Memoria Descriptiva.

Proyecto de Machine Learning.

Predicción de la cantidad de goles marcados por jugadores de futbol en función de sus estadísticas.

Fuentes de datos:

Los datos fueron extraídos de datasets de Kaggle que, a su vez, son resultado de un proceso de web scrapping realizado a la web de Transfermarkt. Contienen registros de tres temporadas diferentes, a razón de aproximadamente 2500 jugadores por fichero. En cada fila hay reflejadas 400 variables estadísticas del jugador.

Proceso de Machine Learning:

Determinación de las variables a utilizar:

De las 400 variables me he quedado con ocho variables iniciales que eran las que consideraba que guardaban una relación más estrecha con los apartados ofensivos de cada jugador. A su vez, de estas variables generé algunas que consideré más acordes a un problema real. Finalmente me quedan siete variables, de las cuales, cuatro demuestran tener una correlación adecuada con mi variable objetivo y las otras tres no. No obstante, en las pruebas posteriores efectué predicciones con los dos casos.

Tratamiento de los datos:

Debido a que en las visualizaciones de mis datos se observa que la distribución de estos no es uniforme, aplico un algoritmo de normalización que finalmente desecho, ya que los resultados obtenidos en las predicciones empeoran a las obtenidas con mis datos tal como están.

Utilizo un escalado estándar para utilizar dichos datos en los modelos que lo requieren.

Entrenamiento de los modelos:

He realizado pruebas con los siguientes modelos:

- Regresión lineal.
- Regresión polinómica.
- Random Forest.

- Regresor SVM.
- Gradient Boosting.
- Redes Neuronales.

Para las pruebas, generalmente he empleado algoritmos de búsqueda de los mejores parámetros (“gridsearch”), aunque siempre intentando comprobar que dichos resultados se cumplieran al utilizar el modelo con mi conjunto de test.

Los resultados han sido similares en casi todos, excepto con la regresión lineal que ha dado resultados muy malos. Los resultados han sido más satisfactorios empleando solamente mis cuatro variables más correladas con la variable objetivo. El modelo de regresión polinómica ha sido el más acertado, pero por una pequeña diferencia.

Principales problemas encontrados:

La escasez de registros en los valores más altos de la variable objetivo ha sido uno de los principales problemas. La pesadez de algunos modelos ha sido un problema, al necesitar mucho tiempo de ejecución hasta poder visualizar los resultados. El hecho de intentar adecuar mis datos a un problema más real, ha provocado que las variables finales tuvieran algo menos de correlación con mi variable objetivo de la que brindaban las variables iniciales.

Conclusiones:

He podido profundizar en las prácticas de Machine Learning, utilizando una buena variedad de los modelos existentes, así como varios métodos para el tratamiento previo de los datos.

En cuanto a los resultados, se han podido obtener precisiones bastante altas, aunque dada la naturaleza del problema planteado, entiendo que dicha precisión podría estar sujeta a varios factores no contemplados en el modelo y que podrían alterar bastante las futuras predicciones.