Comparación de algoritmos de imputación para el parámetro de la precipitación de modelos hidrológicos empleando técnicas de ciencia de datos y Big Data

Lady Johanna Trejos Hernández

Jorge Iván Villada Lizarazo

Universidad Tecnológica de Pereira

Facultad de Ingenierías

Programa de Ingeniería de Sistemas y Computación

Pereira

2019

Comparación de algoritmos de imputación para el parámetro de la precipitación de modelos hidrológicos empleando técnicas de ciencia de datos y Big Data

Lady Johanna Trejos Hernández

Jorge Iván Villada Lizarazo

Trabajo de grado para optar por el título de Ingeniero de Sistemas y Computación

Director:

MSc. Carlos Andrés López

Universidad Tecnológica de Pereira

Facultad de Ingenierías

Programa de Ingeniería de Sistemas y Computación

Pereira

2019

Agradecimientos

A Dios, por ser guía y regalarnos sabiduría para culminar con éxito este proceso académico.

A nuestros padres por su apoyo incondicional y su inmenso esfuerzo para que podamos enriquecer día a día nuestro conocimiento y ser mejores personas, sin olvidar que la calidad humana es prioridad en cualquier aspecto de nuestra vida.

Al semillero SEMDA por depositar su confianza en nosotros y orientarnos para poder adquirir el conocimiento necesario para el desarrollo de este trabajo.

Al director Carlos Andrés López, por ser facilitador y por compartirnos parte de su conocimiento.

A Alexander Rozo y Jhennifer Ruiz, de El Cenit, por la confianza depositada en nosotros para la realización de este trabajo y por permitirnos acceder a los datos del conjunto de Quindío que fortalecieron nuestro proyecto.

A Juan Mauricio Castaño, de la Red Hidroclimatológica de Risaralda, quien desde el primer contacto nos brindó su ayuda y nos proporcionó el conjunto de datos de Nariño, que fue una pieza clave para alcanzar los objetivos. También, a Juliana Vargas, por toda la ayuda ofrecida durante el proceso de implementación de los algoritmos.

A cada maestro que nos compartió sus conocimientos a lo largo de este proceso de pregrado, a cada compañero por sus valiosos aportes y a todas las personas que nos fortalecieron con energías positivas y buenos deseos.

Quedamos eternamente agradecidos con cada uno de ustedes.

Tabla de contenido

Lista de ilustraciones

Lista de tablas

1. Introducción

En los últimos años el cambio climático ha potenciado el riesgo de desastres en Colombia. Para 2018 la cifra de damnificados por un invierno intenso era cerca de 54.000, esto debido a que el incremento de las precipitaciones generó desbordamientos de ríos en diferentes departamentos del país (Revista Semana, 2018). Por esta razón, la prevención de riesgos se hace necesaria y depende del monitoreo de condiciones hidroclimatológicas que determinan el comportamiento de los ríos. Para ello hace uso de un modelo hidrológico que considera variables como humedad relativa, precipitación y temperatura.

Sin embargo, durante la construcción y calibración del modelo se encuentra el problema de los datos faltantes asociados a carencia de lectura o fallas del instrumento, afectando la precisión o el registro de este. En este trabajo, se realiza la comparación de los algoritmos KNN y Curva de Doble Masa, para completar los datos faltantes en series de precipitación.

1.1. Antecedentes de la idea

«El Cenit se consolida como un centro que pueda monitorear, analizar, estudiar, comunicar y emitir alertas a toda la comunidad del Eje Cafetero colombiano, con el objetivo de disminuir la vulnerabilidad, generando comunidades más resilientes» (El Cenit).

Para llevar a cabo el monitoreo del nivel de los ríos, se debe tener un modelo hidrológico que permita analizar el comportamiento y de esta manera poder comunicar, a las organizaciones encargadas de la gestión del riesgo en cada municipio, las situaciones de peligro que se identifiquen.

Sin embargo, el Cenit no cuenta con este modelo y pretende usar el del sector del páramo Romerales en la cuenca alta del río Quindío como base para la construcción del modelo hidrológico para Risaralda.

Por otro lado, la Red Hidroclimatológica de Risaralda en el proceso de actualización del modelo hidrológico de la cuenca del río Otún identifica la necesidad de contar con un conjunto de datos lo más completo posible para que el modelo resultante sea preciso.

Por esta razón, se busca comparar los algoritmos de imputación para que cada entidad pueda tomar la decisión de cuál usar dependiendo de los resultados obtenidos en este trabajo.

1.2. Situación problema

El riesgo de desastres se entiende como las probabilidades de que ocurra un evento identificable y que trae consecuencias a comunidades en general, traducidas en pérdida de vidas y materiales. Teniendo en cuenta esto, debe ser de vital importancia conocer los riesgos de desastres no solo para los tomadores de decisiones sino para la comunidad en general que no está exenta de estos.

El clima es muy cambiante y más en los últimos años por cuestiones del calentamiento global, llevando a que las personas no estén preparadas para sus consecuencias, como las fuertes lluvias que causan avalanchas, deslizamientos o pérdidas en cultivos.

Por esta razón diferentes centros de monitorización hidroclimatológica de la ciudad de Pereira deben estar preparados para cualquier evento, por ejemplo, El Cenit que es el

centro de monitoreo para la gestión del riesgo de desastres del Eje Cafetero, se ha planteado desarrollar un modelo hidrológico para el Eje Cafetero (Caldas, Quindío y Risaralda). Para esto, se tomará como base el modelo hidrológico realizado para el departamento del Quindío en la tesis Modelación hidrológica del sector del páramo romerales en la cuenca alta del río Quindío, hecho por Yeraldín Agudelo Quiñonez y Jhennifer Ruiz Rozo. El modelo usando las variables de precipitación, humedad relativa y temperatura da como resultado el comportamiento del caudal de la cuenca alta del río Quindío.

Este modelo hidrológico se realizó comparando los datos simulados con las mediciones arrojadas por las estaciones hidroclimatológicas ubicadas en la cuenca o cercanas a esta. Sin embargo, estas estaciones se instalaron en diferentes años y no tienen el mismo periodo de registro. Además, se observan vacíos en los datos, que pudieron ser producidos por mal funcionamiento de las estaciones, por falta de lectura debido a problemas de mantenimiento, mal uso de los funcionarios, entre otros, justificando el proceso imputación de datos como condición previa necesaria para la elaboración del modelo.

La imputación consiste en estimar los datos faltantes a través de alguna técnica informática o estadística, para este caso se empleó el método de la Distancia Inversa Ponderada (IDW por sus siglas en inglés) (Ruiz Rozo & Agudelo Quiñonez, 2016) que permite acercarse al comportamiento observado por las estaciones pero que es susceptible a mejoras.

El Cenit desea construir un modelo hidrológico del Eje Cafetero asociado al manejo de riesgos lo que establece como prioridad la cercanía entre el modelo y la realidad.

Por otro lado, la Red Hidroclimatológica de Risaralda quiere actualizar su modelo hidrológico de la cuenca del Río Otún, por lo que necesitan obtener series de datos de precipitación completas y con calidad, para realizar la calibración de los modelos actuales y futuros.

Es así como se hace necesario realizar una comparación entre métodos de imputación para la precipitación e identificar las diferencias entre los mismos, y de esta forma ofrecer opciones para completar los datos faltantes.

1.3. Definición del problema

Se requieren algoritmos que refinen el proceso de imputación de datos en series de precipitación para la creación de modelos hidrológicos.

1.4. Objetivo general

Comparar algoritmos de imputación para el parámetro de la precipitación de modelos hidrológicos empleando técnicas de ciencia de datos y Big Data.

1.5. Objetivos específicos

- Preprocesar datos hidrológicos con herramientas de Big Data.
- Implementar un modelo de gestión de datos para la imputación.
- Identificar la proporción entre los datos faltantes y la totalidad de los datos.
- Implementar los algoritmos KNN y curva de doble masa para la imputación de la variable de precipitación.
- Comparar los resultados obtenidos en la imputación de datos.

1.6. Justificación del estudio

La prevención de los riesgos causados por alteraciones en el comportamiento de los ríos es un tema de vital importancia debido a que personas, flora, fauna e infraestructura se encuentran cerca de estos. La ausencia de atención previa a estos escenarios vulnerables, se pueden traducir en pérdidas materiales o de vidas.

Para conocer el comportamiento del caudal de un río se construyen modelos hidrológicos basado en los datos tomados por diferentes estaciones hidroclimatológicas. Al trabajar con datos generados por sensores o por instrumentos que requieren lectura por parte de una persona encargada, tener registros incompletos es una situación común. Sin embargo, al modelar situaciones reales los valores faltantes pueden alterar el resultado, lo que hace necesaria la imputación, pero para esto hay que tener en cuenta que si se usa un método inapropiado se puede obtener menos precisión en el resultado.

Tanto para la actualización del modelo de la cuenca del río Otún como para la construcción del modelo que realizará El Cenit, la imputación de datos juega un papel muy importante, porque se necesita que cada modelo hidrológico se ajuste a los datos reales con precisión. Por este motivo, en el presente trabajo se busca comparar el desempeño de los algoritmos KNN y Curva de Doble Masa, para determinar cuál algoritmo puede tener un mejor desempeño en cada caso.

Seleccionando una técnica adecuada para la imputación de los datos, aumentará la precisión del modelo hidrológico y este podrá ser utilizado para la predicción temprana de riesgos de desastres naturales en el Eje Cafetero, logrando salvaguardar la vida y la

infraestructura como consecuencia de la alerta generada por el monitoreo del modelo hidrológico.

1.7. Alternativas de solución

Imputación de datos con métodos de *machine learning*.

Los algoritmos de *machine learning* son muy conocidos para predecir los valores de salida de cierto problema con alto grado de certeza, si se le enseña por medio de un entrenamiento previo cómo se comporta el problema, o también «aprenden» a fuerza bruta solo viendo el comportamiento de los datos. Para la imputación de los datos se pueden probar diferentes algoritmos de *machine learning* para ver cuál tiene mayor nivel de precisión y ver si el modelo se ajusta más a la realidad.

Hacer la imputación de datos sacando un promedio de los datos existentes.

Sacar un promedio de los datos existentes para imputar, es una alternativa óptima en algunos casos. El problema planteado se podría solucionar con esta técnica si los datos de la variable de la precipitación son similares.

Hacer la imputación de datos con el método de Listwise.

Eliminar los registros donde falten datos y solo trabajar con los que tengan la información completa, es una posible solución. Aunque se debe tener en cuenta que, si el porcentaje de registros incompletos es grande, no es una opción adecuada (Medina & Galván, 2007).

Hacer la imputación de datos con una regresión.

Hacer una regresión simple de los datos es una solución factible si los datos manejan cierto tipo de relación o patrón. Hay que tener en cuenta que, si el análisis secundario involucra técnicas de análisis de covarianza o de correlación con otros, no se sugiere la utilización de este método ya que sobrestima la asociación entre las variables (Medina & Galván, 2007).

2. Marco referencial

2.1. Estado del arte

El tema de la imputación de datos no es algo nuevo. Al trabajar con grandes cantidades de estos, la información no siempre está completa. Descartar los registros incompletos o inconsistentes, es una solución; pero no es la más recomendable, porque si hay muchos valores faltantes, el resultado de los estudios puede tener gran porcentaje de error. Así que la imputación de los datos es una opción más certera. Comúnmente se busca que esta sea lo más cercana posible a los valores que en realidad deberían estar en los campos vacíos.

El método de imputación más conocido, de origen estadístico, es aquel donde se calcula el promedio de los datos conocidos y se reemplazan los datos faltantes por el valor obtenido. Existen muchas más técnicas, pero el grado de error que estas manejan es bastante alto; aun así, es mejor realizar imputación, que trabajar con poca cantidad de datos, donde el error obtenido sería más grande.

Analizando este problema, existen ya diferentes métodos para mejorar la imputación y que la información, aunque no sea totalmente precisa, sea cada vez más aproximada a la real. Diversas técnicas de *machine learning* han sido adaptadas para generar datos a partir de valores conocidos y así generar una imputación óptima.

A continuación, se darán a conocer diferentes trabajos que han hecho investigación con técnicas de *machine learning* para la imputación de datos y han tenido buenos resultados; donde el porcentaje de error ha bajado considerablemente en comparación a las técnicas estadísticas convencionales.

Imputación de valores de erosividad bajo datos de lluvia incompletos por métodos de aprendizaje automático, 2017, K. Vantas & E. Sidiropoulos.

Como objetivo general se busca presentar una comparación entre las ecuaciones empíricas «se basa en la observación y el estudio experimental de un fenómeno del cual generalmente se desconoce o se tiene poca información de las leyes fundamentales que lo gobiernan» (Ecuaciones empíricas para calcular el gasto volumétrico, 2013), que solo se basan en relaciones exponenciales entre la erosividad y precipitación contra métodos de *machine learning*.

Palabras clave: *Machine learning*, valores faltantes, imputación, erosividad pluvial

Las técnicas de *machine learning* utilizadas fueron las redes neuronales con regularización bayesiana y regresión de cresta con transformación no lineal, para la estimación e imputación de los valores de erosividad de las precipitaciones. Teniendo como conjunto de datos para trabajar y sacar conclusiones la base de datos *Greek Hydroscope*.

También se realizó una prueba de Friedman, para determinar si un algoritmo tiene un rendimiento sistemáticamente mejor o peor.

Se concluyó que los algoritmos de *machine learning* presentados, superaban los métodos clásicos de estimación mediante ecuaciones paramétricas, reduciendo los efectos de la estimación de R a partir de los registros semanales de precipitaciones. (Vantas & Sidiropoulos, 2017)

Construcción de un modelo de imputación para variables de ingreso con valores perdidos a partir de ensamble learning, 2017, German Rosati.

Tiene como objetivo la construcción de un modelo de imputación de datos usando técnicas de *machine learning*. Este se aplica en una encuesta permanente de hogares y la encuesta anual de hogares de la ciudad de Buenos Aires.

Palabras clave: regularización; LASSO; no respuesta

Para la imputación se hace un ensamble de modelos de regresión LASSO *(Least Absolute Shrinkage and Selection Operator*) a través del algoritmo *Bagging*, que es una técnica de *machine learning*. También se usan técnicas estadísticas como *Hot Deck*, que es muy utilizada en encuestas actualmente y ha tenido buenos resultados. Se tienen en cuenta muchos otros métodos para trabajar, pero solo se escoge uno.

Se concluye que la imputación Bagging-LASSO parece funcionar mejor que el método Hot Deck a la hora de calcular el MSE (Error cuadrático medio) con un 20% menos y un 30% menos en el RMSE (Raíz cuadrada del error cuadrático medio). Por lo tanto, los

valores imputados con Bagging-LASSO se acercaron más a los valores reales de la encuesta. (Rosati, 2017)

Imputación de datos faltantes usando algoritmos genéticos para aprendizaje supervisado, 2017, W. Shahzad, Q. Rehman & E. Ahmed.

Como objetivo general proponer una técnica para tratar valores faltantes para integrar mecanismos de búsqueda basada en la población para explorar más espacio de búsqueda junto con la explotación. Los mecanismos basados en la población, para este caso son los algoritmos genéticos los cuales simulan cómo se comporta una determinada población a través de las épocas.

Palabras clave: algoritmo genético, ganancia de información, datos faltantes, aprendizaje supervisado.

Se usa un algoritmo genético el cual tiene todas las fases que simulan cómo se comporta una comunidad, las cuales son la selección de los cromosomas óptimos, el cruzamiento, mutación y nuevamente la selección de los cromosomas que obtuvieron una función de ganancia mayor o en otras palabras, lo que obtuvieron una respuesta más cercana a los datos originales.

Se concluye que la técnica propuesta funciona bien para conjuntos de datos con un mayor porcentaje de valores perdidos, también para conjuntos de datos en los que los atributos tienen una amplia gama de valores distintos, ya que el algoritmo genético entra en juego cuando hay espacio para una combinación cada vez mayor de valores diferentes. (Shahzad, Rehman, & Ahmed, 2017)

Imputación de datos faltantes por aprendizaje supervisado, 2016, Jason Poulos & Rafael Valle.

Tiene como objetivo general comparar tres diferentes métodos clasificadores de *machine learning* para la imputación de datos: Redes Neuronales Artificiales, *Random Forest* y Árboles de Decisión (*Decision Trees*). Estos son muy conocidos, así que se ponen en evidencia la efectividad que tienen a la hora de hacer imputación.

Palabras y frases clave: Redes neuronales artificiales (RNA), árboles de decisión, métodos de imputación, datos perdidos, perturbaciones, bosques aleatorios.

Se entrenan los tres clasificadores diferentes sobre los datos preprocesados: árboles de decisión, *Random Forest* y redes neuronales.

Las redes neuronales implementadas constan de cuatro capas, cada una de las cuales tiene 1024 nodos y se actualizan con el método de tasa de aprendizaje adaptativo Ada delta.

Los intervalos de predicción se obtienen a partir de la desviación estándar de los errores del conjunto de pruebas de las convergencias de retirada entrenadas de las redes neuronales. Los clasificadores *Random Forest* y los árboles de decisión se entrenan con hiperparámetros preseleccionados. Los intervalos de predicción se derivan de la variación creada al variar la profundidad máxima de los árboles de decisión, y para *Random Forest*, el número de árboles y la regla de decisión para el número de características a considerar cuando se busca la mejor división.

Se llega a la conclusión que las redes neuronales artificiales son las que mejor rendimiento obtienen ante los árboles de decisión y *Random Forest*, dando como resultado un porcentaje de error menor en la predicción de valores faltantes. (Poulos & Valle, 2016)

Elevar la precisión de la imputación de datos faltantes utilizando el clasificador Bolzano, 2016, S. Kanchana & A.S. Thanamani.

En este trabajo se evalúa la imputación múltiple de datos faltantes usando métodos tradicionales no supervisados como la media, la mediana, desviación estándar y algoritmos de *machine learning* supervisados como Naïve Bayes usando el teorema de Bolzano como clasificador.

Palabras clave - Clasificador Bolzano, algoritmo de imputación, clasificador NBI, ML supervisado, ML no supervisado.

Se llega a la conclusión que el clasificador Naive Bayes es un método eficaz para el tratamiento de datos perdidos, además se concluye que es superior a la imputación múltiple. (Kanchana & Thanamani, 2016)

Diagnóstico del cáncer de mama basado en el clasificador de aprendizaje automático de Naïve Bayes con imputación de datos perdidos por KNN, 2013, C. Güzel, **M. Kaya, O. Yıldız & H. Ş. Bilge.**

El objetivo general es usar dos algoritmos de *machine learning* que son KNN (k Nearest Neighbor) y Naive Bayes para la imputación de datos para detectar el cáncer de mama y así poder ayudar a los médicos a tomar la decisión de hacer o no una biopsia.

Palabras clave: Cáncer de mama, imputación de datos faltantes, kNN, Naive Bayes, masa mamográfica.

Se usa un conjunto de datos de masas mamográficas basado de características BIRADS (*Breast Imaging Reporting and Data System*) donde muestra si las masas mamográficas eran benignas o malignas. Con base en la edad, forma, margen, densidad y severidad se debía tomar una decisión. Los métodos de *machine learning*, KNN y Naive Bayes fueron aplicados a este conjunto de datos de dos formas. Una fue aplicando primero el algoritmo de KNN y luego de arrojar un resultado, se aplicó el algoritmo de Naive Bayes y la otra forma de hacerlo fue intercambiando el orden en la aplicación de los algoritmos, pero con K=7 en el algoritmo de KNN. Donde se mostró que cuando se aplica primero KNN y luego Naive Bayes se obtiene un 82,60% de precisión en la imputación, y cuando se aplica primero Naive Bayes y luego KNN, el porcentaje bajó a un 77,72%

Se concluye que la exactitud en la imputación aumentó con los dos algoritmos pero que es mejor usar primero el algoritmo KNN y luego Naive Bayes. También sugieren que para hacer un diagnóstico más correcto puede ser posible utilizando diferentes enfoques de imputación. (Güzel, Kaya, Yıldız, & Bilge, 2013)

Uso de la imputación nominal basada en clasificadores para mejorar el aprendizaje automático, 2011, X. Su, R. Greiner, T. M. Khoshgoftaar y A. Napolitano.

El objetivo general es presentar una imputación nominal basada en clasificadores.

Palabras clave: Datos incompletos, imputación, máquinas de vectores de soporte (SVM), aprendizaje basado en instancias, datos nominales

Se usan 10 clasificadores para hacer una comparación entre todos y así llegar a la conclusión de cuál o cuáles de estos tienen un nivel de precisión más alto a la hora de hacer la imputación de datos nominales.

Algunos de los clasificadores usados fueron: las máquinas de vectores de soporte (SVM), árboles de decisión, redes neuronales, entre otros.

Se concluye que la máquina de vectores de soporte (SVM) y el árbol de decisión se desempeñan mejor en la imputación nominal basada en clasificadores. (Su, Greiner, Khoshgoftaar, & Napolitano, 2011)

Metodología para la imputación de datos faltantes en meteorología, 2010, J. A. Urrutia, P. Reiner, H. D. Salazar.

«Presentar la metodología que se debe seguir para la imputación de datos en series de precipitación y/o temperatura. El procedimiento consiste en hacer uso de correlaciones parciales, modelos de regresión, ajustes de los datos por medio del método de doble masa y verificación de la tendencia a través del test de Kendal» (Urrutia, Reiner, & Salazar, 2010)

Palabras clave: Imputación de datos, Correlaciones Parciales, Modelos de regresión, Test de Kendall, Método de Doble Masa.

Se usaron dos métodos estadísticos para hacer la imputación. Estos son el método de doble masa y Test de Mann-Kendall, donde se tuvo como límite trabajar con estaciones que tuvieran como mínimo el 80% de los datos. Se tuvo en cuenta la correlación parcial de los datos para hacer la imputación

Se concluye que el uso de los modelos de regresión lineal y correlaciones parciales, deben ser aplicados cuando se tienen datos faltantes menor o igual al 20% para que la predicción de los datos tenga un error considerablemente bajo, debido que si es mayor al 20% los modelos de regresión lineal no son lo suficientemente robustos para atacar este tipo de situaciones (2014).

Imputación de datos con redes neuronales, 2009, María E. Valesani, O. Quintana y O. Vallejos.

Tiene como objetivo general la aplicación de Redes Neuronales Artificiales (RNA) como métodos de imputación, para ser utilizados sobre una base de datos real.

Palabras clave: Imputación de datos, redes neuronales artificiales, perceptrones multicapa, aprendizaje supervisado, imputación de datos en ganadería.

Se borran datos a propósito de la base de datos para simular la pérdida de información para determinar si las redes neuronales artificiales brindan una herramienta adecuada para la imputación en comparación a los métodos tradicionales de imputación como la media y mediana. En este caso, la imputación es aplicada a datos de censos ganaderos.

Se concluye que las redes neuronales artificiales «se diferencia fundamentalmente lo métodos estadísticos tradicionales, en que los primeros no realizan condicionamiento, ni ninguna hipótesis sobre la distribución de los datos a estudiar» (Valesani, Quintana, & Vallejos, 2009) y que merece ser tenida en cuenta también para otro tipo de censos y encuestas.

Imputación múltiple a través de algoritmos de aprendizaje, 2007, M. B. Richman, T. B. Trafalis e I. Adrianto.

Se busca tener un método que utilice la información disponible en los datos restantes para predecir los valores perdidos y hacer la debida imputación. Estas técnicas incluyen la sustitución de datos cercanos, técnicas de interpolación regresión lineal y técnicas de *machine learning*.

En este trabajo, se prueban diferentes tipos de técnicas de *machine learning*, como las máquinas de vectores de soporte (SVM) y las redes neuronales artificiales (RNA) contra métodos de imputación estándar, regresión simple, sustitución de la media y supresión de casos.

Se llega a la conclusión que las máquinas de vectores de soporte (SVM) tienen el error más bajo en la imputación, por los que se justifica un uso más generalizado de esta técnica en situaciones en las que es importante obtener estimaciones precisas de los datos faltantes. (Richman, Trafalis, & Adrianto, 2007)

2.2. Marco teórico

Curva de doble masa

Según Searcy y Hardison (1960), esta técnica se basa en el hecho que la gráfica del acumulado de una cantidad contra el acumulado de otra cantidad durante el mismo periodo se verá como una línea recta siempre que los datos sean proporcionales. La pendiente de esta línea es la constante de proporcionalidad entre dichas cantidades.

Los métodos para aplicar la curva de doble masa a datos hidrológicos varían dependiendo del tipo de dato que se esté analizando. Si se trata únicamente de datos de precipitación, se puede decir que la relación entre dos cantidades X y Y se puede expresar como una línea con ecuación de la forma Y=bX, donde b es la pendiente de la curva de doble masa.

Una ruptura en la pendiente indica un cambio en la constante de proporcionalidad entre las dos variables o quizás que la proporcionalidad no es una constante en todas las tasas de acumulación. Estos cambios solo pueden deberse a causas diferentes a las meteorológicas (Montealegre B, 1990), como un cambio en la ubicación del instrumento o un cambio del observador que realiza las lecturas.

En estudios hidrológicos, se recomienda graficar el acumulado de una variable contra el acumulado de un patrón compuesto por todos los registros similares en un área dada. Ya que, si se usan solo los acumulados de dos variables medidas, se pueden producir resultados indefinidos porque es posible que no se sepa cuál de ellas causó una ruptura en la pendiente. Para este patrón se debe contar por lo menos con tres estaciones vecinas con registros anuales confiables.

Para probar, por ejemplo, la consistencia en los registros anuales de varias estaciones, primero se deben tabular los datos anuales de precipitación para cada año y luego acumularlos en orden cronológico. Para obtener el patrón contra el que se grafican las estaciones individuales, se calcula el promedio de la precipitación entre todas las estaciones para cada año y después se hace el acumulado. La gráfica se construye como abscisas los valores del patrón y como ordenadas los de la estación problema.

La curva de doble masa se usa para determinar la consistencia de los datos de precipitación especialmente en áreas montañosas, pues en estas el clima cambia con la diferencia en elevación y la precipitación en dos estaciones cercanas que difieren mucho en elevación puede deberse a diferentes eventos meteorológicos.

También puede ser usada para ajustar registros de precipitaciones. Cuando la razón de una ruptura en la curva es determinada, el registro se puede ajustar al que hubiera sido tomado bajo otro conjunto de condiciones.

La mayoría de los análisis relacionados con precipitación son complicados debido a los datos incompletos, periodos donde no se registró ningún valor. La curva de doble masa puede ser usada también para estimar estos valores. Para estimar el valor faltante de una estación A usando los registros de la estación B: se multiplica el dato de la estación B, para el periodo faltante correspondiente, por la razón entre la pendiente de la curva de doble masa de la estación A y la pendiente de la curva de doble masa de la estación B. Si se tienen varias estaciones cercanas, se realiza este cálculo para cada estación y después se promedian los resultados obtenidos.

K-Nearest Neighbors (KNN)

Los k-vecinos más cercanos (k-Nearest Neighbors) es un algoritmo de aprendizaje automático supervisado, usado para:

1. Problemas de clasificación: la salida es la clase a la que pertenece el nuevo dato (predice un valor discreto). Un dato se clasifica por mayoría de votos entre sus vecinos, el objeto pertenece a la clase más común entre sus vecinos más cercanos. (Bronshtein, 2017)

2. Problemas de regresión: la salida es el valor del objeto (predice valores continuos), este valor se calcula con los valores de sus vecinos más cercanos. (Bronshtein, 2017)

El aprendizaje supervisado es un tipo de algoritmo de Machine Learning que usa un conjunto de datos conocido (conjunto de entrenamiento) para realizar predicciones. El conjunto de entrenamiento contiene datos de entrada y salida, a partir de los cuales el algoritmo busca crear un modelo que pueda realizar predicciones acerca de nuevos datos. Y un conjunto de prueba se utiliza para validar el modelo. (MathWorks)

Tanto el conjunto de entrenamiento como el de prueba, son parte de un conjunto de datos conocido que es dividido en dos partes, donde la mayoría de los datos son usados para entrenamiento y una porción más pequeña es utilizada para las pruebas. Los tamaños más usados para estos dos conjuntos son 80:20 o 70:30, es decir, que el 70% de los datos se usa como conjunto de entrenamiento y el 30% como conjunto de prueba, o el 80% y el 20%.

K-NN es un algoritmo de aprendizaje perezoso (lazy learning), es decir, que durante la fase de entrenamiento solo guarda el conjunto de entrenamiento, no construye un modelo (Aler Mur). Las predicciones se hacen cuando llega un nuevo dato, al que le calcula la similitud con cada una de las observaciones del conjunto de entrenamiento, para seleccionar sus k vecinos más cercanos y con estos realizar la clasificación o regresión.

Ruiz (2017) afirma que el K-NN es muy sensible a los valores de:

- La variable K, los resultados obtenidos en la predicción dependen del valor elegido para K. El valor óptimo para K depende de cada problema, así que se elige

empíricamente, probando diferentes números de vecinos cercanos y eligiendo finalmente aquel con el que se haya obtenido un mejor desempeño (Hassanat, Abbadi, Altarawneh, & Alhasanat, 2014)

- La métrica de similitud utilizada, pues de ella dependen las relaciones de cercanía (similitud) que se establezcan. La distancia Euclidiana es la función más usada para calcular la distancia entre datos continuos, pero se pueden usar otras como la distancia Manhattan, Minkowski, Hamming (esta para variables categóricas), entre otras.

Tipos de datos faltantes

Es importante distinguir entre los diferentes tipos de datos faltantes, para identificar correctamente la técnica que debe ser empleada para realizar la imputación. Se tienen tres tipos de datos faltantes (Bennett, 2001):

1. **Faltantes completamente al azar *(Missing Completely At Random, MCAR):*** se refiere a los datos, cuando el proceso de pérdida no depende de las otras variables del conjunto de datos. Cualquier observación tiene igual probabilidad de perderse.
2. **Faltantes al azar *(Missing At Random MAR):*** significa que las observaciones faltantes están condicionadas por otras variables explicativas en el conjunto de datos.
3. **Faltantes no al azar *(Missing Not At Random):*** el patrón de pérdida no es aleatorio y tampoco es predecible de otras variables del conjunto de datos. La pérdida depende de la variable que falta.

Tratamiento de datos faltantes

Existen diversas técnicas para tratar con los datos faltantes, entre ellos se encuentran:

Métodos que ignoran las observaciones perdidas.

Análisis de casos completos. También conocido como Eliminación por lista *(listwise deletion)*, consiste en suprimir un registro completo si le falta uno o más datos.

Análisis de casos disponibles. O eliminación por pares *(pairwise deletion)*, consiste remover del análisis las variables con datos incompletos y continuar analizando las que estén completas. Usa de cada registro la mayor cantidad de datos posible. (Lodder, 2013)

Métodos de imputación únicos.

Bennet (2001), describe algunos como:

Última observación desplazada. *(Last Observation Carried Forward)* este método reemplaza cada valor que falta por el último valor observado del mismo tipo. Se usa sobre todo en datos longitudinales o series de tiempo.

Sustitución por la media. Reemplaza los valores faltantes por el promedio de los valores disponibles de la misma categoría.

Métodos de regresión. Implica el desarrollo de una ecuación de regresión para una variable dada, tratándola como «resultado» y usando las demás variables como predictores.

Hot-Deck. Consiste en reemplazar los datos faltantes por valores tomados de registros similares en términos de datos observados.

Cold-Deck. Es un método similar a Hot-Deck en su enfoque, pero su estrategia para evaluar la similitud se basa en información externa o conocimientos previos en lugar de la información disponible en el conjunto de datos actual.

Otros métodos de imputación.

Imputación múltiple. Reemplaza cada valor faltante por $M \geq 2$ posibles valores para crear M conjuntos de datos completos. M normalmente está entre 5 y 10.

Expectativa-Maximización. Es un tipo de método de probabilidad máxima que se puede usar para crear un nuevo conjunto de datos, en el cual todos los valores faltantes se imputan con valores estimados por los métodos de probabilidad máxima. (Kang, 2013)

Eficiencia de Nash-Sutcliffe (NSE)

Es el criterio más usado para medir la eficiencia de los modelos hidrológicos, es una herramienta para medir la capacidad predictiva del modelo.

Se define como:

$$\mathrm{NSE} = 1 - \left[\frac{\sum_{i=1}^{n} \left(Y_i^{obs} - Y_i^{sim}\right)^2}{\sum_{i=1}^{n} \left(Y_i^{obs} - Y^{mean}\right)^2} \right]$$

Donde Y_i^{obs}es la i-ésima observación para el componente evaluado, Y_i^{sim} es el i-ésimo valor simulado, Y^{mean}es el promedio de los datos observados y $n$es el número total de observaciones.

«El NSE puede tomar valores entre $-\infty$ y 1.0, siendo 1.0 el valor óptimo. Los valores entre 0.0 y 1.0 son generalmente vistos como niveles aceptables de desempeño, mientras que los valores iguales o menores a 0.0 indican que la media de los valores observados es un mejor predictor que el valor simulado, lo cual indica un desempeño insuficiente» (Moriasi, y otros, 2007)

Tabla 1. Valores referenciales del Criterio de Nash-Sutcliffe

Rango de NSE	**Ajuste**
<0.2	Insuficiente
0.2 - 0.4	Satisfactorio
0.4 - 0.6	Bueno
0.6 - 0.8	Muy bueno
> 0.8	Excelente

Fuente: Molnar, P., 2011. Calibration. Watershed Modelling, SS 2011. Institute of Environmental Engineering, Chair of Hydrology and Water Resources Management, ETH Zürich. Switzerland.

Raíz del error cuadrático medio (RMSE)

> La raíz del error cuadrático medio es la desviación estándar de los residuos (errores de predicción). Los residuos son una medida de qué tan lejos están los puntos de datos de la línea de regresión; RMSE es una medida de la dispersión de estos residuos. En otras palabras, dice qué tan concentrados están los datos alrededor de la línea de mejor ajuste y permite cuantificar la magnitud de la desviación de los valores simulados respecto a los observados. (Agrimetsoft)

Se define como:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Q_{sm} - Q_i)^2}{n}}$$

Donde:

Q_{sm}: Valor simulado

Q_i: Valor observado

n: Cantidad de observaciones

El rango que puede tomar es de 0 a ∞, donde el 0 corresponde a un ajuste perfecto y valores más grandes indican un mejor ajuste.

Porcentaje de sesgo (PBIAS)

«Mide la tendencia de los datos simulados a ser más grandes o pequeños que sus homólogos observados, siendo su valor óptimo 0. Los valores positivos indican modelo de sesgo de subestimación y los valores negativos indican modelo de sesgo de sobreestimación» (Autoridad Nacional del Agua, 2016). Se calcula con la expresión:

$$PBIAS = \left[\frac{\sum_{i=1}^{n}(Y_i^{obs} - Y_i^{sim} * (100)}{\sum_{i=1}^{n}(Y_i^{obs})}\right]$$

El PBIAS evalúa la desviación de los datos, expresada como porcentaje.

Error promedio absoluto (MAE)

«Se define como la magnitud promedio de los errores de un ejercicio de pronóstico sin tener en cuenta su signo, es decir, el promedio de los valores absolutos de los errores calculados» (Vélez Correa & Nieto Figueroa, 2016).

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$$

Donde n es el número de muestras y $\hat{y}_t$ es la estimación de y_t.

Correlación

El objetivo de la correlación es examinar la dirección y fuerza de la asociación entre dos variables cuantitativas, para determinar si al aumentar el valor de una de ellas, el valor de la otra aumenta o disminuye y con cuánta intensidad.

Los dos coeficientes de correlación más usados son el de Pearson (paramétrico) y el de Spearman (no paramétrico).

Coeficiente de correlación de Pearson

Evalúa la asociación lineal entre dos variables X y Y, se trata de un índice que mide si los puntos tienen tendencia a disponerse en línea recta. Su rango de valores es entre -1 y 1, donde 0 indica ausencia de relación entre las variables.

Se define como la covarianza entre X e Y dividida por el producto de las desviaciones típicas de cada variable (Laguna)

$$r = \frac{S_{xy}}{S_x S_y}$$

Si r < 0 hay una correlación negativa, las dos variables se correlacionan en sentido inverso, es decir, a valores altos de una variable le corresponden valores bajos de la otra y viceversa.

Si r > 0 hay una correlación positiva, las dos variables se correlacionan en sentido directo. A valores altos de una le corresponden también valores altos de la otra e igualmente con los valores bajos (Coeficiente de correlación, s.f.)

Entre más cercano a -1 o a 1 sea r, más fuerte es la correlación, mientras que si es más cercano a 0 la correlación es débil.

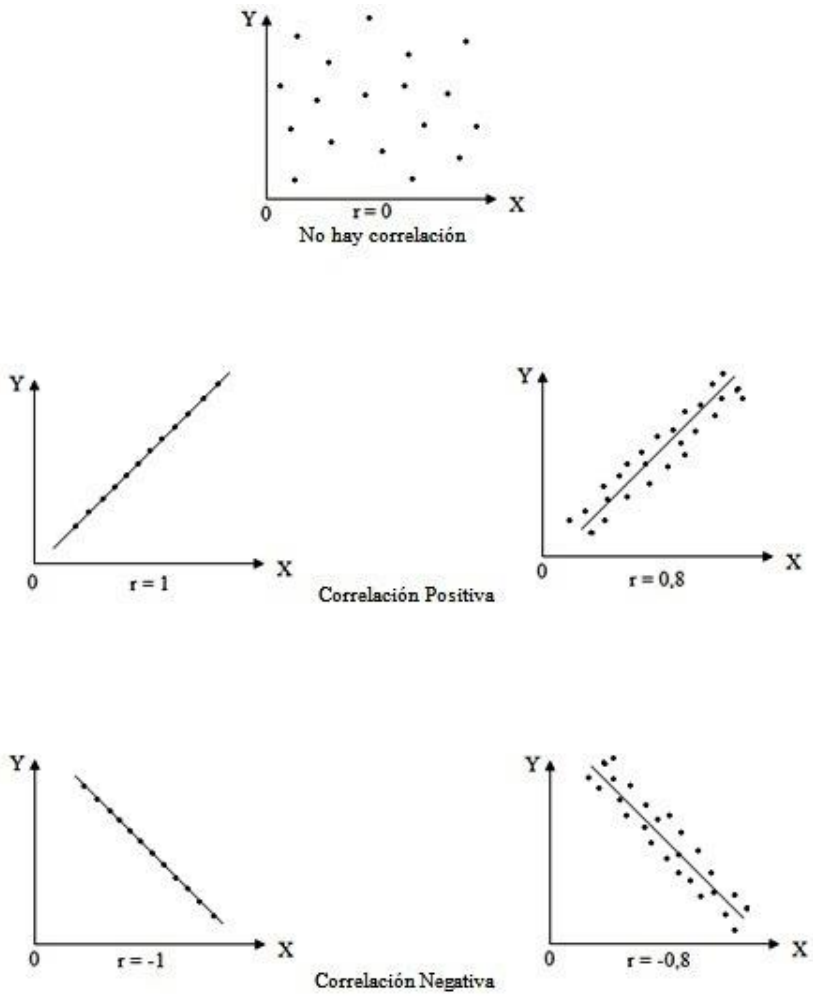

Ilustración 1. Coeficiente de correlación

Fuente: Suárez, M. (s.f.) Coeficiente de correlación de Karl Pearson. Recuperado de https://www.monografias.com/trabajos85/coeficiente-correlacion-karl-pearson/coeficiente-correlacion-karl-pearson.shtml

2.3. Marco conceptual

Apache pig

Es una plataforma para analizar y consultar grandes conjuntos de datos usando un lenguaje de alto nivel. Es un componente del Ecosistema Hadoop y utiliza el lenguaje Pig Latin.

Cuenca hidrográfica

> Una cuenca hidrográfica incluye a todos los caudales pequeños en un territorio que se suman entre ellos para, finalmente, agregarse a un caudal principal y desembocar en una misma salida. Las cuencas altas están generalmente ubicadas en montañas o cabeceras de cerros. (Glosario Ambiental ¿Qué es una cuenca?, 2018)

Imputación de datos

«Es el proceso para asignar valores de reemplazo para datos faltantes, no válidos o inconsistentes en una observación» (Imputation, 2015).

Jupyter Notebook

Jupyter Notebook es una aplicación web de código abierto, que pertenece al Proyecto Jupyter. Permite crear y compartir documentos que contienen código interactivo, ecuaciones, visualizaciones y texto narrativo.

Jupyter ofrece una shell interactiva vía web a la que se puede acceder desde el navegador. La shell está compuesta por bloques en los que cada uno puede contener texto en formato Markdown, código en distintos lenguajes, ecuaciones en LaTeX, gráficas, elementos multimedia, entre otros (Zaforas, 2016). Se puede ir ejecutando cada celda o ejecutar todo de una sola vez, obteniendo resultados parciales.

Es una herramienta ampliamente usada en el campo de la Ciencia de Datos para limpieza y transformación de datos, simulación numérica, modelado estadístico, visualización de datos, aprendizaje automático y mucho más.

Está diseñada para tener compatibilidad entre Python y Markdown, facilitando la toma de anotaciones

Machine learning

> *Machine Learning* es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente. Aprender en este contexto quiere decir identificar patrones complejos en millones de datos. La máquina que realmente aprende es un algoritmo que revisa los datos y es capaz de predecir comportamientos futuros. Automáticamente, también en este contexto, implica que estos sistemas se mejoran de forma autónoma con el tiempo, sin intervención humana. (González, 2014)

Modelo hidrológico

> Es una representación simplificada de un sistema real complejo llamado prototipo, bajo forma física o matemática. De manera matemática, el sistema real está representado por una expresión analítica. En un modelo hidrológico, el sistema físico real que generalmente se representa es la «cuenca hidrográfica» y cada uno de los componentes del ciclo hidrológico. De esta manera un modelo matemático ayudará a tomar decisiones en materia de hidrología, por lo que es necesario tener conocimiento de entradas (*inputs*) al sistema y salidas (*outputs*) a partir del sistema, para verificar si el modelo es representativo del prototipo. (Modelación hidrológica, s.f.)

MongoDB

MongoDB es una base de datos distribuida, orientada a documentos y de código abierto. Almacena datos en forma de documentos tipo JSON, posee un potente lenguaje de consulta y admite agregaciones, búsqueda de gráficos o texto, y búsqueda basada en información geoespacial. Las consultas se realizan en JSON por lo que son más legibles y fáciles de programar. Es una plataforma de datos con un completo paquete de herramientas que facilita el trabajo con los datos. (MongoDB, s.f.)

Pandas

Es una librería de código abierto para Ciencia de datos, que permite la manipulación y el análisis de datos en Python. Proporciona estructuras de datos y herramientas de análisis de datos de alto rendimiento.

Precipitación

> El término precipitación se usa para designar cualquier tipo de forma en que el agua cae desde las nubes a la tierra. Existe una lista hecha por meteorólogos de diez tipos de precipitación, pero sólo se distinguen normalmente tres: lluvia, granizo y nieve. (Precipitaciones, s.f.)

Python

Es un lenguaje de programación de alto nivel interpretado, orientado a objetos, con semántica dinámica. Fue desarrollado a finales de los 80 por Guido van Rossum. (What is Python?, s.f.)

3. Diseño metodológico

Inicialmente se hará el preprocesamiento de los datos hidrológicos, los cuales son obtenidos mediante estaciones hidroclimatológicas que registran información diariamente. Dado que en este trabajo se va a tratar con datos diarios de varias estaciones y con registros de más de 10 años, se genera un volumen de datos que hace necesario emplear técnicas de Big Data.

Posteriormente se tendrá en cuenta la proporción entre los datos faltantes y su totalidad, debido a que, si los porcentajes faltantes son altos, la imputación puede que no sea una opción adecuada ya que no se tiene suficiente información para que los algoritmos puedan trabajar correctamente.

Luego se realizará la implementación de los algoritmos KNN y Curva de Doble Masa. Y para su validación, se tomará un conjunto de prueba (una copia de las muestras), se eliminarán diferentes porcentajes de datos y se le realizará la imputación con cada algoritmo. A cada conjunto resultante se le realizarán mediciones de desempeño usando el RMSE, MAE, NSE y PBIAS.

Por último, se compararán los resultados obtenidos con los algoritmos de imputación, para poder analizar cuál tuvo mejor desempeño.

4. Desarrollo

Para el desarrollo de este trabajo se utilizan dos conjuntos de datos:

- Conjunto de Nariño: compuesto por estaciones hidroclimatológicas ubicadas principalmente en Nariño y algunas estaciones de Cauca y Putumayo. Estos datos son proporcionados por la Red Hidroclimatológica de Risaralda.
- Conjunto de Quindío: compuesto por estaciones hidroclimatológicas de Quindío y entregado por El Cenit.

Para almacenarlos, se configura un servidor con MongoDB que es una base de datos No SQL, elegida para este proyecto por su facilidad para realizar consultas y para manejar datos geográficos como las ubicaciones de las estaciones.

El conjunto de datos de Nariño cuenta con 6.855.960 registros, ya que contiene otras mediciones además de la precipitación, como la temperatura, humedad, entre otros. Por lo cual, se emplea Apache Pig sobre Hadoop, herramienta que permite un procesamiento eficiente para este volumen de datos.

Preprocesamiento de datos

En esta etapa se preparan y organizan los datos que se usan como entrada en los algoritmos de imputación. Aquí se realiza una limpieza y transformación de los datos, para que ambos conjuntos queden con la misma estructura y los algoritmos no tengan inconvenientes al usarlos.

Para el análisis exploratorio se usa Jupyter Notebook, entorno que facilita la ejecución del código por secciones para identificar la estructura de los datos, los cambios que se requieren, la cantidad de datos faltantes, entre otros.

Los datos contienen registros donde el valor de precipitación es elevado, algunos superan los 250 mm. Se decide no eliminar estos datos y dejar esta decisión a los expertos en hidrología cuando usen los algoritmos.

Conjunto de Nariño

El objetivo está enfocado a los datos de precipitación, sin embargo, el conjunto de Nariño también contiene parámetros como temperatura, humedad, entre otros. Por esta razón, se usa la herramienta Apache Pig para filtrar los datos correspondientes a la precipitación y también para etiquetar los mismos indicando cuando falta un valor de precipitación en un registro. El resultado se guarda directamente en la base de datos, quedando un total de 2.146.586 documentos.

Durante la limpieza de datos, en el campo del Nombre de la estación se encuentran errores que no permiten identificar completamente el código y el nombre, como se evidencia en la Tabla 2. Para la corrección, se consulta el Catálogo Nacional de Estaciones del IDEAM y se busca cada estación usando las partes reconocibles en el campo del nombre, manteniendo el formato que tiene el conjunto de datos para este campo.

Tabla 2. Corrección de los nombres de las estaciones

Nombre	Nombre corregido
ESTACION: 5255010:00 a. m.PTOSAN LUIS\\par	ESTACION: 5255230 APTOSAN LUIS\\par
ESTACION: 5280010:00 p. m.ISANDA\\par	ESTACION: 5280010 PISANDA\\par
ESTACION: 5250010:00 p. m.ENOLEL\\par	ESTACION: 5250010 PENOLEL\\par
ESTACION: 5135010:00 a. m.PTOLA FLORIDA\\par	ESTACION: 5135010 APTOLA FLORIDA\\par
ESTACION: 4410110:00 p. m.TO LIMON\\par	ESTACION: 4410110 PTO LIMON\\par
ESTACION: 4730010:00 a. m.NGOSTURAS\\par	ESTACION: 4730010 ANGOSTURAS\\par
ESTACION: 4710110:00 p. m.TO CAICEDO\\par	ESTACION: 4710110 PTO CAICEDO\\par
ESTACION: 4745010:00 p. m.TO LEGUIZAMO\\par	ESTACION: 4745010 PTO LEGUIZAMO\\par
ESTACION: 4710210:00 p. m.ISCICULTURA\\par	ESTACION: 4710210 PISCICULTURA\\par
ESTACION: 4735010:00 p. m.I#U#A BLANCA\\par	ESTACION: 4735010 PINUNA BLANCA\\par
ESTACION: 5240110:00 p. m.ASTO\\par	ESTACION: 5240110 PASTO\\par

Además, teniendo en cuenta que el algoritmo Curva de Doble Masa necesita la ubicación de las estaciones, se revisan las coordenadas proporcionadas en este conjunto y se determina que los valores presentan errores, pues no corresponden al formato de coordenadas geográficas. Por lo tanto, se busca también en el catálogo del IDEAM la ubicación correcta de cada estación, usando el código que cada una tiene.

Sin embargo, se encuentra que los códigos en la base de datos tienen siete dígitos y en el catálogo son de ocho dígitos. Por consiguiente, se realiza la corrección al identificar que el dígito faltante es un 0 en la tercera posición, como se ilustra en la Tabla 3

Tabla 3. Corrección de los códigos de las estaciones

Nombre de la estación	Código	Código del IDEAM
Chalet Guamues	4710040	47010040
Quilinsayaco	4710130	47010130
Sta Isabel	4710230	47010230
Encano El	4715100	47015100
Obonuco	5245010	52045010
Cobenas	5255080	52055080

Por otro lado, se unen los valores de latitud y longitud en un GeoJSON, que es el formato para representar elementos geográficos que usa MongoDB; el resultado se guarda en un nuevo campo llamado UBICACIÓN

Conjunto de Quindío

Los datos del conjunto de Quindío se procesan directamente en Python usando la librería Pandas, pues al ser de menor tamaño, 43.830 registros, su procesamiento es más sencillo.

En primer lugar, se cambia la estructura del documento para que coincida con el formato que se necesita para el algoritmo, pasando de tener un registro para cada fecha y una columna para cada estación, a tener una fila por cada observación individual y una columna por cada variable (fecha, estación, valor). También se realizar el etiquetado que indica si al registro le falta el valor de precipitación.

Porcentaje de datos faltantes

Para el desarrollo de los algoritmos se elige como muestra el año 1990, a la que se le realiza la imputación de los valores de precipitación. Esta muestra cuenta con varias estaciones de las cuales se descartan aquellas que tengan un porcentaje de datos faltantes mayor al 20% del total de los datos.

Para seleccionar las estaciones que se deben descartar, se hace un conteo de las etiquetas que indican dato faltante para identificar la proporción de estos respecto al total. Para el conjunto de Nariño, en 1990 se tienen 127 estaciones de las cuales ocho se descartan por tener más del 20% de datos faltantes para el año 1990, como lo indica la Tabla 4

Tabla 4. Estaciones descartadas

Código de la estación	Nombre de la estación	Número de datos faltantes	Porcentaje de datos faltantes
52045070	Wilquipamba	304	83,29
44120020	Umancia	297	81,37
52055080	Cobenas	246	67,40
47045010	Pto. Leguizamo	224	61,37
26030030	Cabana Inderena	160	43,84
52010030	Rosario El	120	32,88
52055100	Villarosa	120	32,88
44130020	Guaquira	90	24,66

En el conjunto de datos de Quindío no se descartan estaciones para el año de estudio, pues el porcentaje de datos faltantes era bastante bajo. De las seis estaciones que se tienen, solamente dos tienen vacíos: a Navarco le faltan 5 datos (≈1,37%) y a La playa 1 dato (≈0,27%).

Curva de Doble masa

Para realizar este algoritmo primero se llenan los valores faltantes con cero para evitar errores al calcular el acumulado. Luego, se crea la columna ACUMULADO donde para cada estación se va acumulando diariamente los valores de precipitación desde el 1 de enero de 1990 hasta el 31 de diciembre de 1990. Posteriormente se calcula el promedio diario de los acumulados de las diferentes estaciones y se almacena en la columna AVG_ACUMULADO.

Para calcular la pendiente de la curva de doble masa, se usa una regresión lineal con los datos de la columna ACUMULADO de cada estación contra el AVG_ACUMULADO que corresponde al patrón para la imputación.

Para empezar la imputación, se separan los documentos con valores faltantes y para cada uno se identifica la estación a la que pertenece y se buscan los datos de las N estaciones más cercanas a ella para el mismo día del dato ausente. Se estima un valor por cada vecina, multiplicando el valor observado en ella por la razón entre la pendiente de la curva de doble masa de la estación con dato faltante y la pendiente de la curva de doble masa de la estación vecina. Después se suman los resultados y se dividen entre la cantidad de estaciones vecinas usadas para obtener el valor a imputar, es decir, el que se va a guardar en el documento donde no se registró un valor. Este procedimiento se puede representar como:

$$P_e = \frac{1}{n} \sum_{i=1}^{n} \frac{b}{b_o^i} P_o^i$$

Donde

P_e = Precipitación estimada

P_o = Precipitación observada en la estación vecina

b = Pendiente de la curva de doble masa de la estación que se está imputando

b_o = Pendiente de la curva de doble masa de la estación vecina

n = Cantidad de estaciones cercanas a usar

Algoritmo KNN

Para la implementación de este algoritmo se usa como criterio de similitud la correlación entre los valores de precipitación de las estaciones, buscando que las estaciones sean «cercanas» en comportamiento.

Para realizar la imputación se asigna un peso a cada estación, correspondiente a la correlación que tenga con la estación problema, de modo que entre más cercano a 1 sea, tendrá mayores posibilidades de ser elegida como vecina y mayor influencia tendrá en la estimación del valor faltante.

En la etapa de entrenamiento, se calcula el coeficiente de correlación de Pearson entre cada estación y las demás, para determinar la asociación entre sus valores.

Iniciando la imputación, se eligen las K vecinas con mayor correlación con la estación problema y se calcula el valor a imputar multiplicando la correlación y el valor de cada vecina y sumando todos estos resultados, luego dividiéndolos entre la suma de los coeficientes de correlación de las K vecinas usadas. Esto se puede representar como:

$$V_e = \frac{\sum_{i=1}^{k} r_i \cdot V_i}{\sum_{i=1}^{k} r_i}$$

Donde V_e es el valor por estimar, r_i es el coeficiente de correlación entre la i-ésima vecina y la estación problema, V_i es el valor de precipitación registrado en la i-ésima vecina para el mismo día del dato faltante y k es la cantidad de estaciones vecinas a usar.

5. Análisis de resultados

Para medir la precisión obtenida con los algoritmos de imputación se usa la Raíz del Error Cuadrático Medio (RMSE), el índice de eficiencia de Nash-Sutcliffe (NSE), el porcentaje de sesgo (PBIAS) y la Media del Error Absoluto (MAE).

Hay que tener en cuenta que los resultados de los algoritmos se ven afectados por los errores propios de los datos proporcionados, como los presentados en la medición en campo o datos de ubicación erróneos (longitud, latitud), pues la salida de los algoritmos está directamente relacionada con los mismos.

Es importante resaltar, que las estaciones que se encuentran cerca por su ubicación geográfica no necesariamente tienen un comportamiento similar. Por ejemplo, en la Ilustración 2 donde se muestran las vecinas que tienen correlación más alta con los datos de la estación Mocoa Acueducto, y en la Ilustración 3 donde se muestran las vecinas más cercanas por ubicación geográfica, solamente la estación El Pepino es una vecina común.

Ilustración 2. Estaciones vecinas a Mocoa Acueducto por correlación

Ilustración 3. Estaciones vecinas a Mocoa Acueducto por ubicación

Las correlaciones negativas no se tienen en cuenta para esta muestra, puesto que la más alta en el conjunto de Nariño es de -0.1705557 mientras que las correlaciones positivas llegan hasta 0.750644 que son mucho más significativas para estimar el valor a imputar. El conjunto de Quindío solo presenta correlaciones positivas, siendo la más alta de 0.776236.

Para ver el comportamiento de los algoritmos, se eligieron las estaciones con mejor y peor correlación con sus vecinas, las cuales fueron Mocoa Acueducto y Concepción respectivamente. Como lo muestra la Ilustración 4, la estación Mocoa Acueducto tiene correlaciones mayores a 0.5. Mientras que en la Ilustración 5, se comprueba que la estación Concepción la correlación más alta que tiene es de solo 0.152832, lo cual es bastante bajo; teniendo en cuenta que entre más cercana a cero sea la correlación, menos asociación hay entre las estaciones.

ESTACION	VECINA	CORRELACION
ESTACION: 44015040 MOCOAACUEDUCTO\par	ESTACION: 44010020 PEPINO EL\par	0.750644
ESTACION: 44015040 MOCOAACUEDUCTO\par	ESTACION: 47010050 CHUNGACASPI\par	0.676207
ESTACION: 44015040 MOCOAACUEDUCTO\par	ESTACION: 47010080 QUINCHOA\par	0.639866
ESTACION: 44015040 MOCOAACUEDUCTO\par	ESTACION: 47010150 CARRIZAL\par	0.583698
ESTACION: 44015040 MOCOAACUEDUCTO\par	ESTACION: 44010040 MINCHOY\par	0.568057
...	...	...
ESTACION: 44015040 MOCOAACUEDUCTO\par	ESTACION: 26040300 BOCATOMA ALERTAS\par	-0.109459
ESTACION: 44015040 MOCOAACUEDUCTO\par	ESTACION: 26030050 TAMBO\par	-0.116670
ESTACION: 44015040 MOCOAACUEDUCTO\par	ESTACION: 26020100 BUENOS AIRES\par	-0.120376
ESTACION: 44015040 MOCOAACUEDUCTO\par	ESTACION: 26040170 CALERA LA\par	-0.131624
ESTACION: 44015040 MOCOAACUEDUCTO\par	ESTACION: 26035030 APTOG L VALENCIA\par	-0.148680

Ilustración 4. Correlaciones con la estación Mocoa Acueducto

ESTACION	VECINA	CORRELACION
ESTACION: 47030020 CONCEPCION\par	ESTACION: 47010150 CARRIZAL\par	0.152832
ESTACION: 47030020 CONCEPCION\par	ESTACION: 51035010 APTOLA FLORIDA\par	0.124627
ESTACION: 47030020 CONCEPCION\par	ESTACION: 47010170 VICHOY\par	0.123101
ESTACION: 47030020 CONCEPCION\par	ESTACION: 44010040 MINCHOY\par	0.117154
ESTACION: 47030020 CONCEPCION\par	ESTACION: 44010110 PTO LIMON\par	0.116463
...	...	...
ESTACION: 47030020 CONCEPCION\par	ESTACION: 26040310 RIO PALO\par	-0.100590
ESTACION: 47030020 CONCEPCION\par	ESTACION: 52080010 PISANDA\par	-0.112070
ESTACION: 47030020 CONCEPCION\par	ESTACION: 26040210 TRAPICHE EL\par	-0.132452
ESTACION: 47030020 CONCEPCION\par	ESTACION: 52070030 LLANOVERDE\par	-0.134981
ESTACION: 47030020 CONCEPCION\par	ESTACION: 52010060 MAMACONDE\par	-0.139562

Ilustración 5. Correlaciones con la estación Concepción

Al quitar aleatoriamente el 20% de los datos de la estación Mocoa Acueducto y realizar la imputación con los dos algoritmos desarrollados, los resultados fueron los siguientes. En la Ilustración 6 se ilustra el comportamiento del algoritmo KNN para la estación Mocoa Acueducto, y en la Ilustración 7 el algoritmo Curva de Doble Masa con la misma estación.

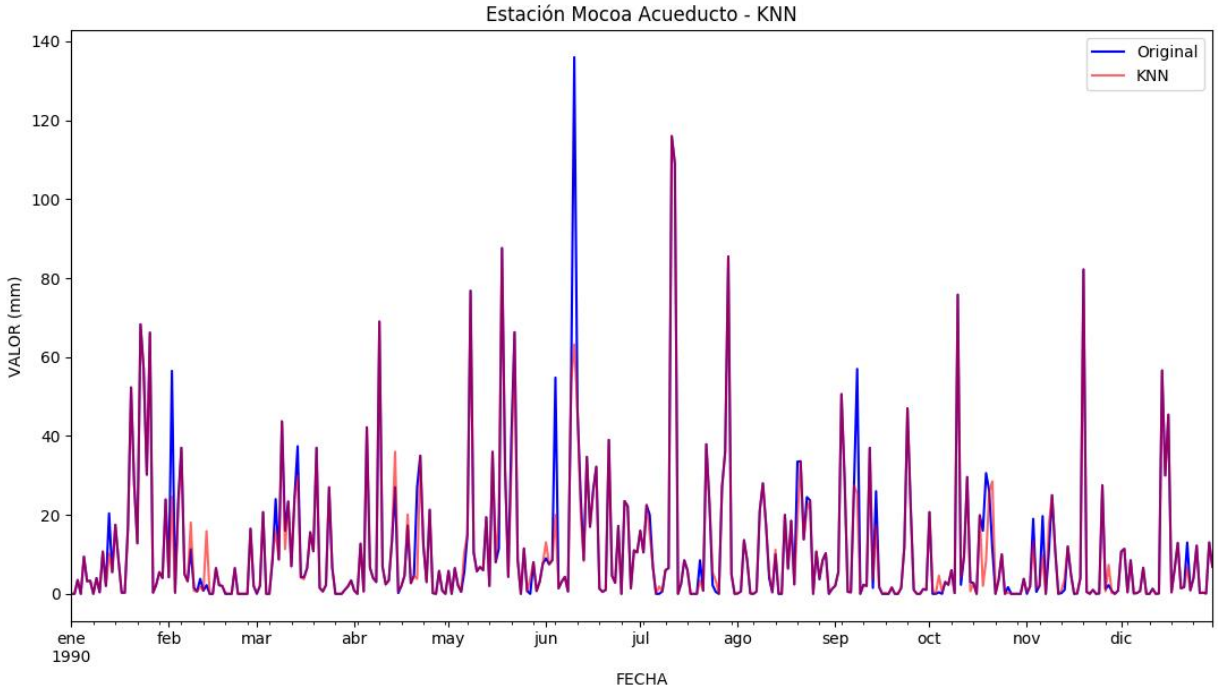

Ilustración 6. Imputación a la estación Mocoa Acueducto con KNN

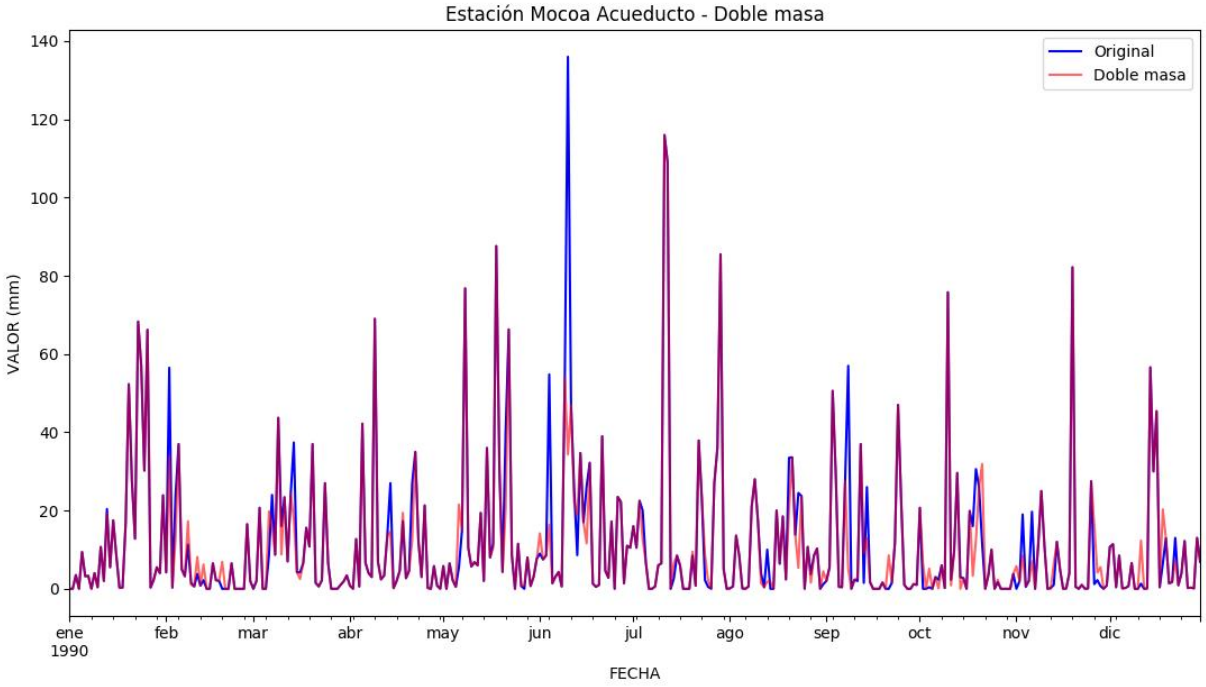

Ilustración 7. Imputación a la estación Mocoa Acueducto con Curva de Doble Masa

El desempeño promedio obtenido para esta estación se muestra en la Tabla 5**,** donde se puede ver que el algoritmo KNN tuvo errores más bajos que el de Curva de Doble Masa, con una diferencia de 4.1020 mm en el RMSE. También es de resaltar la eficiencia obtenida, pues según los valores referenciales del criterio de Nash-Suttclife, un NSE mayor a 0.6 indica un ajuste muy bueno. Según los resultados del PBIAS, ambos métodos subestiman los valores al imputar, esto se puede comprobar en las gráficas X y Y, donde ninguno de ellos logra acercarse al valor real de algunos picos.

Tabla 5. Desempeño de los algoritmos al imputar la estación Mocoa Acueducto

Medida de desempeño	**KNN**	**Doble Masa**	**Diferencia KNN-Doble Masa**
RMSE	12,5484	16,6504	4,1020
NSE	0,6139	0,3201	0,2938
PBIAS	24,5944	28,5207	3,9263
MAE	6,2221	8,9689	2,7468
Tiempo de ejecución	1,1229	10,1625	9,0396

A continuación, se muestran los resultados de los dos algoritmos al hacer la imputación a la estación Concepción del conjunto de datos de Nariño quitando también el 20% de los datos.

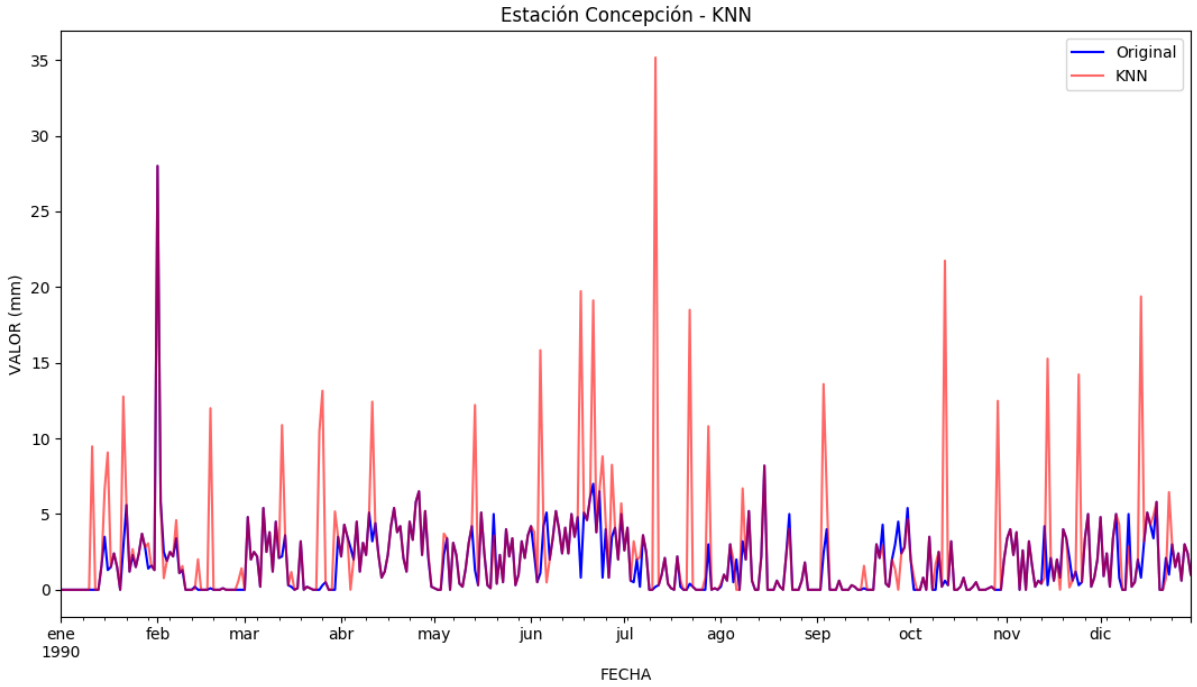

Ilustración 8. Imputación a la estación Concepción con el algoritmo KNN

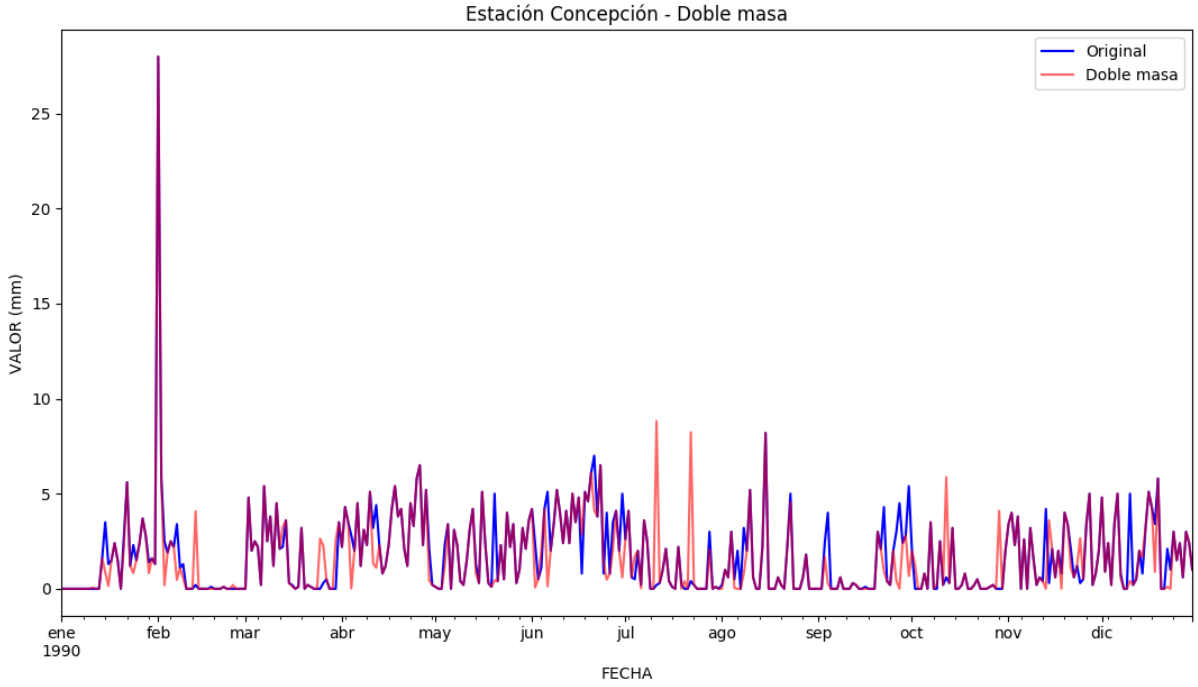

Ilustración 9. Imputación a la estación Concepción con el algoritmo Curva de Doble Masa

Según las gráficas obtenidas, se nota que el algoritmo de Curva de Doble Masa se mantiene dentro del rango de valores originales, no genera resultados tan elevados como los que se observan en la Ilustración 8. Además, en la columna Diferencia de la Tabla 6, se observa que el algoritmo de Doble Masa tiene errores más bajos, pues el RMSE y el MAE del KNN son 5.6659 mm y 3.3188 mm más altos, respectivamente. También el NSE es mejor con Doble Masa pues, aunque también es negativo, es más cercano a 0 que el del KNN; el PBIAS con valor 29.6695 demuestra que el algoritmo tiende a subestimar, mientras el KNN presenta una sobreestimación muy elevada.

Tabla 6. Desempeño de los algoritmos al imputar la estación Concepción

Medida de desempeño	KNN	Doble Masa	Diferencia KNN-Doble Masa
RMSE	8,3406	2,6747	5,6659
NSE	-20,4713	-1,2082	-19,2631
PBIAS	-241,0451	29,6695	-270,7146
MAE	5,2633	1,9445	3,3188
Tiempo de ejecución	1,3243	11,1050	-9,7807

Se llevó a cabo otra prueba donde a la estación Mocoa Acueducto se le eliminaron todos los valores de precipitación de los 12 meses de manera consecutiva y se realizó la imputación haciendo uso de los dos métodos. Los resultados se muestran en la Tabla 7 donde se observa que en la mayoría se obtuvieron mejores resultados con KNN para el RMSE, NSE y MAE; con excepción de febrero, octubre y noviembre donde el desempeño fue mejor con el algoritmo de Curva de Doble Masa.

Tabla 7. Imputación por cada mes a la estación Mocoa Acueducto

	RMSE		NSE		PBIAS		MAE		TIEMPO	
MES	**KNN**	**DOBLE MASA**	**KNN**	**DOBLE MASA**	**KNN**	**DOBLE MASA**	**KNN**	**DOBLE MASA**	**KNN**	**DOBLE MASA**
Enero	15,4745	15,8771	0,3888	0,3565	29,0381	0,2701	10,4287	9,4835	0,4851	4,7845
Febrero	9,7411	7,2519	0,4042	0,6698	7,5647	-16,0645	5,2564	5,0913	0,4457	4,0330
Marzo	7,0986	10,5400	0,6760	0,2857	20,3150	-6,5995	4,8343	7,9047	0,5032	4,3543
Abril	10,6870	13,3872	0,5173	0,2426	15,5704	-14,5962	6,5794	9,2993	0,4820	4,1978
Mayo	11,3617	21,8967	0,7443	0,0502	7,9954	4,4068	6,7030	14,1864	0,4862	4,7022
Junio	16,5609	21,4933	0,6211	0,3618	4,9229	27,3262	9,1636	10,8229	0,4708	4,4368
Julio	10,2302	19,4696	0,8851	0,5837	18,0401	31,1447	5,2042	8,6260	0,4963	4,4602
Agosto	7,2454	8,3899	0,5215	0,3584	9,7547	23,0017	4,7491	5,9961	0,4892	4,5422
Septiembre	11,1356	18,6741	0,5739	-0,1982	29,2274	12,2863	6,5395	12,9042	0,4715	4,2507
Octubre	12,2654	11,3025	0,3556	0,4528	42,5193	0,3265	6,1651	7,2236	0,5025	4,2996
Noviembre	10,9279	8,8827	0,5280	0,6881	26,3124	-16,2399	4,9146	5,8532	0,4738	4,2125
Diciembre	7,5385	11,5238	0,6668	0,2213	-8,6727	-49,3798	4,5864	7,7888	0,4900	4,2675

En la Tabla 8**¡Error! No se encuentra el origen de la referencia.**, se encuentran los valores imputados por ambos algoritmos para el mes de febrero y en la Tabla 9 los de mayo. Se incluyen estos resultados pues, según la primera tabla, para estos dos meses los algoritmos presentan una diferencia considerable en el RMSE y el NSE.

La Tabla 7, expone que en febrero el algoritmo Curva de Doble Masa tiene un RMSE de 7.2519 mm mientras que el de KNN tiene 9.7411 mm, siendo una diferencia de 2,4892 mm. Y en cuestiones de eficiencia, el primer algoritmo tiene muy buen desempeño, pues su NSE es mayor a 0.6. Para mayo, la diferencia es más notoria, el método KNN obtuvo un RMSE de 11.3617 mm y el de Doble Masa 21.8967 mm, 10.5350 mm más de error.

Por otro lado, la Tabla 8 muestra los errores relativos para ambos métodos, donde:

DM: valor estimado por el algoritmo Curva de Doble Masa

KNN: valor estimado por el algoritmo KNN

Original: valor observado

DIF_DM: diferencia entre el valor observado y el estimado por el algoritmo Curva de Doble Masa.

%DIF_DM: error relativo entre el valor observado y el estimado por el algoritmo Curva de Doble Masa

DIF_KNN: diferencia entre el valor observado y el estimado por el algoritmo KNN.

%DIF_KNN: error relativo entre el valor observado y el estimado por el algoritmo KNN

Comparando los métodos por el promedio de los errores obtenidos en la Tabla 8, se ve que la diferencia también es mínima como sucede con el RMSE, pues los valores son muy cercanos. Mientras que, en el mes de mayo, KNN se destaca, mostrando en la Tabla 9 un error relativo de 1.2699 comparado con el error de 4.9099 que tiene el método de Doble Masa.

Tabla 8. Error relativo al imputar todo el mes de febrero en la estación Mocoa Acueducto

MES	DM	KNN	ORIGINAL	DIF_DM	%_DM	DIF_KNN	%_KNN
Febrero	0	0,4112	0	0	0	-0,4112	NaN
	13,3072	19,6592	5	-8,3072	1,6614	-14,6592	2,9318
	0,2410	1,2358	0,3	0,0590	0,1968	-0,9358	3,1192
	7,3222	1,5346	6,5	-0,8222	0,1265	4,9654	0,7639
	0	0	0	0	0	0	0
	0,6998	0	0	-0,6998	NaN	0	0
	10,5395	6,5949	2,2	-8,3395	3,7907	-4,3949	1,9977
	7,8451	11,2975	2,2	-5,6451	2,5659	-9,0975	4,1352
	1,3121	0,9407	1,5	0,1879	0,1253	0,5593	0,3728
	0	0	0	0	0	0	0
	10,3006	3,6145	3,8	-6,5006	1,7107	0,1855	0,0488
	42,8657	19,4212	56,5	13,6343	0,2413	37,0788	0,6563
	7,4886	3,7417	6,5	-0,9886	0,1521	2,7583	0,4244
	14,5942	7,6131	4,2	-10,3942	2,4748	-3,4131	0,8127
	8,3907	2,9929	0	-8,3907	NaN	-2,9929	NaN
	0,6961	2,4652	2	1,3039	0,6519	-0,4652	0,2326
	18,2939	15,0196	37	18,7061	0,5056	21,9804	0,5941
	3,6682	5,0129	16,5	12,8318	0,7777	11,4871	0,6962
	0	0	0	0	0	0	0
	1,0321	5,1100	0	-1,0321	NaN	-5,1100	NaN
	8,6976	0,2167	0	-8,6976	NaN	-0,2167	NaN
	16,0783	22,0013	23,6	7,5217	0,3187	1,5987	0,0677
	21,8221	17,2808	11,3	-10,5221	0,9312	-5,9808	0,5293
	0	2,5737	2	2,0000	1,0000	-0,5737	0,2869
	7,4354	3,7213	0,8	-6,6354	8,2942	-2,9213	3,6516
	12,4195	18,0999	3,2	-9,2195	2,8811	-14,8999	4,6562
	0,4819	0,6826	0,6	0,1181	0,1968	-0,0826	0,1377
	0	0,4112	0	0	0	-0,4112	NaN
Promedio	7,6976	6,1304	6,6321	-1,0654	1,1918	0,5017	1,1354

Tabla 9. Error relativo al imputar todo el mes de mayo en la estación Mocoa Acueducto

MES	DM	KNN	ORIGINAL	DIF_DM	%_DM	DIF_KNN	%_KNN
	1,0184	2,2962	0	-1,0184	NaN	-2,2962	NaN
	22,5190	7,5446	4,3	-18,2190	4,2370	-3,2446	0,7546
	14,4020	13,8580	36	21,5980	0,5999	22,1420	0,6151
	3,0778	3,2809	0,7	-2,3778	3,3969	-2,5809	3,6870
	0,9456	1,5238	8	7,0544	0,8818	6,4762	0,8095
	26,6886	13,8833	5,7	-20,9886	3,6822	-8,1833	1,4357
	52,2521	2,8346	0,5	-51,7521	103,5041	-2,3346	4,6692
	23,7661	4,1115	5,5	-18,2661	3,3211	1,3885	0,2525
	23,2785	28,6367	15	-8,2785	0,5519	-13,6367	0,9091
	15,1842	8,2356	7,7	-7,4842	0,9720	-0,5356	0,0696
	20,3155	34,4098	31	10,6845	0,3447	-3,4098	0,1100
	4,8461	9,5156	11,5	6,6539	0,5786	1,9844	0,1726
	12,0880	5,3414	2	-10,0880	5,0440	-3,3414	1,6707
	4,4888	10,5628	0,7	-3,7888	5,4125	-9,8628	14,0897
	2,0106	5,4113	6,5	4,4894	0,6907	1,0887	0,1675
Mayo	16,8599	67,5877	76,8	59,9401	0,7805	9,2123	0,1200
	22,1801	32,9880	40	17,8199	0,4455	7,0120	0,1753
	17,6382	20,9461	19,4	1,7618	0,0908	-1,5461	0,0797
	10,1177	7,7825	6,8	-3,3177	0,4879	-0,9825	0,1445
	14,7307	19,6544	10,5	-4,2307	0,4029	-9,1544	0,8718
	0,5092	1,8169	2,2	1,6908	0,7685	0,3831	0,1741
	14,8185	16,7650	6	-8,8185	1,4698	-10,7650	1,7942
	11,6124	14,7283	11,5	-0,1124	0,0098	-3,2283	0,2807
	28,6888	40,5436	87,6	58,9112	0,6725	47,0564	0,5372
	0	0,0319	0	0	0	-0,0319	NaN
	14,9061	4,5220	8	-6,9061	0,8633	3,4780	0,4347
	26,2686	45,3148	66,3	40,0314	0,6038	20,9852	0,3165
	30,5892	4,5930	0	-30,5892	NaN	-4,5930	NaN
	9,1320	5,5817	5,8	-3,3320	0,5745	0,2183	0,0376
	7,3906	1,1310	3,1	-4,2906	1,3841	1,9690	0,6351
	13,8848	13,2733	8,6	-5,2848	0,6145	-4,6733	0,5434
Promedio	15,0390	14,4744	15,7323	0,6933	4,9099	1,2579	1,2699

Por otro lado, se eligieron nuevamente las estaciones Mocoa Acueducto y Concepción para eliminarles al azar el 5%, 10%, 15% y 20% y ejecutar los algoritmos diez veces para cada porcentaje. Esto con el fin de observar el desempeño de ambos métodos al aumentar la cantidad de datos faltantes.

Los resultados muestran que, para Mocoa Acueducto, el algoritmo KNN mantiene errores más bajos que el de Curva de Doble Masa, para las cantidades analizadas, pero la diferencia entre el RMSE y el MAE es poca. Mientras que para la estación Concepción, el

algoritmo de Curva Doble Masa conserva un desempeño mejor y una diferencia notable con el KNN.

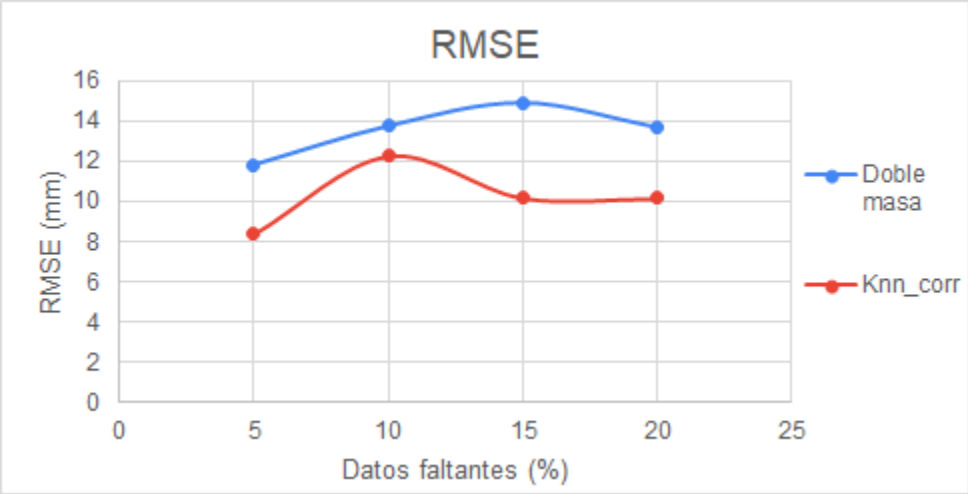

Ilustración 10. RMSE para la estación Mocoa Acueducto

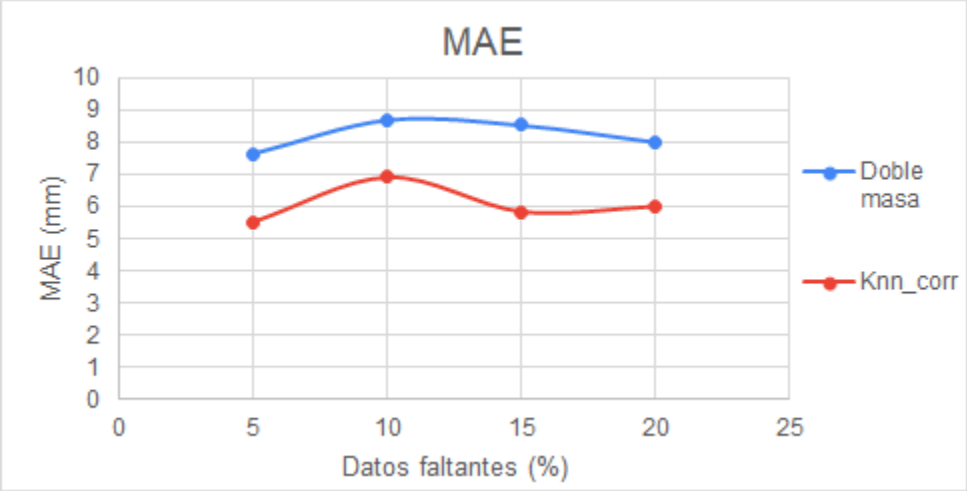

Ilustración 11. MAE para la estación Mocoa Acueducto

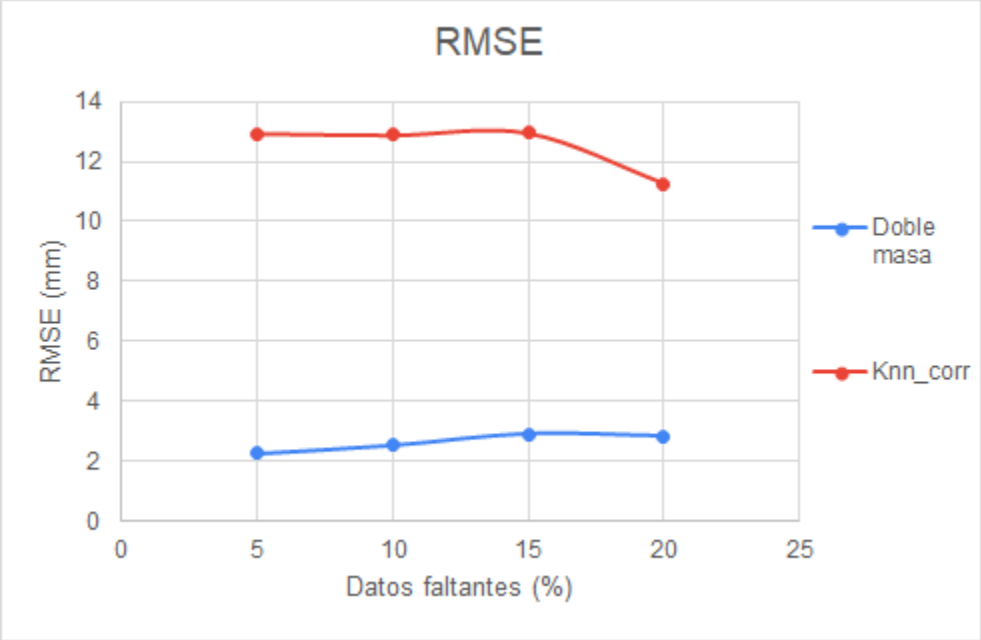

Ilustración 12. RMSE para la estación Concepción

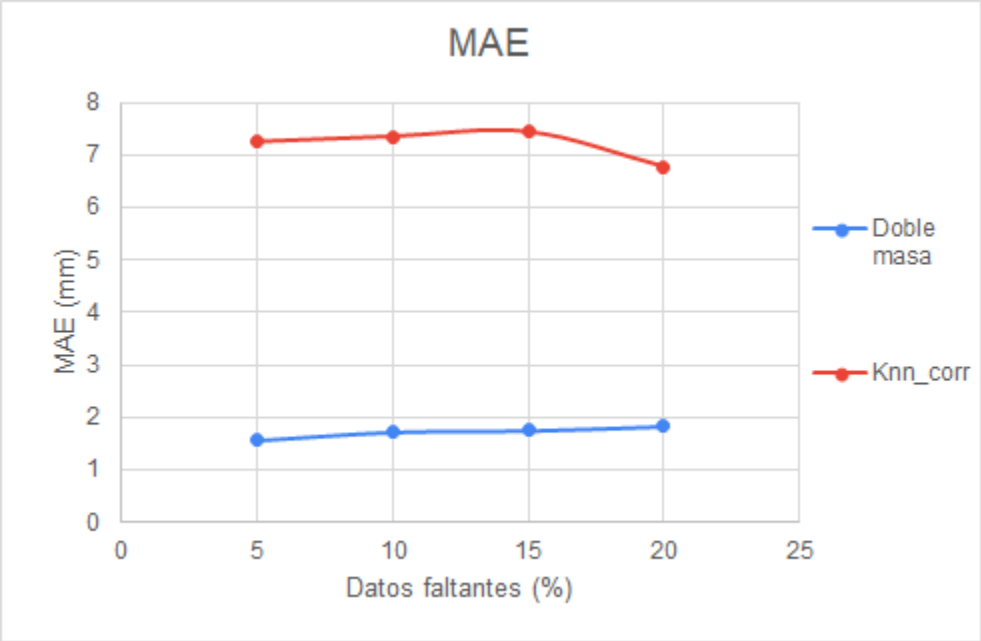

Ilustración 13. MAE para la estación Concepción

Durante el proceso de desarrollo se planteó la hipótesis que el algoritmo KNN tendría mejor desempeño que el de Curva de Doble Masa en aquellas estaciones con mayor correlación con sus vecinas.

Para corroborar esta suposición, a cada estación se le quitaron aleatoriamente 73 datos, correspondientes al 20% del año, mientras las demás estaciones permanecían con sus datos completos para servir como vecinas. Luego, se ejecutan los algoritmos KNN y Curva de Doble Masa para realizar la imputación y comparar los valores estimados con los observados. Los resultados completos para el conjunto de Nariño se encuentran en el Anexo A, y los de Quindío en el Anexo B.

De la información obtenida en la prueba, se evidencia que hay estaciones que tienen correlaciones considerablemente altas con sus vecinas y aun así el algoritmo de Curva de Doble Masa obtiene mejores resultados. Por ejemplo, en la Tabla 10 se muestra la estación Mocoa Acueducto, con un promedio de correlación con sus vecinas de 0.6839**,** para la cual el algoritmo KNN obtuvo, en promedio, errores más bajos (RMSE y MAE) que el de Curva de Doble Masa. Sin embargo, la estación Taminango que tuvo un promedio de correlación con sus vecinas de 0.6231 obtuvo mejores resultados en la imputación con el algoritmo Curva de Doble Masa.

Tabla 10. Desempeño de los algoritmos al imputar dos estaciones con correlaciones altas

		Algoritmo		
Estación	**Medida de desempeño**	**KNN**	**Doble Masa**	**Diferencia KNN-Doble Masa**
Mocoa Acueducto	RMSE	10,3744	14,9509	-4,5765
	NSE	0,6928	0,3621	0,3306
	PBIAS	15,2158	18,5944	-3,3786
	MAE	5,8610	8,7197	-2,8587
	Correlación promedio	0,6839	-	-
	Tiempo de ejecución	1,1361	10,6483	-9,5121
Taminango	RMSE	7,1318	6,4638	0,6680
	NSE	0,4671	0,5603	-0,0932
	PBIAS	15,6790	14,0863	1,5926
	MAE	3,0371	2,8242	0,2129
	Correlación promedio	0,6231	-	-
	Tiempo de ejecución	1,1378	10,9486	-9,8108

En la Tabla 11 se encuentran los valores obtenidos en una de las diez ejecuciones que se realizaron a la estación Mocoa Acueducto para imputar el 20% de los datos que se removieron para la prueba.

Tabla 11. Resultados de la imputación a la estación Mocoa Acueducto (20% de datos faltantes)

DATO IMPUTADO						
DM	**KNN**	**ORIGINAL**	**DIF_DM**	**%DIF_DM**	**DIF_KNN**	**%DIF_KNN**
0,1970	1,2846	0,3	0,1030	0,3432	-0,9846	3,2820
7,6458	2,7957	2,2	-5,4458	2,4754	-0,5957	0,2708
6,9202	1,1607	3,1	-3,8202	1,2323	1,9393	0,6256
13,9572	4,6677	8	-5,9572	0,7446	3,3323	0,4165
5,2964	9,5715	18,5	13,2036	0,7137	8,9285	0,4826
18,9726	27,0385	33,6	14,6274	0,4353	6,5615	0,1953
5,4073	8,9931	3,3	-2,1073	0,6386	-5,6931	1,7252
0	1,9925	0	0	0	-1,9925	N.A.
15,2022	11,0980	2	-13,2022	6,6011	-9,0980	4,5490
4,6822	0,4197	3	-1,6822	0,5607	2,5803	0,8601
14,8114	10,0603	5,5	-9,3114	1,6930	-4,5603	0,8292
9,4327	5,0948	1,5	-7,9327	5,2885	-3,5948	2,3965
29,1146	32,0024	66,2	37,0854	0,5602	34,1976	0,5166

DATO IMPUTADO						
DM	**KNN**	**ORIGINAL**	**DIF_DM**	**%DIF_DM**	**DIF_KNN**	**%DIF_KNN**
2,2038	2,5754	0,5	-1,7038	3,4077	-2,0754	4,1508
19,3979	6,5419	8,6	-10,7979	1,2556	2,0581	0,2393
6,0254	7,1560	2,8	-3,2254	1,1519	-4,3560	1,5557
12,6547	14,4387	1,6	-11,0547	6,9092	-12,8387	8,0242
0,1532	0,3540	0,7	0,5468	0,7811	0,3460	0,4942
26,8635	40,6672	87,6	60,7365	0,6933	46,9328	0,5358
16,8302	13,9769	24	7,1698	0,2987	10,0231	0,4176
25,7014	46,5928	45,4	19,6986	0,4339	-1,1928	0,0263
6,1227	3,7227	6,5	0,3773	0,0581	2,7773	0,4273
8,5821	0,2979	0	-8,5821	N.A.	-0,2979	N.A.
0,4135	6,9426	7	6,5865	0,9409	0,0574	0,0082
4,8773	0,5115	0,8	-4,0773	5,0966	0,2885	0,3606
27,9844	42,6667	32,2	4,2156	0,1309	-10,4667	0,3251
11,9094	9,2167	10,8	-1,1094	0,1027	1,5833	0,1466
7,2672	7,6366	19,7	12,4328	0,6311	12,0634	0,6124
7,9499	7,3573	4	-3,9499	0,9875	-3,3573	0,8393
6,6749	0,5959	0	-6,6749	N.A.	-0,5959	N.A.
18,7007	7,8341	11,4	-7,3007	0,6404	3,5659	0,3128
0,3941	0,6797	0,6	0,2059	0,3432	-0,0797	0,1328
3,4072	15,0935	5	1,5928	0,3186	-10,0935	2,0187
13,1460	21,9436	23,6	10,4540	0,4430	1,6564	0,0702
0,7881	2,1270	1	0,2119	0,2119	-1,1270	1,1270
29,9062	47,1041	36	6,0938	0,1693	-11,1041	0,3084
2,0718	0,4122	0	-2,0718	N.A.	-0,4122	N.A.
35,4465	87,7398	136	100,5535	0,7394	48,2602	0,3549
26,5468	17,9428	52,3	25,7532	0,4924	34,3572	0,6569
8,6841	3,9604	3,4	-5,2841	1,5542	-0,5604	0,1648
0	0,1374	0	0	0	-0,1374	N.A.
0,0876	3,4663	0	-0,0876	N.A.	-3,4663	N.A.
0	0	0	0	0	0	0
15,3528	0,9694	1	-14,3528	14,3528	0,0306	0,0306
20,8138	9,9446	23,9	3,0862	0,1291	13,9554	0,5839
0	0,1374	0	0	0	-0,1374	
11,6947	8,9096	6,5	-5,1947	0,7992	-2,4096	0,3707
7,8810	1,1761	0,6	-7,2810	12,1350	-0,5761	0,9602
7,1111	0,2151	0	-7,1111	N.A.	-0,2151	N.A.
0	0,5728	0	0	0	-0,5728	N.A.
9,0622	3,2865	2,2	-6,8622	3,1192	-1,0865	0,4939

DATO IMPUTADO						
DM	**KNN**	**ORIGINAL**	**DIF_DM**	**%DIF_DM**	**DIF_KNN**	**%DIF_KNN**
9,1052	11,6158	11	1,8948	0,1723	-0,6158	0,0560
0	0	0	0	0	0	0
9,4738	7,8278	6,8	-2,6738	0,3932	-1,0278	0,1511
17,8284	13,3104	24,3	6,4716	0,2663	10,9896	0,4522
33,5014	32,1653	57	23,4986	0,4123	24,8347	0,4357
8,7489	11,7596	19	10,2511	0,5395	7,2404	0,3811
13,9654	7,0138	6	-7,9654	1,3276	-1,0138	0,1690
9,1887	11,8528	10,5	1,3113	0,1249	-1,3528	0,1288
14,7651	8,0827	16	1,2349	0,0772	7,9173	0,4948
0	0,0687	0	0	0	-0,0687	N.A.
0,5035	1,1225	0,2	-0,3035	1,5176	-0,9225	4,6127
0	0,2405	0,6	0,6000	1	0,3595	0,5992
4,5377	9,6485	11,5	6,9623	0,6054	1,8515	0,1610
15,6385	23,2750	37	21,3615	0,5773	13,7250	0,3709
0	2,1886	2	2,0000	1	-0,1886	0,0943
7,4782	4,4516	13	5,5218	0,4248	8,5484	0,6576
2,5848	4,2597	4,2	1,6152	0,3846	-0,0597	0,0142
18,4114	23,3988	37,4	18,9886	0,5077	14,0012	0,3744
4,9756	4,1451	3,1	-1,8756	0,6050	-1,0451	0,3371
0,7332	0,1763	0,5	-0,2332	0,4663	0,3237	0,6475
0	0,2748	0,2	0,2000	1	-0,0748	0,3742
9,2742	6,1308	1	-8,2742	8,2742	-5,1308	5,1308

De las 65 estaciones analizadas, solamente en 21 el algoritmo KNN obtuvo un RMSE más bajo que el de Curva de Doble Masa, como se puede ver en el Anexo A, donde los valores que están en negrita indican que ese método tuvo un RSME menor que el otro.

Sin embargo, al comparar los valores del RMSE de ambos métodos, en 12 de estas 21 estaciones, se evidencia que estos difieren en menos de un milímetro (1 mm). En las otras nueve, la mayor diferencia se da en la estación Mocoa Acueducto donde el algoritmo de Doble Masa tuvo un RMSE de 14.95 y el KNN tuvo 10.37, es decir, que hubo una diferencia de 4.58 milímetros.

Además, se calculó un promedio tanto de los valores estimados, los valores originales y las diferencias, como de los errores relativos en cada estación para las diez ejecuciones. El resumen se muestra en la Tabla 12, en la cual los valores en negrita indican que el error relativo fue menor en el método indicado en el nombre de la columna (%DIF_DM o %DIF_KNN).

Se puede notar que gran parte de los valores de la columna %DIF_DM están en negrita, lo cual indica que en promedio el algoritmo de Doble Masa tuvo un error relativo más bajo que el KNN, para la mayoría de las estaciones. De donde se infiere que los valores estimados por el algoritmo de Doble Masa se aproximaron más a los valores observados que los imputados por el KNN.

Tabla 12. Errores relativos en las estaciones del conjunto de datos de Nariño

ESTACIÓN	DM	KNN	ORIGINAL	DIF_DM	%DIF_DM	DIF_KNN	%DIF_KNN
TAMBO	3,6705	4,4642	4,6589	0,9884	**0,8788**	0,1947	0,9822
CORINTO	3,7359	4,8267	4,7055	0,9696	**0,3376**	-0,1212	0,4474
RIO PALO	4,6936	3,7516	7,1193	2,4257	0,3557	3,3677	**0,3044**
DINDE	5,5118	6,1590	6,9945	1,4827	**0,8840**	0,8356	0,9574
GUACHENE	3,2289	5,8313	3,8041	0,5752	**0,5442**	-2,0272	0,8658
VENTADE CAJIBIO	4,6447	5,5971	5,9326	1,2879	3,1254	0,3355	**2,0131**
BUENOS AIRES	4,3895	5,4902	6,0882	1,6988	**0,2924**	0,5980	0,3367
GUACHICONO	5,0974	4,1179	6,3288	1,2313	**0,2375**	2,2109	0,3093
APTOG L VALENCIA	5,1230	4,6377	6,2019	1,0789	**2,5959**	1,5643	2,9510
JULUMITO ALERTAS	4,5232	5,8147	5,9342	1,4111	**0,6032**	0,1196	0,7232
SATE	5,3024	5,7674	6,6562	1,3538	0,8886	0,8888	**0,8544**
CHURUYACO	12,3737	11,6065	14,9288	2,5550	1,0255	3,3223	**0,8905**
CARRIZAL	5,4744	7,9473	6,2495	0,7751	**1,4644**	-1,6979	2,5233
MAMACONDE	1,7533	3,4047	2,2963	0,5430	**0,8015**	-1,1084	1,3965
SAN JOSE DE TAPAJE	11,5936	4,7813	15,2863	3,6927	0,8813	10,5050	**0,5734**
UNIONLA	3,2145	3,2986	4,6293	1,4148	**0,3889**	1,3307	0,4686
AGUADA LA	3,7855	4,8330	4,7466	0,9611	**0,7652**	-0,0865	1,0121
CHUNGACASPI	7,5358	9,4386	8,4919	0,9561	3,0396	-0,9467	**2,8942**
PEPINO EL	11,9696	9,4736	15,1011	3,1315	3,2960	5,6275	**1,4219**

ESTACIÓN	DM	KNN	ORIGINAL	DIF_DM	%DIF_DM	DIF_KNN	%DIF_KNN
IMUES	1,3131	3,8140	1,6723	0,3593	**0,4595**	-2,1417	1,9524
GUASCA LA	1,1310	3,7668	1,1507	0,0197	**0,1523**	-2,6161	0,4551
PENOLEL	1,7706	4,3476	2,2770	0,5063	**0,4262**	-2,0706	1,2563
PUERRES	2,2546	6,8854	2,3390	0,0845	**0,4689**	-4,5463	1,3221
PICUDO EL	8,5731	13,5403	10,5926	2,0195	**0,8926**	-2,9477	1,3383
BERRUECOS	2,3209	3,3460	3,4421	1,1212	**0,8753**	0,0961	1,1216
PISANDA	2,2809	4,3180	2,8449	0,5641	**0,5713**	-1,4730	0,9297
NARINO	3,2151	3,5932	4,0999	0,8847	**1,2099**	0,5067	1,5829
HIDROMAYOCAMP	1,9354	3,9201	2,7092	0,7738	**2,4914**	-1,2109	4,4173
BALSALA	3,5965	5,6551	4,8315	1,2351	**0,3705**	-0,8236	0,6046
BOCATOMA ALERTAS	3,3483	4,7065	4,4192	1,0709	**0,3777**	-0,2874	0,5729
VIVERO LINARES	2,0148	4,1728	2,9078	0,8930	**0,7587**	-1,2650	1,5623
OVEJAS ABAJO ALERT	5,0535	5,1335	5,5178	0,4643	0,6166	0,3843	**0,5705**
CALERA LA	3,3005	5,9772	4,9479	1,6475	**0,4044**	-1,0293	0,5414
QUINCHOA	5,2531	8,7046	6,0877	0,8346	**1,1551**	-2,6169	1,6367
MOSQUERA	5,2237	7,0781	6,2762	1,0525	**0,8604**	-0,8019	1,4178
SAUCEEL	4,0076	4,0366	4,5201	0,5125	**0,9904**	0,4836	1,6556
PTO CAICEDO	9,8936	6,5402	12,9748	3,0812	3,6533	6,4345	**2,1305**
JUNIN	17,1702	3,3517	20,7774	3,6072	1,6373	17,4257	**0,9247**
GAMBOA	5,5507	6,2794	6,1205	0,5699	**0,7464**	-0,1589	0,8349
VERGEL EL	4,5429	3,2698	5,7489	1,2060	**1,3003**	2,4791	1,4510
LLANOVERDE	3,5689	3,5214	5,1658	1,5969	0,4668	1,6444	**0,4505**
SANDONA	2,0710	6,1003	2,3634	0,2924	**0,5951**	-3,7369	1,3459
CATALINA LA	5,9161	6,1191	6,9164	1,0003	**0,5822**	0,7973	0,6014
OSPINA PEREZ	2,7100	3,4220	3,3292	0,6192	1,2676	-0,0929	**0,9800**
REMOLINO GRANDE	3,9604	7,7492	4,8374	0,8770	**1,2842**	-2,9118	1,9171
TAMINANGO	3,5105	3,4812	4,1618	0,6513	0,9126	0,6805	**0,7643**
TANGUA	1,3718	3,0846	2,1589	0,7871	**0,7129**	-0,9257	0,9207
CONDAGUA	7,4443	10,8350	9,4703	2,0260	**0,6421**	-1,3648	0,7196
MATAJE	6,6708	4,4955	7,6803	1,0095	2,6546	3,1848	**1,5417**
COCOEL	3,8804	5,1064	3,9711	0,0907	**3,6066**	-1,1353	4,5582
CONCEPCION	1,4650	7,5788	1,7786	0,3136	**1,7287**	-5,8002	9,5444
APTOLA FLORIDA	2,8669	4,8912	4,0773	1,2104	**2,7123**	-0,8140	3,9575
PTO LIMON	10,8184	9,0511	12,7216	1,9033	2,1570	3,6706	**1,8628**
MOCOAACUEDUCTO	10,0552	10,4894	12,4104	2,3552	2,4328	1,9211	**1,5869**
VICHOY	3,6683	8,9728	5,2440	1,5757	**1,2067**	-3,7288	2,7051
PTO ASIS	7,5359	13,4272	10,5607	3,0248	**1,2942**	-2,8665	2,1482
CHACHAGUI	2,7098	3,3771	3,3656	0,6559	**1,9928**	-0,0115	2,0638

ESTACIÓN	DM	KNN	ORIGINAL	DIF_DM	%DIF_DM	DIF_KNN	%DIF_KNN
TRAPICHE EL	3,8096	4,4846	4,5192	0,7096	**0,6425**	0,0346	0,7182
BUESACO	3,1734	3,6246	3,4845	0,3111	**0,2919**	-0,1400	0,3601
PRESAHIDROMAYO	2,8909	3,5210	4,3460	1,4551	**0,4657**	0,8250	0,5780
MAGUI	7,0498	4,6693	10,5578	3,5080	1,9069	5,8885	**1,4543**
BUENOS AIRES	2,4677	6,6990	4,1247	1,6570	**0,6252**	-2,5743	0,7350
MINCHOY	6,8910	9,1493	10,2337	3,3427	2,7083	1,0844	**2,4569**
SAN FRANCISCO	2,9871	9,2165	3,7697	0,7826	**0,6996**	-5,4468	1,4042
CUMBAL	1,9607	4,0609	2,2915	0,3308	**0,4829**	-1,7694	0,6251

Este procedimiento también se ejecuta con los datos del conjunto de Quindío, que cuenta solo con seis estaciones, de las cuales solo cuatro tuvieron medidas de RMSE y MAE más bajas con el algoritmo KNN (ver Anexo B). No obstante, en tres de las cuatro estaciones este método tuvo una diferencia, en el RMSE, menor a un milímetro respecto al de Curva de Doble Masa. De modo que el desempeño de ambos algoritmos fue muy parejo.

6. Conclusiones, aportes y recomendaciones

- El algoritmo de los k vecinos más cercanos (KNN) depende principalmente de la correlación que tenga la estación a imputar con sus k vecinas. La imputación tendrá un error más bajo a medida que la correlación entre las estaciones sea mayor, es decir, más cercana a 1, y que además los datos entre las estaciones tengan una diferencia muy baja ya que los valores de las estaciones vecinas influyen directamente en el valor que va a ser imputado.
- Si se desea imputar datos a una estación A con K estaciones vecinas idénticas a ella, pero con sus datos completos, el error será más cercano cero entre menos datos le falten a la estación A. Esto aplica para ambos algoritmos.

- El algoritmo KNN en tiempo de ejecución es mucho más rápido que el algoritmo Curva de Doble Masa, debido a que este último debe hacer una consulta a la base de datos cada vez que se va a imputar un dato, para obtener las vecinas de la estación problema. Este tiempo de petición y respuesta hace que el algoritmo se tarde un poco. Mientras que el algoritmo KNN calcula y guarda las correlaciones localmente así que al consultar las correlaciones de las estaciones vecinas el tiempo de respuesta muy corto permitiendo al algoritmo terminar en mucho menos tiempo.
- Al realizar la imputación con el algoritmo KNN hay que tener en cuenta, además de la correlación, que el rango de los valores de las estaciones vecinas debe ser similar al de la estación problema. Pues si las vecinas con mejores correlaciones presentan valores muy altos y la estación a imputar suele registrar unos muy bajos, el algoritmo puede presentar una sobreestimación alta. También se puede presentar una subestimación, en el caso que la estación a imputar tenga valores superiores a sus vecinas.
- Puede presentarse el caso que, al elegir las K vecinas más cercanas, las que tengan mayor correlación tampoco presenten un valor para el día del dato faltante, por lo que se va a realizar la imputación con otras vecinas de menor correlación y el error podría aumentar.
- No es adecuado asociar un algoritmo a cada estación, sino elegir el que mejor se pueda acoplar a las condiciones de precipitación que se hayan presentado en el mes a imputar.

- Se deja abierta la posibilidad de adaptar el algoritmo de KNN asociando las estaciones por otros parámetros como la temperatura o la humedad, para comparar sus resultados con el desarrollado en este trabajo.

7. Referencias

Agrimetsoft. (s.f.). *Agricultural and Meteorological Software.* Recuperado el 3 de Octubre de 2019, de https://agrimetsoft.com/data-tool

Aler Mur, R. (s.f.). *Universidad Carlos III de Madrid.* Recuperado el 30 de Septiembre de 2019, de http://ocw.uc3m.es/ingenieria-informatica/analisis-de-datos/transparencias/KNNyPrototipos.pdf

Autoridad Nacional del Agua. (2016). *Autoridad Nacional del Agua.* Recuperado el 3 de Octubre de 2019, de https://www.ana.gob.pe/sites/default/files/normatividad/files/memoria_final.pdf

Ayuntamiento de a Coruña. (s.f.). Recuperado el 25 de Junio de 2019, de http://teleformacion.edu.aytolacoruna.es/AYC/document/atmosfera_y_clima/humedad/precipitaciones0.htm

Bennett, D. A. (Mayo de 2001). *Wiley Online Library*. Recuperado el 29 de Junio de 2019, de https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-842X.2001.tb00294.x

Bronshtein, A. (11 de Abril de 2017). *Medium.* Recuperado el 30 de Septiembre de 2019, de https://blog.usejournal.com/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7

El Cenit. (s.f.). *EL CENIT_ Monitoreo de riesgo de desastres.* Pereira.

González, A. (1 de Julio de 2014). *Clever Data*. Recuperado el 25 de Junio de 2019, de https://cleverdata.io/que-es-machine-learning-big-data/

Güzel, C., Kaya, M., Yıldız, O., & Bilge, H. Ş. (16 de Marzo de 2013). *Academia*. Recuperado el 29 de Junio de 2019, de https://www.academia.edu/35469686/Breast_Cancer_Diagnosis_Based_on_Na%C3%AFve_Bayes_Machine_Learning_Classifier_with_KNN_Missing_Data_Imputation

Hassanat, A., Abbadi, M., Altarawneh, G., & Alhasanat, A. (Agosto de 2014). *arXiv.* Recuperado el 23 de Septiembre de 2019, de https://arxiv.org/ftp/arxiv/papers/1409/1409.0919.pdf

IDEAM. (s.f.). Recuperado el 25 de Junio de 2019, de http://www.ideam.gov.co/web/agua/modelacion-hidrologica

Kanchana, S., & Thanamani, A. S. (Febrero de 2016). *ResearchGate*. Obtenido de https://www.researchgate.net/publication/299186135_Elevating_the_Accuracy_of_Missing_Data_Imputation_Using_Bolzano_Classifier

Kang, H. (24 de Mayo de 2013). *National Center for Biotechnology Information*. Recuperado el 29 de Junio de 2019, de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/

Laguna, C. (s.f.). *Instituto Aragonés de Ciencias de la Salud.* Recuperado el 4 de Octubre de 2019, de http://www.ics-aragon.com/cursos/salud-publica/2014/pdf/M2T04.pdf

Lodder, P. (2013). *Paul Twin*. Recuperado el 29 de Junio de 2019, de https://www.paultwin.com/wp-content/uploads/Lodder_1140873_Paper_Imputation.pdf

MathWorks. (s.f.). *MathWorks.* Recuperado el 30 de Septiembre de 2019, de https://es.mathworks.com/discovery/supervised-learning.html

Medina, F., & Galván, M. (Julio de 2007). *Comisión Económica para América Latina y el Caribe.* Recuperado el 22 de Mayo de 2019, de https://repositorio.cepal.org/bitstream/handle/11362/4755/S0700590_es.pdf

MongoDB. (s.f.). *MongoDB.* Recuperado el 25 de Octubre de 2019, de https://www.mongodb.com/es

Montealegre B, J. E. (Mayo de 1990). *IDEAM.* Recuperado el 19 de Octubre de 2019, de http://documentacion.ideam.gov.co/openbiblio/bvirtual/009198/009198.pdf

Moriasi, D., Arnold, J., Van Liew, M., Bingner, R., Harmel, R. D., & Veith, T. (Mayo de 2007). *Research Gate.* Recuperado el 3 de Octubre de 2019, de https://www.researchgate.net/publication/43261199_Model_Evaluation_Guidelines_for_Systematic_Quantification_of_Accuracy_in_Watershed_Simulations

Poulos, J., & Valle, R. (Octubre de 2016). *ResearchGate*. Recuperado el 28 de Junio de 2019, de https://www.researchgate.net/publication/309551513_Missing_Data_Imputation_for_Supervised_Learning

Prezi. (26 de Noviembre de 2013). Recuperado el 28 de Junio de 2019, de https://prezi.com/ydmjpweh9llt/ecuaciones-empiricas-para-calcular-el-gasto-volumetrico/

Python. (s.f.). Recuperado el 25 de Junio de 2019, de https://www.python.org/doc/essays/blurb/

Revista Semana. (21 de Agosto de 2018). *Semana Sostenible.* Recuperado el 10 de Noviembre de 2019, de https://sostenibilidad.semana.com/impacto/articulo/causas-de-las-inundaciones-en-colombia-e-impactos-en-la-biodiversidad/41385

Richman, M. B., Trafalis, T. B., & Adrianto, I. (2007). *Academia*. Recuperado el 29 de Junio de 2019, de

https://www.academia.edu/1189493/Multiple_imputation_through_machine_learning_algorithms

Rosati, G. (Junio de 2017). *ResearchGate*. Recuperado el 28 de Junio de 2018, de https://www.researchgate.net/publication/319263056_Construccion_de_un_modelo_de_imputacion_para_variables_de_ingreso_con_valores_perdidos_a_partir_de_ensamble_learning_Aplicacion_en_la_Encuesta_Permanente_de_Hogares_EPH_-_PREPRINT-

Ruiz Rozo, J., & Agudelo Quiñonez, Y. (2016). *Modelación hidrológica del sector del páramo Romerales en la cuenca alta del río Quindío.* Trabajo de grado para optar por el título de Ingeniero Civil, Armenia. Recuperado el 2019

Ruiz, S. (20 de Julio de 2017). *Analítica Web.* Recuperado el 30 de Septiembre de 2019, de https://www.analiticaweb.es/algoritmo-knn-modelado-datos/

Searcy, J. K., & Hardison, C. H. (1960). *United States Geological Survey.* Recuperado el 23 de Septiembre de 2019, de https://pubs.usgs.gov/wsp/1541b/report.pdf

Shahzad, W., Rehman, Q., & Ahmed, E. (2017). *Semantic Scholar*. Recuperado el 29 de Junio de 2019, de https://pdfs.semanticscholar.org/2ef2/1b22d96d3a254752a2b2c5940e2160701cee.pdf

Statistics Canada. (27 de Noviembre de 2015). Recuperado el 25 de Junio de 2019, de https://www150.statcan.gc.ca/n1/pub/12-539-x/2009001/imputation-eng.htm

Su, X., Greiner, R., Khoshgoftaar, T. M., & Napolitano, A. (2011). *Academia*. Recuperado el 29 de Junio de 2019, de https://www.academia.edu/17129584/Using_Classifier-Based_Nominal_Imputation_to_Improve_Machine_Learning

Universitat de València. (s.f.). Recuperado el 4 de Octubre de 2019, de https://www.uv.es/ceaces/base/descriptiva/coefcorre.htm

Urrutia, J. A., Reiner, P., & Salazar, H. D. (Diciembre de 2010). *Redalyc.* Recuperado el 30 de Septiembre de 2019, de https://www.redalyc.org/pdf/849/84920977010.pdf

Valesani, M. E., Quintana, O. P., & Vallejos, O. A. (Mayo de 2009). *Universidad Nacional de la Plata*. Recuperado el 29 de Junio de 2019, de http://sedici.unlp.edu.ar/bitstream/handle/10915/19739/Documento_completo.pdf?sequence=1&isAllowed=y

Vantas, K., & Sidiropoulos, E. (2017). *Academia*. Recuperado el 28 de Junio de 2019, de https://www.academia.edu/35087365/Imputation_of_erosivity_values_under_incomplete_rainfall_data_by_machine_learning_methods

Vélez Correa, J., & Nieto Figueroa, P. (2016). *Colegio de Estudios Superiores de Administración.* Recuperado el 1 de Octubre de 2019, de

https://repository.cesa.edu.co/bitstream/handle/10726/1577/MFC00491.pdf?sequence=1&isAllowed=y

WWF Colombia. (20 de Febrero de 2018). Recuperado el 25 de Junio de 2019, de http://www.wwf.org.co/?uNewsID=323450

Zaforas, M. (2016). *Paradigma Digital*. Recuperado el 10 de Octubre de 2019, de https://www.paradigmadigital.com/dev/jupyter-data-science-aplicada/

8. Anexos

Anexo A – Conjunto de datos de Nariño

Resultados al quitar el 20% de los datos registrados por cada estación para el año 1990 y realizar la imputación con los algoritmos KNN y Curva de Doble Masa.

	DOBLE MASA					KNN					
NOMBRE	**RMSE**	**NSE**	**PBIAS**	**MAE**	**TIEMPO**	**RMSE**	**NSE**	**PBIAS**	**MAE**	**TIEMPO**	**CORR. PROMEDIO**
MOCOAACUEDUCTO	14,95	0,36	18,59	8,72	10,65	**10,37**	0,69	15,22	5,86	1,14	0,68
CHUNGACASPI	15,99	-0,14	5,88	7,64	10,07	**12,03**	0,39	-13,31	6,88	1,22	0,63
PEPINO EL	17,45	0,24	19,44	10,51	10,13	**14,99**	0,45	37,02	8,61	1,24	0,59
SAN JOSE DE TAPAJE	27,45	-0,51	23,34	17,84	21,34	**25,02**	-0,23	68,48	15,11	1,27	0,20
MINCHOY	14,09	0,01	31,83	8,35	10,68	**11,98**	0,25	9,92	7,38	1,14	0,53
CHURUYACO	22,05	-0,24	14,93	13,60	11,76	**20,15**	-0,02	20,76	12,78	1,19	0,39
QUINCHOA	10,35	-0,10	12,23	5,33	10,36	**8,65**	0,03	-43,66	5,10	1,14	0,64
MATAJE	17,37	-0,36	10,25	8,02	10,45	**15,93**	-0,05	40,20	7,23	1,14	0,26
VENTADE CAJIBIO	10,81	0,06	19,32	5,43	10,26	**9,78**	0,20	5,31	5,12	1,22	0,42
SATE	11,06	0,24	17,07	5,78	10,41	**10,08**	0,37	11,81	5,40	1,23	0,49
LLANOVERDE	10,40	-0,13	23,71	5,28	10,35	**9,47**	0,10	24,68	4,89	1,14	0,44
JULUMITO ALERTAS	9,55	0,32	22,10	4,77	10,30	**9,09**	0,36	0,19	4,83	1,30	0,50
SAUCEEL	7,90	0,46	7,31	3,17	10,44	**7,51**	0,51	6,74	3,01	1,14	0,54
NARINO	8,24	0,17	20,39	4,05	10,46	**7,93**	0,22	11,53	3,93	1,14	0,49
JUNIN	25,74	-0,76	17,04	17,71	11,97	**25,45**	-0,65	83,79	18,09	1,14	0,36
PRESAHIDROMAYO	7,75	0,40	29,03	3,13	10,62	**7,51**	0,41	17,27	3,08	1,14	0,60
GUACHICONO	14,95	0,02	17,90	6,75	11,72	**14,71**	0,05	33,97	6,90	1,22	0,40
PTO LIMON	26,16	-0,16	14,58	14,29	11,00	**25,98**	-0,13	28,60	14,10	1,14	0,20
CORINTO	10,71	0,11	16,56	5,23	11,15	**10,65**	0,13	-6,52	5,72	1,36	0,37
APTOG L VALENCIA	13,22	-0,20	11,69	7,12	10,26	**13,17**	-0,19	22,01	7,14	1,21	0,25
PTO CAICEDO	31,43	-0,13	15,67	14,49	10,88	**31,38**	-0,10	48,15	13,63	1,14	0,21
BUENOS AIRES	**9,97**	0,39	26,85	4,60	10,36	9,98	0,38	8,61	4,82	1,23	0,48
BOCATOMA ALERTAS	**8,87**	0,27	17,91	3,86	10,42	8,88	0,24	-12,81	4,25	1,14	0,52
OSPINA PEREZ	**5,47**	0,32	15,36	2,68	10,29	5,53	0,29	-5,90	2,66	1,14	0,57
BERRUECOS	**6,19**	0,27	27,60	2,81	10,31	6,27	0,23	-0,87	2,93	1,14	0,65
CATALINA LA	**12,23**	0,37	12,00	5,62	10,24	12,31	0,36	9,95	5,71	1,14	0,50
CARRIZAL	**8,82**	-0,17	9,73	4,92	16,05	8,97	-0,20	-27,44	5,49	1,31	0,62
MAGUI	**24,85**	-0,13	30,18	10,60	10,43	25,13	-0,16	53,20	10,57	1,14	0,19
CHACHAGUI	**7,62**	0,35	15,82	3,26	10,72	7,98	0,21	-3,94	3,38	1,17	0,57
BUESACO	**7,47**	0,23	2,20	3,59	10,71	7,85	0,16	-8,62	3,81	1,14	0,54
OVEJAS ABAJO ALERT	**10,22**	0,24	4,58	5,23	10,20	10,63	0,18	4,20	5,49	1,14	0,45
UNIONLA	**6,94**	0,36	28,44	3,29	11,74	7,51	0,24	26,28	3,59	1,22	0,61
RIO PALO	**13,19**	0,30	29,37	6,00	10,52	13,79	0,24	45,65	6,02	1,29	0,50
DINDE	**10,57**	0,12	16,97	5,85	11,00	11,18	0,01	7,15	6,16	1,19	0,44
APTOLA FLORIDA	**10,99**	0,17	9,29	4,19	10,37	11,65	-0,42	-49,98	5,20	1,14	0,34
MAMACONDE	**6,05**	0,08	19,76	2,48	11,33	6,70	-0,34	-55,01	3,18	1,38	0,50
TAMINANGO	**6,46**	0,56	14,09	2,82	10,95	7,13	0,47	15,68	3,04	1,14	0,62
HIDROMAYOCAMP	**5,32**	0,31	20,55	2,43	10,87	6,00	-0,26	-55,22	2,76	1,14	0,63
TAMBO	**8,60**	0,20	13,56	4,18	11,48	9,27	0,02	-3,33	4,73	1,29	0,46
AGUADA LA	**8,53**	0,25	15,94	4,01	10,71	9,29	0,03	-8,59	4,56	1,33	0,47
CALERA LA	**8,11**	0,16	30,85	4,56	10,60	8,87	-0,02	-23,72	5,30	1,14	0,42

	DOBLE MASA					KNN					
NOMBRE	**RMSE**	**NSE**	**PBIAS**	**MAE**	**TIEMPO**	**RMSE**	**NSE**	**PBIAS**	**MAE**	**TIEMPO**	**CORR. PROMEDIO**
TRAPICHE EL	**9,64**	0,20	13,94	4,69	10,46	10,45	0,01	-2,36	4,87	1,14	0,43
VIVERO LINARES	**5,22**	0,36	24,08	2,20	10,01	6,06	-0,29	-57,49	3,04	1,14	0,60
GAMBOA	**8,55**	0,17	8,76	4,59	10,95	9,43	-0,04	-3,23	5,12	1,16	0,45
VERGEL EL	**8,29**	0,35	19,69	4,05	10,53	9,21	0,21	41,00	4,50	1,14	0,58
PTO ASIS	**17,06**	-0,01	26,81	10,00	10,61	18,12	-0,17	-30,03	11,78	1,14	0,35
COCOEL	**10,80**	0,11	-13,48	4,65	10,53	12,03	-0,29	-46,44	5,31	1,14	0,33
GUACHENE	**7,35**	0,30	11,69	3,41	10,65	8,61	-0,07	-57,95	4,73	1,18	0,46
PISANDA	**5,07**	0,31	17,28	2,43	10,40	6,35	-0,19	-57,11	3,22	1,14	0,55
PICUDO EL	**14,76**	0,01	17,24	9,14	11,09	16,21	-0,21	-30,10	10,84	1,14	0,36
BALSALA	**8,28**	0,24	23,55	4,07	10,55	9,73	-0,07	-20,30	5,21	1,14	0,48
MOSQUERA	**16,61**	-0,08	13,05	8,30	10,40	18,09	-0,31	-18,23	9,75	1,14	0,20
TANGUA	**4,53**	0,13	33,37	1,90	10,86	6,14	-0,91	-48,72	2,81	1,14	0,49
BUENOS AIRES	**9,28**	-0,07	37,42	4,22	10,62	11,19	-1,18	-68,75	6,63	1,14	0,26
VICHOY	**10,12**	0,02	28,79	4,94	10,28	12,23	-0,71	-76,38	7,71	1,14	0,30
CUMBAL	**4,99**	0,10	10,68	2,39	10,46	7,13	-1,12	-84,71	3,59	1,14	0,38
CONDAGUA	**11,47**	0,05	19,26	8,10	10,45	13,84	-0,41	-17,39	9,31	1,14	0,44
PENOLEL	**4,64**	0,16	15,90	1,88	10,73	7,02	-1,06	-104,77	3,20	1,14	0,52
IMUES	**3,56**	0,20	14,68	1,69	10,33	6,30	-2,25	-144,95	3,17	1,14	0,49
REMOLINO GRANDE	**11,99**	-0,14	9,58	5,62	10,41	15,07	-0,97	-67,72	8,17	1,13	0,26
GUASCA LA	**3,68**	0,31	-9,77	1,35	10,66	6,81	-1,87	-256,95	3,21	1,14	0,52
SANDONA	**4,22**	0,02	11,28	2,32	10,70	8,07	-3,03	-159,95	5,26	1,14	0,39
PUERRES	**4,93**	-0,17	0,79	2,82	10,55	9,36	-3,25	-200,53	5,73	1,17	0,42
SAN FRANCISCO	**5,75**	0,15	18,69	3,47	10,49	13,02	-4,37	-150,63	7,40	1,13	0,41
CONCEPCION	**2,72**	-0,69	16,24	1,79	11,45	11,24	-28,63	-324,79	6,73	1,14	0,15

Anexo B – Conjunto de datos de Quindío

	DOBLE MASA					KNN					
NOMBRE	**RMSE**	**NSE**	**PBIAS**	**MAE**	**TIEMPO**	**RMSE**	**NSE**	**PBIAS**	**MAE**	**TIEMPO**	**CORR. PROMEDIO**
NAVARCO	8,4531	-0,2584	-25,9649	4,3773	5,0250	7,7496	0,0340	4,8607	3,9950	0,4173	0,4144
BREMEN	13,1408	-0,8549	3,1867	6,7617	4,8310	10,8195	-0,1614	36,2049	6,1072	0,4107	0,4135
LA PLAYA	6,9124	0,4431	14,1599	3,8641	4,9216	6,9786	0,4359	7,3785	4,0291	0,4078	0,4132
LA MONTANA	4,3218	0,5363	11,6281	1,9605	4,9210	4,1662	0,4075	-16,8390	2,1232	0,4102	0,4070
EL BOSQUE	4,3068	0,1831	-16,3160	2,5383	4,9546	4,3169	0,1868	-14,0912	2,4669	0,4002	0,4159
LA PICOTA	4,7596	0,5163	3,1405	2,0556	5,0073	4,5465	0,4641	-14,7736	1,9856	0,4011	0,4132