Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

PREDICCIÓN DEL RIESGO DE PÉRDIDA DE PRODUCCIÓN AGRICOLA A NIVEL CULTIVO Y MUNICIPIO CON BASE EN PRONÓSTICOS DE LLUVIAS

Información general del proyecto

Sector.	**Lenguaje.**	**Fuente de datos.**
Agropecuario	R y Python	IDEAM, Ministerio de agricultura y DANE

Entidades involucradas

- Departamento Nacional de Planeación
  - Dirección de Desarrollo Rural Sostenible
  - Dirección de Desarrollo Digital

Presentación

Teniendo en cuenta la importancia que representa el sector agrícola para la economía del país y la gran influencia de los cambios climáticos en su desarrollo, resulta conveniente detectar las principales problemáticas que se puedan generar a causa de la gran variabilidad del clima en el territorio nacional y de esta forma darles solución. Por esta razón, es necesario caracterizar y clasificar la producción dentro de los municipios y cultivos basados en la información del Ministerio de agricultura y el DANE.

Este proyecto tiene como principal fin, mitigar las adversidades que se puedan presentar en el sector agrícola, detectando el riesgo al que están expuestos los cultivos por cambios principalmente en el nivel de pluviosidad de forma tal que se pueda cubrir con anticipación este riesgo por medio de instrumentos financieros como seguros.

Taking into account the importance that the agricultural sector represents for the economy of the country and the great influence of climate changes in its development, it is convenient to detect the main problems that can be generated due to the great variability of the climate in the national territory and in this way give them solution. For this reason, it is necessary to characterize and classify production within municipalities and crops based on information from the Ministry of Agriculture and DANE.
The main purpose of this project is to mitigate the adversities that may arise in the agricultural sector, detecting the risk to which the crops are exposed by changes mainly in the level of rainfall so that this risk can be covered in advance by means of Financial instruments as insurance.

Objetivo general

Establecer una medida del riesgo de pérdida de producción al que están expuestos los cultivos y municipios por variaciones climáticas, especialmente en los niveles de pluviosidad.

Objetivos específicos

1. Consolidar la información de cantidad de lluvias en los diferentes municipios de Colombia con la información de producción disponible para el proyecto.

2. Categorizar los cultivos y municipios que se encuentran en todo el país y extraer sus características más importantes.
3. Clasificar el riesgo de pérdida de producción agrícola al que están expuestos los cultivos y municipios ante cambios en los niveles de lluvias, basados en los pronósticos de pluviosidad obtenidos previamente.

Metodología

Toda la información proveniente de las evaluaciones agropecuarias municipales (EVA) y del censo nacional agropecuario (CNA), permite observar la ubicación y principales características de los cultivos que se encuentran en el territorio nacional. De la EVA se encuentran datos desde el año 2006 hasta el 2018 y el censo nacional agropecuario que es el tercero de su tipo, fue publicado en el 2014 con los datos recolectados en el 2013. Del mismo modo se tienen los datos de pronósticos de lluvias obtenidos previamente que son cruzados con las otras dos bases de datos.

Si bien la información que se encuentra dentro de estos datos es muy completa, algunas variables no son de interés del proyecto, por lo cual el primer paso fue realizar un preprocesamiento en cuanto a renombrar variables, explorar diccionarios y determinar aquellas variables que puedan aportar al estudio (omitiendo variables redundantes o irrelevantes en cuanto al grado de desagregación necesario).

En cuanto a la EVA, las variables relevantes son: código del municipio y del departamento donde se encuentra el cultivo; año, grupo, subgrupo y cultivo específico; producción ajustada, es decir el rendimiento por área cosechada; rendimiento, área cosechada; área sembrada y se crea un público que será la variable de salida. Ya determinadas las variables a incluir, se adecúan sus formatos para su posterior manipulación, así mismo se crean algunas llaves que resultan necesarias para hacer la unión de esta base con los pronósticos de lluvias. Dado que esta última base posee datos desde 1980 hasta el 2019, se dejan únicamente aquellos del 2006 en adelante para lograr emparejarlos con los de la EVA. Para mejorar los datos, se eliminan aquellos cultivos que aparecen tan solo 3 años o menos dentro un lugar específico pues el análisis de estos no provee hallazgos relevantes y su precisión resulta muy baja.

Para el análisis, se manejaron dos tipos de desagregaciones, el primerio a nivel cultivo y el segundo a nivel municipio, dicha desagregación se utilizó debido a que el segundo es más útil en cuanto a la toma de decisiones para la implementación de un seguro territorial al municipio, sin embargo, el primero es más útil en cuanto a la información que esta pueda dar para futuras investigaciones:

1. A nivel cultivo, que condensa la información de todos los cultivos iguales de un año en un solo objeto, sin discriminar por el municipio en que aparece. Para ello se suman las variables: producción ajustada, rendimiento, área cosechada, área sembrada para tenerlas a nivel general por cultivo en cada año.
2. La otra base es a nivel municipio, en donde se eliminan las variables grupo, subgrupo y cultivo, para establecer una producción ajustada; rendimiento; área cosechada y área sembrada total de todos los cultivos que existen en cada municipio por año.

Consolidadas estas nuevas bases el proceso a seguir es el mismo para cada una de ellas, la variable objetivo es la elasticidad de la producción con respecto a las lluvias, para lo cual es necesario calcular el crecimiento de

la variable producción ajustada y de las lluvias, este cálculo da como resultado, números positivos y negativos lo cual no resulta fácil de comprender pues los modelos a entrenar más adelante son clasificadores binarios, para esto a los valores positivos les es asignado un 0 (que indica que ante un crecimiento en las lluvias, la producción no disminuye) y a los valores negativos un 1 (que indica que ante un crecimiento de las lluvias, la producción se ve afectada con pérdidas). Con el público establecido se realiza el entrenamiento de los modelos, procedimiento realizado en Python.

Los modelos que fueron entrenados son:

- Red Neuronal (Neuronal Network/NN)
- Random Forest (RF)
- Support Vector Machine

Por otro lado, para el censo nacional agrícola se tienen variables más específicas respecto a los cultivos y las variables que se mantienen por ser las más relevantes para el estudio son: código del departamento y municipio; unidad cobertura de observación; código de la vereda; número de orden del lote; cultivo presente o pasado; cultivo; mes de siembra del lote; año de siembra del lote; cultivo solo o asociado; tipo de la semilla; finalidad de la plantación; producción en toneladas; rendimiento de las toneladas por hectárea; si sufrió afectación por fenómenos naturales; área sembrada y área cosechada. Con estas variables se mejoran los datos organizando su formato, asignando categorías numéricas a aquellas que lo requieren y aquellas donde se encuentran NAs son reemplazadas por 0. Ya con esto se puede cruzar la información de lluvias tomando solo el año 2013 que es el que coincide con la información del censo.

Como se hizo con la EVA, se crean dos nuevas bases de datos, a nivel cultivo, no fue posible agregarlo a nivel municipio debido a que muchas de las variables relevantes para el análisis, perdían su valor al realizar dicha agregación. Para la base a nivel cultivo se tienen las variables: código de municipio, unidad de cobertura de observación, código de vereda, numero de orden del lote, cultivo presente o pasado, cultivo, mes siembra del lote, año de siembra del lote, cultivo solo o asociado, tipo de semilla, finalidad de la plantación, afectación natural y se agregan por cultivo y año; producción por toneladas, rendimiento de toneladas por hectárea, área sembrada y área cosechada. Al igual que con la EVA se exportan las bases a formato Excel para ser modeladas en Python. Para el CNA, la variable objetivo es el logaritmo de la producción por toneladas, esto debido a que existen dificultades en el modelamiento a causa de la distribución de los datos, teniendo características de una distribución *long tail distribution,* adicional a ello, los cambios en una variable de entrada, resultaría como cambios porcentuales en la variable de salida, mas no el valor en niveles, lo cual facilita saber si hubo crecimiento o decrecimiento.

Los modelos entrenados para el censo fueron:

- Árbol de decisión
- Red Neuronal (Neuronal Network/NN)

Resultados

Una vez estimados los modelos para la EVA, se calcula para cada uno algunas métricas de rendimiento que permiten saber cuál de los modelos tiene los resultados más acertados para el estudio del proyecto.
Estas métricas son:

- Accuracy (en español exactitud), mide el total de los elementos clasificados correctamente
- Precisión, son los elementos clasificados correctamente como positivos del total de elementos clasificados como positivos
- Recall, son los elementos clasificados correctamente como positivos de un total entre los verdaderos positivos y los falsos negativos (los positivos reales)
- F1, funciona cuando se quiere un equilibrio entre Precision y Recall

Cabe resaltar que dependiendo de los resultados y el enfoque que se espera dar al problema, algunas métricas son más relevantes que otras.

Los cálculos de las métricas arrojan los siguientes resultados para la base a nivel cultivo (Figura 1).

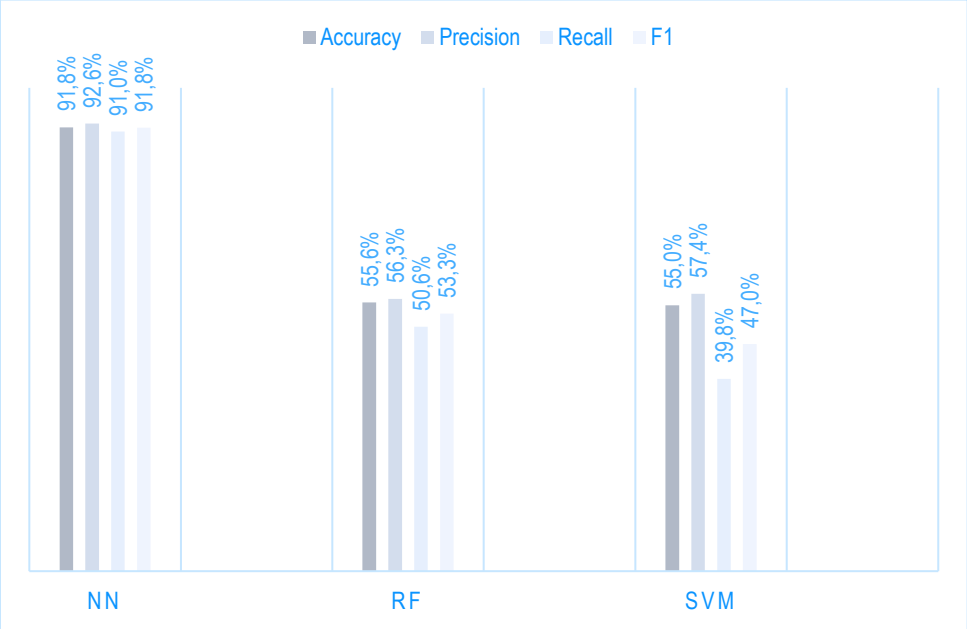

Figura 1: Métricas de rendimiento de los modelos entrenados para la base a nivel cultivo de la EVA

De estos resultados se puede concluir que el modelo con mejor desempeño es la red neuronal (NN), pues tiene las métricas de rendimiento más altas, dichos resultados corresponden a los datos de prueba con una separación de 70-30 para todos los modelos.

Los cálculos de las métricas arrojan los siguientes resultados para la base a nivel municipio (Figura 2)

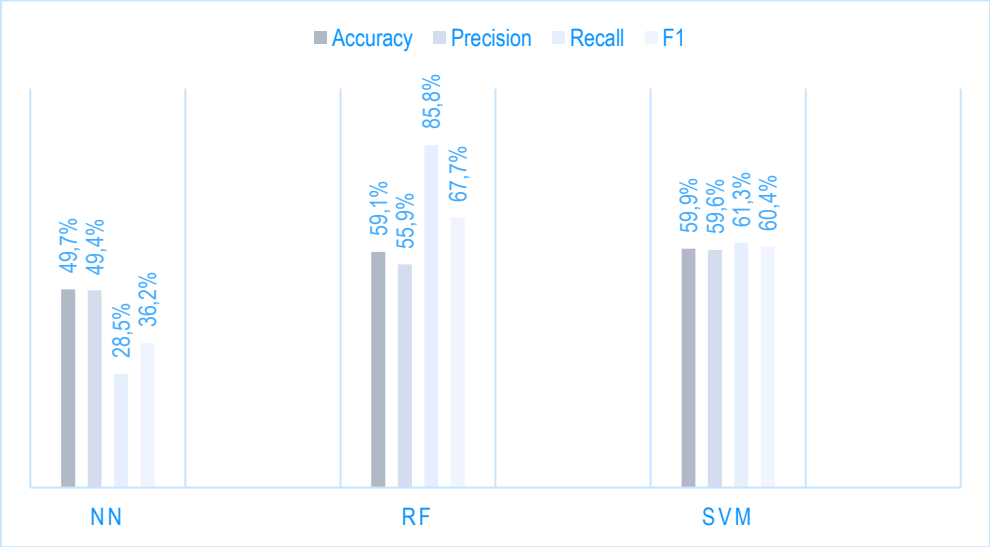

Figura 2: Métricas de rendimiento de los diferentes modelos entrenados para la base a nivel municipio de la EVA.

De estos resultados se puede concluir que el modelo con mejor desempeño es el Random Forest (RF), pues tiene las métricas de rendimiento más altas.

Para los modelos del censo, se tuvieron en cuenta métricas de error a nivel continuo como es el error cuadrático medio (MSE), su raíz cuadrada ((RMSE) y el error absoluto medio (MAE), mostrados en la Figura 3.

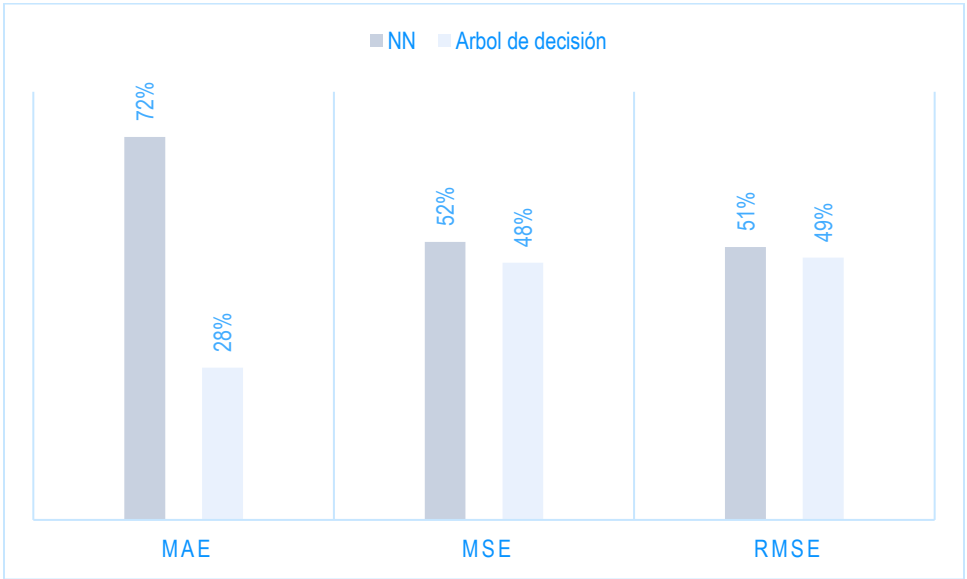

Figura 3: Métricas de error para el modelo del Censo nivel cultivo.

Las métricas de error calculadas para el modelo nivel cultivo y la comparación entre los modelos resulta que el árbol de decisión tiene mejores resultados que la red neuronal, esto debido a que entre menor sea el valor, menor error de pronóstico representa y mejores resultados tiene.

Conclusiones y recomendaciones

1. Se consolidó la información de cantidad de pluviosidad por municipio con las encestas de producción agrícola EVA
2. Se hizo la categorización de los cultivos y municipios identificando las variables de mayor relevancia con sus características para su posterior uso y manipulación.
3. Se desarrollaron métricas que permiten dar una visión general en cuanto a la pérdida de producción al que están enfrentados los municipios ante cambios en la pluviosidad con el fin de dar herramientas para políticas públicas que busquen mitigar dicho riesgo. Además, se hicieron modelos con el Censo Nacional Agrícola (CNA) con el fin de dar más herramientas al experto para la toma de decisiones.
4. Los modelos desarrollados tienen como insumo resultados a triangulaciones y pronósticos de lluvias, por lo cual esta información tiene asociado la suma de error estadístico propios del modelo y de los resultados de los modelos anteriores.

Socialización

El proyecto se socializó y validó con la Dirección de Desarrollo Rural.