Los Data Partners

La UniversityHack 2023 cuenta con la cooperativa La Viña y The Weather Company como data partners.

Cooperativa La Viña

Bodega creada en 1945, cuando un grupo de 38 emprendedores se asociaron bajo una misma cooperativa reuniendo una superficie de cultivo de 47 hectáreas con una producción de 250.000 kg de uva. A lo largo de estos casi 80 años la actividad vitivinícola ha ganado importancia y la bodega ha ido creciendo en número de socios y producción.

Su objetivo es ser la bodega referente del Sector Vitivinícola Valenciano siendo fieles a los principios cooperativos de los cuales se sienten orgullosos. Actualmente distribuye vinos en más de 40 países y cuenta con vinos excelentes como Venta del Puerto, Icono y Casa L'Àngel tal y como lo acreditan los premios obtenidos, así como las valoraciones realizadas por prestigiosos especialistas.

The Weather Company

El clima cumple un papel fundamental en la inteligencia comercial, como guiar la ruta que toma un piloto para evitar turbulencias, cuando un agricultor fertiliza sus cultivos o cómo una compañía de energía moviliza a sus equipos después de un corte de energía.

Como parte del compromiso de IBM, The Weather Company se enfoca en utilizar su conocimiento y tecnología para avanzar en la ciencia del pronóstico meteorológico preciso y ayudar a las personas en todas partes a tomar las mejores decisiones.

El objetivo

España es el tercer productor mundial de vino. Disponer de una previsión precisa de la producción en una campaña agrícola es cada vez más necesario de cara a optimizar todos los procesos de la cadena: recolección, traslado, procesado, almacenamiento y distribución.

Dado lo anterior y partiendo de amplios datasets con histórico de producciones de los viñedos que conforman la cooperativa La Viña, así como histórico de la climatología de los mismos, te retamos a crear el mejor modelo de predicción de producción de una campaña en base al cual se pueda estimar la cosecha que dispondrá la cooperativa meses antes de la recolección, por ello se debe tener en cuenta que no se pueden usar datos meteorológicos posteriores al 30 de junio de cada año para estimar la producción del mismo.

Los datasets

Ponemos a tu disposición tres datasets: **TRAIN**, que contiene información histórica de las fincas que conforman la cooperativa La Viña, **METEO**, que dispone de información meteorológica de estaciones climatológicas de la zona a nivel horario de The Weather Company y **ETO**, que dispone de información meteorológica ampliada y transformada de las mismas estaciones agregada en franjas del día.

El primer dataset, **TRAIN**, contiene información anual del resultado de las campañas de la cooperativa La Viña, mostrando las siguientes variables:

- CAMPAÑA: Año de la campaña.
- ID_FINCA: Identificador de finca.
- ID_ZONA: Identificador de una zona con una tipología de suelo común
- ID_ESTACION: Identificador de estación meteorológica.
- ALTITUD: Altitud media de la finca sobre el nivel del mar en metros.
- VARIEDAD: Código de variedad de la uva que se cultiva en la finca.
- MODO: Código del modo de cultivo.
- TIPO: Tipo de cultivo dentro de la variedad.
- COLOR: Identificador del color de la uva.
- SUPERFICIE: Superficie en hectáreas que ocupa la finca.
- PRODUCCION: Producción en kg. Obtenida en la campaña.

Aunque la serie histórica de producción empieza en 2014, la información de la superficie de las fincas solo está disponible para las campañas 2020, 2021 y 2022. La producción de 2022 para cada finca es el valor a estimar por lo que aparece como 'NA'.

El segundo dataset, **METEO**, contiene información horaria de estaciones climatológicas de The Weather Company de la zona de influencia en el periodo comprendido entre el 29-06-2015 y el 30-06-2022, dispone de las siguientes variables:

Variable	Descripción
validTimeUtc	Fecha
precip1Hour	volumen de lluvia en la última hora
precip6Hour	volumen de lluvia en las últimas 6 horas
precip24Hour	volumen de lluvia en las últimas 24 horas
precip2Day	volumen de lluvia en los últimos 2 días
precip3Day	volumen de lluvia en los últimos 3 días
precip7Day	volumen de lluvia en los últimos 7 días
precipMtd	Volumen de lluvia en el mes en curso
precipYtd	Volumen de lluvia en el año en curso
pressureChange	Variación máxima en la presión atmosférica en las últimas 3 horas
pressureMeanSeaLevel	Diferencia barométrica respecto al nivel del mar
relativeHumidity	humedad relativa

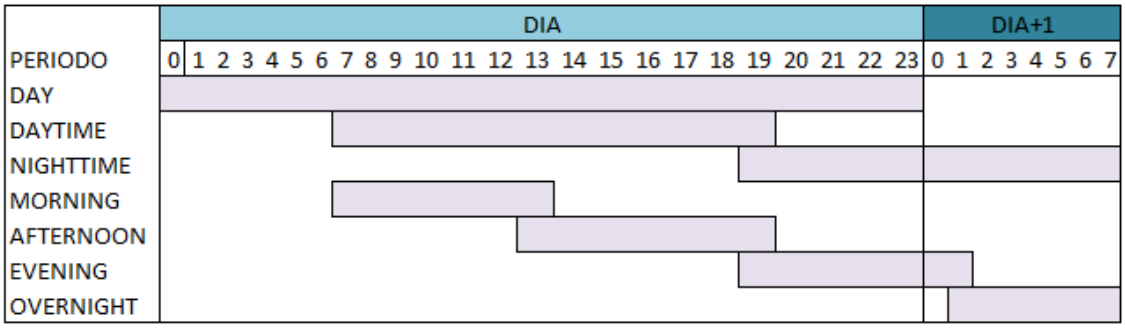
snow1Hour	volumen de nieve en la última hora
snow6Hour	volumen de nieve en las últimas 6 horas
snow24Hour	volumen de nieve en las últimas 24 horas
snow2Day	volumen de nieve en los últimos 2 días
snow3Day	volumen de nieve en los últimos 3 días
snow7Day	volumen de nieve en los últimos 7 días
snowMtd	volumen de nieve en el mes en curso
snowSeason	volumen de nieve trimestral (DIC-FEB, MAR-MAY, JUN-AGO, SEP-NOV)
snowYtd	volumen de nieve en el año en curso
temperature	Temperatura ambiente a 2 metros del suelo
temperatureChange24Hour	variación de temperatura respecto al día anterior
temperatureMax24Hour	temperatura máxima últimas 24 horas
temperatureMin24Hour	temperatura mínima últimas 24 horas
temperatureDewPoint	Punto de rocío, temperatura a la cual el aire debe ser enfriado a presión constante para alcanzar la saturación. El punto de rocío es también una medida indirecta de la humedad del aire.
temperatureFeelsLike	Sensación térmica. Temperatura aparente resultante de combinación de la temperatura, la humedad y la velocidad del viento.
uvIndex	radiación ultravioleta categorizada: -2, -1= No disponible / 0-2 = baja / 3-5 = moderada / 6-7 = alta / 8-10 = muy alta / 11-16 = extrema
visibility	Visibilidad horizontal desde la estación meteorológica, 999 equivale a ilimitada
windDirection	Dirección del viento en grados 0 – Norte, 90 – Este, 180 – Sur, 270 – Oeste
windGust	velocidad máxima de ráfaga de viento registrada durante el período de observación
windSpeed	fuerza del viento
ID_ESTACION	Identificador de la estación meteorológica

El último dataset, **ETO**, contiene información agregada y transformada de las estaciones climatológicas de The Weather Company en el mismo periodo, las variables se construyen con el siguiente patrón (excepto las variables date y ID_Estacion):

Variable + “Local” + periodo + tipo de agregación

DewPointLocalAfternoonAvg

Los periodos



Las variables

Variable	Descripción	Unidad
date	día	**Numérico**
DewPoint	Punto de rocío, temperatura a la cual el aire debe ser enfriado a presión constante para alcanzar la saturación. El punto de rocío es también una medida indirecta de la humedad del aire.	Grados Kelvin
Evapotranspiration	Evapotranspiración del cultivo de referencia. Esta es una tasa que da la cantidad de agua perdida por un cultivo de referencia	mm/h
FeelsLike	Sensación térmica. Temperatura aparente resultante de combinación de la temperatura, la humedad y la velocidad del viento.	Grados Kelvin
GlobalHorizontalIrradiance	Cantidad total de radiación solar recibida en una superficie horizontal	W/m2
Gust	velocidad máxima de ráfaga de viento registrada durante el período de observación	m/s
MSLP	Presión barométrica	Pa
precipAmount	volumen de lluvia por hora	mm/h
relativeHumidity	humedad relativa	%
SnowAmount	volumen de nieve por hora	m/h
Temperature	Temperatura ambiente a 2 metros del suelo	Grados Kelvin
uvIndex	Radiación ultravioleta: 0-2 = baja / 3-5 = moderada / 6-7 = alta / 8-10 = muy alta / 11-16 = extrema	
visibility	Visibilidad horizontal desde la estación meteorológica, 999 equivale a ilimitada	m
windSpeed	fuerza del viento	m/s
ID_ESTACION	Identificador de la estación meteorológica	

Las agregaciones

- Min: Mínimo del periodo
- Avg: Media del periodo
- Max: Máximo del periodo

Todos los dataset tienen extensión txt con la siguiente estructura y formato:

- Nombre de variables: incluidos en la cabecera
- Separador: "|"
- Símbolo decimal: "."
- Codificación: UTF-8

Nombre de fila: No dispone

Dataset respuesta

Es el fichero solicitado con tus predicciones de producción.

Se denominará "Equipo_UH2023.txt" donde 'Equipo' será el nombre del equipo con el que te has inscrito.

Sin cabecera ni nombres de filas.

Constará de 1.075 filas con 2 columnas cada fila:
• *ID_FINCA*: ordenado de forma ascendente
• *PRODUCCION*: Estimación de la producción para la campaña 2022

Separando campos con "|", el valor de la producción en kg, y los decimales con ".".

¿Qué pedimos?

Además del "dataset respuesta", te pedimos:

1. Un script ("script exploración") que contendrá el análisis exploratorio y procesos relevantes testados o ejecutados, pero no aplicados en la solución final.
2. Un script ("script predicción") que contendrá el proceso de extracción, transformación y carga de los datos, el procesado aplicado, así como la generación de predicciones.
3. Una breve descripción donde se expondrá el proceso y la metodología seguida, las técnicas aplicadas y los resultados obtenidos (en formato presentación, pdf o html, máximo 7 páginas con 3 imágenes).

Un valor menor no conllevará explícitamente una mejor clasificación. El "script de predicción" mencionado debe cumplir que sea generalizable y en el caso de métricas equiparables, se tendrán en cuenta los criterios siguientes:

- el Jurado podrá valorar si la documentación interna aportada (código y comentarios) está correctamente estructurada, expresada y es reproducible.
- los scripts de exploración y predicción deben constituir un proyecto de data science con todas sus fases.

Se valorará

FASE LOCAL

La calidad y la técnica utilizada para generar un modelo. Para ello se utilizará como métrica el RMSE que permite comparar objetivamente los valores reales frente a los valores predichos por el modelo, se tendrán que minimizar las desviaciones de los valores obtenidos respecto a los datos reales.

El "error cuadrático medio" o RMSE, definido como:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Siendo:
"n" el número de casos,
"ŷ" el valor estimado,
"y" el valor real

FASE NACIONAL

Los equipos que participen en la fase nacional se valorarán siguiendo el siguiente criterio. Un primer nivel en función exclusivamente de la métrica, se seleccionarán los 10 trabajos que obtengan las mejores métricas.

Los 10 equipos seleccionados se evaluarán con los siguientes criterios:

- 70% de la puntuación dependerá del RMSE obtenido, donde la mejor métrica obtendrá 10 puntos y el resto en función de la diferencia porcentual con dicho valor.
- 30% de la puntuación se constituirá con las puntuaciones obtenidas en el baremo de modelización responsable:

- Explicabilidad (30%)

Para una mejor adopción de la IA los modelos deber ser explicables, debemos evitar hablar de modelos de caja blanca / negra. En el desarrollo de todo modelo debe tenerse en cuenta la explicabilidad desde el diseño, un modelo explicable se integra en la gestión de forma más rápida que uno que no lo es, incluso modelos no explicables pueden llegar a no utilizarse nunca aun teniendo una muy buena precisión.

De esta forma será necesario evaluar el nivel de explicabilidad del modelo vs precisión del modelo, usando el resultado de dicha evaluación como una de las variables a tener en cuenta en la elección del modelo ganador. En la memoria deberá justificarse porqué ha sido elegido el modelo desde el punto de vista de su explicabilidad aportando los datos objetivos (peso de las principales variables en el resultado obtenido, de forma global al modelo y de forma particular para casos concretos) así como subjetivos relativos a dicha selección.

- Transparencia: (25%)

De igual manera a la explicablidad, la transparencia debe estar presente desde el diseño para una mejor adopción. Todo modelo debe ir acompañado de una memoria donde se describa, desde distintos puntos de vista, el funcionamiento del modelo y favorecer así el entendimiento por parte del usuario.

Debido a esto en la memoria del modelo se evaluará que haya quedado documentado:

- Instrucciones de uso
- Tratamientos sobre los datasets de datos
- Elección de la muestra de entrenamiento y validación
- Argumento de la tipología del modelo a desarrollar
- Criterios aplicados para la selección del ganador
- Visualización y explicación de los resultados

- Justicia (25%)

La IA debe usarse de forma justa, por lo que debe de velar por la equidad y evitar sesgos de cualquier tipo. En el desarrollo de cualquier modelo, y desde el diseño, debe revisarse que la muestra es lo suficientemente representativa y que no existe ningún sesgo (ni en los datos utilizados en el entrenamiento ni en el comportamiento del propio modelo).
De esta forma será necesario evaluar que se he velado por el cumplimiento de dicho principio durante el desarrollo del modelo, para ello en la memoria deberán aparecer los análisis llevados cabo para corroborar la suficiente diversidad de la muestra, así como la inexistencia de sesgos.

- Sostenibilidad ambiental (20%)

El desarrollo de los modelos de IA debe velar por la sostenibilidad y ser respetuosos con el medioambiente, por lo que se deberá asegurar la optimización computacional que garantice un menor consumo energético.
De esta forma será necesario evaluar el consumo energético (en base al tiempo de computación) vs precisión del modelo, usando el resultado de dicha evaluación como una de las variables a tener en cuenta en la elección del modelo ganador. En la memoria deberá justificarse porqué ha sido elegido el modelo desde el punto de vista de su consumo energético, aportando datos objetivos y subjetivos relativos a dicha selección.

DEFENSA FINAL

En el evento de Presentación y Fallo, el Jurado Nacional tendrá en cuenta, además de los criterios anteriores, que el Proyecto se trasmita de forma clara y concisa.

Ayudas al desarrollo del reto

Además de los datasets proporcionados, se muestran algunos recursos que podrían ser de interés para la realización del presente reto.

Enlaces de interés

https://www.tecnicoagricola.es/estados-fenologicos-de-la-vid/

https://www.vinetur.com/2019111458471/las-7-etapas-de-un-vinedo-el-ciclo-vegetativo-de-la-vid.html

https://www.mapa.gob.es/es/agricultura/temas/sanidad-vegetal/GUIAUVADETRANSFORMACION_tcm30-57934.pdf

https://utielrequena.org/la-brotacion-la-vendimia-las-fases-del-ciclo-la-vid/

Librerías R y Python

Prophet
Prophet es un procedimiento de pronóstico implementado en R y Python. Es rápido y proporciona pronósticos completamente automatizados que los científicos y analistas de datos pueden ajustar manualmente.

AI Fairness

Conjunto de herramientas extensible de código abierto para examinar, informar y mitigar la discriminación y el sesgo en los modelos de aprendizaje automático a lo largo del ciclo de vida de la aplicación de IA.

Dalex

Herramientas para la explicabilidad de los modelos

Librerías en R

dplyr/data.table
Los paquetes dplyr y data.table son herramientas para la exploración y manipulación de datos.

ggplot2
Completo paquete que nos permite representar una gran galería de gráficos. Mejora las funciones habituales de R para gráficos pudiendo incluir más capas y especificaciones.

caret
Incluye sencillas herramientas para analizar la calidad de los datos, selección de características, optimización de parámetros o construcción de modelos predictivos.

mlr
Otro de los meta paquetes más populares. Presenta un marco completo para acceder a distintos paquetes de estadística y machine learning de una forma integrada y coherente.

Tidyverse
Colección de paquetes de R diseñados para data Science.

Keras
API para redes neuronales de alto nivel.

Tensor Flow

Interfaz para acceder a la biblioteca de software libre TensorFlow™ que utiliza diagramas de flujo de datos realizar cálculos numéricos.

iml

Librería con herramientas de interpretación y explicación de modelos.

Series Temporales

Paquetes de R para el análisis de series temporales.

Librerías en Python

NumPy
Manejo de matrices y la realización de operaciones matriciales y vectoriales de forma sencilla y eficiente.

Matplotlib
Gráficas muy completas para mostrar los resultados de tus pruebas.

Scikit-learn
Librería centrada en machine learning: de clasificadores o regresores, hasta selección automática de modelos y análisis de resultados.

lime

Librería con herramientas de interpretación y explicación de modelos.