ZEKE

INTEGRADORES DE SISTEMAS

Desarrollo de una herramienta de pronóstico de volúmenes y calibres a cosecha de cerezas basado en Inteligencia Artificial (Deep Learning)

FDF

FUNDACIÓN PARA EL DESARROLLO FRUTÍCOLA

Viña del Mar, 1 Diciembre 2019

Introducción 3

Objetivos 3

Etapas del proyecto 4

Red neuronal 4

Número de registros 5

Variables 5

Estimación importancia de Variables 7

Experimentos 7

Experimentos Junio 2019 7

Experimentos Julio 2019 8

Experimentos Noviembre 2019 9

Conclusiones 11

1. Introducción

El contar con una mala estimación de la producción de cerezas genera problemas logísticos y operativos que finalmente se traducen en un perjuicio económico para exportadores y productores. La estimación de mano de obra necesaria, los materiales que se requieren para embalar los productos y la capacidad de los packagings son tres factores que se ven afectados directamente por la estimación de la producción.

El siguiente informe contiene los resultados obtenidos en el desarrollo del proyecto "Desarrollo de una herramienta de pronóstico de volúmenes a cosechas basadas en inteligencia artificial (Deep Learning)". Este proyecto tiene como propósito el estudiar y desarrollar un modelo que permita el predecir la producción por hectárea de un predio, teniendo en cuenta una serie de parámetros ambientales y del cultivo. El presente informe busca entonces resumir el trabajo realizado y también reunir las principales lecciones aprendidas en el desarrollo del proyecto, aplicables a otros proyectos similares.

2. Objetivos

El objetivo de este proyecto es realizar una prueba de concepto que permita validar la factibilidad de realizar una estimación en base a un algoritmo de inteligencia artificial. En la literatura existen iniciativas similares que se han aplicado a otro tipo de cultivos como el arroz. Sin embargo ese tipo de cultivo son mucho menos complejos en su manejo que el caso de las cerezas, lo que no permite asegurar que los resultados se pueden transferir de un dominio a otro.

Los algoritmos basados en Inteligencia Artificial tienen la capacidad de encontrar relaciones entre variables en un volumen alto de datos y cuya relación puede no ser de forma lineal o directa entre sí. Esto es realizado mediante el ajuste de parámetros de una serie de matrices que ponderan el efecto de las variables de entrada, buscando minimizar la diferencia entre un valor y su predicción, de modo supervisado.

Para este tipo de problemáticas, estos algoritmos deben ser alimentados con un conjunto amplio de datos históricos recopilados durante varias temporadas. Estos datos corresponden a diferentes variables que influyen en el proceso de producción (climáticas o productivas) y medibles de forma objetiva y precisa. Debido a que no es clara la relación de cada una de estas variables en la producción final, uno de los resultados esperados en este proyecto es descubrir cuales son la o las variables que influyen de manera importante en la estimación de la producción de huertos de cerezo.

3. Etapas del proyecto

La metodología utilizada en este proyecto tiene dos partes: en la primera se busca entender el problema y las variables relacionadas, y en la segunda entrenar un modelo usando el esquema básico de los proyectos de Machine Learning. Las actividades consideradas son las siguientes:

1) Entendimiento del problema (Visita packing y predios).
2) Investigación del estado del arte.
3) Determinación de variables climáticas y productivas a utilizar.
4) Recopilación de datos.
5) Análisis y limpieza de datos.
6) Definición de arquitectura de red neuronal.
7) Experimentos de predicción.
8) Análisis de resultados.

En la ejecución del proyecto, la etapa más compleja fue la recopilación de datos. Esto se debe a que el volumen de datos recopilados no fue el óptimo a nivel de cantidad y de calidad. Del total de datos capturados, no todos los productores poseían registros confiables, guardan la información en distintos formatos sin mayor estandarización entre productores ni entre registros; y algunos datos se presentaban con registros fuera de rango en comparación con otros productores. El contar con un método estandarizado para almacenar y procesar esta información entre todos los actores se hace necesario para poder enfrentar este tipo de proyectos en el futuro.

Luego de un segundo intento se pudo ampliar el conjunto de datos para poder realizar una mejor prueba, pero aún con un número muy por debajo de lo esperado. Esto fue debido a que los datos añadidos al conjunto poseían los mismos problemas previamente mencionados que tenían los datos en el primer análisis.

4. Red neuronal

En una primera etapa del proyecto se buscó seleccionar una red neuronal apropiada para realizar las predicciones. Se compararon una serie de arquitecturas para la predicción, de las cuales se consideraron modelos de tipo red neuronal recurrente (RNN), modelos de tipo perceptrón multicapa (MLP), modelos matemáticos de regresión en series de tiempo, entre otros. Inicialmente se realizó una comparación del desempeño bruto de estas arquitecturas, descartando aquellas que no lograsen hacer predicciones correctas debido al bajo número de registros. Como resultado de este análisis

preliminar, se seleccionó una arquitectura basada en MLP la cual de forma consistente lograba la menor tasa de error frente a la baja cantidad de datos presentes.

El objetivo del modelo utilizado fue el predecir el volumen de producción (kg/ha) de un predio, utilizando información tanto productiva, climática, de ubicación, y condiciones del predio. De estas variables, al no poseer información previa sobre su relación directa con la producción estimada, se realizó un análisis comparativo de desempeño entre diversas combinaciones de variables, para poder determinar de este modo las variables de mayor relevancia para el modelo.

Número de registros

La cantidad de registros utilizadas en el proyecto corresponde a 538 registros con un grado de completitud variable. El número ideal para una prueba de conceptos es 5.000 datos etiquetados. Lo anterior afecta negativamente en la capacidad de estimación del modelo a entrenar, debido a que el modelo debe inferir sobre información incompleta, impidiéndole reducir el error entre la información real y la predicción.

Variables

Las variables recopiladas se pueden ver en la siguiente tabla:

Nombre Variable	Utilizada en pruebas
Exportadora	No
Razón Social	No
Nombre Predio	No
Región	No
Comuna	No
Latitud	Si
Longitud	Si
Cuartel	Si
Superficie	Si
Variedad	Si
Portainjerto	Si

Número de Plantas / hectárea	Si
Año plantación	Si
Primer año de producción	Si
Dardos por planta	**NO**
Frutos por dardos	**NO**
Temporada	Si
Producción (Kilogramos por hectárea)	Si
Calibre (SP, P, SJ, J,XL,L)	**NO**
Estación climática más cercana	Si
PF	Si
HF7	Si
GD10	Si
Lluvia (Ago-Dic)	Si
Humedad (Ago-Dic)	Si
Tmax (Ago-Dic)	Si
Tmin (Ago-Dic)	Si

Las variables no utilizadas se descartan debido a que no todos los registros cuentan con valores para esa variables, tal como se muestra en la siguiente figura.

Variable	2014	2015	2016	2017	2018
N° de dardos en promedio por planta	23	50	56	63	69
N° de frutos en promedio por dardo	13	39	44	27	55

Estimación importancia de Variables

Con el fin de encontrar las mejores variables se realizaron varios experimentos donde se utilizaban diferentes conjuntos de variables y los modelos.

Cada prueba consiste en seleccionar un conjunto de variables, luego entrenar el modelo utilizando 449 registros, y por último evaluar la predicción sobre un conjunto de 89 registros que no están presentes en el entrenamiento. Los modelos se ordenan de acuerdo a este resultado.

5. Experimentos

En el transcurso del proyecto se generaron varias rondas de experimentos para descubrir el mejor método para tratar los datos.

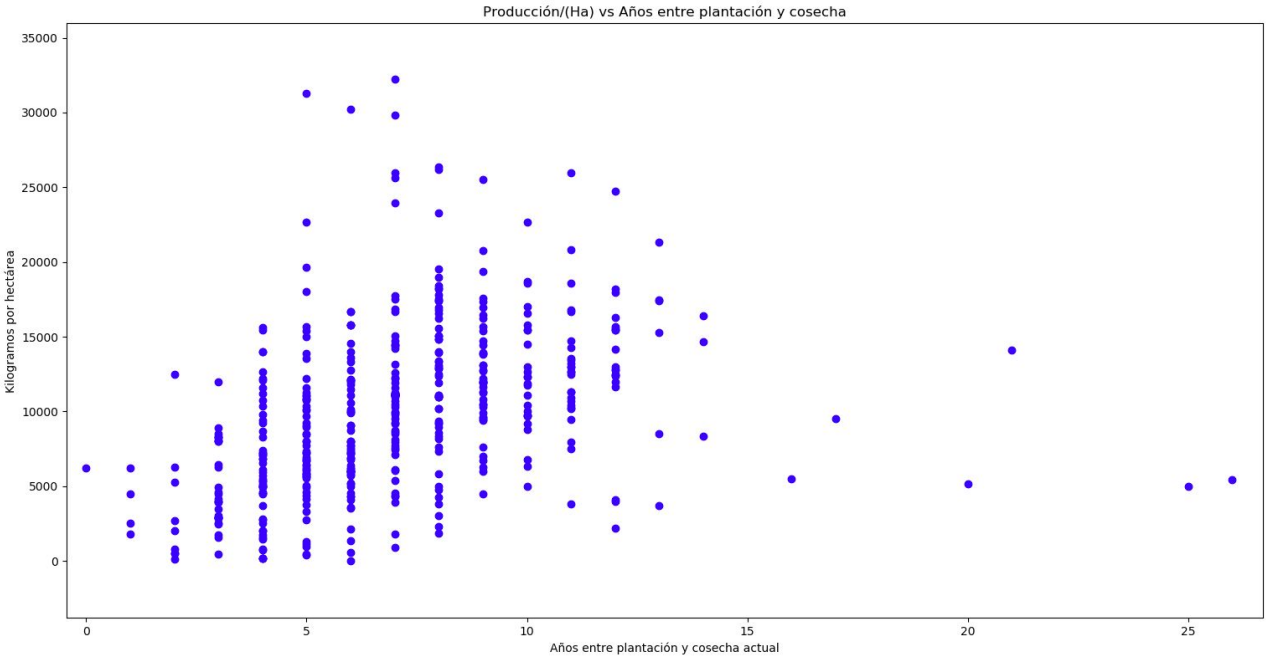
Experimentos Junio 2019

Se centra en probar un enfoque de series de tiempo usando cadenas de tiempo. Los resultados obtenidos se muestran en la siguiente tabla:

Tamaño de la serie	Error promedio en la estimación (kilogramos por hectárea)
3	4968
4	4474

Debido a la forma que tienen los datos se concluye que utilizar este esquema es factible y se decide cambiar un perceptrón multicapas.

Adicionalmente, se buscan relaciones entre variables en forma manual, para corroborar que los resultados obtenidos son correctos, y las predicciones respetan estos patrones. Por ejemplo, en la siguiente figura se ve la relación obtenida entre producción por hectárea, y los años entre la plantación y cosecha, donde se observa una relación positiva.



Experimentos Julio 2019

Se centraron en probar el perceptrón multicapa y la importancia de las variables HF7 y GD10. Algunos datos climáticos de 2014 se deben extraer de las tablas de climas por hora. Entre los problemas con los datos encontrados se encuentra 2 localidades sin datos climáticos (Romeral y Las Cabras).

Adicionalmente se realiza un análisis enfocadamente sólo en la variedad Lapins, debido a que es la que tiene mayor cantidad de datos y se busca evitar el efecto que genera el uso de otras variedades.

La siguiente tabla muestra las combinaciones de variables utilizadas en esta ronda de experimentos, y los valores de error medio en la predicción.

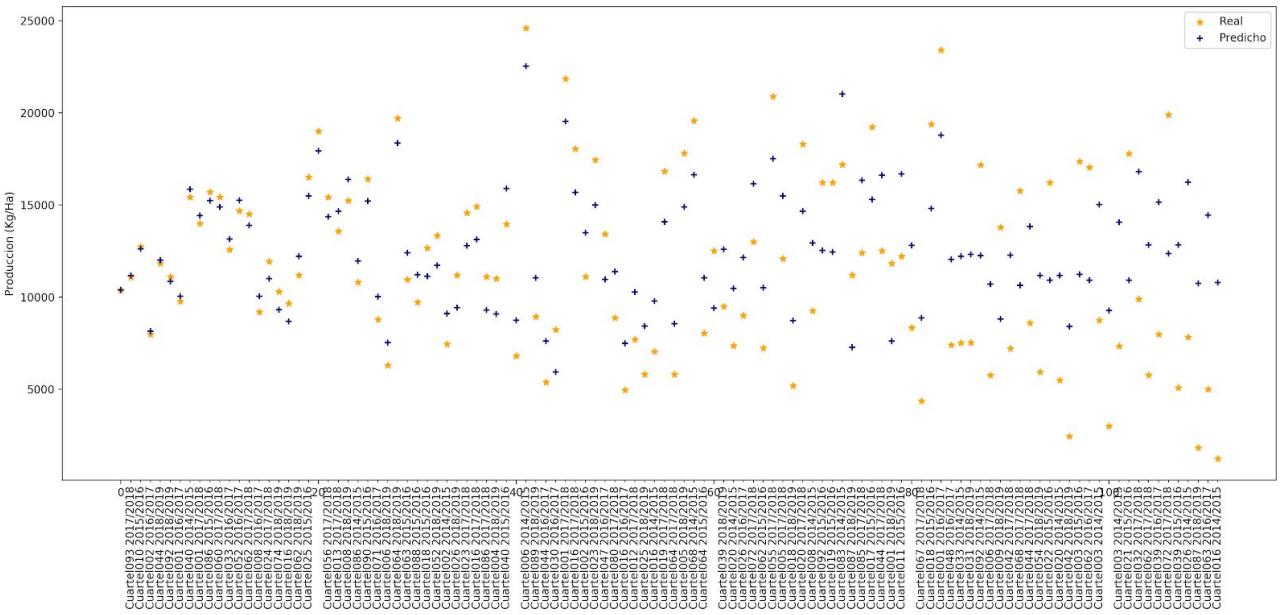
Variables utilizadas	Valor del error medio en la predicción (Kg/Ha)
Horas Frio (base 7)	3657
Porción Frío	3465
HF7 + GD10	3702
GD10 + Porción Frío	3490
Todos	3452

La conclusión de estos experimentos es que las variables de clima afectan positivamente en la predicción, y son incorporadas en el futuro.

Experimentos Noviembre 2019

Estos experimentos son aquellos que tienen los mejores resultados y consolidan todo el aprendizaje previo. Después de cerca de 52 variaciones de modelos, el listado con las mejores variables para realizar las predicciones corresponde a las siguientes: edad de la planta, cantidad de lluvia (ago-dic), variedad, superficie. En una segunda categoría se encuentran las variables: humedad, gd10, temperatura.

El mejor modelo tiene un error en la predicción de un 25.4% que equivale a 2.715 kilos / hectárea. A continuación se puede ver un ejemplo de las predicciones utilizando este modelo.



Se puede notar que gran parte del error se encuentra en la zona de la derecha de las predicciones, donde el modelo sobreestima la producción. Es posible que estas malas predicciones no puedan ser explicadas con las variables existentes en el modelo y se deba investigar individualmente el motivo de las desviaciones.

6. Conclusiones

Al finalizar la ejecución del proyecto existen varias lecciones aprendidas que deben ser consideradas en el futuro.

En primer lugar la calidad y cantidad de los datos impacta fuertemente en el desarrollo de un modelo predictivo para el cultivo de cerezos. Estos datos son administrados por diferentes productores y exportadores para quienes el valor de esta información histórica no es relevante. Contar con una forma estandarizada de registrar y almacenar la información es indispensable para poder generar cualquier tipo de modelo matemático o de inteligencia artificial. Este fue el punto clave que determinó el techo al que podía llegar la predicción. Se utilizó más tiempo del estimado inicialmente en limpiar y validar estos datos.

Entre los modelos que se utilizaron para realizar las predicciones, aquellos que funcionan en base a series de tiempo no dieron los resultados esperados. Esto se debe a que las series de tiempo son demasiado cortas (hasta 5 registros en la serie si se considera como inicio los datos de 2014). Para que una red de este tipo funcione correctamente se requieren una serie de cerca de 600 datos. Este modelo fue rápidamente descartado. En un esquema similar, los modelos del tipo Redes Neuronales Recurrentes (RNNs) tampoco dieron los resultados esperados. Estas redes tienen la capacidad de encontrar relaciones temporales entre variables, pero la ausencia de muchos datos históricos también generó que se obtuvieron malos resultados.

Finalmente se utilizó un modelo de tipo perceptrón multicapa ya que tiene buenos resultados en el caso de las regresiones multivariada. Este modelo logró un error promedio en la estimación cercano a 2.715 kilogramos por hectárea, equivalente a un error del 25.4%. Analizando estos resultados se descubre que el error en la predicción tiene un sesgo hacia la sobre estimación. Este error no se puede explicar con ninguna de las variables presentes en los modelos lo que lleva a la conclusión que existen variables que no han sido consideradas en los modelos y que pueden ayudar a explicar esta baja en la producción.

7. Próximos pasos

- Desarrollar una herramienta de terreno, tipo Apps, que capture y tome datos del huerto, de manera de administrarlos en una base de datos.
- Generar una herramienta que trabaje con el cuaderno de campo digital y permita sistematizar y agrupar los datos para futuros análisis.

Con esto se evitaría perder los datos y se podrían utilizar a futuro, ya sea para desarrollar algún modelo de pronóstico o para trabajar en cualquier indole con inteligencia artificial.