Informe final publicable de proyecto
Integración de datos genómicos y ambientales mediante aprendizaje profundo para selección genómica

Código de proyecto ANII: IA_1_2022_1_173411

Fecha de cierre de proyecto: 01/04/2025

FARIELLO RICO, Maria Ines (Responsable Técnico - Científico)
CROSSA, Jose (Co-Responsable Técnico-Científico)
LECUMBERRY RUVERTONI, Federico (Co-Responsable Técnico-Científico)
MONTESINOS LÓPEZ, Abelardo (Co-Responsable Técnico-Científico)
MONTESINOS LÓPEZ, Osval Antonio (Co-Responsable Técnico-Científico)
BELZARENA, Diego (Investigador)
CASTRO OLMEDO, GRACIANA MARÍA (Investigador)
FERNÁNDEZ MUÑIZ, María Ximena (Investigador)
LONG GROSSO, Micaela (Investigador)
MUSITELLI, Mateo (Investigador)
PARDO PICCONE, Alvaro Daniel (Investigador)
ROSAS CAISSIOLS, Juan Eduardo (Investigador)

UNIVERSIDAD DE LA REPÚBLICA. FACULTAD DE INGENIERÍA (Institución Proponente) \\
INSTITUTO NACIONAL DE INVESTIGACIÓN AGROPECUARIA. INIA TREINTA Y TRES \\ COLEGIO DE POSTGRADUADOS \\
UNIVERSIDAD DE LA REPÚBLICA. FACULTAD DE INGENIERÍA \\ UNIVERSIDAD DE COLIMA \\ UNIVERSIDAD DE GUADALAJARA \\
UNIVERSIDAD CATÓLICA DEL URUGUAY DÁMASO ANTONIO LARRAÑAGA. DEPARTAMENTO DE INGENIERÍA

Resumen del proyecto

La Selección Genómica (SG) es un método para predecir las características de interés de un cultivo o un animal de producción a partir de su genotipo (el ADN) y otras fuentes de información, como, tipo de suelo, lluvia y temperatura, tipo de alimentación, etc. Estas características, conocidas como fenotipos, pueden ser la cantidad de leche que produce una vaca, el peso o calidad de la carne de una ternera, la resistencia a plagas o el peso de producción de un cultivo, entre otros. El procedimiento para hacerlo es, primero recolectar una muestra de referencia que contenga datos genotípicos, fenotípicos y ambientales y entrenar un modelo que a partir de los datos genotípicos y ambientales, prediga los fenotipos de manera satisfactoria. Posteriormente con este modelo entrenado se hacen predicciones para genotipos candidatos para los cuales únicamente se cuenta con información genotípica y ambiental. Esta metodología ha revolucionado el mejoramiento genético ya que incrementa la ganancia genética por unidad de tiempo y ahorra recursos significativos en el fenotipado. Sin embargo, su implementación práctica es todavía compleja ya que requiere alta precisión en las predicciones para que su implementación sea exitosa. Se han explorado varios algoritmos de Aprendizaje Automático (AA) para mejorar estas capacidades predictivas, sin embargo, los resultados obtenidos no son aún suficientes para su implementación exitosa en forma rutinaria, sobre todo en cultivos de granos.
Exploramos métodos de Aprendizaje Profundo, en particular el uso de Transformers, para evaluar la factibilidad de incrementar la capacidad predictiva. Se buscará que esta metodología se pueda implementar en forma rutinaria en programas de mejoramiento genético en la región, con lo cual se pueda coadyuvar a incrementar la ganancia genética de las especies productivas de la región.
Durante este proyectose consolidó la colaboración entre grupos de investigación uruguayos y mexicanos, antecedentes en AA y SG. Se desarrollaron actividades de formación en base a seminarios interdisciplinarios, cursos, posgrados y visitas de profesores.

Ciencias Naturales y Exactas / Matemáticas / Matemática Aplicada / Aprendizaje automático y predicción
Palabras clave: Aprendizaje Profundo / Selección Genómica / Transformers /

Antecedentes, problema de investigación, objetivos y justificación.

La metodología de Selección Genómica (SG) fue propuesta hace más de veinte años por Meuwissen (Meuwissen, 2001). El objetivo de ésta es lograr la mejora de características de interés productivo (fenotipos) en cultivos o ganadería sin necesidad de tener que medirlas directamente, es decir, seleccionar a los mejores candidatos (plantas o animales) sin la necesidad de medir esas características en campo. Para ello, se recolecta una muestra de referencia que contenga datos genotípicos y fenotípicos, se entrena un modelo que a partir de los datos genotípicos que prediga los fenotipos de manera satisfactoria y luego con el modelo entrenado se hacen predicciones para genotipos candidatos para los cuales únicamente se cuenta con información genotípica, eligiendo de ésta manera a la próxima generación a cultivar.
Originalmente los modelos utilizados para esta labor de predicción fueron modelos convencionales de estadística (regresión lineal penalizada y modelos mixtos) y posteriormente versiones Bayesianas de los modelos mixtos (BayesA, BayesB, BayesC, Bayes-Lasso (Gianola, 2013)). En general los modelos Bayesianos y modelos mixtos son los más usados en predicción genómica, ya que por un lado no requieren búsquedas intensivas de parámetros y por otro producen predicciones muy competitivas, que no son fáciles superar con otros modelos de AA y que requieren de un proceso de ajuste de parámetros más cuidadoso y mayor capacidad computacional.
La SG ha revolucionando el plant breeding ya que por ser una metodología predictiva ahorra recursos significativos en el fenotipado y contribuye a incrementar la ganancia genética por unidad de tiempo, aunque su éxito está directamente relacionado a la calidad de las predicciones (Crossa et al., 2017; de los Campos et al., 2009; Gianola et al., 2016). Sin embargo, los métodos antes mencionados tienen la restricción de que no capturan eficientemente patrones no lineales ni interacciones complejas presentes en los datos, dejando un espacio para explorar otras metodologías predictivas que logren robustecer la predicción genómica (PG) y de esta forma poder llevarlas a la práctica del metodología de SG de una forma aún más exitosa y rutinaria. Es en este marco donde los métodos de Aprendizaje Profundo (AP) que han mostrado un gran éxito en áreas tan diversas como el comercio electrónico, procesamientos de imágenes, procesamiento del lenguaje natural, desarrollo de videojuegos, clasificación e identificación de objetos y personas,

tienen un gran potencial para mejorar significativamente las predicciones, ayudando a incrementar aún más, la ganancia genética de las especies de producción de la región. En particular, se sabe que no toda la variación de los fenotipos es genética, sino que se debe a factores ambientales también. En este sentido, el uso de redes neuronales permite integrar fácilmente diferentes tipos de datos, ya que no se requiere a priori un tipo determinado de datos, ni establecer las interacciones entre los genotipos y datos provenientes de otras fuentes, como pueden ser mediciones ambientales o de imágenes (Måløy et al., 2021).

Diferentes arquitecturas de redes neuronales, han contribuido a resolver diferentes tipos de problemas. En particular los modelos de encoder-decoder, es decir, que codifican una señal y luego el mismo modelo la decodifica, han centrado el interés en problemas de predicción entre secuencias de datos o series temporales (sequence-to-sequence). Redes recurrentes (RNN, LSTM) han mostrado un gran desempeño pero presentan problemas en secuencias largas. Para atacar este problema de los encoder-decoder, se ha introducido el uso de los mecanismos de atención temporal/secuencial para identificar información relevante en partes distantes de la secuencia. Los modelos de transformers (Vaswani et al., 2017), conocidos por su gran potencial en modelos que realizan tareas en lenguaje natural, combinan una arquitectura de encoder-decoder con un mecanismo de atención para identificar dependencias en lugares lejanos de la secuencias (Abdollahi-Arpanahi et al., 2020; Ji et al., 2021; Jubair et al., 2021; Måløy et al., 2021; Martinek et al., 2022).

El mejoramiento genético en la producción agropecuaria ha logrado adaptar especies de producción del mundo animal y vegetal a los requerimientos del ambiente y de sus productores, mejorando fenotipos más amigables con el ambiente, rendimientos o características deseadas por los consumidores de los productos, redundando en una mejor producción y más redituable (Astori & Alonso, 1979; Dellazoppa, 2014).

Si bien los métodos de selección clásicos, donde el productor elige a los mejores reproductores, según la característica que desee, siguen siendo utilizados, es con la llegada de la genómica donde se han logrado los mejores saltos en productividad. Esto se debe a que seleccionar a partir de los datos genómicos de los individuos permite seleccionarlos al principio de sus ciclos reproductivos sin tener que esperar las mediciones en campo, que permitan conocer el desempeño del individuo para luego tomar la decisión de si retenerlo o no. Esta metodología toma especial relevancia cuando las características son difíciles o muy costosas de medir. Por lo tanto el impacto esperado de la metodología de SG es mayor en características genéticamente complejas y que además son de difícil y costosa medición, cuyo progreso genético se incrementa por una mayor precisión de estimación del mérito genético (Crossa et al., 2010, 2011, 2017; de los Campos et al., 2009; Gianola et al., 2016; Meuwissen et al., 2001).

Por otro lado, si bien no siempre es claro cómo interactúa el genotipo con el ambiente, sabemos que éste tiene una gran importancia y que un mismo híbrido (variante de un cultivo) puede adaptarse mucho mejor a determinadas condiciones climáticas que a otros, mientras que diferentes híbridos pueden rendir de manera diferente en un mismo ambiente . Por lo tanto, es necesario realizar predicción genómica de cultivos que se han observado en ciertos ambientes y así estimar el efecto del fenómeno de la interacción genómica-ambiente (Jarquín et al., 2014). Estas predicciones son complejas dado la incertidumbre que produce el notable cambio climático; por esto el uso de datos climáticos históricos así como de imágenes colectadas en diferentes épocas del cultivo son datos esenciales para aumentar la precisión de la predicción genómica de nuevos cultivos en futuros ambientes (sitios, años) (Crossa et al., 2021).

Los primeros datos de mejoramiento de plantas y animales han demostrado que la SG aumenta significativamente la precisión de predicción, respecto a la selección basada en pedigrí para rasgos complejos (Bernardo & Yu, 2007; Crossa et al., 2010, 2011, 2013, 2014; de los Campos et al., 2009; De los Campos et al., 2010; de Los Campos et al., 2013; Heslot et al., 2012, 2014; Hickey et al., 2012; Lorenzana & Bernardo, 2009; Meuwissen et al., 2001) Desde entonces, los programas de mejoramiento de cultivos en todo el mundo han estado estudiando y aplicando SG y, simultáneamente, se han realizado amplios estudios de investigación sobre nuevos modelos estadísticos para incorporar covariables, pedigrí, variables genómicas y ambientales, como características del suelo o datos meteorológicos, entre otros.

El problema de integrar los datos genómicos con otro tipo de datos como aquellos de clima, imágenes satelitales, drones y aviones representa un reto científico de alta complejidad (genético, matemático, estadístico, y computacional), pero sumamente importante para modelar la interacción genotipo-ambiente y aumentar la precisión de la predicción genómica. Por esta razón, las redes neuronales profundas se presentan como una gran oportunidad para el estudio de este tipo de problemas, ya que no es necesario determinar a priori las relaciones entre todas las fuentes de datos. Las diversas arquitecturas que se han propuesto los últimos años, permiten incorporar además distintos aspectos de los modelos de predicción. Por ejemplo, los transformers (Vaswani et al., 2017), permiten tener en cuenta la dependencia

local que induce el desequilibrio de ligamiento en los genomas, o la dependencia local que pueden tener los datos del ambiente, como por ejemplo las series de tiempo descriptoras del clima (Shook et al., 2021).
A diferencia de las CNN (redes neuronales convolucionales, muy utilizadas en clasificación de imágenes), en los que se trata a toda la secuencia genómica o la serie temporal de la misma manera, tomando ventanas fijas, los "self-attention maps", centrales en los Transformers, permiten aprender el grado de dependencia de los datos entre sí directamente de éstos. Según el largo de memoria que se permita, la red aprenderá cómo dependen las variables entre sí, en diferentes puntos del genoma.
A lo largo de la propuesta hemos mencionado la ventaja del uso de las redes neuronales, ya que no se necesita modelar a priori las interacciones entre las variables, ya sea en los genotipos debido al desequilibrio de ligamiento, temporal en los datos de clima o imágenes aéreas o entre ciertas regiones de los genotipos y las otras variables medidas. La elección de la arquitectura de la red y algunos parámetros, como por ejemplo la dimensión de los espacios latentes puede afectar el resultado, tanto en tiempo de cálculo, cantidad de parámetros o precisión en las mediciones. En esta toma de decisiones es crucial la interacción entre todos los investigadores, ya que no se puede hacer una búsqueda de arquitecturas, como usualmente se hace con los parámetros, sino que deben de elegirse algunas, entender cómo están funcionando, qué es lo que están captando de los datos y luego modificarlas, esperando que mejore la precisión de las estimaciones. Para esto es necesario, ir hacia atrás, ver en qué individuos se equivocan más, cómo se relacionan éstos con el resto, entender si las variables de clima pueden ser responsables o no, y luego modificar las arquitecturas, para intentar capturar lo que entendamos que no se está capturando. Para esto el buen acceso a los equipos de cómputo y la disponibilidad de recursos humanos que dediquen mucho tiempo al proyecto, serán de especial relevancia.
Los métodos desarrollados buscan contribuir al incremento de la precisio?n en la estimación de diferentes medidas de interés en cultivos integrando datos ambientales y de imágenes con información genómica completa mediante métodos de Aprendizaje Automático. Hemos profundizado en el estudio de los fundamentos y aplicaciones de métodos modernos de Aprendizaje Profundo, explorando diferentes arquitecturas de redes neuronales que puedan tener en cuenta la estructura de los datos. El incremento de la precisión en la Predicción Genómica podrá utilizarse en programas de mejoramiento genético de cultivos como, trigo, maíz y arroz. En un plazo más largo, los resultados de este proyecto podrán contribuir a alcanzar líneas de producción que sean acordes a las necesidades de los productores y consumidores de Uruguay y México, ya que dos de las medidas a predecir refieren al rendimiento y calidad de los granos.

Metodología/Diseño del estudio

El proyecto propuesto presentó un gran desafío desde el punto de vista de los datos, ya que al incluir la variable ambiental, no se agrega sólo un nuevo conjunto de variables, sino que se debe comprender cómo se hicieron los ensayos, en dónde se plantó cada cultivo, cuántas veces se fue repitiendo y cómo fue variando a lo largo de los años qué cultivos se plantaron y en donde. Este desafío llevó por un lado, un par de meses de analizar y comprender la base de datos, para luego establecer el diseño experimental para realizar la validación cruzada de los métodos que se propusieron. Finalmente se llegó a tres diseños diferentes, teniendo en cuenta híbridos y locaciones de la base de datos de maiz. En el primer caso se decidió realizar la validación en híbridos neuvos, es decir, se conocen datos de híbridos en todos los ambientes y lo que se intenta predecir es cómo le irá a nuevos híbridos en ambientes ya conocidos. Este caso sería el típico en mejoramiento genético, donde lo que se busca es elegir nuevos híbridos. En segundo lugar se estableció un esquema de validación, donde se conocían todos los híbridos, pero no se incluían algunos ambientes. El objetivo de este diseño es decidir de un conjunto de híbridos conocidos, cuáles serán los más adecuados para un nuevo ambiente. En este caso, no se trataría de mejora genética en sí, sino de buscar los híbridos que mejor se adaptarán a una nueva locación. Por útlimo, para evaluar la ganancia de incluir datos ambintales, se realizcó un esquema donde se elegía un ambiente y se predecía dentro de ese mismo y luego se incluían los otros ambientes en los que los híbridos fueron también plantados y se volvía predecir. En este caso se buscó cuantificar cuánto aporta, conocer el rendimiento del maíz en otros ambientes, para mejorar la predicción en un ambiente.
Luego partimos de los resultados y desarrollos realizados por cada uno de los grupos: en México, Universidad de Colima, Universidad de Guadalajara, y Colegio de Postgraduados y en Uruguay, INIA, Facultad de Agronomía y Facultad de Ingeniería de la Universidad de la República Oriental del Uruguay y la Universidad Católica. Realizamos una revisión del estado del arte de las diferentes áreas, incluyendo una revisio?n de las te?cnicas de fusión de datos ambientales con información genómica. Se aplicaron te?cnicas modernas de ana?lisis de datos en altas dimensiones como son

Aprendizaje Profundo (AP), en particular redes neuronales convolucionales (CNN) para analizar las bases de datos fenoti?picas, geno?micas y ambientales.
Dada la gran dimendionalidad de los datos genómicos y su peso relativo respectoa a las variables ambientales, se exploraron diversos métodos de reducción de dimensionalidad que luego se combinaron con los métodos propuestos, tanto lineales como de aprendizaje automático.
En una úlitma instancia se evaluó el desempeño de los modelos en los esquemas de validación propuestos en una primera instancia e incorporando las reducciones de dimensionalidad incoporadas.
Respecto a la metodología de trabajo para fomentar la cooperación entre México y Uruugay, se realizaron instancias de discusión remota, se dictaron dos cursos. En el primero parte del equipo de Uruguay viajó a Méjico y en el segundo Osval Montesionos visitó Uruguay. José Crossa logró visitar Uruguay en una tercera instancia donde además participó del tribunal de defensa de grado de Abate, Ledesma y Sarachu.
Como estaba previsto la formación de recursos humanos se dio a través un proyecto de grado y proyectos de maestría que se realizaron durante el proyecto o comenzaron y continuarán en el próximo.

Resultados, análisis y discusión

Este proyecto resulta como continuación del trabajo realizado en el proyecto también financiado por ANII, FSDA 1_2018_1_154364: Predicción genómica con técnicas de aprendizaje profundo, profundizando en algunas arquitecturas de redes neuronales propuestas por dicho proyecto, y agregando además datos de ambientes, como ser clima y características del suelo. La colaboraciones logradas durante este proyecto se plasman en el proyecto FMV_3_2024_1_1805: Integración de datos ambientales y genómicos para la predicción genómica en cultivos mediante redes neuronales con atención cruzada, donde se seguirán optimizando las arquitecturas de las redes neuronales para explotar aún mejor la información y se integrarán al programa de mejoramiento de trigo nacional. También han dado lugar a publicaciones conjuntas en fundamentos del análisis metodológico (A. Montesinos-López et al., 2024; O. A. Montesinos-López et al., 2025).
La combinación de modelos de aprendizaje profundo con procesamiento no lineal, junto con la inclusión de datos ambientales, mejora significativamente la precisión en la predicción del rendimiento de maíz, en conjuntos de datos que contienen múltiples ambientes. La variabilidad proveniente de los diferentes ambientes, así como el uso de métodos adecuados de preprocesamiento que permitan reducir la dimensionalidad de los datos genómicos, son estrategias efectivas para capturar la complejidad de los rasgos agrícolas y las interacciones genotipo-ambiente.
Reducción de dimensionalidad:
Los métodos de reducción de dimensionalidad buscan reducir la dimensión de datos, tratando de perder la menor cantidad de información posible. En el caso de los datos genómicos, cuya dimensión es muy alta, aplicar este tipo de métodos presenta las siguientes ventajas: los métodos a aplicar requieren un ajuste de menos cantidad de parámetros, resultando en menor demanda computacional y menor probabilidad de sobreajuste, la señal puede quedar concentrada en menos variables y, al reducir la dimensión de los datos genómicos, se puede equiparar la dimensión de las distintas fuentes de información, logrando que los métodos capturen señal de todas las fuentes. En el proyecto se realizaron varias pruebas, en particular se usaron diferentes cantidades de vectores luego de realizar un análisis de componentes principales (PCA) que es una reducción de dimensionalidad lineal, y por otro lado t-SNE (Maaten & Hinton, 2008) y autoencoders (AE)(Battey et al., 2021) que realizan operaciones no lineales. En general los modelos basados en autoencoders arrojaron mejores resultados para todos los métodos (regresiones lineales, árboles y redes neuronales) que los basados en PCA. Esto nos permite conjeturar, que los métodos de reducción que captan relaciones no lineales entre SNPs resultan los más adecuados, por más que los métodos que se apliquen luego sean lineales.
Si bien en general para PCA, se eligió quedarse con el 95% de la varianza, se observó que para algunos métodos,por ejemplo gradient boosting con regularización (XGBoost), pruebas aisladas demostraron que reduciendo aún más la dimensionalidad, se logran resultados mejores.
A su vez, se comparó experimentalmente el desempeño de AEs convencionales contra variacionales (Kingma & Welling, 2013) en la tarea de reducción de dimensionalidad. La implementación de las VAEs se basó en (Battey et al., 2021), un trabajo que explora la reducción de dimensionalidad mediante VAEs para datos genómicos poblacionales. Este código, basado en python y tensorflow, se adaptó para su utilización en clusteruy de manera de tener la posibilidad de correr experimentos a gran escala (a parte de ligeros cambios debido a actualizaciones de librerías). Para ambos métodos se realizó una búsqueda de hiperparámetros exhaustiva dónde se propuso minimizar el error de reconstrucción en un

conjunto de validación (en el caso de las VAEs añadiendo la penalización de la divergencia KL entre la distribución de los datos en el espacio latente y la distribución normal).
Luego, para comparar cuantitativamente los resultados de las AEs y VAEs, se entrenó en cada caso una CNN Residual para la tarea de regresión fenotípica a partir de los datos de dimensión reducida. Si bien en ambos casos los resultados obtenidos fueron más bajos que aquellos que se obtuvo utilizando Gradient Boosting sobre los datos originales, se destaca que los resultados alcanzados utilizando los AEs superan a aquellos presentados en (Abdollahi-Arpanahi et al., 2020), donde se entrenó directamente una CNN con los datos sin reducción.
Combinación de datos genómicos con otras fuentes de información:
La combinación de datos de diferentes fuentes planteó esquemas de predicción diferentes, según el objetivo de predicción, si es un nuevo híbrido o un nuevo ambiente, o si se usa un modelo más clásico. En general, siempre fue más beneficioso usar genotipo y ambiente, que usar sólo ambiente o sólo genotipo. De todas las maneras, para algunos modelos los resultados no son del todo malos, pudiendo utilizar los modelos propuestos en bases de datos que no contengan datos genómicos o que no contengan datos ambientales.
Para predecir el rendimiento de un híbrido que el modelo nunca había visto, XGBoost fue el modelo que mejores resultados obtuvo. Este resultado es similar al obtenido para otras especies y otros fenotipos, donde, aunque no se consideraran datos ambientales, el objetivo es predecir un fenotipo a partir de un nuevo genotipo. En este caso, las redes neuronales, pese a su capacidad de aprender patrones, no pudieron superar a los modelos lineales mixtos. Además se continuó trabajando sobre arquitecturas tipo transformer, pero todavía los resultados no son del todo satisfactorios.
Para la predicción de rendimiento en híbridos que el modelo ya había visto, pero en nuevas condiciones ambientales, las redes fueron las que mejores resultados obtuvieron, pudiendo explotar la estructura intrínseca de las series temporales de los datos climatológicos.
En este proyecto no se evaluó la capacidad predictiva de los modelos para predecir variables climáticas, ya que se consideran conocidas, dado que estaban presentes en la base de datos. Para utilizarlos en producción, será necesario evaluar cómo predecir este tipo de datos, si se tomarán predicciones hechas por otros, o si usando promedios de observaciones históricas del clima, alcanza para realizar buenas predicciones, al menos en el ranking de híbridos más adecuados.
Cooperación, materiales generados y rol model:
Se realizaron actividades para fortalecer la cooperación entre el grupo de investigación local y un grupo de investigación mexicano con amplia tradición en predicción genómica. Además de las instancias de discusión virtuales, se realizaron cuatro estancias de intercambio con los investigadores. Por un lado, se dictó un seminario sobre los avances del grupo en predicción genómica en el Colegio de Postgraduados de Montecillo, Texcoco, coordinado por el Dr. Crossa y se realizó una visita al Centro Internacional de Mejoramiento de Maíz y Trigo. Dictamos el curso «PredGenIA: Métodos de Aprendizaje Profundo para Predicción Genómica», en el Centro Universitario de Ciencias Exactas e Ingenierías (CUCEI), Guadalajara coordinado con Abelardo Montesinos. Dictamos el curso «Predicción Genómica: desde regresiones lineales a redes neuronales» en la Facultad de Ingeniería con la participación de Osval Montesinos en Montevideo y José Crossa desde México. Realizamos una jornada de charlas «Innovaciones en el mejoramiento genético: herramientas y estrategias cuantitativas», en Facultad de Agronomía con la presencia de José Crossa. Todo el material de ambos cursos está disponible en Google Classroon (https://classroom.google.com/), a los que se pueden acceder como estudiantes utilizando los códigos h6gl56t (curso dictado en México) y grnxg2c (curso dictado en Uruguay, con videos de las clases disponibles).
Durante 2024, la responsable del proyecto participó del día de las chicas en TIC en Facultad de Ingeniería y en una actividad organizada por la empresa Digital Sense en la Escuela Experimental de Malvín, fue panelista en la sesión Mujeres en IA y Educación en el congreso EduIA, Conferencia REgional en Inteligencia Artificial en Educación organizado por Ceibal y brindó una conferencia en la inauguración de la exposición ArtFutura titualada IA, educaión género. También participó de Ingeniería deMuestra, acompañada de otras mujeres del Instituto de Matemática y Estadística "Prof. Ing. Rafael Laguardia", mostrando la presencia de mujeres en el mundo de las matemáticas. Graciana Castro participó en actividades de Techy x el día.
Se está diseñando un video en el que se verá a la mayor parte del equipo, incluso los investigadores mexicanos por vía remota. En el video cuidaremos que los minutos en los que se muestran hombres y mujeres, sean lo más parejos posible.

Conclusiones y recomendaciones

Este proyecto de investigación ha representado un avance significativo en el estudio de métodos de Aprendizaje Profundo (AP) aplicados a la Selección Genómica (SG), un método crucial para la mejora de cultivos y animales de producción. Hemos logrado demostrar la viabilidad de incorporar métodos de AP, particularmente con Transformers, para aumentar la precisión en la predicción de características de interés a partir de datos genéticos y ambientales. Esto es fundamental, ya que una mayor precisión en las predicciones permite optimizar los programas de mejoramiento genético, lo que se traduce en una mayor ganancia genética por unidad de tiempo y un uso más eficiente de los recursos.
Un hallazgo clave ha sido la confirmación de que la integración de datos genómicos con información ambiental, como clima y características del suelo, mejora considerablemente la capacidad predictiva, especialmente para el rendimiento en cultivos como el maíz. La aplicación de técnicas de reducción de dimensionalidad no lineales, como los autoencoders, ha demostrado ser superior a los métodos lineales tradicionales (PCA) para manejar la alta complejidad de los datos genómicos. Aunque los modelos de redes neuronales mostraron un desempeño excepcional en la predicción de híbridos en nuevos ambientes, los modelos basados en árboles de decisión destacaron en la predicción de híbridos completamente nuevos. Esto subraya la importancia de elegir la arquitectura adecuada según el objetivo específico de la predicción.
Además de los avances técnicos, el proyecto ha fortalecido la colaboración entre grupos de investigación de Uruguay y México, fomentando un intercambio de conocimientos enriquecedor y la formación de recursos humanos capacitados en la intersección de la genética y el aprendizaje automático. La creación de materiales de formación accesibles y la participación en actividades de divulgación demuestran el compromiso con la difusión del conocimiento y el fomento de la participación de mujeres en el ámbito de las TIC y la ciencia. Estos logros sientan las bases para futuras investigaciones y la implementación rutinaria de la SG en programas de mejoramiento genético en la región, contribuyendo a una producción agropecuaria más eficiente y rentable.
Trabajo futuro:
Finalmente, mencionar que al momento de optimizar los hiperparámetros que definieron la arquitectura de la red, nos vimos limitados por la capacidad computacional. Es por esto que si hubiésemos dispuesto de más capacidad computacional, o en su defecto, más tiempo, hubiéramos refinado más la búsqueda de estos hiperparámetros.
En \ref{sec:desglose} hablamos de los \textit{metadatos}, una serie de registros que describen las condiciones en que se realizaron los cultivos. Estos datos no los utilizamos pero podrían haber aportado información, sobre todo habiendo visto la importancia que tienen los factores externos por sobre los genéticos.
Pensando en trabajos futuros, este proyecto deja varios frentes de investigación abiertos. En los siguientes puntos detallamos algunos de ellos.

Productos derivados del proyecto

Tipo de producto	Título	Autores	Identificadores	URI en repositorio de Silo	Estado
Tesis de grado/monografías	Predicción genómica multimodal : Integración de datos genómicos y ambientales para estimar el rendimiento en maíz mediante Aprendizaje Automático	Abatte, Iván - Ledesma, Joaquín - Sarachu, Santiago		https://hdl.handle.net/20.500.12008/48451	Finalizado
Tesis de maestría	Integración de datos genómicos y ambientales utilizando aprendizaje profundo para predicción genómica	Ximena Fernández			En proceso
Tesis de maestría	Reducción de dimensionalidad de datos genómicos basada en grafos.	Camilo Borba			En proceso
Artículo científico	Refining penalized Ridge regression: a novel method for optimizing the regularization parameter in genomic prediction	Abelardo Montesinos-López , Osval A Montesinos-López , Federico Lecumberry , María I Fariello , José C Montesinos-López , José Crossa		https://hdl.handle.net/20.500.12008/46936	Finalizado

Tipo de producto	Título	Autores	Identificadores	URI en repositorio de Silo	Estado
Publicación de trabajo en evento (artículo de conferencia)	Transformers for Genomic Prediction	Fariello, M.I., Castro, G., Hoffman, R., Musitelli, M., Belzarena, D., Lecumberry, F.		https://hdl.handle.net/20.500.12008/49923	Finalizado
Otro	Métodos de Aprendizaje Profundo para Predicción Genómica (PredGenIA)	Federico Lecumberry, Graciana Castro, María Inés Fariello			Finalizado
Póster	Deep learning for genomic prediction and tasks learned on the way	María Inés Fariello, Ignacio Hounie, Juan Elenter, Guillermo Etchebarne, Leonardo de los Santos, Lucía Arboleya, Diego Belsarena, Graciana Castro, Romina Hoffman, Mateo Musitelli, Federico Lecumberry		https://hdl.handle.net/20.500.12008/51573	En proceso
Presentación en evento	Predicción genómica mediante aprendizaje profundo	Federico Lecumberry, Graciana Castro, Mateo			En proceso

Tipo de producto	Título	Autores	Identificadores	URI en repositorio de Silo	Estado
		Musittelli, María Inés Fariello,			
Póster	TRANSFORMERS FOR GENOMIC PREDICTION: working with Yeast and Wheat traits	Graciana Castro , Romina Hoffman, Mateo Musitelli, María Inés Fariello, Federico Lecumberry		https://hdl.handle.net/20.500.12008/51574	En proceso
Póster	PredGenIA: ? Transformers para Predicción Genómica	Graciana Castro - Romina Hoffman - Mateo Musitelli - María Inés Fariello - Federico Lecumberry		https://hdl.handle.net/20.500.12008/51571	En proceso

Referencias bibliográficas

Abdollahi-Arpanahi, R., Gianola, D., & Peñagaricano, F. (2020). Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genetics, Selection, Evolution: GSE, 52(1), 12.

Astori, D., & Alonso, J. M. (1979). La evolución tecnológica de la ganadería uruguaya, 1930-1977.

Battey, C. J., Coffing, G. C., & Kern, A. D. (2021). Visualizing population structure with variational autoencoders. G3 , 11(1). https://doi.org/10.1093/g3journal/jkaa036

Bernardo, R., & Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. Crop Science, 47(3), 1082–1090.

Crossa, J., Beyene, Y., Kassa, S., Pérez, P., Hickey, J. M., Chen, C., de los Campos, G., Burgueño, J., Windhausen, V. S., Buckler, E., Jannink, J.-L., Lopez Cruz, M. A., & Babu, R. (2013). Genomic prediction in maize breeding populations with genotyping-by-sequencing. G3 , 3(11), 1903–1926.

Crossa, J., Campos, G. de L., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., Makumbi, D., Singh, R. P., Dreisigacker, S., Yan, J., Arief, V., Banziger, M., & Braun, H.-J. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics, 186(2), 713–724.

Crossa, J., Fritsche-Neto, R., Montesinos-Lopez, O. A., Costa-Neto, G., Dreisigacker, S., Montesinos-Lopez, A., & Bentley, A. R. (2021). The modern plant breeding triangle: Optimizing the use of genomics, phenomics, and enviromics data. Frontiers in Plant Science, 12, 651480.

Crossa, J., Pérez, P., de los Campos, G., Mahuku, G., Dreisigacker, S., & Magorokosho, C. (2011). Genomic Selection and Prediction in Plant Breeding. Journal of Crop Improvement, 25(3), 239–261.
Crossa, J., Pérez, P., Hickey, J., Burgueño, J., Ornella, L., Cerón-Rojas, J., Zhang, X., Dreisigacker, S., Babu, R., Li, Y., Bonnett, D., & Mathews, K. (2014). Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity, 112(1), 48–60.
Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de Los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., & Varshney, R. K. (2017). Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends in Plant Science, 22(11), 961–975.
Dellazoppa, R. (2014). Agro: la revolución sorprendente.
De los Campos, G., Gianola, D., Rosa, G. J. M., Weigel, K. A., & Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genetics Research, 92(4), 295–308.
de Los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., & Calus, M. P. L. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics, 193(2), 327–345.
de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., & Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics, 182(1), 375–385.
Gianola, D. (2013). Priors in whole-genome regression: the bayesian alphabet returns. Genetics, 194(3), 573–596.
Gianola, D., Fariello, M. I., Naya, H., & Schön, C.-C. (2016). Genome-wide association studies with a genomic relationship matrix: A case study with wheat and Arabidopsis. G3 (Bethesda, Md.), 6(10), 3241–3256.
Heslot, N., Akdemir, D., Sorrells, M. E., & Jannink, J.-L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik, 127(2), 463–480.
Heslot, N., Yang, H.-P., Sorrells, M. E., & Jannink, J.-L. (2012). Genomic selection in plant breeding: A comparison of models. Crop Science, 52(1), 146–160.
Hickey, J. M., Crossa, J., Babu, R., & de los Campos, G. (2012). Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Science, 52(2), 654–663.
Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Pérez, P., Calus, M., Burgueño, J., & de los Campos, G. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theoretical and Applied Genetics, 127(3), 595–607.
Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics (Oxford, England), 37(15), 2112–2120.
Jubair, S., Tucker, J. R., Henderson, N., Hiebert, C. W., Badea, A., Domaratzki, M., & Fernando, W. G. D. (2021). GPTransformer: A Transformer-based deep learning method for predicting Fusarium related traits in barley. Frontiers in Plant Science, 12, 761402.
Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. In arXiv [stat.ML]. arXiv. http://arxiv.org/abs/1312.6114
Lorenzana, R. E., & Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik, 120(1), 151–161.
Maaten, L., & Hinton, G. E. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research, 9(86), 2579–2605.
Måløy, H., Windju, S., Bergersen, S., Alsheikh, M., & Downing, K. L. (2021). Multimodal performers for genomic selection and crop yield prediction. In Smart Agricultural Technology (Vol. 1, p. 100017). https://doi.org/10.1016/j.atech.2021.100017
Martinek, V., Cechak, D., Gresova, K., Alexiou, P., & Simecek, P. (2022). Fine-tuning Transformers for genomic tasks. In bioRxiv. https://doi.org/10.1101/2022.02.07.479412
Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4), 1819–1829.
Montesinos-López, A., Montesinos-López, O. A., Lecumberry, F., Fariello, M. I., Montesinos-López, J. C., & Crossa, J. (2024). Refining penalized ridge regression: a novel method for optimizing the regularization parameter in genomic prediction. G3 (Bethesda, Md.). https://doi.org/10.1093/g3journal/jkae246
Montesinos-López, O. A., Barajas-Ramirez, E. A., Montesinos-López, A., Lecumberry, F., Fariello, M. I., Montesinos-López, J. C., Ramirez Alcaraz, J. M., Crossa, J., & Howard, R. (2025). Tuning matters: Comparing lambda optimization approaches for ridge regression in genomic prediction. Genes, 16(6). https://doi.org/10.3390/genes16060618

Shook, J., Gangopadhyay, T., Wu, L., Ganapathysubramanian, B., Sarkar, S., & Singh, A. K. (2021). Crop yield prediction integrating genotype and weather variables using deep learning. PloS One, 16(6), e0252402.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, ?., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper/7181-attention-is-all