Aplicación de minería de datos sobre un repositorio de variables fitofenológicas de cultivos cítricos para la extracción de conocimientos

Martín Ehman[1], Gabriel Surraco[1], Karina Eckert[1], Sergio Garrán[2], Vanesa Hochmaier[2], y Armando Taie[3]

[1] Universidad Gastón Dachary. Posadas, Misiones
{martinehman90, gabrielsurraco1, karinaeck}@gmail.com
[2] INTA EEA Concordia. Concordia, Entre Ríos
sergiomariogarran@gmail.com hochmaier.vanesa@inta.gob.ar
[3] INTA EEA Corrientes. El Sombrerito, Corrientes
taie.armando@gmail.com

Resumen El desarrollo sustentable y eficiente de los cultivos implica un seguimiento sobre los factores que afectan a los mismos. Esta investigación busca determinar las características que influyen en el desarrollo de los cultivos cítricos a través de las variables fitofenológicas que son almacenadas en el sistema FruTIC, mediante técnicas de minería de datos. La metodología utilizada es CRISP-DM y la implementación en el lenguaje R. Los modelos de clasificación construidos fueron evaluados con las métricas *Kappa* y área bajo la curva ROC; y los de regresión, con el RMSE y R^2. La predicción de las variables tuvieron una precisión superior al 76 % para la mayoría de los modelos, excepto para minador y *Diaphorina*; aun así se pudieron identificar y cuantificar la importancia de los atributos predictores con respecto a cada una de las variables objetivos definidas. De los modelos implementados *random forest* y *xgboost* obtuvieron mejor desempeño.

Palabras Clave: Minería de datos · Metodología CRISP-DM · Cultivos cítricos · Modelos de predicción de rendimiento · FruTIC · Manejo integrado de cultivos.

1. Introducción

El manejo integrado de cultivos define estrategias que abarcan varios conceptos a tener en cuenta en el desarrollo de cultivos: la sustentabilidad del medio ambiente, el manejo de plagas y enfermedades, y el uso eficiente de los recursos con el objetivo de lograr el mejor rendimiento posible.

Una herramienta desarrollada por el INTA para el manejo integrado de cultivos es el FruTIC [**1**] [**2**]. La misma es utilizada para generar reportes y alertas en tiempo real y operativo a los productores y técnicos, e informar los momentos óptimos para realizar las diferentes intervenciones en el cultivo, permitiendo así un uso más eficiente de los recursos disponibles. Este sistema almacena información sobre las variables meteorológicas, las etapas de desarrollo de los cultivos y la incidencia de plagas y enfermedades.

Si bien esta herramienta provee de varias funcionalidades de gran utilidad, se ha podido observar que al tratarse de una herramienta de manejo integrado de cultivos no se evalúan las distintas variables dentro de las bases de datos del FruTIC de una manera conjunta. Por ello se ha propuesto realizar un análisis integrado de las variables presentes por medio de técnicas de minería de datos para descubrir patrones y correlaciones en los datos. El desarrollo de modelos de predicción permite contar con conocimiento anticipado a la hora de tomar decisiones, basado en una fuente de información sólida de carácter histórico. Asimismo, estos modelos pueden permitir un uso más eficiente de los recursos al realizar acciones puntuales sobre los lotes que resultan menos invasivos para el ambiente y permitan un ahorro en los recursos económicos. Por otro lado, el uso de modelos predictivos pueden reducir las cantidades de monitoreo que tienen que realizar los técnicos y productores si se contara con buenos niveles de precisión. Finalmente, a través de las técnicas desarrolladas se cuantifica la importancia de cada una de las variables predictoras en los modelos permitiendo conocer en mayor detalle el grado de influencia de las mismas sobre los atributos de respuesta.

1.1. Trabajos relacionados

Teniendo en cuenta antecedentes relevantes asociados a esta investigación, se puede citar en primer lugar a *Knowledge Discovery and Data Mining to Identify Agricultural Patterns* [**3**] en donde los autores proponen la utilización de un proceso de descubrimiento del conocimiento sobre datos provenientes de actividades agrícolas de cultivos de trigo y arroz. Se basan especialmente en algoritmos de agrupamiento, utilizando la herramienta *WEKA* para su implementación.

Por otra parte Kaur y Singh proponen en *Knowledge Discovery on Agricultural Dataset Using Association Rule Mining* [**4**] el uso de reglas de asociación sobre *agrodatos* y el desempeño de distintos algoritmos de este tipo.

Finalmente, se puede citar el trabajo realizado por Bombelli, "Modelado para la predicción de enfermedades en cultivos de alto valor comercial" [**5**] en

donde propone la utilización de modelos matemáticos para predecir enfermedades en los cultivos de arándanos. Esta investigación es importante en referencia al presente trabajo dado que utiliza el concepto de triángulo de las enfermedades.

2. Materiales y métodos

2.1. Repositorios de datos

La base de de datos con la que se trabajó fue provista por el INTA EEA Concordia. La misma proviene del sistema de manejo integrado de cultivos FruTIC, con datos referentes a variables meteorológicas, información fenológica (estadios de floración y brotación de la planta), estado de cultivos, color y calibre de frutos, información sobre cantidades y especie de moscas de los frutos encontradas en las trampas, y cantidades de ramas infestadas con minador y *Diaphorina*. Estos repositorios presentaban datos provenientes de varios lotes observados: 042-Salustiana, 058-INTA-NovaR, Don Tito Nova, Don Tito Valencia Late y Lote II Nova. Los registros de datos analizados están comprendidos en el periodo enero del 2006 a diciembre del 2015.

Los conjuntos de datos fenológicos y de plagas presentaban una frecuencia de muestreo semanal, mientras que los datos meteorológicos una frecuencia por hora. Cada uno de los *datasets* debieron ser analizados en términos de revisión de calidad de datos para luego ser pre procesados.

2.2. Metodología CRISP-DM

La metodología de minería de datos utilizada fue CRISP-DM, la cual es un modelo de procesos jerárquicos estandarizado que agrupa las tareas del proyecto en cuatro niveles de abstracción (de generales a específicas): fases, tareas generales, tareas específicas e instancias de procesos. En la Figura 1 puede observarse la arquitectura de niveles o jerarquías que presentan las tareas en esta metodología.

Las fases definidas en CRISP-DM son: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación, despliegue [**6**]. Esta metodología es un proceso con retroalimentación permitiendo que en cualquiera de las etapas pueda volverse hacia atrás para realizar los ajustes necesarios ante el surgimiento de nuevos requerimientos [**7**].

En la presente investigación se realizaron todas las fases a excepción de la fase de despliegue, que consiste en que los modelos seleccionados en la evaluación sean puestos en producción.

Entendimiento del negocio
Inicialmente se realizó un relevamiento de los tópicos relacionados a la minería de datos como los modelos que se iban a aplicar y la forma de evaluar la precisión de estos modelos predictivos. Por otra parte, se tuvo que investigar acerca del contexto agrícola: triángulo de las enfermedades, manejo

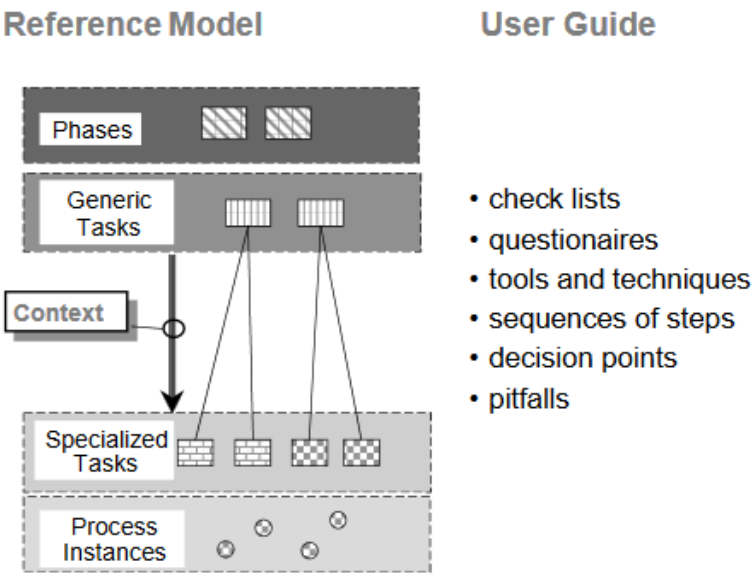

Figura 1. Fases de la metodología CRISP-DM [**6**].

integrado de cultivos, fenología de los cultivos, así como plagas y enfermedades que afectan a estos cultivos. Esta fase de entendimiento del negocio es esencial para conocer el contexto y como los datos toman relevancia en el mismo.

Entendimiento de los datos
Una vez conocido el contexto se procedió a realizar un análisis inicial de los datos. Los procesos involucrados fueron: análisis de la estructura de datos, descripción de atributos, conteo de frecuencias de atributos categóricos y exploración de distribuciones de atributos numéricos.

Preparación de los datos
La preparación de los datos está relacionado con la limpieza y puesta en calidad de los datos. En esta etapa se realizaron procesos de transformación de tipos de datos, creación de nuevos atributos, recodificación de categorías, exploración y corrección de *outliers* [**8**], exploración e imputación de datos faltantes.

Tratamiento de outliers
Inicialmente se procedió a explorar, a través de visualizaciones, los *datasets* en busca de valores que no sigan el comportamiento normal de los datos. Se pudo notar que solamente los datos referidos a meteorología presentaban, en ciertos periodos, valores fuera de los parámetros normales. En la Figura 2 se puede observar la existencia de *outliers*. Los mismos fueron generados por fallas en las estaciones meteorológicas en contextos conocidos. Se puede percibir que en

algunos casos se presentan valores aislados, aunque también las fallas en las estaciones pueden durar varios días. Cuando la estación meteorológica presenta problemas de medición, los valores asignados son -99, sin importar la variable medida.

Se procedió a convertir estos casos a valores *NA* en el lenguaje R, y posteriormente tratarlos en conjunto con el procesamiento de datos faltantes.

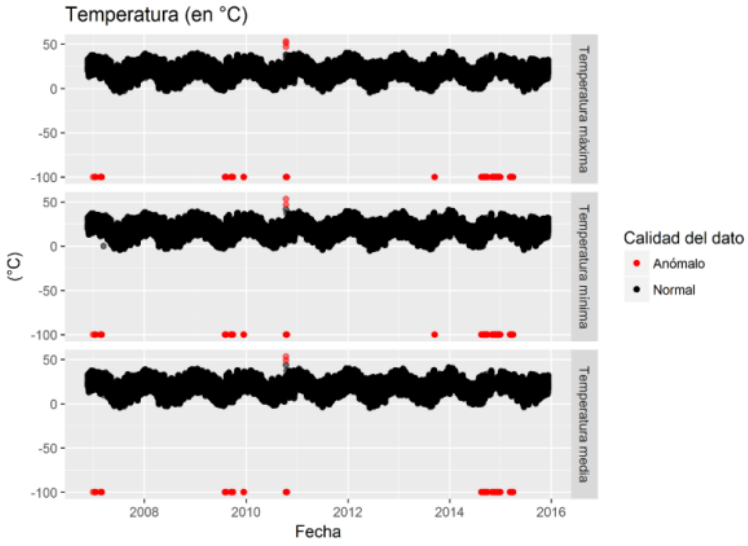

Figura 2. Exploración de *outliers* en la variable temperatura.

Tratamiento de valores faltantes

En el caso de valores faltantes, para inferirlos se recurrió a diversas técnicas como el Promedio Móvil Ponderado [**9**], Modelo Logístico Ordenado [**10**] y el Ajuste de la Media Predictiva [**11**]. Luego de que los datos fueron imputados se evaluó que los mismos se encuentren dentro de los parámetros normales de los atributos imputados, para evitar sesgos en el análisis. En la Figura 3 puede avistarse la variable *brotación* antes y luego del proceso de imputación. En la misma puede observarse las proporciones de cada categoría de brotación teniendo en cuenta la semana del año. Con las visualizaciones se pudo confirmar que aún luego de imputar los datos se siga cumpliendo el ciclo fenológico de la planta en el transcurso del año.

Los datos faltantes en los *datasets* fenológicos y de plagas se produjeron según alguna de las siguientes causas:

- Falta de personal para realizar monitoreos.
- Se prevé que no hayan cambios importantes de una semana a otra.

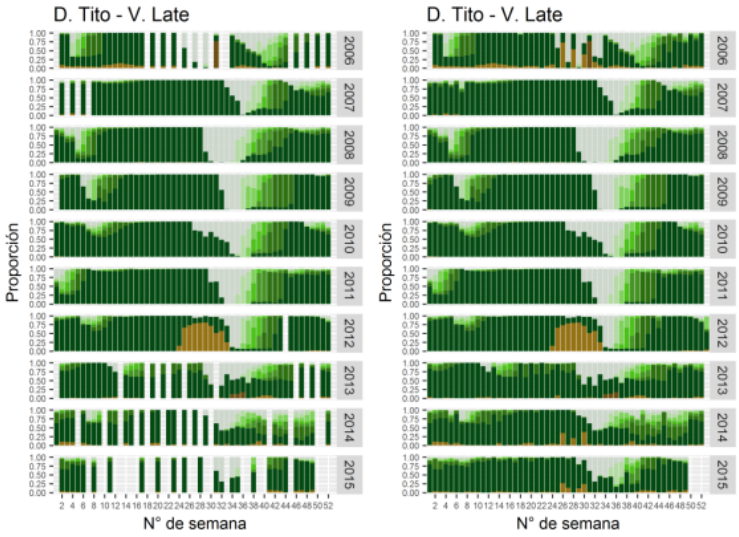

Figura 3. Distribución del atributo brotación antes y después de la imputación.

En la Tabla 1 se presentan los porcentajes de datos faltantes según atributo para el *dataset* de variables meteorológicas. Se consideró descartar aquellos atributos con un porcentaje de datos faltantes mayor al 10 % e imputar aquellos atributos con un porcentaje menor o igual al 10 %, para disminuir la introducción de sesgo en los datos.

Consolidación de datos
Dado que se estudiaron varios *datasets*, antes de modelar se procedió a integrar los mismos en un solo *dataset* consolidado. Dado que los *datasets* referidos a fenología y los de plagas poseen un muestreo semanal (los monitoreos son semanales), la integración de las mismas se realizó de manera directa. Por otro lado, los datos de meteorología se encontraban con una frecuencia por hora, por lo que fue necesario reducir el *dataset* a una frecuencia semanal, y de esta manera se procedió a converger este *dataset* con los demás. Si bien existe una pérdida de información, se crearon atributos acumulativos, como por ejemplo, las precipitaciones acumuladas semanales, para conservar la máxima cantidad de información posible. El conjunto de datos final presentó 34 atributos y 13.620 observaciones.

Antes de proceder a la fase de modelado se realizó un análisis del conjunto de datos consolidado para observar el comportamiento y las interrelaciones entre las distintas variables. Entre estas tareas se realizaron visualizaciones y análisis de correlaciones.

Modelado

Las técnicas supervisadas de minería de datos pueden clasificarse según el tipo

Tabla 1. Porcentajes de datos faltantes por atributo en el *dataset* Meteorología.

Variable	**Número de registros con datos faltante**	**Proporción de datos faltantes**
radiacion.max	47.004	59,25 %
humedad.hoja	19.125	24,11 %
viento.max	3.520	4,44 %
viento.media	3.520	4,44 %
temp.min	2.967	3,74 %
temp.max	2.967	3,74 %
temp.media	2.940	3,71 %
humedad.max	2.581	3,25 %
precipitacion	2.573	3,24 %
radiacion.media	2.558	3,22 %
humedad.media	2.558	3,22 %
humedad.min	2.558	3,22 %
fecha	0	0 %

de problema a resolver, estos pueden ser problemas de clasificación o problemas de regresión. En las técnicas de minería de datos de clasificación la variable de respuesta es de tipo categórica o discreta. Por otro lado, en las técnicas de regresión la variable objetivo que se intenta predecir corresponde a un valor continuo.

Además, los modelos pueden combinarse para obtener mejores desempeños. Los modelos combinados llevan el nombre de *modelos ensamblados* y su objetivo principal es reducir la variancia y el sesgo para lograr un desempeño global mejorado [**12**]. Algunos ejemplos de modelos ensamblados son los modelos basados en *boosting* y *bagging*.

El *dataset* consolidado consistió en los siguientes atributos: *fecha, lote, id.planta, total.minador, total.diaphorina, estado, brotacion, floracion, brot.inicial.minador, brot.inicial.diaphorina, calibre, color, mtd.jackson.mediterraneo, mtd.mcphail.mediterraneo, mtd.mcphail.americana, vel.max.viento, vel.media.viento, humedad.media, humedad.min, humedad.max, temp.media, temp.max, temp.min, grados.acum, duracion.heladas, precip.promedio, precip.acumuladas, radiacion.media, duracion.precip, moscas.temp, moscas.humedad, minador.temp, diaphorina.temp, mes, anio* y *sem.del.anio*.

Una vez que se obtuvo el conjunto de datos consolidado se procedió a desarrollar los modelos de predicción para cada variable objetivo. Las variables de respuesta seleccionadas para generar los modelos fueron: *estado, brotación, floración, calibre*, atributos referidos a cantidades de moscas (*mtd.jackson.mediterraneo, mtd.mcphail.mediterraneo, mtd.mcphail.americana*), cantidades de ramas con incidencia de minador y *Diaphorina*. Los tipos de datos y valores que pueden tomar estos atributos se describen en la Tabla 2.

Tabla 2. Descripción de las variables de respuesta seleccionadas.

Atributo	**Descripción**
Estado	Estado general de la planta. Valores posibles: Malo, Regular, Bueno, Muy Bueno.
Brotación	Estadio de brotación de la planta observada. Valores posibles: B1, B2, B3, B3.4, B4, B5, B6, B7, B8.
Floración	Estadio de floración de la planta observada. Valores posibles: F0, F1.0, F1.1, F2, F3, F4, F5, F6, F7, F8.
Calibre	Diámetro ecuatorial del fruto. Valores mayores a cero.
MTD *Jackson* Mediterráneo	Cantidad de moscas del Mediterráneo por día en trampas *Jackson* Valores mayores o iguales a cero.
MTD *McPhail* Mediterráneo	Cantidad de moscas del Mediterráneo por día en trampas *McPhail* Valores mayores o iguales a cero.
MTD *McPhail* Americana	Cantidad de moscas americanas por día en trampas *McPhail* Valores mayores o iguales a cero.
Total minador	Cantidad total de ramas con presencia de minador en la planta observada.
Total *Diaphorina*	Cantidad total de ramas con presencia de *Diaphorina* en la planta observada.

Una vez determinadas las variables objetivo se definió el tipo de problema en cuestión para asignar qué modelos deberían generarse para las predicciones. Para los atributos de respuesta del tipo numérico se crearon modelos de minería de datos basados en las siguientes técnicas: *k-nearest neighbors* [**13**][**14**], *random forest* [**15**][**16**], *gradient boosting machine* [**17**] [**18**], *extreme gradient boosting machine* [**16**] [**18**]. Por otro lado, para los modelos de predicción de atributos categóricos se seleccionaron las siguientes técnicas: CART [**20**][**21**], C5.0, *k-nearest neighbors*, *random forest*, *gradient boosting machine*, *extreme gradient boosting machine* y redes neuronales. Las técnicas mencionadas fueron implementadas con el paquete *caret* [**22**] en el lenguaje R.

Durante el modelado se utilizó la configuración por defecto, y consiste en dar valores al azar a los parámetros y evaluar distintas combinaciones de estas para minimizar el error. Si bien por fuera de estos parámetros pueden haber otros que se desempeñen mejor en el espacio de búsqueda, se consideró una posición optimista utilizando solamente los parámetros asignados por defecto para evitar un incremento en el tiempo de procesamiento.

Evaluación

Para evaluar los modelos previamente construidos se utilizó una división del *dataset* de 75-25. Para el entrenamiento y la validación cruzada se utilizó el 75 % de los datos y el 25 % restante para la evaluación del modelo. En relación a las variables categóricas se utilizó un muestreo estratificado debido a la existencia de clases no balanceadas, esto permitió reproducir la proporción de cada categoría en los conjuntos de entrenamiento y prueba. El método de *resampling* utilizado fue validación cruzada con $k = 3$.

Para evaluar los modelos de clasificación se utilizaron las métricas la precisión, el coeficiente *Kappa* [**23**], y el área bajo la curva ROC [**24**]. Para los modelos de regresión se tuvieron en cuenta el RMSE [**25**] y el coeficiente de determinación [**25**].

Para indicar que un modelo obtuvo un nivel de precisión aceptable, se han establecido umbrales de precisión que estos modelos debieron superar. Los modelos de clasificación se evaluaron teniendo en cuenta los valores de área bajo la curva ROC con un umbral mínimo de 0,7. Para los modelos de regresión se utilizó el coeficiente de determinación con un umbral mínimo de 0,8.

Se puede observar que en la mayoría de casos los modelos con mayor poder de predicción fueron *random forest* y *extreme gradient boosting machine* tanto en modelos de clasificación como de regresión. Estos modelos se conocen como modelos ensamblados y poseen una complejidad superior a modelos como CART y *knn*. Esta complejidad se ve reflejada en el número de parámetros utilizados en los modelos y en el grado de complejidad que puede adquirir su interpretación [**24**]. Las precisiones de los modelos seleccionados sobre el conjunto de prueba se indican en la Tabla 3.

Tabla 3. Precisión de los modelos seleccionados evaluados sobre el conjunto de prueba.

Atributo	**Modelo**	**Métrica**	**Valor**
Estado	*xgboost*	AUC	0,7910
Brotación	*xgboost*	AUC	0,7723
Floración	*xgboost*	AUC	0,7889
Calibre	*random forest*	AUC	0,97
MTD *Jackson* Mediterráneo	*xgboost*	R^2	0,9709
MTD *McPhail* Mediterráneo	*xgboost*	R^2	0,8891
MTD *McPhail* Americana	*random forest*	R^2	0,9292
Total minador	*random forest*	R^2	0,4234
Total *Diaphorina*	*random forest*	R^2	0,3157

En la Tabla 4 se establecieron qué modelos tuvieron un mejor desempeño en relación a la métrica evaluada.

Importancia de atributos predictores y precisiones de modelos

En la predicción del estado de las plantas el modelo que obtuvo mejor desempeño fue *xgboost* con 0,7910 para el área bajo la curva ROC, seguido por *gbm* con

0,7582; *random forest* con 0,7441; *C5.0* con 0,6797; *rpart* con 0,6686; *knn* con 0,6289; y finalmente *nnet* con un área bajo la curva ROC de 0,6055 y *Kappa* igual a cero. En función a los resultados mencionados se seleccionó el modelo basado en *xgboost*.

Tabla 4. Modelo seleccionado según atributo y métrica evaluada.

Atributo	Métrica	Modelo ganador	Modelo seleccionado
Estado	Precisión	*random forest*	*xgboost*
	Kappa	*random forest*	
	AUC	*xgboost*	
Brotación	Precisión	*xgboost*	*xgboost*
	Kappa	*xgboost*	
	AUC	*xgboost*	
Floración	Precisión	C5.0	*xgboost*
	Kappa	C5.0	
	AUC	*xgboost*	
Calibre	RMSE	*random forest*	*random forest*
	R^2	*random forest*	
MTD *Jackson* Mediterráneo	RMSE	*xgboost*	*xgboost*
	R^2	*xgboost*	
MTD *McPhail* Mediterráneo	RMSE	*xgboost*	*xgboost*
	R^2	*xgboost*	
MTD *McPhail* Americana	RMSE	*xgboost*	*xgboost*
	R^2	*xgboost*	
Total minador	RMSE	*random forest*	*random forest*
	R^2	*random forest*	
Total *Diaphorina*	RMSE	*random forest*	*random forest*
	R^2	*random forest*	

Algunas de las variables con mayor poder predictivo en este modelo fueron: humedad máxima, calibre, brotación B7, temperatura mínima, temperatura máxima, lote D. Tito Nova, humedad media, velocidad media del viento, velocidad máxima del viento, temperatura media, humedad mínima, velocidad máxima del viento, MTD de moscas del Mediterráneo en trampas *McPhail*, lote INTA Salustiana y radiación global media. Entre los atributos que mas impactaron en las predicciones se encuentra la humedad, ésta puede estar indirectamente relacionada con el contenido del agua en el suelo impactando en el estado de las plantas. El estadio de brotación B7 representa el predominio de ramas con hojas

enfermas o senescentes, este atributo contiene un nivel fuerte de incidencia sobre las predicciones de estado. Con respecto a la relación del calibre en las predicciones del atributo estado, no se tenía presente una relación específica, aunque puede darse el caso particular del lote D. Tito – V. Late, en que se manifiesta en forma creciente una infección de CVC (clorosis variegada de los cítricos, enfermedad causada por la bacteria sistémica *Xylela fastidiosa*) y uno de los síntomas más destacados de esta enfermedad es el crecimiento reducido de los frutos. Los datos de este lote son predominantes en el total, dado que son varios años de registro, y pesan en el modelo predictivo.

En la predicción de los estadios de brotación el mejor desempeño lo obtuvo *xgboost* con un área bajo la curva ROC 0,7723. Las variables con mayor impacto en este modelo fueron principalmente las variables meteorológicas como radiación media, temperatura media y temperatura mínima. Por otro lado, los estadios de floración F1.0 y F1.1 impactaron fuertemente en la predicción del estadio de brotación.

En las predicciones del estadio de floración se obtuvo un área bajo la curva ROC de 0,7889 para el modelo *xgboost*. Entre los atributos con mayor importancia en el modelo seleccionado resultaron: calibre, humedad máxima, lote D. Tito - V. Late, radiación media, brotación B1, velocidad máxima del viento, humedad media, lote INTA Salustiana, humedad mínima, temperatura mínima y lote D. Tito Nova.

En la predicción del calibre de los frutos el modelo con mejor desempeño fue *random forest* y obtuvo un coeficiente de determinación de 0,97. El modelo seleccionado fue *random forest* por presentar el mejor coeficiente de determinación. Entre los atributos que tuvieron mayor impacto en la precisión de este modelo se encuentran: color N0(0 % naranja - 100 % verde), lote INTA Salustiana, humedad máxima, MTD de mosca del Mediterráneo en trampas *Jackson*, MTD de mosca americana en trampas *McPhail*, radiación media, MTD de mosca del Mediterráneo en trampas *McPhail*, moscas.humedad (atributo creado en el pre procesamiento de datos), velocidad media del viento, lote D. Tito Nova y la humedad media. El alto grado de precisión de este modelo es importante dado que el calibre es un componente clave en la calidad comercial del producto, y está ligado fuertemente a los estándares y demandas del mercado.

En la predicciones de moscas se obtuvieron modelos con precisiones de entre 88 % y 97 % para las tres especies. En todos los casos se obtuvo el mejor desempeño con el algoritmo *xgboost*. Los atributos con mayor importancia resultaron: grados.acum, moscas.temp, lote.dtitonova, calibre, temp.media y humedad.media. Cabe destacar que el atributo moscas.temp fue creado a partir de las temperaturas favorables para el crecimiento de moscas, y resultó siendo un buen predictor para el modelo. Por lo tanto, hay que tener en cuenta la relevancia de estudiar el contexto para la construcción de nuevos atributos.

Para la predicción de minador se obtuvo un R^2 de 42,34 % con el modelo *random forest*. Los atributos mas significativos para este modelo resultaron ser: lote.dtitovlate, humedad.max, calibre, temp.max, precip.acumuladas, vel.media.viento.

En los modelos de predicción de *Diaphorina* se obtuvo un coeficiente de determinación de 31,57 % con el modelo *random forest*. Los atributos que tuvieron una mayor influencia en la precisión de este modelo fueron las variables meteorológicas como la humedad máxima, la duración de las precipitaciones, la radiación global media y la temperatura máxima. Por otro lado, otras variables que impactan fuertemente en las predicciones son el calibre del fruto y el lote observado.

3. Conclusiones y recomendaciones

Teniendo en cuenta los umbrales establecidos se puede observar que solamente los modelos de predicción de minador y *Diaphorina* no obtuvieron resultados aceptables. En consecuencia, estos modelos no se consideran aptos para implementarlos en producción, pero aun así son importantes porque pudieron identificar y cuantificar la importancia de los atributos predictores con respecto a la variable de respuesta.

Sin embargo, para los demás atributos de respuesta (estado de cultivos, estadios de brotación y floración, niveles de moscas y calibre del fruto) se obtuvieron modelos predictivos con niveles aceptables de precisión capaces de ser utilizados dentro del sistema FruTIC. Esto permitiría que productores y técnicos no tengan que monitorear los atributos en campo, o puedan reducir la frecuencia de los mismos. Hay que destacar que en varias ocasiones la falta de datos se debía a que no había personal disponible para realizar los monitoreos.

Es importante aclarar que los resultados obtenidos son válidos en el marco de la muestra, los modelos deberían ser actualizados periódicamente para reflejar cambios del contexto y en el caso de que se incorporen nuevas variables que se quieran incluir en los modelos predictivos.

Finalmente, los resultados obtenidos permitieron conocer y cuantificar el nivel de importancia de las variables del triángulo de las enfermedades y las relaciones que se dan entre los atributos del repositorio. De esta manera técnicos e investigadores y otras partes interesadas en el sistema FruTIC tienen un conocimiento más detallado de las variables y asociaciones que existen en los repositorios de datos y que forman parte del triángulo de las enfermedades.

3.1. Líneas futuras de investigación

En primera instancia, sería de gran valor ampliar la etapa de despliegue de la metodología CRISP-DM, ya que permitiría añadir los modelos construidos a las funcionalidades ya existentes dentro del sistema FruTIC. Además, esto permitiría la actualización constante de los modelos con la información que se va agregando al sistema, reflejando el estado actual del escenario.

Por otro lado, sería práctico implementar nuevas variables en los *datasets* que reflejen con una mayor realidad el triángulo de las enfermedades. En particular el registro de intervenciones realizadas en los cultivos, y el monitoreo

de otras plagas y enfermedades.

Dado que el concepto del triángulo de las enfermedades es válido para otros tipos de cultivos, se puede utilizar las técnicas de minería de datos como las propuestas (supervisadas), o bien, realizar diversos análisis con técnicas no supervisadas (como asociación y agrupamiento), para la extracción de conocimiento de repositorios de otras especies. En estos casos, la diferencia estaría en la fenología del cultivo, las condiciones meteorológicas de la región estudiada y las plagas y enfermedades que afectan a estos lotes; y en caso de seleccionar otras técnicas, adecuarse a las restricciones de las mismas.

Referencias

1. Hochmaier V., Garrán S., Mika R., Mousques J., Zaballo E., Burdyn L., Taie A., Freixas A., Cerrudo J., Blanco G. FruTIC y MEF: Herramientas básicas para poder implementar un manejo fitosanitario integrado en el cultivo cítrico de la región del rio Uruguay. XV Jornadas Fitosanitarias Santa Fé (2015)
2. Stablum A., Franco S., Ibarrola S., Milera S., Garrán S., Mika R., Marnetto S. FruTIC: Sistema interactivo que permite un manejo integrado del cultivo cítrico. II Congreso de Agroinformática - XXXIX Jornadas Argentinas de Informática Buenos Aires (2010) 680-695
3. Kaur K., Singh M. Knowledge Discovery and Data Mining to Identify Agricultural Patterns. Vol. 3 No. 3. International Journal of Engineering Sciences and Research Technology (2014) 1337-1345
   Vol. 2 No. 3. R News (2002) 18-22 Vol.37 No. 5. Family Medicine (2005) 360-363
4. Khan F., Divakar S. Knowledge Discovery on Agricultural Dataset Using Association Rule Mining. Vol. 4, No. 2. International Journal of Emerging Technology and Advanced Engineering, (2014) 925-930
5. Bombelli E. Modelado para la predicción de enfermedades en cultivos de alto valor comercial. Buenos Aires, Editorial de la Universidad Tecnológica Nacional (2011)
6. Wirth R., Hipp J.: CRISP-DM: Towards a Standard Process Model for Data Mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining (2000)
7. Chapman P, Clinton J., Kerber R., Khabaza T., Reinartz T., Shearer C., Wirth R.: CRISP-DM 1.0: Step by step data-mining guide, https://www.the-modeling-agency.com/crisp-dm.pdf, último acceso 20 jul 2017
8. Aggarwal C.C.: An Introduction to Outlier Analysis. Outlier Analysis. Springer Cham. (2017)
9. Milton A.: Simple, Exponential and Weighted Moving Averages, https://www.thebalance.com/simple-exponential-and-weighted-movingaverages-1031196, último acceso 10 oct 2017
10. Michalos, Alex C.: Encyclopedia of quality of life and well-being research. Ordered Logit Model. Springer Dordrecht Países Bajos (2014)
11. Castro Cacabelos, M.: Imputación de datos faltantes en un modelo de tiempo de fallo acelerado. Universidade de Santiago de Compostela España (2014)
12. Zaki M., Meira W.: Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press Nueva York (2014)
13. Peterson, L.: K-nearest neighbor. Scholarpedia. Vol. 4. (2009) 1883

14. Venables, W., Ripley, B.: Modern applied statistics with S-Plus. 4ta ed., Springer Nueva York (2002)
15. Breiman, L.: Random Forests. Machine Learning. Vol. 45. Springer (2001) 5-32
16. Liaw A., Wiener M.: Classification and Regression by randomForest. Vol. 2 No. 3. R News (2002) 18-22
17. Friedman J. H.: Greedy function approximation: A gradient boosting machine. The Annals of Statistics. Vol. 29 No.5. Institute of Mathematical Statistics (2001) 1189-1232
18. Ridgey G.: gbm: Generalized Boosted Regression Model, https://CRAN.R-project.org/package=gbm, último acceso 5 abril 2018
19. Chen T., He T., Benesty M., Khotilovich V., Tang Y.: xgboost: Extreme Gradient Boosting, https://CRAN.R-project.org/package=xgboost, último acceso 16 abril 2018
20. Breiman L., Friedman J. H., Olshen R. A., Stone C. J.: Classification and Regression Trees. Wadsworth Publishing Co Inc (1983)
21. Therneau T., Atkinson B.: rpart: Recursive Partitioning and Regression Trees, https://CRAN.R-project.org/package=rpart, último acceso 22 abril 2018
22. Kuhn M., Wing J., Weston S., Williams A., Keefer C., Engelhardt A., Cooper T., Mayer Z., Kenkel B., R Core Team, Benesty M., Lescarbeau R, Ziem A., Scrucca L., Tang Y., Candan C., Hunt T.: caret: Classification and Regression Training, https://CRAN.R-project.org/package=caret, último acceso 25 abril 2018
23. Viera A. J., Garret J. M.: Understanding interobserver agreement: the kappa statistic. Vol.37 No. 5. Family Medicine (2005) 360-363
24. James G., Witten D., Hastie T., Tibshirani R.: An Introduction to Statistical Learning: with Applications in R. Springer, Nueva York (2013) 24-26
25. James G., Witten D., Hastie T., Tibshirani R.: An Introduction to Statistical Learning: with Applications in R. Assessing the accuracy of the model. Springer, Nueva York (2013) 68-71