En la predicción de datos sin importar el modelo que se use siempre hay pasos que no pueden faltar en la construcción de un modelo, los cuales describo a continuación y dejo un repositorio con un ejemplo clásico de Kaggel de sobreviviente del Titanic el cual es un modelo de regresión logística ya que solo tiene dos respuestas 1=sobrevive, 0= No sobrevive (clasificación) aplicando los pasos descritos:

1. Exploración de la información: importante conocer la información del data set así como los valores promedio, máximos y mínimos de cada variable
2. Tratamiento de información faltante: Importante tomar decisiones sobre que hacer con lo valores “na/faltantes”, ya sea completarlos con algún valor estadístico o simplemente eliminarlos consideremos que no aportan al modelo o que no afectan.
3. Tratamiento de columnas categóricas: al tener columnas con datos categóricos debemos convertir a un valor numérico dichas categorías ya que los modelos funcionan con números para sus entrenamientos y predicciones y también mejoran su rendimiento
4. Definir la variable a explicar/dependiente o predecir y las variables explicativas/independientes.
5. Definición y entrenamiento del modelo: definir el modelo a utilizar dependiendo que se quiere predecir es importante, en este caso por ser un modelo de clasificación se utiliza Regresión Logística de la librería “sklearn”. Así como también el tamaño del dataset que se usa para entrenar y el que se usa para prueba (generalmente 70% para entregar y 30% para prueba)
6. Predicciones y métricas: Luego de hacer las predicciones con el modelo entrenado es importante medir su precisión para ver si el modelo tiene el % deseado o si necesitamos ajustarlo para lograr una mayor precisión a la obtenida