

Construcción del modelo con Random Forest

Este proyecto se centra en la construcción de un modelo de aprendizaje automático para predecir los ganadores de los partidos de fútbol de la Premier League. El objetivo es utilizar datos históricos de partidos para entrenar un modelo que pueda predecir el resultado de futuros partidos.

Integrantes:

- Benjamin De La Torre Rojas
- Mauricio David Correa Hernandez
- Vladlen Shatunov

Lectura y carga de datos

El primer paso es leer y cargar los datos de partidos de fútbol utilizando la biblioteca pandas. Los datos se almacenan en un archivo CSV y se cargan en un DataFrame de pandas. Se eliminan las columnas no relevantes, como "comp", "notes", "round", "day", "captain", "referee", "match report", "formation", "opp formation".

Número de Partidos

Se juegan 38 partidos por temporada.

Número de Equipos

Hay 20 equipos en cada liga por temporada.

Temporadas

Se incluyen datos de las temporadas 2018-2019, 2019-2020, 2021-2022, 2022-2023 y la actual 2023-2024.

Limpieza de datos

Se limpia y prepara la columna "date" convirtiéndola a un tipo de dato datetime en pandas. Esto permite añadir más propiedades al modelo. Se realiza un one-hot encoding para las columnas "opponent" y "team", convirtiendo las variables categóricas en variables numéricas.

1 Conversión de Fecha

La columna "date" se convierte a datetime.

2 One-Hot Encoding

Se aplica one-hot encoding a las columnas "opponent" y "team".





Creación de predictores

Se crean nuevos predictores para el modelo a partir de los datos existentes. Se utiliza la columna "venue" para crear un código numérico "venue_code". Se extrae la hora de la columna "time" y se crea un código numérico "hour". Se utiliza la columna "date" para crear un código numérico "day_code" que representa el día de la semana.

Predictor	Descripción
venue_code	Código numérico para el estadio.
hour	Hora del partido.
day_code	Día de la semana del partido.



Escalado de datos

Se utiliza un MinMaxScaler para escalar los datos numéricos, asegurando que todos los predictores estén en la misma escala. Esto mejora el rendimiento del modelo de aprendizaje automático.

1

Datos sin escalar

Los datos numéricos tienen diferentes escalas.

2

Escalado de datos

Se aplica MinMaxScaler para escalar los datos.

3

Datos escalados

Todos los predictores están en la misma escala.

Creación del modelo inicial

Se crea un modelo inicial utilizando un RandomForestClassifier. Se define una malla de parámetros para la búsqueda de hiperparámetros utilizando GridSearchCV. Se divide el conjunto de datos en conjuntos de entrenamiento y prueba para evaluar el rendimiento del modelo.

RandomForestClassifier

Se utiliza como modelo base.

GridSearchCV

Se utiliza para encontrar los mejores hiperparámetros.

Train-Test Split

Se divide el conjunto de datos en conjuntos de entrenamiento y prueba.



Evaluación del modelo

Se evalúa el rendimiento del modelo utilizando métricas como la precisión, la exhaustividad y la puntuación F1. Se muestra la matriz de confusión para visualizar las predicciones correctas e incorrectas.



Accuracy

0.97



precision

0.98



Puntuación F1

0.97



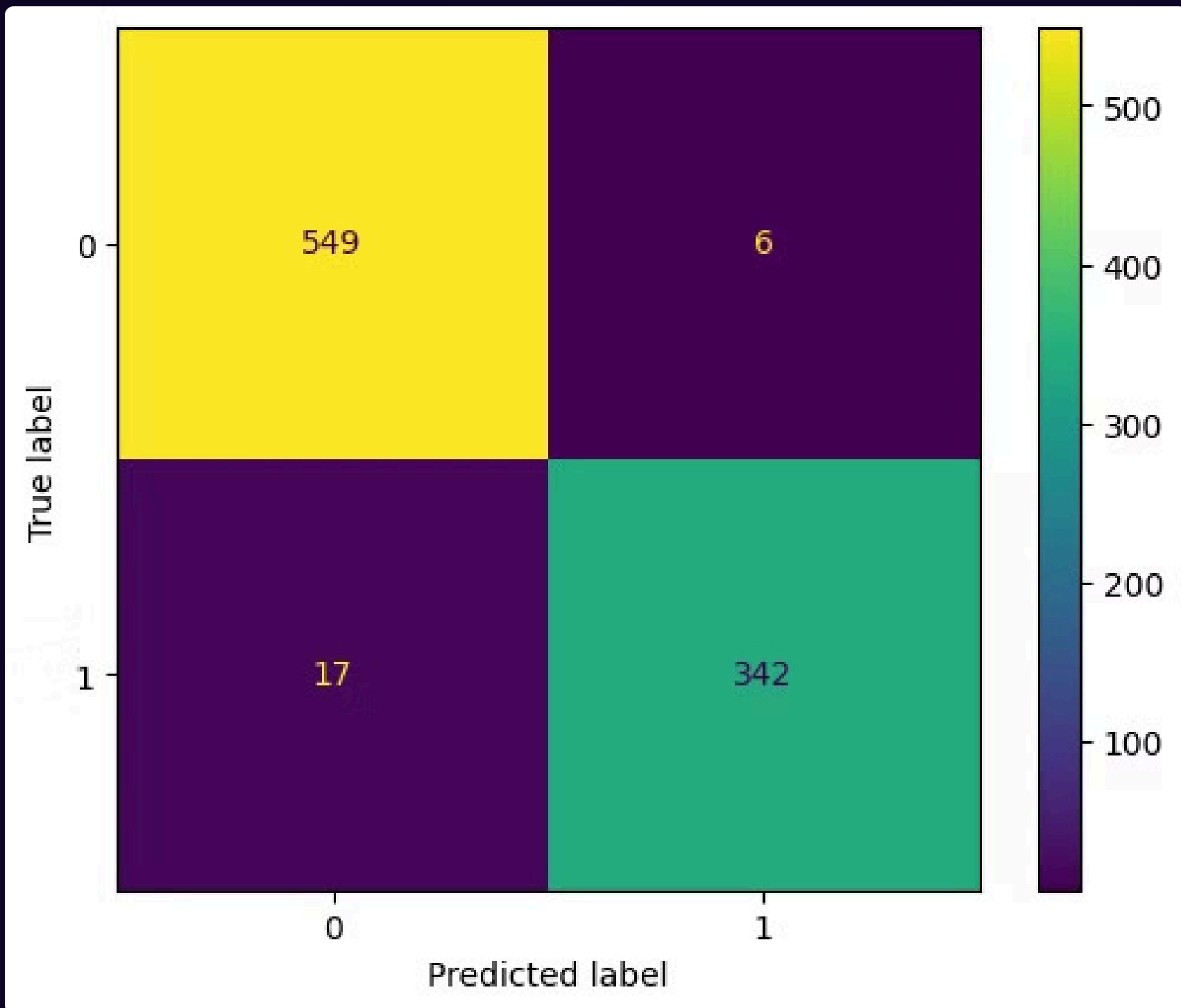
Recall

0.95



Matriz de Confusion

```
cm = confusion_matrix(y_test, rf.predict(X_test))  
disp = ConfusionMatrixDisplay(cm)  
disp.plot()  
unknown link()
```



Preparación de datos para la predicción

Se preparan los datos para realizar una predicción utilizando el modelo entrenado. Se crea un nuevo DataFrame con los datos de un partido específico y se utiliza el modelo para predecir el resultado.

1

Nuevo DataFrame

Se crea un DataFrame con los datos de un partido específico.

2

Predicción

Se utiliza el modelo entrenado para predecir el resultado del partido.



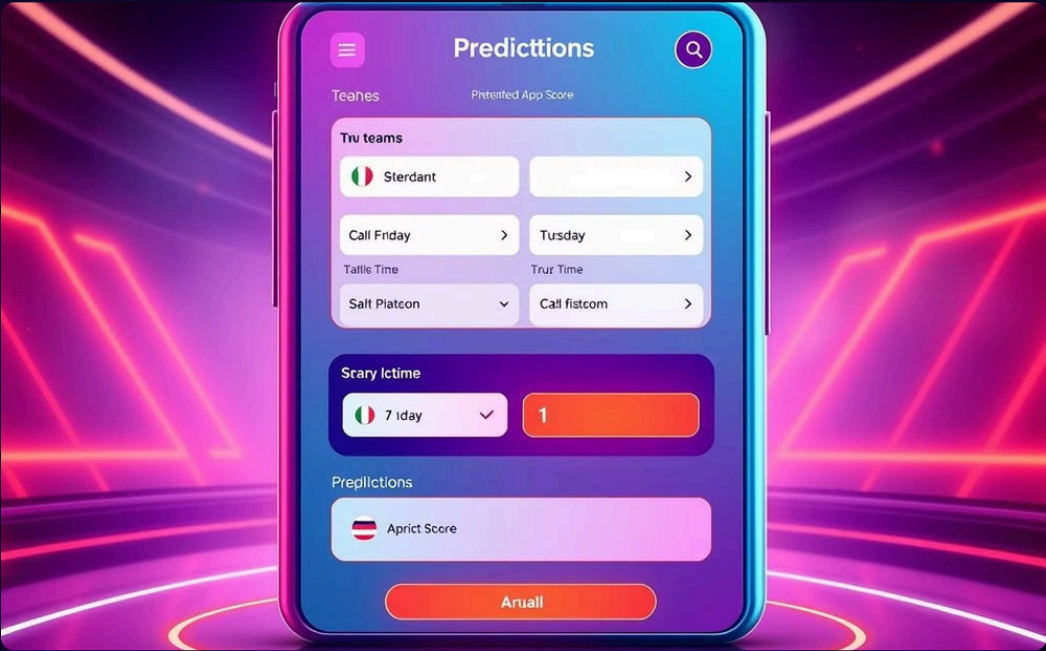
Creación de una interfaz gráfica

Se crea una interfaz gráfica para que los usuarios puedan interactuar con el modelo de predicción. La interfaz permite a los usuarios introducir los datos de un partido y obtener una predicción del resultado.



Lista de partidos

Se muestra una lista de partidos próximos con predicciones para cada partido.



Formulario de entrada

Se proporciona un formulario para que los usuarios introduzcan los datos de un partido específico.