UniversityHack 2023 Datathon

Equipo ***Unity***

Irina Filimonova Sevcenco Moisés Barrios Torres Elena Marrero Castellano

Resumen

El presente documento recoge las memorias del equipo Unity en la competencia UniversityHack 2023 Datathon, celebrada en Valencia, en los días 23 de febrero de 2023, al 11 de mayo de 2023. Se explica el proceso de elaboración de un dataset con información histórica de las fincas que conforman la cooperativa La Viña, junto con dos datasets con información meteorológica de estaciones climatológicas de las fincas. Estos datasets son tratados en un primer momento para obtener datos limpios, útiles para poder establecer un set experimental de *machine learning*. Seguidamente, se describe la utilización de algoritmos de inteligencia artificial para la introspección de estos datos y la captura del conocimiento de patrones implícitos dentro de ellos, cuyo mejor resultado es obtenido con la técnica de *Random Forest*. Haciendo uso de la explicabilidad, transparencia y sostenibilidad, se ofrece al lector una dimensión socialmente más útil de este proceso matemático-computacional. Es posible leer al final del documento las valoraciones finales y recomendaciones hechas a los organizadores del evento para su utilización en servicios reales que ayuden a la sociedad valenciana, española e internacional en la preservación de la producción vinícola.

Tabla de contenido:

0 Transparencia

Este documento tiene el propósito de favorecer el entendimiento por parte del usuario, describiendo detalladamente todas las partes involucradas en la elaboración de este proyecto. En él, se expone el proceso y la metodología seguida para el tratamiento de los datos y la selección de variables, las técnicas aplicadas para el entrenamiento del modelo propuesto, y los resultados obtenidos del mismo.
Para garantizar que los ejecutables sean generalizables, se pone a disposición del usuario la carga de paquetes y librerías previa necesaria para la ejecución de los mismos (ver apartado 1.1). Además, los ejecutables generan la información adicional necesaria si se desea reproducir este trabajo.

1. Introducción

Nuestro propósito, y objetivo de UniversityHack 2023, es plantear una previsión precisa de la producción de uvas de la campaña de 2022 mediante algoritmos de predicción.

1.1. Instrucciones de uso

Este proyecto se ha realizado en los lenguajes de programación de **Python** y **R**. El análisis exploratorio y el tratamiento de los datos se ha hecho en R utilizando librerías como *tidyr*, *dplyr* o *ggplot2*, y la selección de modelos se ha realizado en Python utilizando paquetes como *sklearn*.

En total el proyecto se compone de un script en rmd, llamado **exploratory.rmd,** y un script en un Notebook de Jupyter, llamado **exploratory.ipynb,** para el análisis exploratorio y pruebas realizadas durante todo el concurso. Por otro lado, un script en rmd para el proceso de extracción, transformación y carga de los datos llamado **prediction.rmd**, y un Notebook de Jupyter llamado **prediction.ipynb** para la predicción del modelo propuesto.

Antes de comenzar, se deben instalar las librerías y los paquetes necesarios. Para los .rmd se debe ejecutar al comienzo la instalación de las librerías necesarias, y para los Notebooks de Jupyter se dispone de un archivo **requirements.txt** con los paquetes necesarios. Para disponer de estos paquetes se debe ejecutar el siguiente comando en la terminal:

```
pip install -r requirements.txt
```

Todos los pasos y archivos para la ejecución del proyecto también pueden ser encontrados en este repositorio de Github: link.

1.2. Conjunto de datos

Para la realización del concurso se dispone de un conjunto de datos con un histórico de producciones de los viñedos que conforman la cooperativa La Viña, así como un histórico de la climatología de los mismos.

- **train**: contiene información histórica de las fincas que conforman la cooperativa La Viña.
- **meteo**: dispone de información meteorológica de estaciones climatológicas de la zona a nivel horario de The Weather Company. Son numerosas variables meteorológicas detalladas de la última hora, las últimas 6 horas, las últimas 24 horas, etc.
- **eto**: dispone de información meteorológica ampliada y transformada de las mismas estaciones agregada en franjas del día. Son la información horaria agregada y transformada de las estaciones climatológicas contenidas en `meteo`. Es un resumen de los periodos con el siguiente patrón: 'Variable + “Local” + periodo + tipo de agregación’.

En total se dispone de 8526 observaciones y 11 variables en la colección de ‘train’ desde el año de campaña 14 a la 22. En la colección ‘meteo’ tenemos 1.223.660 observaciones y 34 variables desde el 29-06-201 al 30-06-2022. Y por último 51.180 observaciones y 275 variables en ‘eto’ para el mismo rango de fechas.

1.3. Análisis exploratorio

Previamente al tratamiento de los datos se realiza un análisis exploratorio. En él, se comprueba el tipo de clases de nuestros datos, estadísticos de las variables como la media o la desviación estándar de las variables numéricas. Además, se observan los valores faltantes. En *eto* y *meteo* existen valores faltantes en varias de sus variables, que se solucionan en el Apartado 2.2.

También se exploran las distribuciones de las variables y la correlación de las mismas. Se observa en el conjunto *eto* que las variables de mínimos, máximos y promedios del están muy correlacionadas entre ellas, al igual que las diferentes variables entre sí.

2. Tratamientos sobre los datasets de datos

Para mejorar la calidad de los conjuntos de datos, se realiza un preprocesado de los mismos. Se realizan transformaciones a las variables, además de la limpieza de valores faltantes y una normalización a los conjuntos adecuada. A continuación se encuentran detalladas las técnicas de preprocesamiento utilizadas.

2.1. Preprocesado de las variables

Tener un conjuntos de datos correcto asegura la accesibilidad y disponibilidad de los datos en cualquier instante. Nos permite un mayor incremento en la productividad y una mayor seguridad de nuestros datos.

2.1.1. Selección de variables de meteo

El conjunto ‘eto’ contiene información agregada del conjunto ‘meteo’, pero no de todas las variables. Por lo tanto, se seleccionarán de ‘meteo’ únicamente aquellas variables que no estén recogidas en ‘eto’. Estas son: la variación máxima en la presión atmosférica en las últimas 3 horas, la diferencia barométrica respecto al nivel del mar y la dirección del viento.

2.1.2. Selección de variables de eto

Como en lo comentado en la sección 1.2, la media, el máximo y el mínimo de las variables de ‘eto’ están muy correlacionadas entre ellas. Además, los estadísticos de máximo y mínimo no son estadísticos robustos, de modo que solamente se seleccionan las medias (Avg).

Por otra parte, se selecciona en específico la variable Day, que es la media resultante de las 24 horas del día, y por tanto es la media de la información recogida del resto de franjas horarias (Daytime, Nighttime, Morning, Afternoon, Evening y Overnight). También se seleccionan las variables Nighttime, ya que podrían ser importantes porque durante la noche es cuando ocurren los cambios meteorológicos más radicales. Por ejemplo, si ocurren fuertes nevadas durante la noche, la producción podría verse perjudicada.

2.1.3. Variable de la campaña en eto y meteo

Ambos conjuntos disponen información temporal de las variables, ya sea diaria o por hora. Por otro lado, se sabe que la campaña de una producción comienza el 1 de Julio de un año, y termina el 30 de Junio del año siguiente: el invierno de un año afectará al año de la campaña siguiente, por lo tanto los datos relativos a partir de junio de ese año formarán parte de la campaña del año siguiente. Para

determinar a qué campaña corresponde cada observación y, para unir posteriormente los tres conjuntos, se crea la variable Campaña indicando la campaña a la que pertenece cada observación: el propio año de la observación si el mes es inferior a Junio, o el año siguiente en el caso contrario.

Tras crear Campaña, al existir los datos para los días 29 y 30 de junio de 2015, esos días se considerarán como los únicos de la campaña 15. Al no ser representativo únicamente 2 días en todo un año, se eliminan esas observaciones.

2.1.4. Variable de la dirección del viento

Cuando se trabaja con la dirección del viento, es común representarlo como un ángulo medido en grados. Sin embargo, cuando se usa la dirección del viento como una característica para un modelo de aprendizaje automático, es conveniente transformarlo usando la función seno y coseno. La razón de esto es que la dirección del viento es de naturaleza circular, lo que significa que los valores más altos y más bajos están conectados. Por ejemplo, una dirección del viento de 360 grados es equivalente a una dirección del viento de 0 grados. Esta circularidad puede crear problemas al usar la dirección del viento como una característica en un modelo lineal, ya que asume que los valores están relacionados linealmente. Al transformar la dirección del viento usando la función seno y coseno, se captura la circularidad de la variable. mejorando el rendimiento del modelo haciéndolo resistente a dicha naturaleza circular.

2.2. Limpieza de valores faltantes

Para asegurar la calidad de los datos que se van a procesar y favorecer la correcta utilización de la información se realiza una limpieza de los datos faltantes depurando datos incompletos o poco relevantes para el estudio.

Los conjuntos ‘eto’ y ‘meteo’ contienen variables meteorológicas que solo se han tomado durante un año. Algunas solamente han sido tomadas durante el 2022, mientras que otras solo en el año 2014. Al contener en su mayoría valores faltantes se eliminan esas variables. Para reemplazar los valores faltantes de las demás variables, se sustituyen por la media agrupada por estación meteorológica y mes, a excepción de las variables del índice de rayos ultravioleta, que al ser de tipo ordinal se reemplazan por la moda.

Para el conjunto de ‘train’, la variable superficie tiene valores 0 en los seis primeros años de campaña. Ya que la superficie de los años que sí se dispone presenta pocas diferencias entre campañas, se asumirá que la superficie es la misma durante todos los años y se completarán los valores faltantes con la media agrupada por finca. Además, existen fincas que no tienen valor de superficie en ningún caso, es decir, no existe ninguna observación para la cual ese identificador de finca tenga al menos un valor de superficie. Por lo tanto, se deben completar esos valores faltantes. Para ello, como presenta una fuerte relación lineal con la producción, se predicen los valores faltantes mediante una regresión lineal. Cabe destacar también que en train los valores faltantes de producción son de aquellas observaciones que debemos predecir del año 2022.

3. Adquisición del conjunto de datos final

Una vez se tienen las variables seleccionadas y sin valores faltantes, se unen los tres datasets por la media agrupada por campaña y estación meteorológica, a excepción de las variables de la radiación ultravioleta que de nuevo se agrupa por la moda. De este modo, el conjunto resultante es las observaciones de train junto con las medias agrupadas de todas las mediciones meteorológicas.

Tras la unión de los datos, como los datos meteorológicos comienzan para las campañas posteriores a la del año 16 (ver apartados 1.2 y 2.1.3), se eliminan las observaciones para las campañas 14 y 15.

4. Normalización

Se realiza una normalización a los datos, ya que los modelos de regresión se basan en supuestos sobre la distribución de los datos y las relaciones entre las variables. Permite comparar y analizar diferentes variables en una misma escala y minimiza el impacto de la variabilidad de las unidades de medida en las diferentes variables.

Tras analizar las distribuciones de las variables, aquellas que siguen una distribución normal se realiza la normalización z-score. Aquellas que no la siguen o que están sesgadas se le realiza la transformación logarítmica y, después z-score. Las variables a las que no siguen una distribución normal, y que por lo tanto se les hace el logaritmo, son superficie y humedad relativa.

Por otro lado, las distribuciones nos permiten comprobar que, tras la unión agrupada datasets, las variables de la radiación ultravioleta tienen varianza 0, ya que la mayoría de sus valores toman el mismo valor. Estas dos variables no aportarán información, y por lo tanto se eliminan del conjunto de datos. Además, visualizando las distribuciones de la variable UV Index se observa que tiene varianza cero por lo que se ha decido eliminar dicha variable.

5. Selección de características

Para la selección de características se analizan las variables numéricas y categóricas por separado. Por un lado, de las variables numéricas se estudia la correlación y se seleccionan las características mediante las regresiones lineales. Por otro lado, de las variables categóricas se ha decidido eliminar la altitud y la campaña. Esto es debido a que existe una gran cantidad de fincas para las cuales no se tiene altitud de algunas zonas, y al ser una variable que representa rangos es complicado completar los valores faltantes y un mal reemplazamiento de los mismos puede perjudicar significativamente a los resultados. Por otro lado, una de las razones principales por la cual se elimina campaña es que, al no ser correcto tratarla como una variable categórica, a la hora de predecir, el conjunto de test no habría visto nunca observaciones con ese tipo de campaña y estaría introduciendo ruido al modelo.

5.1. Correlación

Tras haber normalizado las variables, se hace la correlación al conjunto de datos resultante, concretamente la correlación de Pearson, ya que tras la normalización los datos se distribuyen normalmente (ver apartado 4.), y se explora con la correlación la relación lineal entre variables. El propósito es determinar cuáles son las variables que tienen mayor relación lineal, y seleccionar aquellas que expliquen una mayor cantidad de variables.

Para seleccionar variables se utiliza la función ‘findCorrelation’, que busca a través de la matriz de correlación las columnas altamente correlacionadas las variables que se eliminarán para reducir las correlaciones por pares. De este modo, se eliminan las variables que aportan información redundante.

En la correlación se observa que las variables con mayor número de correlaciones altas con otras variables son el punto de rocío, la humedad relativa, la sensación térmica, la velocidad del viento, el MSLP, todas ellas del día y de la noche, y el volumen de lluvia por hora del día. A continuación, a partir del conjunto de datos reducido, se han seleccionado las variables que más explican la producción de uva ordenando la correlación de los datos por la variable de producción.

Tras haber reducido el conjunto de datos, se ha vuelto a comprobar la correlación entre las variables y seguían existiendo correlaciones entre ellas. Por tanto, se ha repetido el mismo procedimiento para volver a descartar las variables más correlacionadas. El conjunto de variables resultante está

conformado por las variables categóricas de train, y las variables numéricas de temperatura y velocidad del viento.

5.2. PCA

Cabe destacar que una de las pruebas realizadas ha sido hacer modelos locales. El propósito es entrenar modelos locales sobre conjuntos de datos que tengan características comunes. Para ello, se le ha realizado una reducción de dimensionalidad con PCA al conjunto de datos obtenido tras la selección de variables (ver apartado 5.1), y se han clusterizado las observaciones con el algoritmo K-means.

A pesar de que en el resultado de las predicciones del modelo propuesto no se utilice la PCA, se menciona en este apartado para contextualizar esta prueba realizada para la posterior comparación de los 3 mejores modelos (ver apartado 9.).

6. Elección de la muestra de entrenamiento y validación

Finalmente, el conjunto total obtenido aplicando las técnicas antes mencionadas está compuesto de 6262 observaciones con 12 variables. Entre las variables finales cabe destacar la temperatura y velocidad del viento, que contienen los mayores niveles de importancia dentro del conjunto de datos preprocesados, además de las variables sobre la información histórica de las fincas.

Una vez obtenido el conjunto de datos final, se dividen los datos en entrenamiento y validación para ajustar y evaluar los modelos. Esta tarea la realiza internamente la función Grid Search CV de la biblioteca sklearn en Python mediante validación cruzada k-fold. Esta técnica implica la división de los datos en k bloques. En cada iteración, se utiliza uno de los bloques como conjunto de validación mientras que el resto se utilizan como conjunto de entrenamiento. Esta operación se repite k veces, utilizando cada bloque una vez como conjunto de validación, lo que permite obtener una estimación más robusta del rendimiento del modelo.

7. Argumento de la tipología del modelo a desarrollar

En esta investigación se ha comparado el rendimiento de tres modelos para la predicción de la producción de uva: Decision Tree (DT), Support Vector Regressor (SVR) y Random Forest (RF).

Tras evaluar los modelos utilizando técnicas de validación cruzada (ver apartado 6.), se ha llegado a la conclusión de que RF es el mejor para predecir la producción de uva. Obtuvo un mayor coeficiente de determinación R^2 en comparación con los otros dos modelos, lo que indica que explica mejor la variabilidad en los datos. Además, RF superó a los otros modelos en términos de precisión y estabilidad. La precisión se ha evaluado mediante la raíz del error cuadrático medio (RMSE), donde RF obtuvo la menor RMSE, lo que indica que sus predicciones son más precisas. También se evaluó la estabilidad del modelo a través de la variabilidad en los resultados de la validación cruzada, donde se observó que el modelo RF presentó una menor variabilidad frente a los otros modelos.

Por otro lado, aunque la relación entre las variables pueda seguir una regresión lineal, esto no significa necesariamente que un modelo lineal sea mejor para predecir. El modelo de RF puede ser más preciso que uno de SVR o DT porque es capaz de manejar relaciones no lineales entre las variables. En lugar de ajustarse a una línea recta o una superficie de decisión fija, RF utiliza múltiples árboles de decisión para capturar relaciones complejas entre las variables, conduciendo así a una mayor precisión en la predicción.

8. Criterios aplicados para la selección del modelo

A continuación se explican los diferentes criterios de evaluación, y la justificación del modelo propuesto elegido. Estos son: la explicabilidad del modelo, la justicia durante el desarrollo del modelo, y la sostenibilidad ambiental.

8.1. Explicabilidad

RF es explicable gracias a la selección de características importantes, la limitación de la profundidad de los árboles, el número limitado de árboles y el uso de variables predictoras intuitivas. Esto nos permite entender cómo se están utilizando las características en el modelo, y por qué se están realizando ciertas predicciones, lo cual aporta un gran valor comercial a la solución.

El RF es un modelo de aprendizaje supervisado que combina múltiples árboles de decisión para mejorar la precisión de las predicciones. En cada árbol, se toman decisiones basadas en una serie de reglas que dividen los datos en subconjuntos cada vez más pequeños. Estos subconjuntos son utilizados para predecir el valor de la variable objetivo. Nuestro RF se compone de varios árboles de decisión, concretamente 50. Este es un dato importante a conocer porque cuanto más grande sea el número de árboles, más compleja será la interacción entre las variables predictoras y la variable objetivo. Como en nuestro caso tenemos un número de árboles de 50 no se dificulta la explicabilidad del modelo.

Por otro lado, la profundidad de un árbol de decisión determina el número de niveles de decisiones que se toman antes de llegar a una hoja. En este caso, la profundidad del modelo propuesto RF está entre 15 y 25. Este rango no es lo suficientemente grande como para que esté dificultando significativamente la interpretabilidad del modelo.

Por último, RF nos proporciona la información sobre la importancia de las variables predictoras y cómo se están utilizando para hacer predicciones. Además, el modelo RF no es una caja negra porque cada árbol de decisión puede ser interpretado y examinado por separado, visualizando que es lo que hace el modelo en cada una de sus ramas.

8.3. Justicia

En el análisis exploratorio se comprueban las distribuciones de las variables (ver apartado 1.3), y en la normalización se asegura que éstas están distribuidas de manera uniforme y equitativa (ver apartado 4.), lo que significa que no hay sesgos en los datos que puedan afectar la precisión y justicia del modelo. Esto se traduce en que las predicciones del modelo para diferentes subgrupos definidos por características meteorológicas son precisas. Por lo tanto, el conjunto de datos utilizado se puede considerar diverso y representa a la población objetivo, lo que garantiza que el modelo no favorezca a un grupo específico de la población.

Además, se estudia el sesgo del modelo elegido. El modelo utilizado se considera justo: con 50 árboles de decisión y 6262 observaciones, el modelo de RF es lo suficientemente robusto como para manejar datos complejos y proporcionar resultados justos y precisos para la población objetivo.

8.4 Sostenibilidad ambiental

El modelo de RF presentado se considera sostenible debido a su capacidad para procesar grandes conjuntos de datos y ajustarse a nuevos datos a medida que se presentan. El tiempo de entrenamiento del modelo es de 1.27 segundos y tiene 50 árboles, tratándose de un modelo eficiente en el uso de los recursos computacionales. Además, su complejidad se estima del orden de:

$$O(a \cdot L \cdot n \cdot log(n))$$

n = 6262 observaciones; a = 50 árboles; L = 11 variables

Finalmente, si se desea hacer un servicio de predicción de la uva, apenas se necesitan recursos para poner el servicio en funcionamiento, ya que es muy ligero y rápido.

9. Visualización y explicación de los resultados

Partiendo del modelo de RF propuesto en la fase local, y con intención de mejorarlo se propone mejorar la selección de características de los datos realizando una combinación mejorada de los datos utilizando la misma elección de modelo: RF.

Las tres combinaciones con mejores resultados han sido:

- Modelo RF al conjunto de datos de train con las variables de temperatura y velocidad. Se seleccionan las variables que explican el mayor número de variables y se obtienen los resultados utilizando el modelo de RF.
- Modelo RF al conjunto de datos agrupados tras la correlación. Se selecciona el conjunto de datos resultante tras realizar la correlación (ver apartado 5.1) y se entrena utilizando un RF.
- Modelos locales haciendo clusterización con k-Means y PCA. Tras visualizar los datos de las dos componentes principales de la PCA, se observa que los datos resultantes de la correlación están agrupados en 2 clusters. Se realiza una agrupación de dichos clusters utilizando k-Means y se entrenan modelos locales con RF para ambos clusters con los datos de train junto con las variables de temperatura y velocidad del viento. Se ha decidido entrenar con este conjunto, ya que han sido los mejores resultados obtenidos con RF con el propósito de mejorarlo.

Estos han sido los mejores resultados obtenidos:

Prueba	**Best_score**
Train + Temperatura y velocidad del viento	2857.655
Train + Eto + Meteo + Correlación	3071.922
Train + Eto + Meteo + Correlación + Modelos Locales + Temperatura y velocidad del viento	3521.915

Finalmente, se ha decidido que el mejor modelo es el RF entrenado con los datos de train con las variables de temperatura y velocidad del viento a la vista de los resultados y basándonos en los criterios descritos en el apartado 9. Por lo tanto, es el propuesto para la competición de Cajamar UniversityHack 2023 por el equipo Unity.

10. Conclusiones

En este trabajo se ha diseñado una aproximación para resolver el objetivo del concurso de Cajamar UniversityHack 2023. El propósito es plantear una previsión precisa de la producción de uvas de la campaña de 2022 mediante algoritmos de predicción.

Tras varias pruebas con distintos conjuntos de datos, los resultados más notables son los obtenidos después de la selección de variables en función de las correlaciones. Además, cabe destacar que el uso de modelos locales, a pesar de no resultar en la mejor métrica, mejora los resultados en la mayoría de experimentos en comparación al uso de un modelo global.

Debido a las limitaciones de tiempo del concurso, se plantean como trabajo futuro varias mejoras. En primer lugar, es fundamental la realización de un adecuado tratamiento de los datos, así como una selección de características acertada. Es por esto que se pueden explorar las ventajas de la utilización de las características de las series temporales presentes en los datos meteorológicos. Un ejemplo es la información que estas características podrían aportar para rellenar correctamente los valores faltantes en las campañas 14 y 15. Asimismo, en lo referente a la elección de modelos, un ajuste más exhaustivo de los hiperparámetros puede mejorar los resultados significativamente.