Cólicos en Caballos

¿SE PUEDE DETERMINAR LA SUPERVIVENCIA?

Alumno: Juan Polo
Tutor: Francisco Di Palma
Profesor: Jorge Ruiz

CODER HOUSE

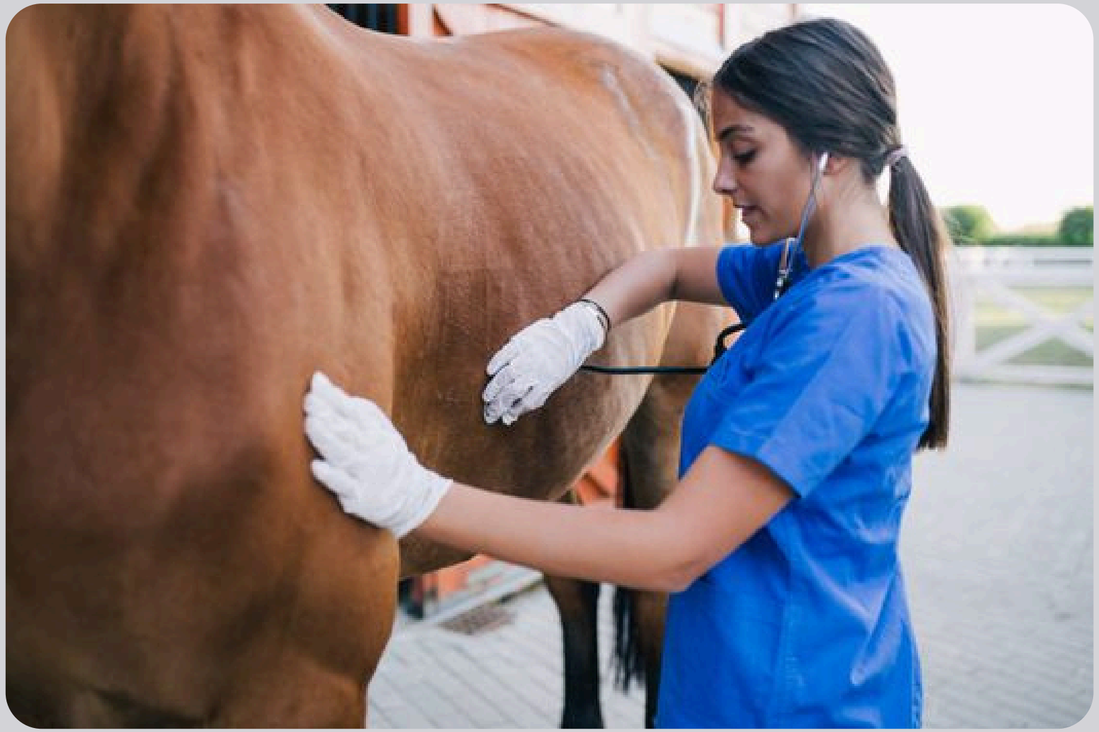
2024

Cólicos en caballos

Conjunto de síntomas asociados con el dolor abdominal. Es una condición común y puede variar desde un malestar leve hasta una emergencia médica grave que requiere intervención inmediata.

Sobre los datos

Fueron extraidos de kaggle y agregamientos sinteticos. Contiene 28 variables (numéricas continuas, discretas y categoricas) y 1299 filas. Variable *target* columna *'resultado'.*



OBJETIVO

Predecir si un caballo vive o muere según sus condiciones médicas.

PROPUESTA

Desarrollar un modelo de ***clasificación*** para estimar la supervivencia de los caballos basado en un conjunto de variables relevantes. Se utilizara técnicas de machine learning (*Random Forest, Regresión Logística*) para predecir la supervivencia de los caballos. El modelo será entrenado, validado y ajustado para maximizar su rendimiento, facilitando la toma de decisiones críticas en contextos veterinarios o de manejo de animales.

PARTE 1

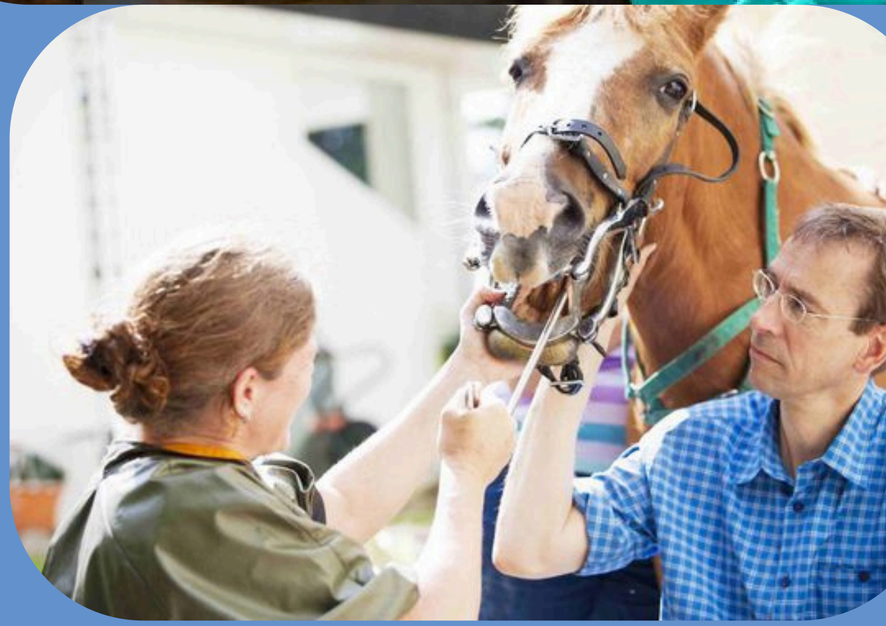
Limpieza y transformacion de los datos

- Concatenacion de los datos para formar un solo dataframe.
- Renombre de columnas para facilitar su entendimiento.
- Remplazo de valores *nulls* por la moda en variables categoricas y la mediana en variables numéricas.
- Descripción de variables numéricas y categoricas.
- Los valores de la columna 'número_hospital' no fueron considerados ID y por eso no se efectuo la eliminación de sus valores duplicados.

PARTE 2

ANÁLISIS Y VISUALIZACIÓN DE DATOS

- Análisis univariados de distribuciones y comportamientos de variables numéricas y categoricas.
- Visualización de *outliers* con gráficos de *boxsplot* y *violinplot*.
- Visualización de correlación de variables numéricas en gráfico de calor, *heatmap*.
- Transformación de columna númerica de *‘pulso’* en una columna categorica *‘categoria_pulso’*.
- **IQR** de columna *‘pulso’* y creación de *features ‘umbral_inferior’* y *‘umbral_superior’.*
- Visualización de grafícos multivaridos de columnas categoricas.
- Frecuencia de distribución de variable *target*.



PARTE 2

ANÁLISIS ESTADÍSTICOS

- **Test Chi Cuadrado:** El resultado nos revela 98 filas de p-valor < 0.05 que rechazan la hipótesis nula, lo que sugiere que hay una relacion significativa entre las variables.
- **Coheficiente Pearson:** Los resultados no nos muestran que haya una relacion lineal significativa entre las variables numéricas continuas.
- **Isolation Forest:** Utilizado para buscar anomalias, con una contaminación de un 0,05 nos muestra 65 filas que podrian ser datos anomalos. Se tomaron solo los datos sin anomalias (*df_horse_sin_anomalias*), para probar luego de ser encodeado, en mi modelo de *Random Forest*.
- **Coheficiente Biserial:** Los resultados nos muestran que algunas variables rechazan la hipótesis nula y muestran una relacion entra la variable dicotómica (‘resultado’) y algunas varibales numericas continuas.
- **MCA:** Se uso para observar cuales eran las variables mas influyentes en los componentes que creo el análisis.

PARTE 5

MODELOS DE MACHINE LEARNING

1° vistazo de los modelos. Sin utilizar tecnicas de mejoramiento de hiperparámetros. Previamente se encodearon los datos con *OHE* para su utilización en *ML*. Tabien fue probado el dataframe sin anomalias proporcionado por IsolateForest, el cual no presento mejorias en las métricas.

Random Forest

0.78 Accuracy

0.0/0.98
0.1/0.43 Recall

0.0/0.75
0.1/0.93 Precisión

0.0/0.85
0.1/0.58 F1-score

Regresión Logística

0.67 Accuracy

0.0/1.00
0.1/0.02 Recall

0.0/0.66
0.1/1.00 Precisión

0.0/0.80
0.1/0.03 F1-score

PARTE 6

OPTIMIZACION DE MODELOS

- **Curva ROC:** El *AUC* (Randomforest=0.89, RegresionLogística=0.54) nos indica que el modelo que mejor rendimiento tiene es el de RandomForest
- **RandomOverSampler/SMOTE/TomekLinks/SMOTETomek:** Se utilizo con el modelo de Random Forest para resamplear las muestras, pero ninguno tuvo mejoras significativas en el modelo.
- XGBoots: El modelo es secuencial y trabaja tratando de minorizar errores, los resultados de las metricas son: accuracy: , precision: , recall: , f1-score: .
- Bootstrap: Su utilizacion en el rendimiento no fue significativa
- Stratified K-Fold con parametros optimizados de GridSearchCV: Se utilizo Stratified K-Fold para que el modelo se entrenado correctamente en equilibrio de las clases, y cargue los parametros en GridSercghCV para ver que pruebas consiguian el mejor paramentro, los resultados de las iteraciones nos dieron que en el modelo de Random Forest (El mejor de entre las dos pruebas) los siguientes parametros: max_depth: 15, n_estimators: 150.

Conclusion

La medida F1-score nos hace un balance entre la precisión y el recall. Me quedo con esta medida mas armonica sobre las clases de mi modelo y seria importante conseguir mas datos de campo para trabajar mejor. Hay mucho por hacer y seguro con mas datos se puede mejor muchos mas la calidad del trabajo.

Utilizacion

Va dirigido a academicos veterinarios que esten interesados en la agilizar y mejorar su trabajo y en la implementacion de nuevas tecnologias para la mejora de sus procesos de trabajo.

Muchas gracias por su atencion!

A sido un placer poder exponer un trabajo y aportar a temas de esta indole.