**Proceso de ciencia de datos, análisis de la resistencia del concreto utilizando diferentes agregados**

***Definición de metas***

**Objetivo**

Utilizando métodos de aprendizaje automático determinar la resistencia del concreto dependiendo del tipo y cantidad de elementos agregados para mejorar la resistencia del mismo, para de esta forma predecir el comportamiento del concreto sin tener que utilizar ensayos destructivos.

**¿Porque los interesados quieren hacer el proyecto?**

Debido a la gran cantidad de tiempo y recursos que se utilizan para obtener resultados sobre la resistencia del concreto por medio de métodos de laboratorio destructivos, sería muy beneficioso contar con un método para predecir el comportamiento del mismo sin tener que realizar pruebas destructivas.

**¿Que necesitan los interesados del proyecto?**

Una manera de poder predecir el comportamiento del concreto variando los componentes del mismo.

**¿Porque la solución actual es inadecuada?**

Para obtener la resistencia a la compresión del concreto se necesita realizar pruebas destructivas en laboratorio, con el fin de determinar la variación de la resistencia variando los componentes y su cantidad, lo que consumo mucho tiempo y recursos, especialmente a la hora de desarrollar nuevas mezclas

**¿Qué recursos se necesitan?**

Para este caso en particular se necesitará la última versión de la plataforma de programación de Python, utilizando en particular un cuaderno de Jupyter para crear el código de análisis y predicción utilizando los datos brindados.

***Recoger y analizar los datos***

**Datos disponibles**

La descripción de los datos disponibles se puede encontrar en el documento “Variables de la resistencia del concreto” y los datos se pueden observar en el formato de Excel o CSV bajo el nombre de “Concrete\_Data”.

**¿Pueden los datos brindados ayudar a resolver el problema? ¿Son suficientes?**

Los datos brindados pueden contribuir a cumplir con el objetivo planteado, siempre y cuando se realice un análisis previo de los mismos a profundidad para comprenderlos y buscar relaciones útiles.

La cantidad es suficiente, ya que, en el campo de ensayos de laboratorio para materiales, tener mas de 1000 datos es una cantidad estadísticamente suficiente

**¿Es la calidad de los datos suficientemente buena?**

***Construcción del modelo***

**Que técnicas se utilizarán para el modelo**

Se utilizará métodos de regresión, probando con tres modelos para ver cuál se ajusta mejor, una vez escogido el modelo se utilizará el que se ajuste mejor para realizar las predicciones y determinar la validez de los resultados con los valores estadísticos obtenidos.

Los modelos a utilizar son los siguientes

* Linear regression (Ln)
* Support vector regression
* Random forest regression

Una vez aplicados los modelos al conjunto de datos, se escogerá el que brinde el mejor valor general de puntuación para utilizarlo como método predictivo, posterior a la aplicación del modelo predictivo, se obtendrán los valores de la raíz cuadrada y la raíz cuadrática media, con el fin de determinar la confiabilidad de las predicciones

**Evaluación del modelo**

Una vez revisados todos los modelos planteados anteriormente se determino que el que tenia la mejor puntuación fue el Random Forest, como se muestra en el siguiente cuadro

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelos** | | | |
| **Modelos** | **Random Forest** | **Support Vector** | **Linear Regression** |
| **Overall** | 0,98 | 0,11 | 0,62 |

Después se aplicó el modelo predictivo utilizando Random Forest, con los siguientes resultados

|  |  |
| --- | --- |
| **Prediction RF** | |
| **Estimator** | **Value** |
| **r2** | 0,879 |
| **RMSE** | 5,677 |

Ambos valores son aceptables y demuestran la validez de los resultados obtenidos.

**Presentación de los resultados**

Se presentará un informe con los resultados obtenidos, incluyendo el comportamiento general de los datos, así como las relaciones obtenidas entre las diferentes variables independientes y dependientes.

**Mantenimiento del modelo**

El modelo es aplicable solo para las variables mostradas, si se busca ingresar otra variable se deben de realizar nuevos ensayos, ya que las relaciones obtenidas dependen directamente unas de otras.

Si se desea ingresar valores nuevos de las mismas variables es recomendable volver a correr el modelo con los nuevos datos, para verificar que las métricas de rendimiento se mantienen constantes.