total de carbono en compuestos orgánicos en agua pura. Los valores estándar van desde <2 mg/L como TOC en agua tratada / potable y <4 mg/L en el agua de origen que se utiliza para el tratamiento. |
| 3. Solids (Total dissolved solids - TDS): [320.942611 - 61227.196] | 8. Trihalomethanes: [0.738 - 124] |
| El agua disuelve una gran cantidad de solidos como potasio, calcio, sodio, bicarbonatos, cloruros, magnesio y sulfatos. Estos minerales producen un cambio no deseado de sabor, color y apariencia. Los altos niveles de TDS (Total dissolved solids) indica que el agua está altamente mineralizada. El nivel deseado de TDS es de 500 mg/l y el límite máximo es de 1000 mg/l para ser apta para beber. | Los THM son químicos que se encuentran en el agua tratado con cloro. La concentración de THMs en el agua bebible varía de acuerdo con el nivel de material orgánico en el agua, la cantidad de cloro en el agua y la temperatura en la que el agua es tratada. El agua con niveles hasta 80 ppm se considera como segura para beber |
| 4. Chloramines: [0.352 - 13.127] | 9. Turbidity:[1.45 - 6.739] |
| El Cloro y las Cloraminas son usadas mayormente para desinfectar el agua pública. Las cloraminas se forman normalmente al añadir amonio al momento de tratar el agua para que esta pueda ser consumida. Los niveles de cloro hasta 4 miligramos por litro son considerados como agua segura para ser consumida. | La turbidez del agua depende de la cantidad de materia solida presente en el agua suspendida. El valor recomendado por la OMS de 5,00 NTU. |
|  |  |
|  |  |
| 5. Sulfate: [129 - 481.030642] | 10. Potability:[0-1] |
| Los sulfatos son sustancias naturales que se encuentran en minerales, tierra y rocas. están presentes en el aire, agua, plantas y en la comida. Por ejemplo, el agua de mar tiene una concentración de alrededor 2700 miligramos por litro de agua. Por otro lado, las concentraciones en el agua dulce van desde 3 a 30 mg/L en la gran mayoría de casos. | Indica si el agua es segura para el consumo humano donde 0 significa no potable e 1 potable. |

# Descripción de la arquitectura de Red Neuronal



Para nuestra arquitectura de red neuronal del modelo Perceptrón Multicapa (MLP), hemos decidido utilizar 2 capas ocultas para obtener una clasificación más versátil de las regiones en las que se encontrarán nuestros datos. Para lograr determinar la cantidad de neuronas ocultas por capa, nos guiamos de la investigación realizada en comunidades de científicos de datos, estas recomendaban que sea una cantidad que se encuentre entre la cantidad de neuronas de entrada y las de salida, aunque no existe una regla definitiva para esto. Por esto, y luego de realizar un gran número de pruebas, decidimos utilizar 7 neuronas ocultas en la primera capa oculta y 5 neuronas ocultas en la segunda capa oculta, para así también contar con resultados de buena precisión manteniendo un bajo nivel de consumo de recursos computacionales.

# Detalles del código fuente de la aplicación

A continuación, describiremos el uso especifico que se le dio a cada biblioteca y las funcionalidades principales del programa:

pandas: Facilitar la lectura y manipulación de data set, principal función usada fue la de los data frames.

pandas\_profiling: Crear un reporte detallado de las variables para su mejor entendimiento e identificar la relación entre estas.

sklearn: Implementar algoritmos de lenguaje supervisado, los módulos principales usados fueron los siguientes:

* model\_selection.train\_test\_split para dividir data set en datos de entrenamiento y datos de prueba
* preprocessing.StandardScaler para escalar los datos de entrada y facilitar su comparación
* neural\_network.MLPClassifier para el uso del modelo MLP de las redes neuronales
* naive\_bayes.GaussianNB para el uso del algoritmo de Naive Bayes en nuestras comparaciones
* svm.LinearSVC para el uso del algoritmo de Support Vector Classifier en nuestras comparaciones
* neighbors.KNeighborsClassifier para el uso del algoritmo K-Nearest Neighbor en nuestras comparaciones
* metrics.accuracy\_score para obtener la precisión de todos los algoritmos mencionados anteriormente, una de las bibliotecas más importantes y usadas en este proyecto

**matplotlib.pyplot:** Utilizado para la creación de gráficas de medición de los algoritmos

**Seaborn:** Creación de gráficas comparativas entre nuestros algoritmos y/o sus funciones de activación

# Pruebas de uso, ejecución y descripción de las funcionalidades

## Comparación funciones de activación:

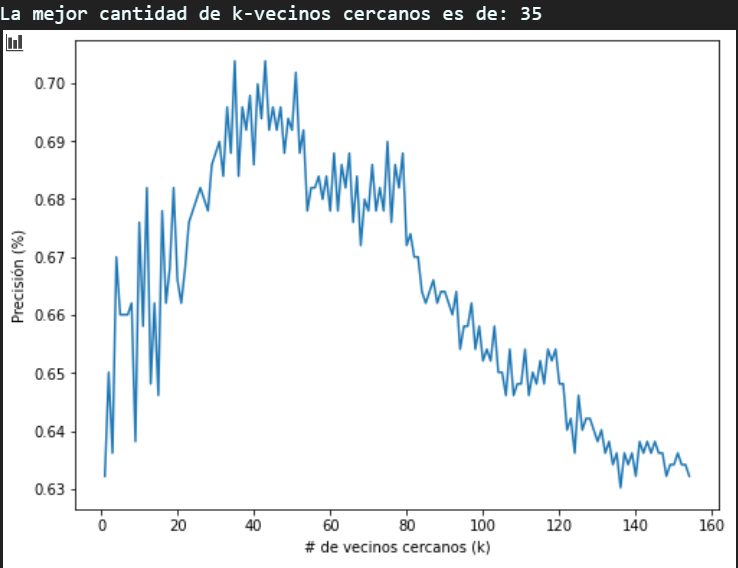
En esta gráfica se puede apreciar una comparación del modelo MLP de las redes neuronales utilizando sus diferentes funciones de activación, se recolectaron datos en 4 instancias para cada función de aceptación y se promediaron los resultados.

Gráfico, Gráfico de barras

Descripción generada automáticamente

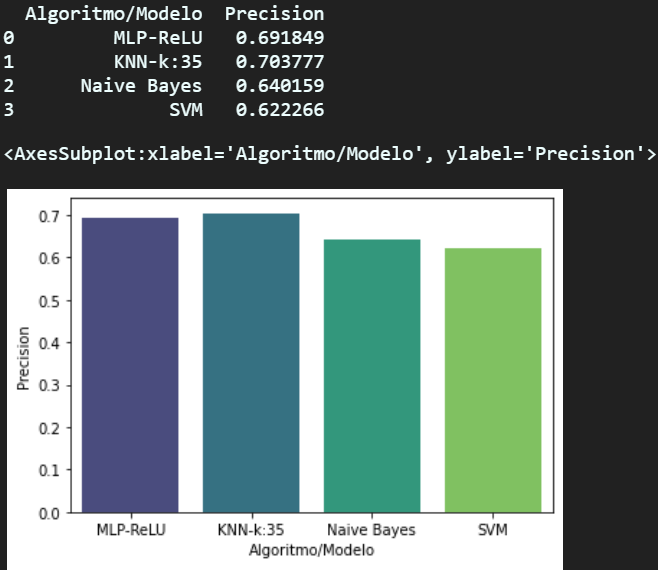
## Comparación k-vecinos más cercanos:

Antes de comparar el K-NN con el modelo MLP, se halló la cantidad de vecinos más cercanos que nos devolvía la mejor precisión al analizar data set.



## Comparación de algoritmos - menor MLP:

Para la comparación se utilizaron 4 instancias nuevamente para los algoritmos, de donde se utilizó la menor precisión lograda por el MLP y se comparó con los demás. En este caso podemos observar que el K-NN logró el mejor resultado respecto a la precisión de todos.



## Comparación de algoritmos - mayor MLP:

Para este caso, se tomó la instancia donde el modelo MLP obtuvo la mayor precisión dentro de sus otras instancias, se puede observar que en este caso supera en precisión al K-NN, logrando ser el algoritmo con mejor precisión para la mayoría de los casos; aunque no para todos ellos. Cabe resaltar que por la complejidad y lo reciente de la data, la mayor precisión lograda hasta el momento (viernes 11 de junio del 2021) es de 73.85%.

Gráfico, Gráfico de barras

Descripción generada automáticamente

# Referencias bibliográficas

Recuperación de data set:

- Aditya, K. (2021). Water Quality-Drinking water potability. Recuperado el 10 de junio de 2021, de Kaggle website: <https://www.kaggle.com/adityakadiwal/water-potability>

Estándares, definiciones y normas del agua potable:

- Ministerio de Salud (2010). Reglamento de la Calidad del Agua para Consumo Humano-Decreto Supremo N° 031-2010-SA. Recuperado el 10 de junio de 2021, de gob.pe website: <https://www.gob.pe/institucion/minsa/normas-legales/244805-031-2010-sa>

- Fawell, J (2007). pH in Drinking water. Recuperado el 10 de junio de 2021, de who.int website: <https://www.who.int/water_sanitation_health/dwq/chemicals/ph_revised_2007_clean_version.pdf>

- Water Science School (2020). Hardness of Water. Recuperado el 10 de junio de 2021, de usgs.gov website: <https://www.usgs.gov/special-topic/water-science-school/science/hardness-water?qt-science_center_objects=0#qt-science_center_objects>

- OMS (2006). Guías para la calidad del agua potable. Recuperado el 10 de junio de 2021, de who.int website: <https://www.who.int/water_sanitation_health/dwq/gdwq3_es_fulll_lowsres.pdf>

Recursos de aprendizaje:

- Scikit-learn developers (2020). API Reference. Recuperado el 10 de junio de 2021, de scikit-learn.org website: <https://scikit-learn.org/stable/modules/classes.html>

- Rob, H (2021). How to choose the number of hidden layers and nodes in a feedforward neural network?. Recuperado el 10 de junio de 2021, de stackexchange.com website: <https://stats.stackexchange.com/questions/181/how-to-choose-the-number-of-hidden-layers-and-nodes-in-a-feedforward-neural-netw>