VNiVERSiDAD
Đ SALAMANCA

Aplicación de Machine Learning en la predicción de resultados de partidos de Rugby

Tecnologías de Información Emergentes

Grado en Ingeniería Informática

Diego Borrallo Herrero 49367527M

Índice

Introducción

En este informe se desarrolla un proyecto de análisis de datos aplicado al rugby internacional. La idea principal es aprovechar datos históricos de resultados de partidos y utilizar modelos de Machine Learning para analizar el desempeño de equipos, predecir resultados futuros y evaluar las fortalezas relativas de los equipos. El trabajo incluye la implementación de modelos de regresión y clasificación para cubrir diversos objetivos dentro del análisis deportivo.

Descripción del proyecto

Este proyecto combina la recopilación y procesamiento de datos deportivos con técnicas de Machine Learning. Utiliza un conjunto de datos con resultados de partidos de rugby desde 1871 hasta 2022, al que se ha añadido información de ranking mundial. Los datos se procesan para construir variables útiles como el rendimiento del equipo en los últimos partidos, diferencias de puntuación y otras características relevantes. Posteriormente, se entrenan los modelos tanto para predicciones continuas (regresión) como para clasificaciones discretas (clasificación). Los objetivos incluyen la predicción de resultados y la clasificación de los equipos como ganadores, perdedores o en empate.

Modelos empleados

Modelos de regresión

Los modelos de regresión se utilizan para predecir valores continuos, como la diferencia de puntos entre los equipos en un partido. Esto permite analizar el rendimiento esperado de los equipos y proporciona una base para predicciones cuantitativas.

Regresión Lineal

Es un modelo estadístico que busca encontrar la relación lineal entre las variables de entrada (como el rendimiento del equipo o el ranking) y el objetivo (la diferencia de puntuación). Este modelo es sencillo de interpretar y eficiente para relaciones lineales.

Random Forest Regressor

Este modelo utiliza un conjunto de árboles de decisión para realizar predicciones continuas. Cada árbol contribuye con una predicción, y el resultado final es el promedio de todas. Es eficaz para modelar relaciones no lineales y manejar datos complejos con múltiples características.

Support Vector Regression (SVR)

Es una extensión del algoritmo de SVM para problemas de regresión. Busca encontrar una función que ajuste los datos dentro de un margen tolerable de error, optimizando la generalización del modelo. Es ideal para conjuntos de datos de tamaño mediano y con relaciones no lineales.

Modelos de clasificación

Los modelos de clasificación predicen etiquetas discretas, como la clasificación de un partido como victoria, derrota o empate. Estos modelos son fundamentales para categorizar eventos deportivos en clases específicas.

Regresión Logística

Es un modelo estadístico que estima la probabilidad de que ocurra un evento (por ejemplo, una victoria). Utiliza una función sigmoide para transformar predicciones continuas en probabilidades, lo que facilita la clasificación binaria o multiclase.

Decision Tree

Un árbol de decisión divide iterativamente los datos en ramas basadas en condiciones de las características. Es fácil de interpretar y puede capturar relaciones complejas entre las variables. Sin embargo, es susceptible al sobreajuste si no se regula adecuadamente.

Random Forest Classifier

Combina múltiples árboles de decisión para mejorar la precisión y reducir el sobreajuste. Cada árbol clasifica el evento, y el resultado final es determinado por el voto mayoritario de todos los árboles. Este enfoque es robusto frente a datos ruidosos y distribuciones desbalanceadas.

Resultados

Modelos de regresión

Regresión Lineal

```
Model: Linear Regression
Mean Squared Error: 212.81
R^2 Score: 0.45
```

Imagen 1 - Resultado Regresión Lineal

La regresión lineal muestra un desempeño aceptable, con un coeficiente de determinación moderado que indica que el modelo explica el 45% de la variabilidad en los datos. Sin embargo, el error cuadrático medio es relativamente alto, por lo que las predicciones podrían ser mejoradas.

Random Forest Regressor

```
Model: Random Forest
Mean Squared Error: 231.74
R^2 Score: 0.40
```

Imagen 2 - Resultado Random Forest Regressor

Este modelo tiene un rendimiento algo inferior al de Regresión Lineal, con un R^2 más bajo y un MSE más alto. Esto podría deberse a que los datos no tienen relaciones no lineales fuertes que este modelo pueda explotar.

Support Vector Regression (SVR)

```
Model: Support Vector Regression
Mean Squared Error: 262.35
R^2 Score: 0.32
```

Imagen 3 - Resultado Support Vector Regression

Este modelo tiene el peor desempeño de los modelos de regresión evaluados, con el mayor error cuadrático medio y el menor R^2. Esto sugiere que no logra capturar las relaciones necesarias para predecir correctamente los valores objetivo.

En general, los resultados de los modelos de regresión no han sido tan buenos como podríamos esperar y reflejan la dificultad del problema, debido a la complejidad y la influencia de múltiples factores en los resultados deportivos.

Modelos de clasificación

Regresión Logística

```
Logistic Regression Results
Accuracy: 0.75
Classification Report:
              precision    recall  f1-score   support

    away_win       0.68      0.69      0.69       157
        draw       0.00      0.00      0.00         6
    home_win       0.79      0.80      0.79       245

    accuracy                           0.75       408
   macro avg       0.49      0.50      0.49       408
weighted avg       0.73      0.75      0.74       408
```

Imagen 4 - Resultado Regresión Logística

La regresión logística ofrece el mejor desempeño en términos de precisión general, especialmente para las clases "home_win" y "away_win". Sin embargo, al igual que Random Forest, no predice correctamente los empates debido al desequilibrio en las clases.

Decision Tree

```
Decision Tree Classifier Results
Accuracy: 0.63
Classification Report:
              precision    recall  f1-score   support

    away_win       0.56      0.60      0.58       157
        draw       0.00      0.00      0.00         6
    home_win       0.72      0.67      0.70       245

    accuracy                           0.63       408
   macro avg       0.43      0.42      0.42       408
weighted avg       0.65      0.63      0.64       408
```

Imagen 5 - Resultado Decision Tree

El clasificador basado en árboles de decisión tiene el peor desempeño entre los modelos de clasificación evaluados. Aunque tiene métricas aceptables para las clases principales, su precisión general es limitada y enfrenta los mismos desafíos con la clase "draw".

Random Forest Classifier

```
Random Forest Classifier Results
Accuracy: 0.71
Classification Report:
              precision    recall  f1-score   support

    away_win       0.64      0.63      0.64       157
        draw       0.00      0.00      0.00         6
    home_win       0.75      0.78      0.77       245

    accuracy                           0.71       408
   macro avg       0.46      0.47      0.47       408
weighted avg       0.70      0.71      0.70       408
```

Imagen 5 - Resultado Random Forest Classifier

Este modelo obtiene una precisión aceptable, con un *F1-score* consistente para las clases principales ("home_win" y "away_win"). Sin embargo, tiene problemas significativos para manejar empates ("draw"), con métricas nulas en esta clase. Esto se debe a la baja cantidad de muestras en dicha categoría, lo que dificulta la generalización del modelo.

En resumen, los modelos de clasificación muestran un desempeño decente, con la regresión logística liderando en precisión general. Sin embargo, todos los modelos enfrentan dificultades para manejar las clases menos representadas, destacando la importancia de abordar el desequilibrio en los datos.

Conclusiones

El proyecto nos ha permitido comparar varios modelos de Machine Learning para la predicción de resultados de partidos de rugby, con las siguientes observaciones principales:

- **Regresión**: La Regresión Lineal fue el modelo más eficaz, alcanzando un R^2 de 0.45. Los modelos más avanzados, como Random Forest Regressor y Support Vector Regression, no lograron superar su desempeño, dando visibilidad a la complejidad del problema y las posibles limitaciones en el preprocesamiento.
- **Clasificación**: La Regresión Logística consiguió la mejor precisión (75%), seguida por el Random Forest Classifier (71%). El Decision Tree mostró el peor desempeño (63%), mientras que todas las técnicas enfrentaron dificultades para predecir empates debido al desequilibrio de clases.
- **Limitaciones**: El desbalance en los datos y las métricas moderadas de R^2 resaltan la necesidad de incluir más variables relevantes y técnicas para manejar datos desbalanceados.

- **Futuras mejoras**: Incrementar el tamaño del dataset, aplicar reamostrado para clases minoritarias y explorar modelos más avanzados podrían mejorar significativamente los resultados.

Este análisis pretende demostrar el potencial del Machine Learning en deportes y permite tener una base sólida desde la que partir para futuros desarrollos en el campo de la predicción deportiva.