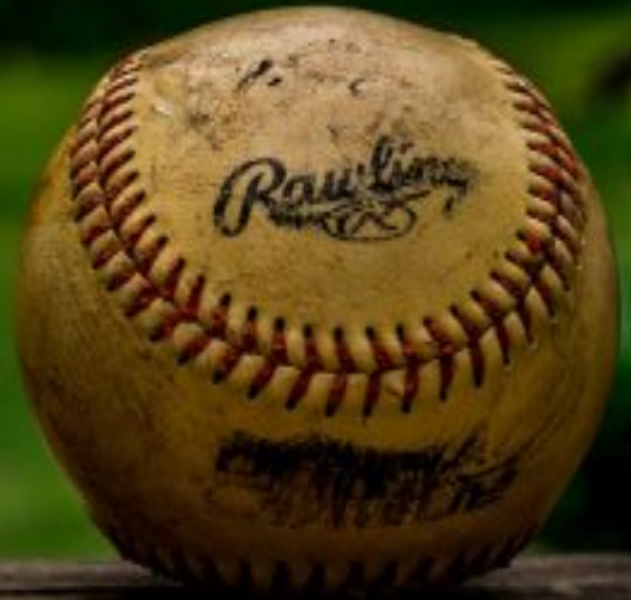
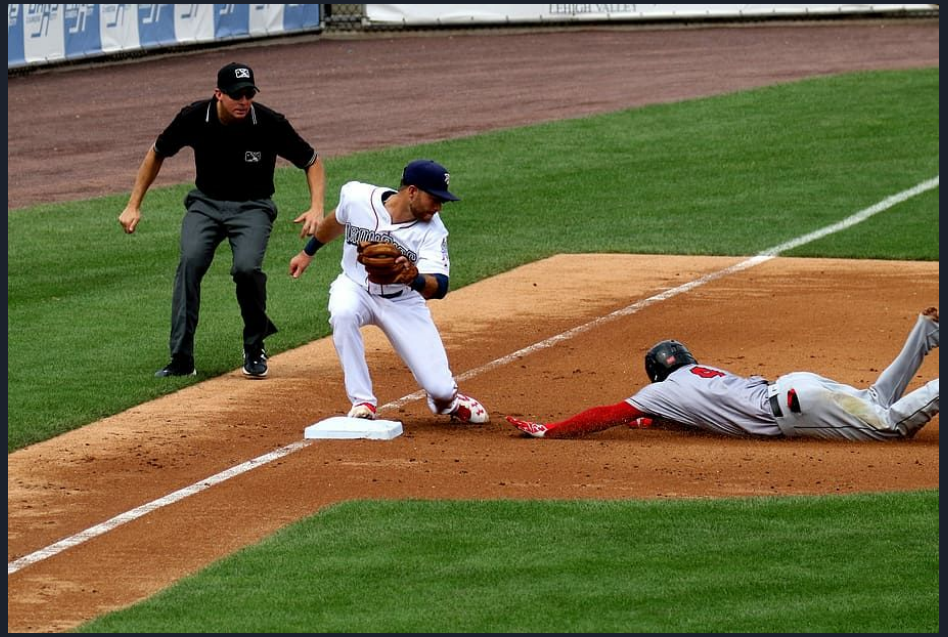
Predicción de robos en el baseball

Juan Sebastian Vanegas Rico - 2182071
Juan José Bayona Sepúlveda - 2183200



Planteamiento

El objetivo principal de este proyecto es descifrar las señales que se utilizan en el juego de baseball, esto con el fin de predecir cuando el otro equipo se dispone a robar. En el baseball las señales se usan como lenguaje no verbal de los coaches hacia los jugadores, esto con el fin de confundir a los rivales sobre las posibles jugadas que se vayan a realizar.



Dataset

	Acc1	Acc2	Acc3	Acc4	Acc5	Acc6	Acc7	Acc8	Acc9	Acc10	Acc11	Acc12	Acc13	Acc14	Acc15	Acc16	Acc17	Acc18	Acc19	Acc20	Resp
2	g	b	a	a	f	c	a	g	b	g	h	d	j	i	c	i	i	i	d	d	1
3	a	j	f	i	g	h	i	j	g	b	e	c	i	b	b	d	e	d	f	i	0
4	a	g	h	b	g	j	c	f	a	a	j	a	g	c	f	g	h	g	f	i	1
5	e	e	b	a	d	f	c	e	c	j	i	j	h	e	j	h	c	c	e	c	0
6	h	f	h	i	b	a	b	i	f	b	i	f	g	e	f	b	d	d	b	b	1

Para este proyecto se realizó un dataset con 20 diferentes señales para 100 jugadas. Estas señales se codificaron como letras para luego asignarles un código numérico y poderlas trabajar con los métodos y funciones vistas en clase.

Tratamiento de los datos

	Acc1	Acc2	Acc3	Acc4	Acc5	Acc6	Acc7	Acc8	Acc9	Acc10	Acc11	Acc12	Acc13	Acc14	Acc15	Acc16	Acc17	Acc18	Acc19	Acc20	Resp
2	6	1	0	0	5	2	0	6	1	6	7	3	9	8	2	8	8	8	3	3	1
3	0	9	5	8	6	7	8	9	6	1	4	2	8	1	1	3	4	3	5	8	0
4	0	6	7	1	6	9	2	5	0	0	9	0	6	2	5	6	7	6	5	8	1
5	4	4	1	0	3	5	2	4	2	9	8	9	7	4	9	7	2	2	4	2	0
6	7	5	7	8	1	0	1	8	5	1	8	5	6	4	5	1	3	3	1	1	1

Después de codificar cada una de las 10 señales el dataset queda como se observa en la imagen; De esta forma es posible trabajar el dataset y empezar a probar con distintos modelos.

Partición de los datos

Particionamiento de datos

```
[40]  1 #@title **Code:** Se separan los datos del ground truth y se establecen las particiones de train y de test
      2 from sklearn.model_selection import train_test_split
      3 data = df.iloc[:,:-1]
      4 data_y = df.Resp
      5 X_train, X_test, y_train, y_test = train_test_split(data, data_y, test_size = 0.2, shuffle = False)
      6 print("DONE!")
```

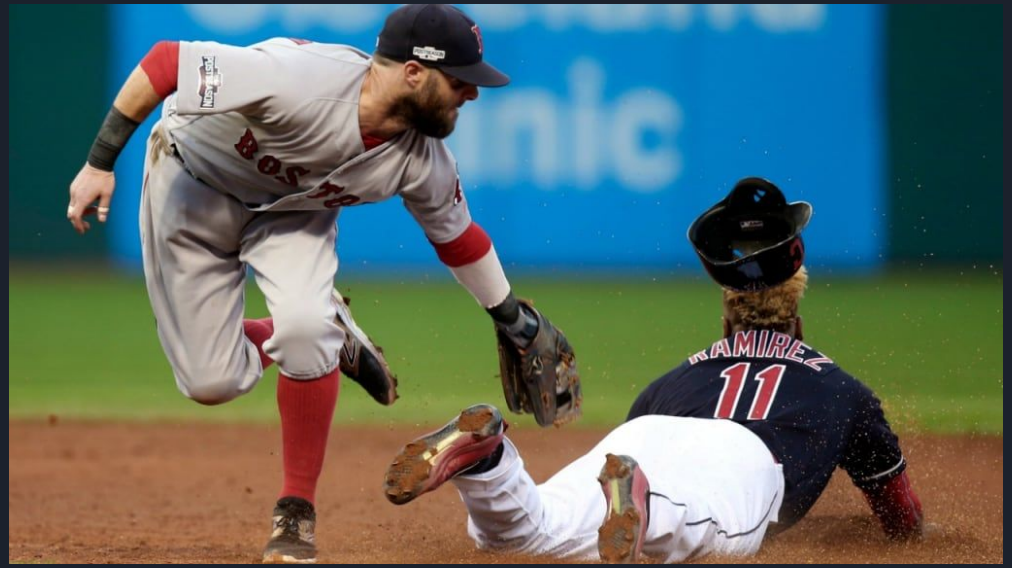
```
DONE!
```

Modelos Utilizados

- Decision Tree Classifier
- Support vector machine classification (SVMC)
- DNN
- Gaussian Naive Bayes
- Random Forest
- Cross-Validation (DT, NB, RF, SVMC)

Conclusión

Según los resultados obtenidos se observa que el mejor resultado fue arrojado por el modelo de Random Forest Classifier con un 75% de efectividad. Ya con esto en cuenta, es posible utilizar dicho algoritmo a la hora de hacer las predicciones.



Gracias por su atención.