**PROYECTO FINAL**

CURSO:

**COMUNICACIÓN DE DATOS Y REDES**

TEMA:

**3 UCI Bike Sharing data set**

**Conjunto de datos de bicicletas compartidas**

DOCENTE:

**CESAR JESUS LARA AVILA**

INTEGRANTES:

**AMES ANAPÁN JOSÉ MANUEL**

**MORE AYAY DAHAYRA XIOMARA**

**LIMA QUISPE ALEXANDRA NANCY**

**CASTRO PICHIHUA, VICTORIA BEATRIZ**

**2023**

* **INTRODUCCIÓN**
* Presentación del problema general sobre el que versará el trabajo y cómo se integra dentro del campo del aprendizaje automático.
* Objetivo del estudio
* **DISEÑO DEL EXPERIMENTO**

-Descripción del conjunto de datos

- Número y tipo de características (binarias, discretas, continuas, etc.).

- Número de muestras en los conjuntos de entrenamiento y prueba. En caso aplique,número de muestras por clase.

- Metodología

- De ser el caso, estrategia para el manejo de datos faltantes.- Selección y extracción de características.

- Selección y justificación de la medida de calidad.

- Algoritmos que serán empleados y estrategia para su ajuste.- Estrategia de validación a emplear para el ajuste de hiper parámetros si fuese necesario

* **EXPERIMENTACIÓN Y RESULTADOS**
* **DISCUSIÓN**
* ¿ Cuál es el rendimiento comparativo de los diferentes modelos de regresión utilizados para predecir la demanda de alquiler de bicicletas?
* **CONCLUSIONES Y TRABAJOS FUTUROS**
* **CÓDIGOS Y/O SCRIPTS (JUPYTER NOTEBOOKS o COLAB)**

3 UCI Bike Sharing data set

<https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset>

GLM y stacking, Random Forests y Extra Trees, XGBoost y LightGBM 4

**1.1 Introducción**

El problema general que se aborda en este trabajo es predecir la demanda de bicicletas compartidas utilizando el conjunto de datos UCI Bike Sharing. Este problema se integra dentro del campo del aprendizaje automático gracias a la utilización de diferentes modelos de aprendizaje automático como “GLM (Regularized Linear Models), stacking, Random Forests, Extra Trees, XGBoost y LightGBM”

El problema de la predicción de la demanda de bicicletas compartidas es de gran relevancia en diferentes sectores, como el transporte urbano y la planificación de la movilidad. Conocer la demanda esperada de bicicletas en diferentes momentos del día y en diferentes situaciones climáticas puede ayudar a optimizar la logística y la gestión de las estaciones de bicicletas, así como mejorar la satisfacción de los usuarios y reducir la congestión del tráfico. Dentro del campo del aprendizaje automático, existen diferentes enfoques y técnicas que se pueden utilizar para abordar este problema. Además de los modelos mencionados anteriormente, como GLM, Random Forests, Extra Trees, XGBoost y LightGBM, también se pueden emplear otros algoritmos, como redes neuronales, support vector machines (SVM), k vecinos más cercanos (KNN), entre otros.

El objetivo principal es desarrollar modelos que puedan analizar los datos históricos de las estaciones de bicicletas compartidas y realizar predicciones precisas sobre la cantidad de bicicletas que serán utilizadas en un momento determinado. En este sentido, se emplean varios modelos de aprendizaje automático para abordar este problema. Uno de los modelos utilizados es el modelo GLM (Regularized Linear Models), que se utiliza como una técnica de regresión para modelar la relación entre las variables predictoras y la demanda de bicicletas compartidas. Se busca encontrar una función lineal que mejor se ajuste a los datos históricos, teniendo en cuenta factores como la hora del día, el clima, el día de la semana, entre otros. Además, se emplea el concepto de stacking, que implica combinar la salida de múltiples modelos para obtener una predicción final más precisa. En este caso, se utilizan modelos como Random Forests, Extra Trees, XGBoost y LightGBM. Los modelos de Random Forests y Extra Trees son algoritmos que se basan en árboles de decisión, y se utilizan para capturar relaciones complejas entre las variables predictoras y la demanda de bicicletas compartidas. Por otro lado, XGBoost y LightGBM son algoritmos basados en Gradient Boosting, que brindan un mayor rendimiento y eficiencia en el modelo y son conocidos por su capacidad para manejar grandes conjuntos de datos y realizar predicciones precisas. En la práctica, estos modelos se integran en el problema de predecir la demanda de bicicletas compartidas utilizando los conjuntos de datos UCI Bike Sharing. Se utilizan los datos históricos disponibles, que incluyen información como la hora del día, el clima, el día de la semana, entre otros, para entrenar los modelos y ajustar sus parámetros. Una vez entrenados, se pueden utilizar para realizar predicciones sobre la demanda de bicicletas compartidas en momentos futuros.

Esto tiene como finalidad mejorar la gestión de las estaciones de bicicletas compartidas y optimizar la movilidad urbana. Asimismo, el estudio busca comparar y evaluar diferentes modelos de aprendizaje automático para determinar cuál es el más adecuado para este problema.

* **Diseño del experimento**
* **Descripción del conjunto de datos**

El conjunto de datos “Bike Sharing Dataset” es un conjunto ampliamente utilizado en el campo del aprendizaje automático y la predicción de la demanda de alquiler de bicicletas. Este conjunto de datos captura información sobre el alquiler de bicicletas en una empresa de bicicletas compartidas con sede en Washington, D.C., durante el periodo de dos años.

* **Número y tipo de características (binarias, discretas, continuas, etc.).**

La data que estamos utilizando consta de 12 caracteristicas las cuales son:

* temp: temperatura normalizada en grados Celsius.
* hum: humedad normalizada.
* windspeed: velocidad del viento normalizada.
* laborable: si el día no es ni fin de semana ni festivo, es 1; de lo contrario, es 0.
* temporada: estación (1: primavera, 2: verano, 3: otoño, 4: invierno).
* día laborable: día de la semana.
* weathersit: Situación meteorológica. Tiene los siguientes valores:

1: Despejado, Pocas nubes, Parcialmente nublado, Parcialmente nublado.

2: Niebla + Nublado, Niebla + Nubes dispersas, Niebla + Pocas nubes, Niebla.

3: Nieve ligera, Lluvia ligera + Tormenta + Nubes dispersas, Lluvia ligera + Nubes dispersas.

4: Lluvia intensa + Granizo + Tormenta + Niebla, Nieve + Niebla.

* feriado: Si el día es festivo o no.
* Mes: Mes (1 a 12).
* atemp: Temperatura percibida normalizada en grados Celsius.
* año: Año (0: 2011, 1: 2012).
* hr: Hora del día (solo presente en "hora.csv").

* **Número de muestras en los conjuntos de entrenamiento y prueba. En caso aplique,número de muestras por clase.**

El número de muestras en el conjunto de entrenamiento es: 13903, asimismo el número de muestras en el conjunto de prueba consta de 3476.

* **Metodología**
* **Selección y extracción de características**

En el código se definen conjuntos de características específicos para los conjuntos de datos 'hour.csv' y 'day.csv'. Estas características se seleccionaron como variables independientes para predecir la variable objetivo 'cnt' (conteo de bicicletas alquiladas). La selección de estas características se basa en el conocimiento del dominio y puede requerir análisis exploratorio de datos para identificar las variables más relevantes.

**- Selección y justificación de la medida de calidad**

La medida de calidad utilizada en el código proporcionado es el Error Cuadrático Medio (MSE, por sus siglas en inglés). Esta medida evalúa la diferencia cuadrada promedio entre los valores predichos y los valores reales de la variable objetivo. El MSE se utiliza comúnmente en problemas de regresión para evaluar la precisión del modelo. Su elección puede deberse a su interpretación intuitiva y a su capacidad para penalizar los errores grandes.

**- Algoritmos que serán empleados y estrategia para su ajuste**

En el código se utilizan varios algoritmos de regresión para predecir el número de bicicletas alquiladas. Estos algoritmos incluyen Linear Regression (GLM), Random Forest, Extra Trees, XGBoost y LightGBM.

**-Estrategia de validación a emplear para el ajuste de hiper parámetros si fuese necesario**

* **EXPERIMENTACIÓN Y RESULTADOS**
* Experimentación:

1. Se utilizaron diferentes modelos de aprendizaje automático, incluyendo GLM, Random Forests, Extra Trees, XGBoost y LightGBM, para predecir el recuento de alquiler de bicicletas basado en los conjuntos de datos 'hour.csv' y 'day.csv'.
2. Se seleccionaron características específicas para cada modelo, como la temperatura, humedad, velocidad del viento, día laborable, estación, día de la semana, situación climática, día festivo, mes, hora, temperatura aparente y año.
3. Los conjuntos de datos se dividieron en conjuntos de entrenamiento y prueba utilizando una proporción del 80% y 20% respectivamente, y una semilla aleatoria de 42.

* Resultados:

1. El modelo de GLM obtuvo un error cuadrático medio (MSE) de 19379.83 en el conjunto de pruebas.
2. El modelo de Random Forests obtuvo un MSE de 1762.81 en el conjunto de pruebas.
3. El modelo de Extra Trees obtuvo un MSE de 1661.08 en el conjunto de pruebas.
4. El modelo de XGBoost obtuvo un MSE de 1636.13 en el conjunto de pruebas.
5. El modelo de LightGBM obtuvo un MSE de 1645.66 en el conjunto de pruebas.
6. El modelo de Stacking, utilizando Random Forests, Extra Trees, XGBoost y LightGBM como estimadores base y GLM como estimador final, obtuvo el mejor rendimiento con un MSE de 1439.40 en el conjunto de prueba.

* **DISCUSIÓN**

Estos resultados demuestran que el modelo de Stacking, que combina varios modelos de aprendizaje automático, logra una mejor precisión en la predicción del recuento de alquiler de bicicletas en comparación con los otros modelos evaluados.

**¿ Cuál es el rendimiento comparativo de los diferentes modelos de regresión utilizados para predecir la demanda de alquiler de bicicletas?**

Basado en los resultados obtenidos en la experimentación, se puede observar que los modelos de Random Forests, Extra Trees, XGBoost y LightGBM tuvieron un rendimiento comparativamente mejor en comparación con el modelo GLM. Estos modelos, al ser más complejos y poderosos, fueron capturados relaciones no lineales y patrones más complejos en los datos, lo que les permitió realizar predicciones más precisas de la demanda de alquiler de bicicletas. En términos de métricas de evaluación, se observaría que estos modelos tienen un error cuadrado medio (MSE) y un error absoluto medio (MAE) más bajos en comparación con el modelo GLM. Además, es posible que estos modelos tengan un coeficiente de determinación (R^2) más alto, lo que indica que son capaces de explicar una mayor proporción de la variabilidad en los datos. Sin embargo, es importante destacar que el rendimiento de los modelos puede variar según el conjunto de datos utilizado, el procesamiento de datos, los hiperparámetros ajustados y otros factores. Por lo tanto, es recomendable realizar una evaluación exhaustiva de los modelos en el contexto específico del problema antes de tomar decisiones finales sobre cuál modelo es el más adecuado para predecir la demanda de alquiler de bicicletas en un escenario particular.

- Los resultados muestran los errores cuadráticos medios (MSE) obtenidos para cada modelo de regresión. Cuanto menor sea el valor del MSE, mejor será el rendimiento del modelo en términos de ajuste a los datos y capacidad de predicción.

- Al analizar los resultados del gráfico de barras, podemos observar que el modelo de Stacking (StackingRegressor) muestra el menor MSE, seguido por XGBoost, Extra Trees, LightGBM, Random Forest y GLM (Linear Regression) en ese orden.

- El modelo de Stacking ha logrado obtener el menor MSE, lo que indica que tiene un mejor rendimiento en comparación con los otros modelos evaluados. Esto se debe a que el StackingRegressor combina las predicciones de múltiples modelos base (Random Forest, Extra Trees, XGBoost, LightGBM) utilizando un modelo final (GLM) para mejorar la precisión y capacidad de generalización.

- Por otro lado, el modelo GLM (Linear Regression) muestra el MSE más alto entre todos los modelos evaluados. Esto podría indicar que la regresión lineal simple no es suficiente para capturar la complejidad de los datos y obtener un buen ajuste..

* **CONCLUSIONES Y TRABAJOS FUTUROS**

En conclusiones:

* En este estudio, se analizaron conjuntos de datos de alquiler de bicicletas a nivel horario y diario. Se aplicarán estrategias para el manejo de datos faltantes, utilizando la imputación por interpolación para preservar la integridad de los datos. Luego, se realizó un análisis de calidad de los datos y se evaluaron diferentes modelos de regresión para predecir la cantidad total de bicicletas alquiladas.
* En cuanto al manejo de datos faltantes, se identificaron valores faltantes en ambos conjuntos de datos y se aplicó la imputación por interpolación para estimar los valores faltantes. Esto permitió mantener la integridad de los datos y evitar la eliminación de observaciones valiosas.
* En cuanto a la selección de características, se eligieron características relevantes para la predicción de la cantidad total de bicicletas alquiladas. Se utilizaron diferentes modelos de regresión, como Linear Regression, Random Forest, Extra Trees, XGBoost y LightGBM, y se aplicó una técnica de stacking para combinar las predicciones de varios modelos.
* Los resultados mostraron que el modelo de stacking, que combinó los resultados de los modelos base, obtuvo el mejor rendimiento en términos de error cuadrático medio (MSE), seguido de XGBoost, Extra Trees, LightGBM y Random Forest. Por otro lado, el modelo de regresión lineal simple mostró el peor desempeño.

**En trabajos futuros:**

El estudio también ha identificado áreas de mejora y posibles direcciones para futuros trabajos.

* Exploración de otras variables: Además de las variables utilizadas en este estudio, se pueden explorar otras características importantes que pueden afectar la demanda de alquiler de bicicletas, como la ubicación geográfica y la disponibilidad de estaciones de bicicletas.
* Mejora del preprocesamiento de datos: El preprocesamiento de datos es una etapa clave en la construcción de modelos precisos. Se pueden explorar técnicas más avanzadas de preprocesamiento de datos, como la eliminación de outliers y el manejo de valores faltantes, para mejorar aún más la calidad de los datos utilizados en el modelo.
* Incorporación de datos en tiempo real: Aunque este estudio se basó en datos históricos, sería interesante explorar la incorporación de datos en tiempo real, como las condiciones meteorológicas actuales y los eventos especiales, para mejorar las predicciones de la demanda.
* Optimización de hiperparámetros: Los modelos de aprendizaje automático tienen varios hiperparámetros que deben ser ajustados. En este estudio, se realizaron pruebas de validación cruzada para seleccionar los mejores hiperparámetros, pero se puede explorar técnicas de búsqueda de hiperparámetros más avanzadas, optimización bayesiana, para encontrar la combinación óptima de hiperparámetros.

El estudio ha proporcionado una base sólida para predecir la demanda de alquiler de bicicletas utilizando modelos de aprendizaje automático. Sin embargo, hay espacio para futuros trabajos que amplíen y mejoren el enfoque utilizado en este estudio, incorporando variables adicionales, mejorando el preprocesamiento de datos, utilizando datos en tiempo real y optimizando los hiperparámetros de los modelos.

* **CÓDIGOS Y/O SCRIPTS (JUPYTER NOTEBOOKS o COLAB)**