Métodos de selección de predictores para la construcción de modelos de riesgo de enfermedad en cultivos a partir de variables climáticas

Suarez Franco[1], Franca Giannini Kurina[2], Cecilia Bruno[1,2], Patricia Rodríguez Pardina[2], María de la Paz Giménez[2], Pablo Gastón Reyna[2], Karina Torrico[2] y Monica Balzarini[1,2]

[1] Estadística y Biometría. Facultad de Ciencias Agropecuarias. Universidad Nacional de Córdoba
[2] Grupo de Estadística Asociado. UFyMA-INTA-CONICET

franmarce.3@gmail.com ; francagianninikurina@gmail.com; cebruno@agro.unc.edu.ar; rodriguez.patricia@inta.gob.ar; gimenez.mariadelapaz@inta.gob.ar; reyna.pablo@inta.gob.ar; torrico.karina@inta.gob.ar; mbalzari@agro.unc.edu.ar

Resumen. La alta dimensionalidad y la correlación entre las múltiples variables candidatas a predictoras para la estimación de un modelo estadístico capaz de predecir la enfermedad de un cultivo en función del ambiente determina la necesidad de recurrir a herramientas metodológicas estadísticas que permitan reducir la dimensionalidad. El objetivo de este trabajo fue comparar el desempeño de métodos de selección de variables en su capacidad para detectar variables climáticas relevantes para la construcción de un modelo logístico que será usado para la predicción de probabilidad de presencia de enfermedad en un patosistema. En este trabajo se compararon tres métodos de selección de variables: Método de Filtrado (F), algoritmo genético (AG) y Boruta (B), en tres patosistemas (MRCV en maíz, Begomovirus en poroto y en soja). Las variables seleccionadas por cada método fueron sometidas a un análisis de componentes principales (ACP) para una nueva reducción de dimensión y obtención de variables sintéticas no correlacionadas. El desempeño de los métodos comparados se evaluó mediante la estimación de la precisión, especificidad y sensibilidad para un modelo lineal predictivo. B y F fueron más eficientes en la predicción. La combinación de estos con el ACP aumentó la eficiencia del modelo de predicción.

Palabras Claves: Boruta, Algoritmo genético, Filtrado, Análisis componentes principales, Patosistema.

1 Introducción

En fitopatología, la predicción de la probabilidad de desarrollo de una enfermedad en un cultivo es característica de cada patosistema y de su relación con el ambiente en el cual se desarrolla el cultivo, como así también de la genética y del manejo del cultivo, entre otros factores. En este trabajo nos interesa modelar las variaciones en el desarrollo de enfermedades de cultivos agrícolas debida a realizaciones aleatorias de las temperaturas, precipitaciones y otras variables meteorológicas en determinados periodos de tiempo previo y durante el ciclo del cultivo. Dada la alta dimensionalidad y la correlación presente entre las múltiples variables candidatas a predictoras para la estimación de un modelo estadístico capaz de predecir la enfermedad en función del ambiente resulta indispensable recurrir a herramientas metodológicas estadísticas que permitan reducir la dimensionalidad seleccionando un subconjunto de variables con capacidad predictiva. El monitoreo de cultivos es una labor intensiva que consume tiempo y a menudo requiere de métodos destructivos. Por ello, contar con modelos predictivos de la probabilidad de desarrollo de una enfermedad según el ambiente en el cual se desarrollará, permite planificar técnicas de manejo, selección de cultivares o líneas tolerantes/resistentes para evitar o disminuir pérdidas por enfermedades [1].

Los modelos de predicción de variables clasificatorias (presencia/ausencia de la enfermedad) que incluyen variables de más resultan en predicción de baja precisión. El sobreajuste disminuye el desempeño de los predictores con “nuevos” datos. El impacto de la no monotonicidad de los modelos clasificatorios en relación al número de predictoras (agregar nueva variable no asegura mejora del desempeño del modelo) así como el sobreajuste son problemas que conducen a la necesidad de seleccionar variables para inducir modelos clasificatorios precisos. En la construcción de modelos estadísticos para respuesta binarias como los modelos logit [2] además de los problemas mencionados anteriormente se suma el impacto negativo que tiene la presencia de correlación entre predictores, fenómeno conocido como multicolinealidad. Independientemente del tipo de modelo que se pretende ajustar en un contexto de alta dimensionalidad el principio de parsimonia siempre es válido. Este principio establece que, en igualdad de condiciones, la explicación más sencilla suele ser la más probable. Aplicado en un contexto predictivo, nos lleva a preferir modelos de menor dimensionalidad cuando no existen diferencias estadísticas en la predicción. La selección de variables, también conocida como selección de características en el entorno del aprendizaje automático [3], es una práctica usada para reducir la dimensionalidad de los datos, eliminar variables irrelevantes (el conocimiento de la misma no aporta a la clasificación) y variables redundantes (su valor puede ser determinado a partir de otras), acortar tiempos de ajuste del modelo y mejorar la capacidad predictiva del modelo o rendimiento del aprendizaje [4].

La forma directa más popular de seleccionar variables es la regresión paso a paso, que es una técnica de “wrapper”, donde se adiciona la mejor variable (o elimina la peor) en cada paso del proceso de selección. Un problema asociado a este método es decidir cuándo parar

el algoritmo. Además, puede volverse computacionalmente intensivo para conjuntos de datos grandes. El método de Filtrado (F) es otra estrategia de selección de variables, que se considera indirecta, donde se utiliza una medida de relevancia para puntuar y ordenar las variables y un umbral para eliminar aquellas de menor relevancia [5]. Esta medida de relevancia no tiene en cuenta el paradigma usado para inducir el modelo clasificatorio de interés y por eso la estrategia se dice indirecta. Para evitar que la selección de variables esté excesivamente influenciada por los datos de entrenamiento (sobreajuste) se repite el proceso varias veces mediante remuestreo y la decisión final se soporta con validación cruzada. Otros algoritmos simplificados, en términos computacionales, permiten una búsqueda secuencial de variables. Entre éstos se destaca el algoritmo genético (AG) [6], un algoritmo de tipo evolutivo cuyo funcionamiento es inspirado por el proceso de selección de genotipos en la naturaleza. Así, el método de selección busca la combinación de predictores que consigue maximizar la capacidad predictiva de un modelo. Puede producir resultados óptimos aunque a veces locales [5]. En el AG, una población de soluciones candidatas (llamadas individuos o fenotipos) a un problema de optimización evoluciona hacia mejores soluciones. Cada solución candidata tiene un conjunto de propiedades (genotipos) que pueden ser mutados y alterados aleatoriamente en cada paso de la evolución hacia la mejor solución. Otro algoritmo secuencial, nacido en el campo del aprendizaje automático, es Boruta (B) [7]. Este método de selección de variables elige un modelo de conveniencia, capaz de capturar relaciones e interacciones no lineales, por ejemplo, un bosque aleatorio o random forest [8]. Luego, lo ajusta en las variables explicativas y variables respuesta. Finalmente, extrae la importancia de cada variable de este modelo y conserva solo aquellas que están por encima de un umbral de importancia determinado. En Boruta, las variables no compiten entre sí sino con una versión aleatoria de las mismas (sombra). En este método se supone que una función es útil solo si es capaz de funcionar mejor que la mejor combinación aleatoria de variables. Por otro lado, los métodos de selección citados pueden no resultar suficientes para reducir la dimensión o seleccionar un conjunto de variables no correlacionadas como haría falta para el ajuste de un modelo lineal. En estos casos, la técnica de reducción conocida como Análisis de Componentes Principales [9] puede aplicarse al conjunto de variables seleccionadas para identificar un subconjunto de nuevas variables no correlacionadas (variables sintéticas) convenientes para el ajuste del modelo.

Estos métodos de selección de variable de alta eficiencia computacional, conocidos en aprendizaje automático como métodos de selección de características, fueron desarrollados suponiendo que los modelos de predicción en los que entran estas variables son generados desde el aprendizaje de máquina. Sin embargo, el desempeño de éstos en la construcción de modelos lineales de base probabilística, como el modelo logístico, es menos conocida. El objetivo de este trabajo fue comparar el desempeño de métodos de selección de variable, optimizados computacionalmente para trabajar en alta dimensionalidad, en su capacidad para detectar variables climáticas que son irrelevantes y/o redundantes para la construcción de un modelo logístico que será usado para la predicción de probabilidad de presencia de enfermedad en un patosistema.

2 Materiales y métodos

2.1 Bases de datos

Se usaron tres bases de datos para ilustrar el comportamiento de los métodos de selección de variables predictoras. Cada base de datos pertenece a diferentes patosistemas como se describe a continuación:

1- *MRCVmaiz*: patosistema asociado al virus del mal de Rio Cuarto (MRCV) en maíz (*Zea mays L.*) con 1296 registros de presencia/ausencia de MRCV correspondientes a lotes agrícolas muestreados durante el período comprendido entre los años 2001 y 2020 en la zona maicera argentina en las provincias de La Pampa, San Luis, Mendoza, Tucumán, Catamarca, Jujuy, Corrientes, Chaco, Misiones, Formosa, Entre Ríos, Santa Fe, Santiago del Estero, Córdoba y Buenos Aires. Se calcularon 65 variables biometeorológicas a partir de la Temperatura máxima promedio, Temperatura media promedio, Temperatura mínima promedio, Temperatura de punto de rocío promedio, Humedad relativa promedio, Humedad relativa mínima, Humedad relativa máxima, Velocidad del viento promedio, Presión atmosférica promedio, Precipitación acumulada, Nubosidad promedio, Nubes bajas promedio y Visibilidad mínima, para una ventana de tiempo mensual en cada campaña agrícola, desde septiembre a enero, que corresponde al periodo de mayor presencia poblacional de vectores alados que transmiten MRCV [10]. De esta manera, esta base de datos contiene una dimensión de 1296 filas (n observaciones) por 67 columnas (p variables) que indican las variables biometeorológicas más la variable referida a la presencia/ausencia de MRCV y la identificación del lote.

2- *BGVporoto:* patosistema asociado a incidencia de Begomo virus en poroto (*Phaseolus vulgaris L.*) con 2026 registros asociados a plantas muestreadas en lotes agrícolas durante las campañas 2001 a 2018 en las provincias de Salta, Tucumán, Jujuy, Santiago del Estero y Córdoba. En cada lote se tomaron entre 3 y 17 muestras de plantas sintomáticas cuya confirmación de la presencia de Begomovirus se realizó por sonda de hibridación [11]. Luego, se calculó la incidencia relativa de Begomivirus como el total de plantas positivas sobre el total de planta muestreadas. Cada uno de los 168 lotes fue clasificado en incidencia alta (incidencia relativa >50%) o incidencia moderada (≤50%). Las variables bioclimáticas fueron calculadas para el periodo de mayor susceptibilidad del cultivo a la infección, por ello la ventana temporal fue en periodos decádicos y abarcó desde 10 días antes de la siembra hasta floración (40 días posterior a la fecha de siembra en periodo estival), incluyendo los meses de invierno previos a la siembra (Junio a Septiembre) por ser el periodo de mayor desarrollo de poblaciones del vector [12]. Las variables biometeorológicas consideradas en este patosistema fueron las mismas que las calculadas en la base *MRCVmaiz,* incluyendo además variables biometeorológicas calculadas mensualmente (Temperatura máxima promedio, Temperatura media, Humedad relativa máxima, Humedad relativa promedio, Precipitaciones acumuladas y Presión atmosférica

promedio), totalizando 51 variables biometeorológicas. Dimensión de la base de datos n=168×p=53.

3- *BGVsoja:* patosistema asociado a incidencia de Begomo virus en Soja (*Glycine max (L) Merr*)), contiene un registro de 1303 plantas muestreadas en 114 lotes agrícolas durante las campañas 2001-2018 que abarcaron las provincias de Tucumán, Catamarca, Salta, Santiago del Estero, Jujuy, Córdoba, Buenos Aires, Santa Fe, Chaco y Entre Ríos. Con la confirmación por sonda de hibridación [11] de presencia de Begomovirus en cada planta muestreada, se estimó la incidencia relativa de Begomivirus para cada uno de los 114 lotes que fueron clasificados en incidencia moderada si menos del 50% de las plantas eran positivas y en incidencia alta con más del 50% de las plantas positivas en el lote. Las variables biometeorológicas fueron las misma que las usadas en la base *MRCVmaiz,* pero calculadas en periodos decádicos desde junio hasta marzo (periodo inverno-estival) que incluye tanto el periodo de mayor desarrollo poblacional del vector como el de mayor susceptibilidad del cultivo. Adicionalmente, se incluyeron variables biometeorológicas calculadas cada 20 días para el periodo de septiembre a noviembre (Temperatura máxima promedio, Temperatura media promedio, Humedad relativa máxima, Humedad relativa promedio y Precipitaciones acumuladas) totalizando 382 variables biometeorológicas. Dimensión de la base de datos n=114×p=385.

2.2 Métodos de selección de variables comparados

Filtrado

Se aplicó un ANOVA (Balzarini et al., 2015) a cada variable para identificar aquellas que varían dependiendo de la variable respuesta (clasificada en incidencia ≤50% y >50% o como presencia/ausencia de enfermedad). Luego, se incorporaron en un algoritmo predictivo de bosque aleatorio aquellos predictores con un valor-p inferior a 0.05. Se implementó usando el paquete *caret* [14], mediante la función *sbf*() que devuelve un vector lógico con la selección de las variables, basado el modelo predefinido *random forest* (rfSBF) del software R [15] . Para cuantificar la relación entre las variables predictoras y la variable respuesta la función *sbf* emplea modelos de ANOVA y GAMS [16][17] con un nivel de significancia de 0.05. El método de entrenamiento usado fue validación cruzada repetida k fold (k=10) con 5 repeticiones.

Algoritmo Genético

Se implementó mediante la función *gafs* del paquete *caret* de R configurando 10 generaciones, un tamaño poblacional de 10 individuos, un modelo *random forest* como método de evaluación y validación cruzada con 5 particiones.

Boruta

Se implementó usando el paquete **boruta** de R [18]. Se utilizó un valor de *maxRuns* igual a 3000, el valor-p fue 0.01. En cuanto al parámetro *getImp* el valor predeterminado es *getImpRfZ* que ejecuta un *random forest* desde el paquete *ranger* y reúne puntuaciones Z del cambio (disminución) de medidas de precisión predictiva.

2.3 Análisis de Componentes Principales

Sobre el conjunto de variables biometeorológicas seleccionadas por cada método comparado en este trabajo, se realizó un análisis de componentes principales (ACP) mediante el paquete *FactoMineR* (Lê et al., 2008) de R para obtener un conjunto de variables ortogonales (no correlacionadas) que serán incluidas en las regresiones logísticas evitando la multicolinealidad. Las variables fueron previamente estandarizadas. El criterio de selección del número de componentes principales (CP) a retener en cada base de datos fue aquel cuyo autovalor (λ) fuera igual o mayor a 1 y que la varianza acumulada explicada por el conjunto de CP retenidas fuera como mínimo el 70 % de la varianza total.

2.4 Ajuste de modelo predictivo

Se ajustaron modelos de regresión logística para la variable respuesta (y) dicotómica que indica alta o moderada incidencia relativa de Begomovirus en el sitio *i*. Se ajustaron en total 18 modelos de regresión logística, uno por cada patosistema y conjunto de variables regresoras seleccionadas por cada método de selección (F, AG, B, F+ACP, AG+ACP, B+ACP). El modelo de regresión logística permite predecir la probabilidad (p_i) de incidencia relativa alta en el sitio *i* de muestreo de cada patosistema. El modelo de regresión logística múltiple donde el vector X_i contiene las variables climáticas explicativas se expresa como:

$$Logit(p_i\ ) = \log\left(\frac{p_i}{1-p_i}\right) = X_i\ \beta$$

donde p_i es la probabilidad de incidencia alta dado las variables regresoras en X_i y es un vector de parámetros o coeficientes de regresión lineal que indican el efecto de cada variable biometeorológicas sobre la probabilidad de alta incidencia o presencia de la enfermedad. Todos los modelos fueron evaluados en su capacidad predictiva mediante validación cruzada. En esta etapa se usó la función *glm* del paquete *stats* de R.

2.5 Criterios de comparación

Se construyeron matrices de confusión a partir de las cuales se calcularon medidas de precisión de la predicción, de sensibilidad y de especificidad. La precisión describe el porcentaje de presencia/alta incidencia que fue correctamente clasificado del total de observaciones. La sensibilidad estima el porcentaje de valores positivos (presencia/alta incidencia) que son clasificados como positivos y la especificidad es el porcentaje de valores negativos (ausencia/baja incidencia) que son clasificados como negativos. Estas medidas fueron estimadas mediante la función *confusionMatrix* del paquete *caret*, generada con una muestra aleatoria del 20% de los casos de cada base. El procedimiento se repitió 30 veces y se obtuvieron los promedios de las medidas de capacidad predictiva del modelo. Las medidas de precisión asociada a los distintos métodos de selección de variables se compararon con un modelo lineal mixto de clasificación contemplando el patosistema como efecto aleatorio y controlando la heteroscedasticidad con la función *varIden* en el procedimiento *lme* de R.

3 Resultados y Discusión

Para la base de MRCVmaiz utilizando el método F se identificaron 55 variables como óptimas, con el AG 33 y B 65 variables. Para la base de BGVporoto el método F seleccionó 14 variables, AG 36 y B 19. Mientras que, para BGVsoja las variables seleccionadas fueron 101, 195 y 20, respectivamente (Fig. 1). En el caso particular del cultivo de maíz, el conjunto de variables seleccionadas por los tres métodos corresponde principalmente a la Temperatura media promedio, la Humedad Relativa promedio y la Velocidad del viento del mes de noviembre, donde se produce el momento de máxima población de vectores alados que buscan alimento y encuentran las plantas de maíz emergente o en un estado vegetativo según sea un maíz tardío (zona centro-norte del país) o temprano. Mientras que, para los cultivos de poroto y soja, las variables seleccionadas estuvieron relacionadas con las Temperaturas medias de los meses de invierno, la Nubosidad, la Humedad relativa y la Velocidad del viento. La selección de estas variables biometeorológicas concuerdan con lo expuesto por Morales y Jones (2004) donde indican que, si los inviernos anteriores a la siembra son cálidos, es decir presentan temperaturas entre 15 y 33°C, la población del vector transmisor (mosca blanca) no se ve afectada y comienza a aumentar la cantidad de individuos conforme aumenta la temperatura con el correr de los meses del año.

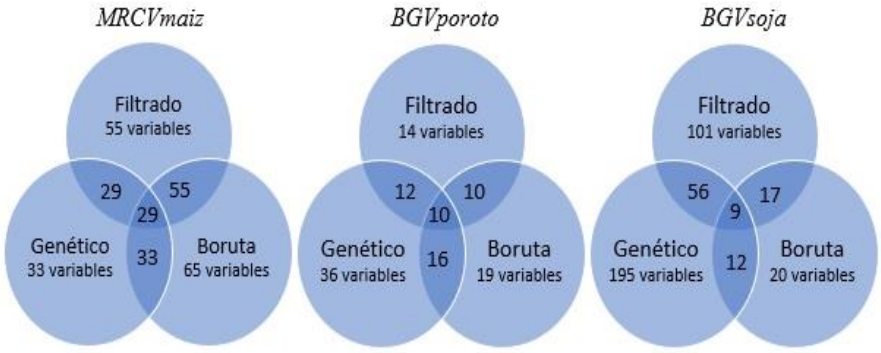

Fig. 1. Diagrama de Venn para representar la concordancia en las variables seleccionadas por tres métodos de selección de variables (filtrado ANOVA-random forest, algoritmo genético y boruta) aplicados en tres bases de datos: MRCVmaiz con 65 variables y 1296 observaciones, BGVporoto con 50 variables y 168 observaciones y BGVsoja con 382 variables y 114 observaciones.

Las medidas de la capacidad predictiva de cada modelo se muestran en la Tabla 1. Para los valores de precisión, el modelo predictivo construido a partir de variables seleccionadas mediante B no presentó diferencias estadísticamente significativas con el construido mediante el método F, presentando valores de 66.9 y 66.0 % respectivamente, mientras que el método AG obtuvo el menor valor precisión (64.2%) con diferencias estadísticamente significativa respecto a los generados con los otros dos métodos. La especificidad asociada al modelo logístico mantuvo las mismas relaciones entre los métodos y para la sensibilidad del modelo predictivo B no se diferenció con el método F. (Tabla 1).

Tabla 1. Medias (± error estándar) de precisión, especificidad y sensibilidad del modelo predictivo construido a partir de tres métodos de selección de variables.

Método	Precisión		Especificidad		Sensibilidad		
Boruta	66.9 (± 4.94)	A	0.73 (± 0.04)	A	0.59 (± 0.05)	A	
Filtrado	66.0 (± 4.94)	A	0.72 (± 0.04)	A	0.58 (± 0.05)	A	B
Genético	64.2 (± 4.94)	B	0.70 (± 0.04)	B	0.56 (± 0.05)		B

Medias con letras distintas indican diferencias estadísticamente significativas ($p \leq 0.05$)

El número de componentes principales con los cuales se alcanzó como mínimo el 70% de la variabilidad explicada se puede observar en la Fig. 2, para cada método (Fig. 2 – columna) en cada base de datos (Fig. 2 - filas)

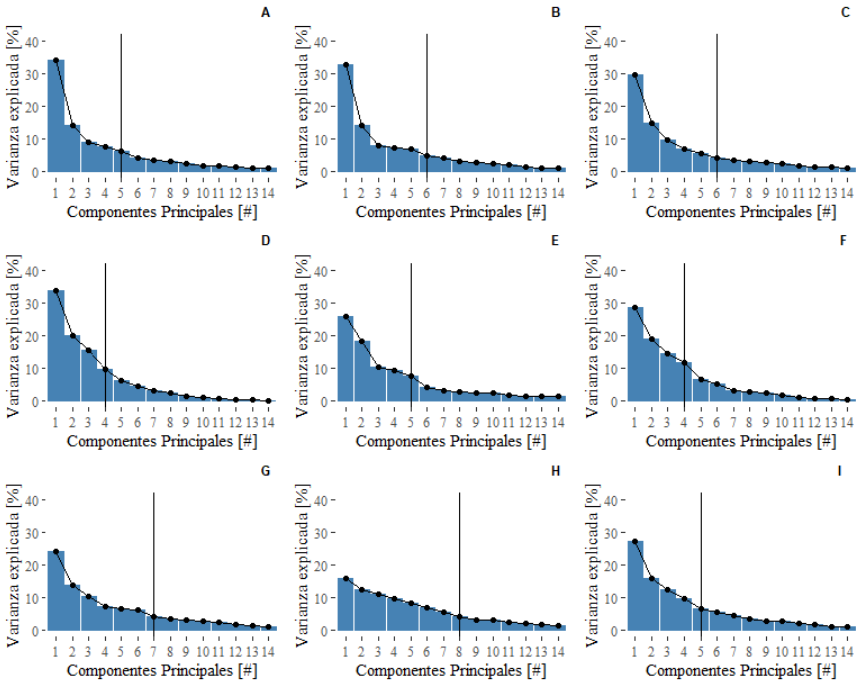

Fig. 2. Varianza explicada, en porcentaje, por las componentes principales obtenidas de un Análisis de componentes principales realizado con las variables seleccionadas por cada método (columna) en cada base de datos (fila). La línea vertical indica el número de componentes principales donde se acumula como mínimo el 70% de la variabilidad total explicada.
A: MRCVmaiz – Variables obtenidas por F. B: MRCVmaiz – Variables obtenidas por AG. C: MRCVmaiz -Variables obtenidas por B. D: BGVporoto – Variables obtenidas por F. E: BGVporoto - Variables obtenidas por AG. F: BGVporoto - Variables obtenidas por B. G: BGVsoja - Variables obtenidas por F. H: BGVsoja - Variables obtenidas por AG. I: BGVsoja - Variables obtenidas por B.

En la Tabla 2 se presentan las medidas de precisión, especificidad y sensibilidad obtenidos por los modelos de regresión logísticos utilizando las componentes principales como variables regresoras. El modelo predictivo ajustado a partir de las variables seleccionadas por el método F+ACP no presentó diferencias estadísticamente significativas con el modelo predictivo B+ACP presentando una media de 70.8% y 70.4%, respectivamente. El modelo de las variables seleccionadas por el método AG+ACP presentó la menor precisión (69.1%) siendo este valor estadísticamente significativo respecto a

F+ACP y B+ACP. Estas relaciones se mantuvieron igual en cuanto a la especificidad y sensibilidad obtenidas por los modelos.

Tabla 2. Medias (± error estándar) de precisión, especificidad y sensibilidad del modelo predictivo construido a partir de las componentes principales que como mínimo alcanzan el 70% de la variabilidad explicada, obtenidas a partir de las variables seleccionadas por tres métodos de selección de variables.

Método	Precisión		Especificidad		Sensibilidad	
Filtrado+ACP	70.80 (±2.42)	A	0.74 (±0.02)	A	0.66 (±0.04)	A
Boruta+ACP	70.40 (±2.42)	A	0.73 (±0.02)	A	0.66 (±0.04)	A
Genético+ACP	69.15 (±2.42)	B	0.72 (±0.02)	B	0.64 (±0.04)	B

Medias con letras distintas indican diferencias estadísticamente significativas ($p \leq 0.05$).

Comparando los resultados obtenidos por las regresiones utilizando las variables seleccionadas por los diferentes métodos y los obtenidos por las regresiones que utilizan las componentes principales obtenidas de las variables seleccionadas vemos que estas últimas son mejores en el caso de precisión y sensibilidad, no se notó un cambio en cuanto a los valores de la especificidad.

4 Conclusión

En contextos de alta dimensión de variables regresoras que además presentan una alta correlación entre ellas, la selección de variables previo al ajuste de un modelo logístico predictivo mejora su desempeño. La eficiencia de los métodos de selección de variables F y B superó el rendimiento del método AG.

5 Bibliografía

[1] M.M. Raza, C. Harding, M. Liebman, L.F. Leandro, Exploring the Potential of High-Resolution Satellite Imagery for the Detection of Soybean Sudden Death Syndrome, Remote Sens. 2020, Vol. 12, Page 1213. 12 (2020) 1213. https://doi.org/10.3390/RS12071213.

[2] W.W. Stroup, Generalized linear mixed models: modern concepts, methods and applications, CRC press, 2012.

[3] Y. Takahashi, M. Ueki, M. Yamada, G. Tamiya, I.N. Motoike, D. Saigusa, M. Sakurai, F. Nagami, S. Ogishima, S. Koshiba, K. Kinoshita, M. Yamamoto, H. Tomita, Improved metabolomic data-based prediction of depressive symptoms using nonlinear machine learning with feature selection, Transl. Psychiatry. 10 (2020) 1–12. https://doi.org/10.1038/s41398-020-0831-9.

[4] R. Sheikhpour, M.A. Sarram, S. Gharaghani, M.A.Z. Chahooki, A Survey on semi-supervised feature selection methods, Pattern Recognit. 64 (2017) 141–158. https://doi.org/10.1016/j.patcog.2016.11.003.
[5] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (2014) 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.
[6] M. Mitchell, Genetic algorithms: An overview., Complexity. 1 (1995) 31–39.
[7] H. Gholami, A. Mohammadifar, S. Golzari, D.G. Kaskaoutis, A.L. Collins, Using the Boruta algorithm and deep learning models for mapping land susceptibility to atmospheric dust emissions in Iran, Aeolian Res. 50 (2021) 100682. https://doi.org/10.1016/j.aeolia.2021.100682.
[8] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32. https://doi.org/10.1023/A:1010933404324.
[9] M. Balzarini, C. Bruno, M. Córdoba, I. Teich, Herramientas en el análisis estadístico multivariado, Córdoba, Argentina. (2015).
[10] J. Ornaghi, G. Boito, G. Sánchez, G. March, J. Beviacqua, Studies on the populations of Delphacodes kuscheli Fennah in different years and agricultural areas, J. Genet. Breed. 47 (1993) 277–281.
[11] P.E. Rodríguez-Pardina, K. Hanada, I.G. Laguna, F.M. Zerbini, D.A. Ducasse, Molecular characterisation and relative incidence of bean- and soybean-infecting begomoviruses in northwestern Argentina, Ann. Appl. Biol. 158 (2011) 69–78. https://doi.org/10.1111/J.1744-7348.2010.00441.X.
[12] F.J. Morales, P.G. Jones, The ecology and epidemiology of whitefly-transmitted viruses in Latin America, Virus Res. 100 (2004) 57–65. https://doi.org/10.1016/J.VIRUSRES.2003.12.014.
[13] M. Balzarini, J. Di Rienzo, M. Tablada, L. Gonzalez, C. Bruno, M. Córdoba, F. Casanoves, Estadística y biometría: Ilustraciones del uso de Infostat en problemas de agronomía [Statistics and biometrics: Illustrations of the use of Infostat in agronomic problems], Córdoba, Argentina Editor. Brujas. (2015).
[14] M. Kuhn, The caret package.–R Foundation for Statistical Computing, Vienna, Austria, URL Https//Cran. r-Project. Org/Package= Caret. (2012).
[15] R.C. Team, R: A language and environment for statistical computing, (2020).
[16] B. Efron, T. Hastie, Computer age statistical inference: Algorithms, evidence, and data science, 2016. https://doi.org/10.1017/CBO9781316576533.
[17] T.J. Hastie, R.J. Tibshirani, Modelos Aditivos Generalizados. Chapman e Hall, (1990).
[18] M.B. Kursa, W.R. Rudnicki, Feature Selection with the Boruta Package, 2010. http://www.jstatsoft.org/ (accessed May 24, 2021).