



UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

FACULTAD DE POSGRADO

MAESTRÍA EN CIENCIA DE DATOS

Proyecto de Investigación previo a la
obtención del Grado Académico de
Magíster en Ciencia de Datos.

TEMA

Pronóstico del rendimiento de cultivo de maíz forrajero basado en variables
agronómicas en la ciudad de Quevedo

AUTOR

Ing. Ricardo Rafael Villarroel Molina

DIRECTOR

Ing. Cristian Zambrano Vega, PhD.

Quevedo – Los Ríos – Ecuador

2024

CERTIFICACIÓN

El **Ing. Cristian Zambrano Vega, PhD.**, Director del Proyecto de Investigación previo a la obtención del Grado Académico de Magíster en Ciencia de Datos.

CERTIFICA:

Que el **Ing. Ricardo Rafael Villarroel Molina**, ha cumplido con la elaboración del Proyecto de Investigación titulado: **“Pronóstico del rendimiento de cultivo de maíz forrajero basado en variables agronómicas en la ciudad de Quevedo”**; el mismo que se encuentra apto para la presentación y sustentación respectiva.

Quevedo, 27 de julio de 2024.

Ing. Cristian Zambrano Vega, PhD.

DIRECTOR

AUTORÍA

Yo, Ricardo Rafael Villarroel Molina con C.I. 0503679631 declaro que el presente proyecto de investigación titulado “**Pronóstico del rendimiento de cultivo de maíz forrajero basado en variables agronómicas en la ciudad de Quevedo**”, es de exclusiva autoría y extendiendo los derechos a la Universidad Técnica Estatal de Quevedo, hacer uso del documento como material académico y de consulta.

Quevedo, 27 de julio de 2024.

Ing. Ricardo Rafael Villarroel Molina

AUTOR

DEDICATORIA

Este trabajo está dedicado con todo cariño y gratitud a mis padres Freddy Villarroel y Marisel Molina, que me enseñaron el valor de la perseverancia y la importancia de seguir adelante incluso frente a las adversidades. A Annelys Kaiser quien con su apoyo incondicional hizo posible este logro y a Lourdes Villarroel, Lisbeth Villarroel, Lucy Villarroel y Esmeralda Alvarado cuyo amor, motivación y ejemplo incansable han sido la luz que ha guiado cada paso de mi camino. A toda la familia Kaiser que su ayuda y cariño ha sido una fortaleza para culminar esta etapa académica. Juntos, han sido un pilar en el cual he construido mis sueños y aspiraciones. A todos ellos, mi eterno agradecimiento por ser mi fuente constante de inspiración y apoyo, reflejando la verdadera esencia de mi éxito.

Ricardo Villarroel Molina

AGRADECIMIENTO

Extiendo mi más grande agradecimiento a Annelys Kaiser que me ha estado acompañando durante todo este proceso, gracias a quien con sus conocimientos e inmenso apoyo he logrado culminar este proyecto de investigación y muchos otros logros. Al investigador Mgs Javier Maiguashca Guzmán del Instituto Nacional de Investigaciones Agropecuarias (INIAP) y a mi director de tesis el Ing. Cristian Zambrano Vega, PhD, quienes con su guía y ayuda se pudo desarrollar con éxito esta investigación. A Berenice Kaiser por su invaluable ayuda y apoyo, mis compañeros de maestría, y a la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT) por su programa de “Becas Nacionales de Posgrado Fortalécete 2022”, el cual me ha brindado una invaluable oportunidad para continuar con mis estudios de posgrado. Esta beca ha sido fundamental para fortalecer mi formación académica y profesional.

PRÓLOGO

En la presente investigación “Pronóstico del rendimiento de cultivo de maíz forrajero basado en variables agronómicas en la ciudad de Quevedo”, bajo la colaboración del Instituto Nacional de Investigaciones Agropecuarias (INIAP), se centró en el desarrollo de un modelo de pronóstico utilizando técnicas de aprendizaje automático supervisado.

La investigación se sitúa en el cruce de la ciencia de datos y la ciencia agronómica, dos áreas de creciente importancia en la actualidad. La aplicación de predicciones basados en algoritmos de aprendizaje automático presenta un avance tecnológico y también una necesidad urgente para mejorar la eficiencia y sostenibilidad de la producción de alimentos tanto para consumo humano como animal, por ende, se aplicaron metodologías científicas y rigurosas para abordar un problema real.

Además, se realizó un análisis descriptivo detallado de los datos agronómicos, y se llevó a cabo experimentos meticulosos para ajustar hiperparámetros de los algoritmos seleccionados, con el objetivo de minimizar la Raíz Cuadrada del Error Cuadrático Medio (RMSE) en las predicciones del rendimiento.

Mgs. Javier Maiguashca Guzmán

Investigador

INSTITUTO NACIONAL DE INVESTIGACIONES AGROPECUARIAS

RESUMEN

Este proyecto de investigación se centró en la construcción de modelos predictivos para el rendimiento del maíz forrajero en Quevedo utilizando técnicas de aprendizaje automático supervisado. Se analizaron variables agronómicas como altura de la planta, diámetro del tallo, inserción de mazorca, variedad de semilla y distancia de siembra, destacando que mayores distancias entre plantas favorecen un aumento en el diámetro del tallo, correlacionándose positivamente con las variables de rendimiento estudiadas (peso total, peso de hojas, peso de tallo y peso de mazorcas). Se entrenaron y optimizaron modelos independientes para cada variable de rendimiento mediante el ajuste de hiperparámetros para minimizar el RMSE con los algoritmos de Linear Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Multi-layer Perceptron (MLP), Deep Neural Network (DNN), Gradient Boosting Tree (GBT) y XGBoost (XGB). Se evaluó el desempeño de los modelos mediante las pruebas estadísticas de significancia de Friedman y la prueba post-hoc de Nemenyi, centradas en las métricas de RMSE y R^2 . Los modelos construidos por el algoritmo de RF obtuvieron los mejores valores de RMSE y de R^2 .

Palabras claves: maíz, agricultura de precisión, aprendizaje automático, pronóstico de rendimiento.

ABSTRACT

This research project focused on the construction of predictive models for forage maize yield in Quevedo using supervised machine learning techniques. Agronomic variables such as plant height, stalk diameter, ear insertion, seed variety and planting distance were analyzed, highlighting that greater distances between plants favor an increase in stalk diameter, correlating positively with the yield variables studied (total weight, leaf weight, stalk weight and ear weight). Independent models for each yield variable were trained and optimized by fitting hyperparameters to minimize RMSE with Linear Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Multi-layer Perceptron (MLP), Deep Neural Network (DNN), Gradient Boosting Tree (GBT) and XGBoost (XGB) algorithms. Model performance was evaluated using Friedman's statistical significance tests and Nemenyi's post-hoc test, focusing on RMSE and R^2 metrics. The models built by the RF algorithm obtained the best RMSE and R^2 values.

Keywords: maize, precision agriculture, machine learning, yield prediction.

ÍNDICE

CERTIFICACIÓN	ii
AUTORÍA	iii
DEDICATORIA	iv
AGRADECIMIENTO	v
PRÓLOGO.....	vi
RESUMEN	vii
ABSTRACT.....	viii
ÍNDICE.....	ix
INTRODUCCIÓN	xix
CAPÍTULO I	1
MARCO CONTEXTUAL DE LA INVESTIGACIÓN	1
1.1 CONTEXTUALIZACIÓN Y UBICACIÓN DE LA PROBLEMÁTICA	2
1.2 SITUACIÓN ACTUAL DE LA PROBLEMÁTICA.....	2
1.3 PROBLEMA DE INVESTIGACIÓN	3
1.3.1 Problema General	3
1.3.2 Problemas Derivados	3
1.4 DELIMITACIÓN DEL PROBLEMA.....	3
1.5 OBJETIVOS	4
1.5.1 Objetivo General.....	4
1.5.2 Objetivos Específicos	4
1.6 JUSTIFICACIÓN	5
CAPÍTULO II.....	6
MARCO TEÓRICO	6
2.1. FUNDAMENTACIÓN CONCEPTUAL	7
2.1.1. Agricultura.....	7

2.1.2.	Maíz Forrajero	8
2.1.3.	Fenología del Maíz	8
2.1.3.1.	Subdivisión de las Etapas Vegetativa y Reproductiva.....	10
2.1.3.2.	Requerimientos Edafoclimáticos	11
2.1.4.	Agricultura de Precisión	13
2.1.5.	Beneficios Potenciales de la Agricultura de Precisión	14
2.1.6.	Rendimiento de Cultivos	14
2.1.7.	Cross Industry Standard Process for Data Mining.....	16
2.1.8.	Aprendizaje Automático Supervisado	17
2.1.9.	Regresión	18
2.1.10.	Regresión Lineal	19
2.1.11.	Support Vector Machines	20
2.1.12.	Random Forest.....	20
2.1.13.	Gradient-Boosting Regressor.....	20
2.1.14.	XGBoost	22
2.1.15.	Redes Neuronales Artificiales	23
2.2.	FUNDAMENTACIÓN TEÓRICA	25
2.3.	FUNDAMENTACIÓN LEGAL.....	27
2.3.1.	Constitución de la República	28
2.3.2.	Ley Orgánica del Régimen de la Soberanía Alimentaria	29
CAPÍTULO III.....		31
METODOLOGÍA DE LA INVESTIGACIÓN		31
3.1.	TIPOS DE INVESTIGACIÓN	32
3.1.1.	Investigación Documental	32
3.1.2.	Investigación Correlacional	32
3.2.	MÉTODO DE INVESTIGACIÓN	32

3.2.1.	Método Exploratorio.....	32
3.3.	CONSTRUCCIÓN METODOLÓGICA DEL OBJETO DE INVESTIGACIÓN	33
3.3.1.	Técnicas de Investigación.....	33
3.3.2.	Instrumentos de Investigación	34
3.4.	ELABORACIÓN DEL MARCO TEÓRICO	35
3.5.	RECOLECCIÓN DE LA INFORMACIÓN.....	36
3.6.	PROCESAMIENTO Y ANÁLISIS.....	39
3.6.1.	Comprensión del Negocio	40
3.6.2.	Comprensión de los Datos	40
3.6.3.	Preparación de los Datos.....	40
3.6.4.	Modelado	42
3.6.5.	Evaluación	43
3.6.6.	Despliegue	45
3.6.7.	Diagrama de flujo de datos	45
	CAPÍTULO IV	46
	RESULTADOS Y DISCUSIÓN	46
4.1.	ANÁLISIS DESCRIPTIVO DE LOS DATOS AGRONÓMICOS	47
4.1.1	Análisis Descriptivo.....	49
4.1.1.1	Días Después de la Siembra.....	50
4.1.1.2	Diámetro del Tallo	50
4.1.1.3	Altura de la Planta.....	51
4.1.1.4	Etapa Vegetativa	52
4.1.1.5	Presencia de Inflorescencia Masculina	54
4.1.1.6	Presencia de Inflorescencia Femenina.....	55
4.1.1.7	Frecuencia de la Etapa Reproductiva.....	56

4.1.2	Análisis Descriptivo de Registros del 21/08/2023.....	57
4.1.2.1	Diámetro del Tallo	57
4.1.2.2	Altura de la Planta.....	58
4.1.2.3	Número de Mazorcas	59
4.1.2.4	Altura de la Inserción de la Mazorca	61
4.1.2.5	Etapas Vegetativa	62
4.1.2.6	Presencia de Inflorescencia Masculina	62
4.1.2.7	Presencia de Inflorescencia Femenina.....	63
4.1.2.8	Etapas Reproductiva.....	64
4.1.3	Análisis Multivariante de Variables Agronómicas Numéricas.....	65
4.1.3.1	Análisis Multivariante de Variables Agronómicas Numéricas por Variedad de Maíz	65
4.1.3.2	Análisis Multivariante de Variables Agronómicas Numéricas por Distancia de Siembra	66
4.1.4	Análisis Descriptivo de Datos de Rendimiento	67
4.1.4.1	Matriz de Correlación de Spearman de Variables de Rendimiento del Maíz	68
4.1.4.2	Peso Total	69
4.1.4.3	Peso de Hojas.....	69
4.1.4.4	Peso de Tallo.....	70
4.1.4.5	Peso de Mazorca	70
4.1.5	Matriz de Correlación de Spearman de Datos Agronómicos y de Rendimiento del Maíz.....	70
4.1.6	Discusión	71
4.2.	AJUSTE DE HIPERPARÁMETROS DE LOS ALGORITMOS	73
4.2.1	Selección de Algoritmos	74
4.2.2	Selección de Rango de Valores para Hiperparámetros de Algoritmos...	76

4.2.3	Preprocesamiento de Datos.....	77
4.2.4	Tiempo de Ajuste de Hiperparámetros	78
4.2.4.1	Tiempo de Ajuste para Peso Total	78
4.2.4.2	Tiempo de Ajuste para Peso Hojas	80
4.2.4.3	Tiempo de Ajuste para Peso Tallo	81
4.2.4.4	Tiempo de Ajuste para Peso Mazorca	83
4.2.5	Evaluación de Métricas en Ajuste de Hiperparámetros.....	85
4.2.5.1	Métricas de Evaluación para Peso Total.....	85
4.2.5.2	Métricas de Evaluación para Peso Hojas	87
4.2.5.3	Métricas de Evaluación para Peso Tallo	89
4.2.5.4	Métricas de Evaluación para Peso Mazorca	91
4.2.6	Valores de Hiperparámetros	93
4.2.7	Discusión	96
4.3	SELECCIÓN DEL MODELO PARA LA PREDICCIÓN DEL RENDIMIENTO DEL MAÍZ FORRAJERO.....	99
4.3.1	Evaluación de Modelos para Peso Total.....	99
4.3.1.1	Métricas de Evaluación con Datos de Entrenamiento	99
4.3.1.2	Métricas de Evaluación con Datos de Prueba.....	101
4.3.1.3	Tiempos de Entrenamiento	103
4.3.1.4	Tiempos de Evaluación.....	104
4.3.2	Evaluación de Modelos para Peso Hojas	105
4.3.2.1	Métricas de Evaluación con Datos de Entrenamiento	105
4.3.2.2	Métricas de Evaluación con Datos de Prueba.....	107
4.3.2.3	Tiempos de Entrenamiento	109
4.3.2.4	Tiempos de Evaluación.....	110
4.3.3	Evaluación de Modelos para Peso Tallo.....	111

4.3.3.1	Métricas de Evaluación con Datos de Entrenamiento	111
4.3.3.2	Métricas de Evaluación con Datos de Prueba.....	113
4.3.3.3	Tiempos de Entrenamiento	115
4.3.3.4	Tiempos de Evaluación.....	116
4.3.4	Evaluación de Modelos para Peso Mazorca	117
4.3.4.1	Métricas de Evaluación con Datos de Entrenamiento	117
4.3.4.2	Métricas de Evaluación con Datos de Prueba.....	119
4.3.4.3	Tiempos de Entrenamiento	121
4.3.4.4	Tiempos de Evaluación.....	122
4.3.5	Discusión	123
CAPÍTULO V.....		125
CONCLUSIONES Y RECOMENDACIONES		125
5.1	CONCLUSIONES	126
5.2	RECOMENDACIONES.....	128
REFERENCIAS BIBLIOGRÁFICAS		130
ANEXOS		142

Índice de Ilustraciones

<i>Ilustración 1: Representación gráfica de una neurona artificial.</i>	<i>23</i>
<i>Ilustración 2: Arquitectura de una red neuronal multicapa no cíclica.</i>	<i>25</i>
<i>Ilustración 3: Ajustes de hiperparámetros mediante función GridSearchCV.</i>	<i>43</i>
<i>Ilustración 4: Diagrama de flujo de datos de las etapas de la metodología.</i>	<i>45</i>
<i>Ilustración 5: Diámetro del tallo</i>	<i>51</i>
<i>Ilustración 6: Altura de la planta</i>	<i>52</i>
<i>Ilustración 7: Etapa vegetativa.....</i>	<i>53</i>
<i>Ilustración 8: Frecuencia de la etapa vegetativa</i>	<i>54</i>
<i>Ilustración 9: Frecuencia de la presencia de inflorescencia masculina</i>	<i>55</i>
<i>Ilustración 10: Frecuencia de la presencia de inflorescencia femenina.....</i>	<i>56</i>
<i>Ilustración 11: Frecuencia de la etapa reproductiva</i>	<i>57</i>
<i>Ilustración 12: Diámetro del tallo para el 21/08/2023</i>	<i>58</i>
<i>Ilustración 13: Altura de la planta para el 21/08/2023.....</i>	<i>59</i>
<i>Ilustración 14: Número de mazorcas para el 21/08/2023.....</i>	<i>60</i>
<i>Ilustración 15: Altura de la inserción de la mazorca para el 21/08/2023</i>	<i>61</i>
<i>Ilustración 16: Etapa vegetativa para el 21/08/2023.....</i>	<i>62</i>
<i>Ilustración 17: Presencia de inflorescencia masculina para el 21/08/2023</i>	<i>63</i>
<i>Ilustración 18: Presencia de inflorescencia femenina para el 21/08/2023.....</i>	<i>64</i>
<i>Ilustración 19: Etapa reproductiva para el 21/08/2023.....</i>	<i>65</i>
<i>Ilustración 20: Gráfica por pares de variables agronómicas numéricas por variedad de maíz.....</i>	<i>66</i>
<i>Ilustración 21: Gráfica por pares de variables agronómicas numéricas agronómicas numéricas por distancia de siembra</i>	<i>67</i>
<i>Ilustración 22: Correlación de variables de rendimiento del maíz.....</i>	<i>69</i>
<i>Ilustración 23: Correlación de datos Agronómicos y de rendimiento del maíz</i>	<i>71</i>
<i>Ilustración 24: Tiempos de ajuste de hiperparámetros para modelos de peso total.....</i>	<i>79</i>
<i>Ilustración 25: Tiempos de ajuste de hiperparámetros en ms para modelos de peso total</i>	<i>79</i>
<i>Ilustración 26: Tiempos de ajuste de hiperparámetros para modelos de peso hojas</i>	<i>80</i>
<i>Ilustración 27: Tiempos de ajuste de hiperparámetros en ms para modelos de peso hojas</i>	<i>81</i>

<i>Ilustración 28: Tiempos de ajuste de hiperparámetros para modelos de peso tallo.....</i>	<i>82</i>
<i>Ilustración 29: Tiempos de ajuste de hiperparámetros en ms para modelos de peso tallo</i>	<i>83</i>
<i>Ilustración 30: Tiempos de ajuste de hiperparámetros en para modelos de peso mazorca</i>	<i>84</i>
<i>Ilustración 31: Tiempos de ajuste de hiperparámetros en ms para modelos de peso mazorca.....</i>	<i>84</i>
<i>Ilustración 32: RMSE en ajuste de hiperparámetros para peso total</i>	<i>86</i>
<i>Ilustración 33: R^2 en ajuste de hiperparámetros para peso total.....</i>	<i>87</i>
<i>Ilustración 34: RMSE en ajuste de hiperparámetros para peso hojas.....</i>	<i>88</i>
<i>Ilustración 35: R^2 en ajuste de hiperparámetros para peso Hojas.....</i>	<i>89</i>
<i>Ilustración 36: RMSE en ajuste de hiperparámetros para peso tallo</i>	<i>90</i>
<i>Ilustración 37: R^2 en ajuste de hiperparámetros para peso tallo.....</i>	<i>91</i>
<i>Ilustración 38: RMSE en ajuste de hiperparámetros para peso mazorca.....</i>	<i>92</i>
<i>Ilustración 39: R^2 en ajuste de hiperparámetros para peso mazorca</i>	<i>93</i>
<i>Ilustración 40: Métrica RMSE de datos de entrenamiento para Peso Total.....</i>	<i>100</i>
<i>Ilustración 41: Métrica R^2 de datos de entrenamiento para Peso Total</i>	<i>101</i>
<i>Ilustración 42: Métrica RMSE de datos de prueba para Peso Total</i>	<i>102</i>
<i>Ilustración 43: Métrica R^2 de datos de prueba para Peso Total.....</i>	<i>103</i>
<i>Ilustración 44: Tiempo de entrenamiento para peso total</i>	<i>104</i>
<i>Ilustración 45: Tiempo de evaluación para peso total</i>	<i>105</i>
<i>Ilustración 46: Métrica RMSE de datos de entrenamiento para Peso Hojas</i>	<i>106</i>
<i>Ilustración 47: Métrica R^2 de datos de entrenamiento para Peso Hojas.....</i>	<i>107</i>
<i>Ilustración 48: Métrica RMSE de datos de prueba para Peso Hojas</i>	<i>108</i>
<i>Ilustración 49: Métrica R^2 de datos de prueba para Peso Hojas.....</i>	<i>109</i>
<i>Ilustración 50: Tiempo de Entrenamiento para peso hojas</i>	<i>110</i>
<i>Ilustración 51: Tiempo de evaluación para peso hojas.....</i>	<i>111</i>
<i>Ilustración 52: Métrica RMSE de datos de entrenamiento para Peso Tallo.....</i>	<i>112</i>
<i>Ilustración 53: Métrica R^2 de datos de entrenamiento para Peso Tallo</i>	<i>113</i>
<i>Ilustración 54: Métrica RMSE de datos de prueba para Peso Tallo</i>	<i>114</i>
<i>Ilustración 55: Métrica R^2 de datos de prueba para peso tallo</i>	<i>115</i>
<i>Ilustración 56: Tiempo de entrenamiento para peso tallo</i>	<i>116</i>

<i>Ilustración 57: Tiempo de evaluación para peso tallo</i>	<i>117</i>
<i>Ilustración 58: Métrica RMSE de datos de entrenamiento para Peso Mazorca.....</i>	<i>118</i>
<i>Ilustración 59: Métrica R^2 de datos de entrenamiento para Peso Mazorca</i>	<i>119</i>
<i>Ilustración 60: Métrica RMSE de datos de prueba para peso mazorca</i>	<i>120</i>
<i>Ilustración 61: Métrica R^2 de datos de prueba para peso mazorca</i>	<i>121</i>
<i>Ilustración 62: Tiempo de entrenamiento para peso mazorca</i>	<i>122</i>
<i>Ilustración 63: Tiempo de evaluación para peso mazorca.....</i>	<i>123</i>

Índice de Tablas

<i>Tabla 1: Descripción y tipo de datos de las variables agronómicas.....</i>	<i>37</i>
<i>Tabla 2:Descripción y tipo de datos de las variables de rendimiento</i>	<i>38</i>
<i>Tabla 3: Periodos de recolección de datos de las variables agronómicas y de las variables de rendimiento.</i>	<i>39</i>
<i>Tabla 4: Variables agronómicas del maíz forrajero</i>	<i>47</i>
<i>Tabla 5: Estadística descriptiva de las variables numéricas</i>	<i>49</i>
<i>Tabla 6: Estadística descriptiva de las variables de rendimiento.....</i>	<i>68</i>
<i>Tabla 7: Estudios relacionados con lista de algoritmos y métricas de evaluación utilizadas.....</i>	<i>74</i>
<i>Tabla 8: Algoritmos rango de valores de hiperparámetros</i>	<i>76</i>
<i>Tabla 9: Valores de hiperparámetros por modelo y algoritmo.....</i>	<i>94</i>

Índice de Ecuaciones

<i>Ecuación 1: Regresión lineal.....</i>	<i>19</i>
<i>Ecuación 2: Función de riesgo empírico minimizado</i>	<i>21</i>
<i>Ecuación 3: Método de Boosting.....</i>	<i>21</i>
<i>Ecuación 4: Gradient Boosting</i>	<i>22</i>
<i>Ecuación 5: Función objetivo de XGBoost.....</i>	<i>22</i>
<i>Ecuación 6: Función de entrada de una neurona artificial.</i>	<i>23</i>
<i>Ecuación 7: Funciones de activación de redes neuronales.....</i>	<i>24</i>
<i>Ecuación 8: Técnica de normalización mim-max.....</i>	<i>41</i>

<i>Ecuación 9: Formula para cálculo de la métrica de RMSE.....</i>	<i>44</i>
<i>Ecuación 10: Formula para cálculo del R^2.</i>	<i>44</i>

Índice de Anexos

Anexo 1: Informe de COMPILATIO	142
Anexo 2: Oficio de aceptación para colaboración en proyecto de investigación	143
Anexo 3: Capturas de pantalla de hojas de cálculo con datos agronómicos.....	145
Anexo 4: Capturas de pantalla de código utilizado para obtención de resultados del objetivo 1	146
Anexo 5: Capturas de pantalla de código utilizado para obtención de resultados del objetivo 2	147
Anexo 6: Capturas de pantalla de código utilizado para obtención de resultados del objetivo 3	148

INTRODUCCIÓN

La agricultura representa un pilar fundamental en la economía mundial. El *Zea mays* (maíz) destaca como uno de los cultivos más relevantes a nivel mundial, tanto en términos sociales como económicos. Junto con el arroz y el trigo, contribuye significativamente a la alimentación global siendo el tercer cultivo más importante del mundo (López, Ortigoza, & González, 2019). Estados Unidos y China se erigen como los principales productores, siendo responsables de más del 50 % de la superficie destinada al cultivo de maíz a nivel mundial. China lidera con 42 millones de hectáreas sembradas (FAO, 2018).

En Ecuador, el cultivo de maíz forrajero desempeña un rol crucial en la seguridad alimentaria y la economía del país. La agricultura es un sector esencial en numerosas naciones y su contribución es vital para el bienestar de la población. El maíz forrajero, en particular, se destaca por su importancia en la producción de alimentos para el ganado, lo que contribuye significativamente a la seguridad alimentaria (Guamán et al., 2020). La región costera de Ecuador concentra la mayor producción de maíz. Se estima que se siembran aproximadamente 365 mil hectáreas de maíz en esta zona, con un rendimiento de 4.8 toneladas por hectáreas y la producción anual alcanza los 1.6 millones de toneladas métricas (Caviedes, Carvajal-Larenas, & Zambrano, 2020).

La siembra de *Zea mays*, se caracteriza por sus altos rendimientos en diferentes provincias del país. Manabí lidera la producción, seguido de Los Ríos y Guayas (Cedeño & Javier, 2022). Los Ríos, ocupando el segundo lugar, destaca con el 38 % de la superficie sembrada a nivel nacional, equivalente a 129 mil hectáreas. Además, su contribución a la

producción total del grano seco es de 44 % es decir 648 mil toneladas (INEC, 2020). Sin embargo, los agricultores se enfrentan a desafíos considerables debido a la variabilidad climática y las condiciones agronómicas, que afectan directamente el rendimiento de los cultivos. Las fluctuaciones en la temperatura y las precipitaciones son factores críticos que influyen en la productividad de los cultivos (Navarro-Racines et al., 2021).

En la provincia de Los Ríos, la ciudad de Quevedo está en una región agrícola de Ecuador. El cultivo de maíz forrajero es muy relevante para la comunidad agrícola local. La necesidad de prever de manera precisa y anticipada el rendimiento de este cultivo se ha convertido en un objetivo primordial para garantizar el éxito de las operaciones agrícolas y la seguridad alimentaria de la población. La predicción del rendimiento de los cultivos es esencial para tomar decisiones de gestión estratégica y táctica en la agricultura (PDOT, 2019).

En este contexto, surge la necesidad de desarrollar modelos de pronóstico que empleen técnicas de aprendizaje automático supervisado y variables agronómicas para estimar con precisión el rendimiento del cultivo de maíz forrajero en Quevedo. Este proyecto tiene como objetivo el desarrollo de dichos modelos, con el propósito de brindar a los agricultores de la región una herramienta efectiva para tomar decisiones informadas en sus prácticas agrícolas.

Las metas alcanzables de este proyecto comprenden un análisis minucioso de los datos disponibles, la realización de experimentos para determinar la configuración óptima de

los parámetros de los algoritmos y, finalmente, la selección del modelo que ofrezca los pronósticos más precisos, evaluados mediante métricas como el error cuadrático medio.

Este estudio representa un esfuerzo por integrar el aprendizaje automático para abordar un desafío crucial en el sector agrícola de Quevedo. Además, el trabajo se realizó con el Instituto Nacional de Investigaciones Agropecuarias (INIAP), lo que garantiza el acceso a datos reales y confiables y la posible aplicación de los resultados en la comunidad agrícola local.

En esta investigación, se evaluaron exhaustivamente las características agronómicas del maíz forrajero para optimizar los modelos predictivos de rendimiento utilizando técnicas avanzadas de aprendizaje automático. Se realizó un análisis descriptivo detallado, que reveló la importancia de la distancia de siembra y el diámetro del tallo como factores críticos en el crecimiento de las plantas. La optimización de hiperparámetros de los algoritmos demostró ser crucial, en los resultados obtenidos RF mostró un equilibrio entre precisión y eficiencia en tiempos de ajuste y prueba. Finalmente, la evaluación de modelos basada en las métricas de RMSE y R^2 identificó a RF como el modelo más adecuado para predecir el rendimiento del cultivo de maíz (peso total), destacándose por su alta precisión y buen ajuste a los datos, tanto en entrenamiento como en pruebas. Este estudio subraya la eficacia de combinar análisis agronómico detallado con metodologías estadísticas avanzadas para mejorar la precisión y la aplicabilidad de los modelos predictivos en la agricultura de precisión.

CAPÍTULO I

MARCO CONTEXTUAL DE LA INVESTIGACIÓN

1.1 CONTEXTUALIZACIÓN Y UBICACIÓN DE LA PROBLEMÁTICA

El proyecto de titulación se desarrolló en Quevedo, parte de la provincia de Los Ríos, la segunda provincia del Ecuador con mayor cantidad de área dedicada al cultivo de maíz. El rendimiento por hectárea de cualquier cultivo agrícola sea para el consumo humano o animal tiene gran dependencia con diferentes variables como lo es el clima, las características del suelo o la semilla en sí misma. Cada localidad y siembra tienen cualidades únicas que hacen que un cultivo determinado sea más o menos productivo.

Los modelos estadísticos o de aprendizaje automático no son ampliamente usados por parte de los agricultores de la zona y difícilmente forman parte de la tecnificación en la agricultura, por lo que los agricultores no cuentan con herramientas para pronosticar la producción de los cultivos o de ayuda en la toma de decisiones. Ya que el bajo rendimiento de los cultivos causa pérdidas económicas es imprescindible contar con herramientas que ayuden a disminuir este riesgo.

1.2 SITUACIÓN ACTUAL DE LA PROBLEMÁTICA

El cambio climático, las semillas poco resistentes y la falta de tecnificación en la agricultura son temas de suma importancia para los agricultores y gobiernos ya que causarían la reducción del rendimiento del cultivo de maíz y por ende se tendrían pérdidas económicas y se vería comprometida la seguridad alimentaria del Ecuador.

1.3 PROBLEMA DE INVESTIGACIÓN

1.3.1 Problema General

¿Cómo predecir el rendimiento del cultivo de maíz forrajero en base a variables agronómicas en la ciudad de Quevedo?

1.3.2 Problemas Derivados

¿Cómo determinar que variables agronómicas tienen mayor relación con el rendimiento del maíz forrajero?

¿Cómo determinar la mejor configuración de los parámetros que se necesitan para predecir el rendimiento del maíz forrajero?

¿Cómo pronosticar el rendimiento del cultivo de maíz forrajero con base en las variables agronómicas?

1.4 DELIMITACIÓN DEL PROBLEMA

- **CAMPO:** Ciencias de la Ingeniería y Tecnología, TIC's
- **ÁREA:** Tecnología de la información y la comunicación.
- **LÍNEA:** Ciencia de Datos
- **LUGAR:** Cantón Quevedo
- **TIEMPO:** Octubre 2023 a Julio 2024

1.5 OBJETIVOS

1.5.1 Objetivo General

Desarrollar un modelo de pronóstico utilizando técnicas de aprendizaje automático supervisado y variables agronómicas para determinar el rendimiento de cultivo de maíz forrajero en la ciudad de Quevedo.

1.5.2 Objetivos Específicos

- Realizar un análisis descriptivo de los datos agronómicos del maíz forrajero utilizando herramientas estadísticas para la comprensión de los datos proporcionados por el INIAP.
- Determinar la configuración óptima de hiperparámetros de los algoritmos de aprendizaje automático supervisado mediante la técnica de Grid Search para mejorar la precisión de las predicciones del rendimiento del maíz forrajero.
- Comparar los modelos entrenados para evaluar sus predicciones de rendimiento del maíz forrajero y seleccionar el modelo más preciso.

1.6 JUSTIFICACIÓN

La tecnificación de la agricultura es fundamental para mejorar la competitividad de los agricultores y garantizar la seguridad alimentaria del país. En este sentido, el uso de tecnologías innovadoras puede brindar un soporte invaluable en la toma de decisiones antes, durante y después de la siembra, permitiendo optimizar los procesos agrícolas y maximizar los rendimientos.

La predicción mediante modelos pronósticos precisos y confiables pueden proporcionar valores que se ajustan de manera más cercana a la realidad, permitiendo a los agricultores anticipar y planificar de manera más efectiva sus actividades agrícolas.

Además, la tecnificación agrícola no solo beneficia a los agricultores, sino que también contribuye al desarrollo económico del país en su conjunto. Al mejorar la productividad y eficiencia en el sector agrícola, se generan mayores ingresos y se fomenta el crecimiento económico sostenible. Asimismo, la adopción de tecnologías innovadoras en el agro puede impulsar la creación de empleo y promover la transferencia de conocimientos y habilidades técnicas.

Por ende, la tecnificación de la agricultura, respaldada por modelos de pronósticos y el uso de tecnologías innovadoras, es esencial para brindar a los agricultores mayor competitividad, ingresos y seguridad alimentaria. Estos avances en el sector agrícola no solo benefician a los agricultores, sino que también contribuyen al desarrollo económico del país en su conjunto.

CAPÍTULO II
MARCO TEÓRICO

2.1. FUNDAMENTACIÓN CONCEPTUAL

2.1.1. Agricultura

La agricultura es una actividad económica crucial que representa aproximadamente el 3 % del Producto Interno Bruto (PIB) global y es esencial para la producción de alimentos. Sin embargo, las prácticas agrícolas son muy sensibles a las condiciones climáticas cambiantes, lo que puede afectar negativamente el crecimiento de las plantas y la producción de cultivos. El cambio climático ha causado cambios en los patrones de precipitación, aumento de la temperatura, fenómenos meteorológicos extremos y aumento en la concentración de carbono en la atmósfera, lo que afecta aún más la agricultura al modificar las dinámicas de crecimiento de las plantas y aumentar los problemas con plagas y enfermedades en los cultivos (Salas & Maldonado, 2020).

El crecimiento, desarrollo, rendimiento y calidad de los cultivos están fuertemente influenciados por variables climáticas como la temperatura, radiación solar y precipitación. Estas variables son fundamentales para determinar las épocas de siembra, la aplicación de fertilizantes y productos fitosanitarios, así como la planificación del riego. Se cuenta con pruebas que respaldan la que las modificaciones en los patrones de precipitación han ocasionado una disminución en la productividad de cultivos como el maíz, arroz, trigo, papa y hortalizas. Este decrecimiento persistirá y se prevé que sea considerable de acuerdo con las estimaciones hasta el año 2050. el maíz y la caña de azúcar aumentarían su tasa fotosintética hasta en un 20 % (Villagran Munar et al., 2023).

2.1.2. Maíz Forrajero

El maíz forrajero es una variedad específica de *Z. mays* (maíz) que se cultiva principalmente con el propósito de ser utilizado como alimento para el ganado, vacuno, ovino, caprino y otros animales de granja. El objetivo es obtener una planta que sea rica en materia vegetal, hojas y tallos que son adecuados para la alimentación animal (Granados Niño et al., 2022).

El potencial de rendimiento y calidad del maíz forrajero pueden ser modificados por el efecto que tiene el ambiente sobre la fisiología de la planta. Este cultivo muestra una alta tasa fotosintética al incrementar la temperatura e intensidad de luz. Si existe escasez de radiación solar en la planta de maíz desde que el cultivo tiene una cobertura total hasta su etapa reproductiva, el rendimiento de forraje por hectárea será limitado. Entonces, la temperatura ambiental es un factor clave para el crecimiento y desarrollo del maíz forrajero. Todo esto es debido a que la temperatura promueve la fotosíntesis, la respiración y el crecimiento celular en las plantas (Granados et al., 2022).

2.1.3. Fenología del Maíz

Los primeros estudios fenológicos en maíz estuvieron basados en fenómenos visibles externos como la aparición de hojas. La fenología del maíz se refiere al estudio de las etapas de crecimiento y los procesos biológicos de los cultivos de maíz, como la emergencia, la floración y la senescencia, que están influenciados por las condiciones ambientales (Guo et al., 2022). La capacidad de detectar de manera confiable la fenología

de los cultivos y su variabilidad espacio-temporal es fundamental para los agricultores, los formuladores de políticas y las agencias gubernamentales, ya que tiene implicaciones para toda la cadena alimentaria (Nieto et al., 2022).

La fenología del maíz se enfoca en el ritmo de desarrollo de la planta, considerando los cambios morfológicos y fisiológicos que ocurren en respuesta a las condiciones ambientales. Una de las formas más comunes de estudiar cómo el cultivo de maíz se desarrolla en respuesta a las condiciones ambientales es utilizando la escala de Ritchie. Esta escala se utiliza para resumir el ciclo de crecimiento del maíz y se divide en dos grandes períodos: el vegetativo y el reproductivo (Nieto et al., 2022).

- El período vegetativo del maíz se inicia con la emergencia de las plántulas, abreviado como VE, y se extiende hasta la formación del panojo, identificado como VT. Durante esta fase, se observa el crecimiento de nuevas hojas, que se numeran como Vn, donde "n" indica el número de hojas visibles en la planta (Beserra, Dias, & Parra, 2002).
- Por otro lado, el período reproductivo del maíz comienza con la floración femenina, conocida como R1, que se caracteriza por la aparición de estigmas en las espigas, también denominada "silking" en inglés. Este período culmina cuando los granos de maíz alcanzan la madurez fisiológica, marcada por la formación de una capa negra en la base de los granos en el centro de la espiga. Esta capa se produce debido a la necrosis de los haces vasculares que conectan los granos con el marlo (Zeng et al., 2016).

En resumen, la escala de Ritchie es una valiosa herramienta para monitorear el progreso del cultivo de maíz, estableciendo una clara distinción entre dos fases principales: la fase vegetativa, caracterizada por el crecimiento de las hojas, y la fase reproductiva, que abarca desde la floración femenina hasta la madurez fisiológica de los granos. Este enfoque proporciona una guía crucial para los agricultores al tomar decisiones críticas en la gestión de sus cultivos (Zeng et al., 2016).

2.1.3.1. Subdivisión de las Etapas Vegetativa y Reproductiva

(Koca & Erekul, 2016) mencionan que las etapas vegetativas y reproductiva se subdividen en cuatro fases clave que son: siembra-emergencia, emergencia-panoja, panoja-espiga y espiga-maduración, lo que proporciona una visión más detallada de la evolución del cultivo.

- La primera etapa, conocida como el establecimiento del cultivo, abarca desde la siembra inicial hasta que el cultivo se ha arraigado en el campo, cubriendo aproximadamente el 10 % de la superficie.
- La segunda etapa se denomina período de rápido desarrollo del cultivo, que sigue inmediatamente a la etapa de establecimiento y continúa hasta que el cultivo cubre la mayoría de la superficie del suelo, al menos un 70-80 %.
- La tercera etapa, conocida como la etapa de mediados de temporada, se extiende desde el final del rápido desarrollo hasta el inicio de la maduración del cultivo.

Esta fase se caracteriza por el envejecimiento de las hojas y el máximo consumo de agua.

- La cuarta y última etapa corresponde a la maduración y cosecha, culminando con la maduración completa del cultivo y una disminución gradual en el consumo de agua. Esta etapa desempeña un papel crucial en la planificación de riegos y la gestión agrícola.

La fenología de los cultivos ofrece información crucial sobre las diferentes etapas de crecimiento de las plantas, desempeñando un papel esencial en la gestión de los agroecosistemas y la evaluación de los rendimientos agrícolas. Sin embargo, la agricultura de precisión demanda datos fenológicos más detallados de los cultivos para lograr una gestión más precisa de los agroecosistemas. Esta necesidad puede abordarse mediante observaciones a largo plazo y de alta resolución proporcionadas por satélites como Landsat. La fenología, al centrarse en el estudio de las etapas de desarrollo de las plantas y su relación con las condiciones climáticas y ambientales, se convierte en un componente fundamental en la toma de decisiones agrícolas, permitiendo a los agricultores elegir el momento adecuado para la siembra, cosecha y la aplicación de insumos agrícolas (Niu et al., 2022).

2.1.3.2. *Requerimientos Edafoclimáticos*

Hasang-Moran et al. (2021) señalan que los requerimientos edafoclimáticos propicios para los cultivos de maíz en Ecuador pueden variar según la región. A continuación, se mencionan algunos aspectos importantes:

- **Temperatura:** El maíz es un cultivo que se adapta a una amplia gama de temperaturas. En general, se desarrolla bien en temperaturas entre 24°C a 28°C. Sin embargo, es importante tener en cuenta que las variedades de maíz pueden tener diferentes tolerancias a temperaturas extremas, por lo que es recomendable seleccionar variedades adecuadas para las condiciones específicas de cada región.
- **Precipitación:** El maíz requiere una cantidad adecuada de agua para su crecimiento y desarrollo. En general, se recomienda un rango de precipitación entre 550 mm y 2000 mm/ año.
- **pH:** El suelo debe encontrarse en el rango de 5.5 a 7.3 para un óptimo crecimiento del maíz. Además, es importante asegurarse de que el suelo tenga una adecuada disponibilidad de nutrientes, especialmente nitrógeno, fósforo y potasio.
- **Radiación solar:** Presentan alta incidencia de luz solar, entre 750 a 1000 Horas luz/año.

Estos son solo algunos de los requerimientos edafoclimáticos generales para el cultivo de maíz en Ecuador. Es importante tener en cuenta que las condiciones específicas pueden variar según la región y es recomendable consultar con expertos locales o instituciones agrícolas para obtener información más precisa y adaptada a cada zona de cultivo en Ecuador (Hasang-Moran et al., 2021).

2.1.4. Agricultura de Precisión

Las tecnologías agrícolas de precisión (AP), como el sistema de posicionamiento global, han estado en el mercado desde la década de 1990 y se considera que ofrecen ventajas tanto ambientales como económicas (Gardezi & Bronson, 2020). La agricultura de precisión está transformando la gestión agrícola a nivel mundial mediante la incorporación de tecnologías avanzadas como el sistema global de navegación por satélite (GNSS), la robótica, sensores y aplicaciones de tasa variable. Herramientas como los sistemas de información geográfica (SIG) y la teledetección, por ende, desempeñan un papel esencial al proporcionar una amplia gama de datos, incluyendo la distribución de cultivos, el estado de salud de las plantas y el seguimiento de las diferentes etapas de crecimiento de los cultivos (Furukawa, Maruyama, Saito, & Kaneko, 2020).

La agricultura de precisión, también llamada agricultura de sitio específico se enfoca en la gestión eficiente de recursos agrícolas para satisfacer las necesidades específicas de los cultivos, aprovechando avances tecnológicos para abordar las condiciones particulares de cada campo. Este enfoque permite la gestión más precisa de nutrientes, agua y cultivos, lo que conduce a una reducción de la variabilidad en el campo y, en última instancia, a una disminución del uso de fertilizantes y un aumento en la rentabilidad (Kumari & Bhavanishankar, 2023).

Además, las (AP) busca maximizar los recursos con el objetivo de minimizar pérdidas y costos de insumos. Utiliza tecnologías de información para recopilar datos del terreno, lo que permite a los agricultores identificar variaciones temporales y espaciales en los

recursos de producción, facilitando la aplicación precisa de tratamientos (Cisternas, Velásquez, Caro, & Rodríguez, 2020).

2.1.5. Beneficios Potenciales de la Agricultura de Precisión

Las ventajas de la agricultura de precisión incluyen la reducción de costos al aplicar fertilizantes solo en áreas necesarias, lo que se basa en muestreos de suelo y análisis de datos de rendimiento. También mejora la gestión de los recursos hídricos y optimiza el rendimiento a través de prácticas de cosecha automatizadas. La implementación de tecnologías de la información contribuye a una toma de decisiones más efectiva en términos de gestión técnico-productiva, económica y ambiental (Cisternas et al., 2020).

2.1.6. Rendimiento de Cultivos

La predicción del rendimiento de los cultivos es una tarea compleja influenciada por diversos factores, como las condiciones climáticas, la calidad del suelo y el tipo de cultivo. Para abordar este desafío, se utilizan modelos de aprendizaje automático, como el Árbol de Decisión, Random Forest, la Regresión Lineal y las Redes Neuronales, basándose en datos históricos y condiciones ambientales (Raskar, Abhang, Pethe, & Attar, 2022).

En el contexto de la agricultura de precisión, se emplean técnicas de monitoreo, sistemas de información de gestión y tecnologías de tasa variable para mejorar el rendimiento de los cultivos y reducir el impacto ambiental. Además, las simulaciones del rendimiento de

los cultivos permiten comprender cómo las deficiencias de agua, nutrientes y las condiciones del campo afectan el crecimiento de las plantas durante la temporada (Nyéki & Neményi, 2022).

La tasa de rendimiento de los cultivos depende de los parámetros del suelo, geografía del área, el tipo de suelo, los nutrientes y las condiciones climáticas o meteorológicas. Además, el rendimiento de forraje en híbridos maíz se encuentra fuertemente correlacionado con características específicas de la planta, como el diámetro del tallo, el peso de las mazorcas, el número de granos por hilera, la longitud de las mazorcas, los días que transcurridos hasta que aparezcan los estigmas en las flores y los días hasta que el cultivo alcanzaba la madurez fisiológica. Esto proporciona información valiosa para la optimización de la producción agrícola (Mahore & Bardekar, 2021).

Las relaciones entre el clima y el rendimiento pueden cambiar con el tiempo debido a la influencia del calentamiento global y las intervenciones humanas. Los cambios simultáneos en la relación precipitación-rendimiento y temperatura-rendimiento podrían plantear desafíos para el desarrollo de medidas de adaptación para la planificación y gestión agrícola (Feng, Hao, Zhang, & Hao, 2021). El rendimiento se ve influenciado por el clima y los insumos aplicados durante la temporada de cultivo, y la modelización de estos factores proporciona una herramienta para predecir el rendimiento antes de la cosecha. Para comprender los impactos del cambio climático en los cultivos, se utilizan modelos estadísticos. Las predicciones de estos modelos pueden variar de acuerdo con las condiciones climáticas y su impacto en los cultivos (Mahore & Bardekar, 2021).

2.1.7. Cross Industry Standard Process for Data Mining

Cross Industry Standard Process for Data Mining (CRISP-DM) es una metodología utilizada para ejecutar proyectos de ciencia de datos donde se apliquen algoritmos de aprendizaje automático ya que proporciona un flujo de trabajo fácil de entender (Saltz, 2021; Stühler, Klau, Zöller, Beiderwellen-Bedrikow, & Tutschku, 2024; Steinberg, Burggaef, Wagner, & Heinbach, 2022; Sierra-Forero, Baron-Velandia, & Vanegas-Ayala, 2024).

Schröer, Kruse, & Gómez (2021) indica que a metodología CRISP-DM consta de seis fases iterativas

- **Comprensión del negocio:** Debe evaluarse la situación de la empresa para obtener una visión general de los recursos disponibles y necesarios.
- **Comprensión de los datos:** Recopilar datos de fuentes de datos, explorarlos y describirlos y comprobar la calidad de los datos son tareas esenciales en esta fase.
- **Preparación de los datos:** La selección de datos debe realizarse definiendo criterios de inclusión y exclusión.
- **Modelado:** La fase de modelización de datos consiste en seleccionar la técnica de modelización, construir el caso de prueba y el modelo.

- **Evaluación:** En esta fase se contrastan los resultados pueden ser las métricas de evaluación de los modelos con los objetivos empresariales definidos.
- **Despliegue:** La fase de despliegue puede tratarse de un informe final o de un componente de software.

2.1.8. Aprendizaje Automático Supervisado

El aprendizaje automático supervisado se destaca como una de las categorías más comunes y esenciales en el ámbito del aprendizaje automático. En este enfoque, se capacita a un modelo utilizando un conjunto de datos de entrenamiento que está etiquetado, lo que significa que cada punto de datos en este conjunto de entrenamiento está vinculado a una etiqueta o resultado conocido. El objetivo principal es permitir que el modelo aprenda a relacionar de manera efectiva las características de entrada (variables) con las etiquetas o resultados correctos, de modo que pueda realizar predicciones precisas en nuevos datos no etiquetados (Nokeri, 2022).

Además, presenta diversas aplicaciones, como la clasificación, la regresión (predicción de valores numéricos) lo que permite automatizar tareas basadas en datos y mejorar la toma de decisiones. Kumar (2019) considera que las características del aprendizaje automático supervisado son:

- **Características de Entrada con las Etiquetas de Salida:** Esto se logra mediante diversos algoritmos, como máquinas de soporte vectorial, regresión logística y redes neuronales.
- **Evaluación del Modelo:** Una vez entrenado, se evalúa el modelo utilizando un conjunto de datos de prueba que no se usó en el entrenamiento. Se comparan las predicciones del modelo con las etiquetas reales para medir su rendimiento, utilizando métricas como la precisión o el error cuadrático medio.
- **Predicciones en datos Nuevos:** Después de la evaluación y ajustes necesarios, el modelo se puede utilizar para realizar predicciones en nuevos datos no etiquetados.

2.1.9. Regresión

Es una de las técnicas en el aprendizaje automático supervisado, se emplea para modelar y analizar las variables. La regresión es el proceso de estimar un valor de salida basándose en múltiples factores. A diferencia de la clasificación, donde el valor de salida es discreto, en la regresión es continuo, y son aplicables los algoritmos de aprendizaje supervisado. Y dado que las entradas se presentan en diversas escalas, es necesario normalizar los valores de entrada como de salida entre cero y uno, y posteriormente desnormalizar el valor de salida generado por el proceso de regresión a su escala original. Para Taejo, (2021), se distinguen dos tipos de regresión:

- **Regresión Univariante:** Estima un solo valor de salida y es común en la aproximación de funciones no lineales con una sola variable de entrada, y en la predicción de series del tiempo.
- **Regresión Multivariante:** Genera múltiples valores de salida a partir de datos de entrada proporcionados como un vector, aplicándose en contexto como la predicción de múltiples precios de acciones o tasas de cambio.

2.1.10. Regresión Lineal

Según Abbas, Afzaal, Farooque, & Tang, (2020) la regresión lineal es un método estadístico diseñado para modelar la relación entre una variable dependiente y una o más variables independientes. Utiliza la ecuación para predecir los valores de y basándose en x , donde a es el intercepto con el eje b y la pendiente de la línea, y se define por la ecuación 1:

$$y = a + bx$$

Ecuación 1: Regresión lineal

En el aprendizaje automático, se busca minimizar las diferencias entre los valores reales y los predichos, a menudo a través del método de mínimos cuadrados ajustando a y b para reducir esta pérdida, típicamente medida por RMSE o MSE. Este proceso se optimiza con algoritmos como el descenso de la gradiente, mejorando la precisión de las predicciones (Abbas et al., 2020).

2.1.11. Support Vector Machines

Se trata de un clasificador dentro del aprendizaje supervisado, conocido como red de vectores de soporte. Está basado en la minimización del riesgo estructural que se utiliza para maximizar la precisión de las predicciones y reducir el problema del sobreajuste. Su importancia radica que realiza de manera eficiente la clasificación de problemas lineales como no lineales, la idea básica de SVM es generar un hiperplano que separe los datos en clases (Madhu et al., 2021).

2.1.12. Random Forest

Random forests es una técnica de machine learning que se usa para solucionar problemas de clasificación y regresión. Este algoritmo se fundamenta en la construcción de un ensamble de árboles de decisión a través de la técnica de agregación por Bootstrap, también conocida como baggin. La estrategia consiste en generar múltiples submuestras del conjunto de datos de manera aleatoria con reemplazo, y construir un árbol de decisión para cada una de estas, Luego las predicciones se promedian en el caso de regresión, o se toma la moda, en el caso de clasificación, para obtener una predicción precisa. Este método mejora la predicción al reducir la varianza y el sesgo, disminuye el riesgo de sobreajuste (Dhaliwal & Williams, 2024).

2.1.13. Gradient-Boosting Regressor

Los árboles de decisión con gradiente boosting (GBDT) se han utilizado ampliamente en El aprendizaje automático. Esta técnica optimiza una función de perdida diferenciable arbitraria mediante un enfoque aditivo por etapas en la construcción del modelo. Cada

iteración en el proceso de construcción del modelo implica ajustar un árbol de clasificación y regresión (CART) sobre el gradiente negativo (es decir sobre el error residual entre la estimación y la salida del objetivo) de una función de pérdida arbitraria. Además, produce resultados altamente competitivos especialmente en modelar datos, los dos principales hiperparámetros para un GBR óptimo, son el número de etapas boosting y el parámetro de contracción, también conocido como tasa de aprendizaje (Mustapha et al., 2024).

Mustapha et al., (2024) indica que el modelo se inicia con un valor constante γ que minimiza la pérdida sobre todas las muestras como en la Ecuación 2 detalla:

$$F^0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^n L(y_i \gamma)$$

Ecuación 2: Función de riesgo empírico minimizado

Mustapha et al., (2024) indica que esto es seguido por varias iteraciones de cálculo de gradiente negativo de la función de pérdida y su uso subsecuente para ajustar un árbol de decisión y la adición de un nuevo modelo al conjunto, como se indica en la ecuación 3:

$$F(x) = F^{m-1}(x) + v f^m(x)$$

Ecuación 3: Método de Boosting

2.1.14. XGBoost

Tarwidi, Pudjaprasetya, Adytia, & Apri, (2023) menciona que XGBoost es una implementación de gradiente boosting, o variante optimizada que combina las predicciones de varios árboles de clasificación y regresión (CART) “débiles” para desarrollar un modelo “fuerte” utilizando estrategias de entrenamiento aditivo. Además, es conocido por prevenir sobreajustes de manera eficiente a través de función objetivo simplificada que combina la perdida y los términos de regularización. Como se muestra en la siguiente ecuación 4:

$$Obj = \sum_m^n l(y_m, \hat{Y}_m) + \sum_k^k \Omega(f_k)$$

Ecuación 4: Gradient Boosting

Según, Tarwidi et al., (2023) l es la función de perdida que mide la diferencia entre la salida experimental, y_m , y la salida estimada $\hat{Y}_m$; Ω es el termino de regularización dado por la ecuación 5:

$$\Omega(f) = \gamma^T + \frac{1}{2} \lambda \sum_{i=1}^T w^2$$

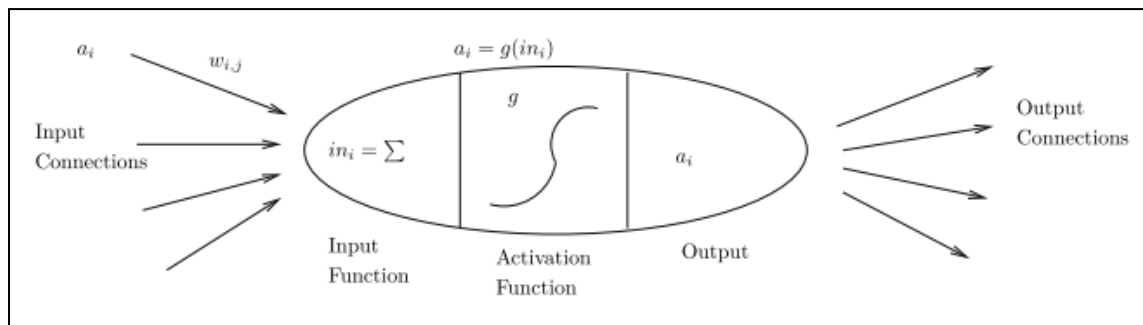
Ecuación 5: Función objetivo de XGBoost

Donde T es el número de hojas y w es el puntaje en cada hoja; γ y λ son constantes para controlar el grado de regularización. XGBoost es adecuado para todo tipo de problemas de aprendizaje supervisado (Tarwidi et al., 2023).

2.1.15. Redes Neuronales Artificiales

Una red de neuronas artificiales está compuesta de un conjunto de unidades de procesamiento (neuronas) que están conectadas entre sí, con cada de estas conexiones tienen un peso asociado (conexión ponderada). Cada de estas unidades tiene un nivel de activación y realiza un cálculo lineal que provee medios para actualizar este nivel de activación (Torgo, 2017). En la Ilustración 1 se representa una neurona artificial.

Ilustración 1: Representación gráfica de una neurona artificial.



FUENTE: Torgo (2017).

El cálculo lineal de la función de entrada es esencialmente una suma ponderada de las entradas a la neurona como lo detalla la ecuación X, donde k es el número de neuronas entrantes a la neurona i , $w_{x,y}$ es el peso de la conexión entre las neuronas x y la neurona y , y a_r es la salida de la neurona r (Torgo, 2017).

$$in_i = \sum_{j=1}^k [(w_{j,y})(a_j)]$$

Ecuación 6: Función de entrada de una neurona artificial.

El cálculo no lineal en la neurona i es aplicado al valor de in_i . Ésta consiste en una función de activación que podría determinar la salida de la neurona i . Según la función de activación es el comportamiento de la red de neuronas (Torgo, 2017). Entre las funciones de activación más comunes indicadas en la Ecuación 7, se tiene a:

✓ Función de Sigmoid: $sigmoid(x) = \frac{1}{1 + exp^{-x}}$

✓ Función ReLU: $ReLU(z) = \max(0, z)$

✓ Función de tanh: $f(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$

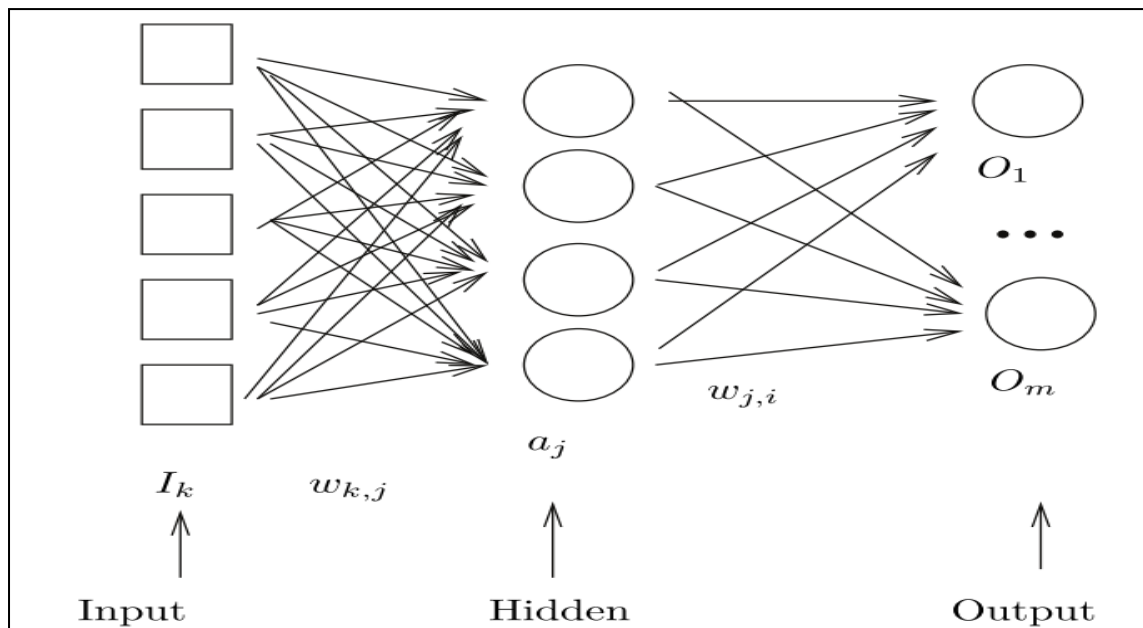
Ecuación 7: Funciones de activación de redes neuronales

Existen diferentes clases de algoritmos que utilizan este método como medio de aprendizaje, utilizando una estructura simple de una sola capa, o múltiples capas (Bishop, 1995). Las redes neuronales de una sola capa son un conjunto de neuronas que están conectados unas con otras (Torgo, 2017). En cambio con una red de multicapas están estructuradas por sucesivas capas que contienen unidades de procesamiento, con conexiones que corren desde todas las unidades de una capa a todas las unidades de la capa siguiente, pero sin otras conexiones permitidas (Bishop, 1995).

Una red neuronal multicapa puede ser según la dirección de estas conexiones, de dos tipos: feed-forward (por su nombre en inglés) y recurrentes. La primera consiste en conexiones unidireccionales entre las neuronas desde las entradas hasta las neuronas de salida (conexiones no cíclicas). Las de tipo recurrente en cambio podrían permitir conexiones arbitrarias entre neuronas (conexiones cíclicas) (Torgo, 2017).

Dentro de una red neuronal multicapa existen tres tipos de capas: la capa de entrada, la capa oculta y la capa de salida. La capa de entrada está compuesta por tantas neuronas como atributos predictores tenga el problema. La capa de salida en el caso de un problema de regresión estaría conformada por una sola neurona y en el caso de una clasificación tendría tantas neuronas como clases. Entre estas dos capas anteriores esta la capa oculta, la arquitectura más común es la de una sola capa oculta (Torgo, 2017). En la Ilustración 2 se muestra una arquitectura de una red multicapa no cíclica (feed-forward).

Ilustración 2: Arquitectura de una red neuronal multicapa no cíclica.



FUENTE: Torgo (2017).

2.2. FUNDAMENTACIÓN TEÓRICA

En el estudio sobre la predicción del rendimiento de cultivos agrícolas mediante el uso de algoritmo de aprendizaje automático, se evaluaron varios modelos para predecir dos variables principales: hectáreas cosechadas y producción en toneladas, los modelos

aplicados incluyeron Redes Neuronales Densas (DNN), Random Forest (RF), Árboles de Decisiones (DT) y Maquinas de Soporte Vectorial (SVM). Para la predicción de hectáreas cosechadas, la arquitectura DNN presentó mejor resultado con un MSE de 0.0081, seguido cerca por RF con 0.0104. En contraste, al predecir la producción en toneladas, RF demostró ser superior con MSE de 0.0550, mostrando una notable capacidad predictiva del 95 % (García-Arteaga, Zambrano-Zambrano, Alcivar-Cevallos, & Romero, 2020).

En la investigación enfocado en la predicción del rendimiento de grano de maíz, se propusieron varios modelos estadísticos. La base de datos se construyó con información sobre los rendimientos máximos de grano de maíz, incluyendo datos sobre rendimientos máximos del grano, genética y manejo agronómico. La investigación concluyó que los factores agronómicos tienen un impacto significativo en el rendimiento, y el modelo propuesto que combinó las variables agronómicas tuvo un coeficiente de determinación (R^2 de 0.73) la cual pudo representar la influencia de estos factores, reflejando el menor error estándar, el cual demostró ser el más efectivo (Reyes, Galvis, Almaraz, & Hernández-Mendoza, 2021).

En el estudio predicción del rendimiento de los cultivos mediante sensores y algoritmos de aprendizaje automático, se evaluaron cuatro algoritmos, que fueron; regresión lineal (LR), red elástica (EN), y regresión vectorial de soporte (SVM), para predecir el rendimiento de tubérculos a partir de suelo y cultivos recolectados mediante sensores proximales. Entre estos, el modelo SVM, superó a los demás en todos los conjuntos de datos evaluados con un RMSE de 5.97 y 4.62 t/ha. Por eso es importante combinar

técnicas de detección proximal con algoritmos de aprendizaje automático para mejorar predicciones del rendimiento de cultivos en agricultura de precisión (Abbas et al., 2020). La investigación sobre predicción del rendimiento de los cultivos mediante aprendizaje automático, se realizaron dos revisiones sistemáticas de la literatura para identificar cuáles son los más empleados. La primera revisión se enfocó en estudios basados en aprendizajes automáticos convencional, la cual reflejó que el algoritmo más utilizado es el de redes neuronales artificiales, y en aprendizaje profundo fue; las redes neuronales convolucionales (CNN), seguido por la memoria a corto y largo plazo (LSTM) y las redes neuronales profundas (DNN). Esta amplia revisión evidencia la evolución y diversidad de métodos y características empleadas para la predicción del rendimiento de cultivos (Van, Kassahun, & Catal, 2020).

En el estudio; comparación de modelos de aprendizaje automático la para la predicción del rendimiento agrícola. Se compararon los algoritmos de máquina de soporte vectoriales (SVM), modelos ensamblados (Random Forest, Gradient Boosting) y el modelo de regresión LASSO. El modelo Gradient Boosting en términos de precisión demostró ser mejor ofreciendo precisión en las predicciones con los menores valores de Error Medio Absoluto (MAE) y Raíz de Error Cuadrático Medio (RMSE), y el mayor coeficiente de determinación R^2 (Lamos-Díaz, Puentes-Garzón, & Zarate-Caicedo, 2020).

2.3. FUNDAMENTACIÓN LEGAL

Para comprender la estructura técnico-legal que respalda el sector agrícola en Ecuador, es fundamental conocer la estructura actual, regulación y marco legal del país, incluyendo tanto

la Ley Orgánica del Régimen de la Soberanía Alimentaria como la Constitución. La Ley Orgánica del Régimen de la Soberanía Alimentaria establece los lineamientos para garantizar la seguridad y soberanía alimentaria en el país, promoviendo la producción agrícola sostenible y la protección de los recursos naturales. Estas leyes brindan el marco legal y normativo necesario para respaldar el sector agrícola en Ecuador, impulsando la producción sostenible, la seguridad alimentaria y el derecho a una alimentación adecuada (Lexis S.A, 2023).

2.3.1. Constitución de la República

La Constitución de la República del Ecuador (2021), establece dentro del marco legal diversos principios y derechos que son fundamentales:

Capítulo Segundo: Derechos del buen vivir - Sección segunda: Ambiente sano

Art. 14.- Se reconoce el derecho de la población a vivir en un ambiente sano y ecológicamente equilibrado, que garantice la sostenibilidad y el buen vivir, *sumak kawsay*.

Título VI Régimen de Desarrollo - Capítulo Tercero: Soberanía Alimentaria

Art. 281.- La soberanía alimentaria constituye un objetivo estratégico y una obligación del Estado para garantizar que las personas, comunidades, pueblos y nacionalidades alcancen la autosuficiencia de alimentos sanos y culturalmente apropiado de forma permanente. Para ello, será responsabilidad del Estado:

Asegurar el desarrollo de la investigación científica y de la innovación tecnológica apropiadas para garantizar la soberanía alimentaria.

Fortalecer el desarrollo de organizaciones y redes de productores y de consumidores, así como las de comercialización y distribución de alimentos que promueva la equidad entre espacios rurales y urbanos.

Sección Octava Ciencia tecnología, Innovación y Saberes Ancestrales

Art. 387.- Será responsabilidad del Estado:

Facilitar e impulsar la incorporación a la sociedad del conocimiento para alcanzar los objetivos del régimen de desarrollo.

Promover la generación y producción de conocimiento, fomentar la investigación científica y tecnológica, y potenciar los saberes ancestrales, para así contribuir a la realización del buen vivir, al sumak kawsay.

2.3.2. Ley Orgánica del Régimen de la Soberanía Alimentaria

La Ley Orgánica del Régimen de la Soberanía Alimentaria (2010) promueve la producción sostenible y sustentable de alimentos, esta ley busca fomentar la investigación y el desarrollo de tecnologías agrícolas que mejoren la productividad y la resiliencia de

los cultivos frente a condiciones climáticas adversas (FAO, 2023). La Ley Orgánica de la Soberanía Alimentaria (LORSA) fue aprobada en marzo de 2009, dentro del marco legal se contempla lo siguiente:

El pleno de la comisión legislativa y de fiscalización considerando:

Que, el **Art. 133**, numeral 2 de la Constitución de la República, establece que las leyes orgánicas son las que regulan el ejercicio de los derechos y garantías constitucionales y que, en concordancia, el Art. 13 de la Constitución consagra que las personas y colectividades tienen derecho al acceso seguro y permanente a alimentos sanos, suficientes y nutritivos, para lo cual el Estado promoverá la soberanía alimentaria.

Título I - Capítulo III: Investigación, asistencia técnica y diálogo de saberes

Art. 9.- Investigación y extensión para la soberanía alimentaria. - El Estado asegurará y desarrollará la investigación científica y tecnológica en materia agroalimentaria, que tendrá por objeto mejorar la calidad nutricional de los alimentos, la productividad, la sanidad alimentaria, así como proteger y enriquecer la agrobiodiversidad.

Además, asegurará la investigación aplicada y participativa y la creación de un sistema de extensión, que transferirá la tecnología generada en la investigación, a fin de proporcionar una asistencia técnica, sustentada en un diálogo e intercambio de saberes con los pequeños y medianos productores, valorando el conocimiento de mujeres y hombres.

CAPÍTULO III
METODOLOGÍA DE LA INVESTIGACIÓN

3.1. TIPOS DE INVESTIGACIÓN

Para cumplir con los objetivos planteados, se emplearon los tipos de investigación documental y correlacional.

3.1.1. Investigación Documental

Fue útil para conocer los diversos enfoques que se han aplicado al pronóstico del rendimiento de cultivos de maíz usando aprendizaje automático. Así como también, que algoritmos obtuvieron los mejores resultados aplicado a este problema de predicción de rendimiento de cultivos de maíz forrajero.

3.1.2. Investigación Correlacional

Se centró en analizar las relaciones entre variables agronómicas y el rendimiento del maíz forrajero, en como estos distintos factores se relacionan entre sí. Mediante el uso de técnicas estadísticas avanzadas.

3.2. MÉTODO DE INVESTIGACIÓN

3.2.1. Método Exploratorio

El método de investigación exploratorio se utilizó para investigar las posibles relaciones entre las variables agronómicas y sus efectos en el rendimiento del maíz forrajero. Este

enfoque permitió identificar y comprender los factores clave que influyen en la calidad de las predicciones de los modelos construidos con base a los datos proporcionados por el INIAP.

3.3. CONSTRUCCIÓN METODOLÓGICA DEL OBJETO DE INVESTIGACIÓN

3.3.1. Técnicas de Investigación

Para el presente proyecto de investigación se aplicaron técnicas para el análisis y modelado de los datos, detallados a continuación:

Estadística Descriptiva: Se realizó un análisis exploratorio para comprender las características fundamentales y las relaciones entre variables. Para ello se realizó un análisis univariante para explorar las características individuales de las variables para comprender y obtener medidas de tendencia central y dispersión, como media, mediana, rango, varianza y desviación estándar. Paralelamente, el análisis bivariante evaluó las relaciones entre pares de variables utilizando técnicas como el coeficiente de correlación de Spearman y gráficos de dispersión.

Aprendizaje de Automático Supervisado: Se aplicaron técnicas de aprendizaje automático para predecir el rendimiento del cultivo del maíz forrajero en base a los datos previamente preprocesados. También se refinaron los modelos mediante ajustes iterativos de hiperparámetros, buscando la configuración que resultaron en las predicciones más precisas y fiables. Este proceso involucró iterar sobre múltiples combinaciones de

parámetros para cada algoritmo, evaluando su efecto en la precisión de las predicciones mediante las métricas de evaluación seleccionadas.

Visualización de Datos: Se emplearon diversas técnicas para facilitar el análisis e interpretación de los resultados. Se utilizaron gráficos de líneas para ilustrar tendencias y cambios y en las variables agronómicas a lo largo del tiempo. Las gráficas de barra proporcionaron una comparación entre diferentes grupos o categorías. Igualmente, se empleó gráficos de distribución que ayudaron a visualizar la dispersión de los valores de las variables. Finalmente, los mapas de calor se emplearon para representar la matriz de correlación de Spearman entre variables, destacando las relaciones existentes mediante gradientes de color.

3.3.2. Instrumentos de Investigación

Para aplicar las técnicas utilizadas para el pronóstico del rendimiento de cultivo de maíz forrajero basado en variables agronómicas, fueron necesarios los siguientes instrumentos:

- **Python:** Se seleccionó este lenguaje de programación por su amplia aplicabilidad en el análisis de datos, aprendizaje automático, el desarrollo web y la informática científica. Además, proporciona un amplio conjunto de bibliotecas y herramientas que lo hacen ideal para tareas de análisis de datos y aprendizaje automático. Las principales bibliotecas utilizadas fueron jupyter, pandas y numpy. El computador utilizado en el que se realizó el procesamiento y entrenamiento de los modelos

tenía Python 3.12.2, una CPU Intel i7-870H@2.2Ghz, RAM de 32 GB DDR4 de 2667 MHz, GPU Nvidia GTX 1050 y disco SSD Nvme de 500 Gb.

- **Seaborn:** Se utilizó esta biblioteca de visualización de datos en Python para crear gráficos estadísticos atractivos e informativos. Basada en Matplotlib, proporcionó una interfaz de alto nivel esencial para la exploración visual de complejas relaciones entre variables, identificación patrones y tendencias estadísticas dentro del conjunto de datos analizados.
- **Scikit-learn:** Esta biblioteca representó una herramienta esencial en la fase de modelado predictivo de la investigación. Se empleó para la implementación y evaluación de algoritmos de aprendizajes automáticos, facilitando la optimización de hiperparámetros, la selección del modelo y la validación de los resultados predictivos.

3.4. ELABORACIÓN DEL MARCO TEÓRICO

Se realizó la búsqueda de información referente a la investigación en documentos científicos como libros, artículos, tesis, etc. Para ello se siguieron los siguientes pasos:

- Se usaron las siguientes palabras claves para la búsqueda tanto en español como en inglés: Maíz, Cultivos, Agricultura de Precisión, Aprendizaje Automático, Pronóstico de Rendimiento y Regresión.

- Estas palabras claves se buscaron en Google Scholar y las bases de datos Scopus y SciELO.
- Se filtró y se analizó la información encontrada para poder escoger los algoritmos de aprendizaje automático a utilizar para la predicción del rendimiento del maíz forrajero, esto se lo realizó dependiendo del ámbito de la investigación, la relevancia de sus resultados y que hayan sido de los últimos 5 años.
- De la información recopilada se agregó a la fundamentación conceptual, el marco referencial, la fundamentación legal y sirvió como fuente fiable para la selección de las métricas de evaluación y de los algoritmos a utilizar.

3.5. RECOLECCIÓN DE LA INFORMACIÓN

La información primaria fue proporcionada por el INIAP para ello se realizó la solicitud formal (Anexo 2) para poder hacer uso de los datos en este caso sobre datos agronómicos del ciclo de maíz forrajero incluido datos de rendimientos. La información fue proporcionada en formato de hojas de cálculos de Excel (Anexo 3), las cuales son el resultado de un estudio agronómico de maíz forrajero. En ese estudio agronómico se tomaron tres diferentes distancias de siembra y cuatro variedades de semilla, con cada una de estas combinaciones de distancia y semilla se realizó la siembra de 4 parcelas (repeticiones). En cada parcela de repetición, se seleccionaron diez plantas para la toma de datos de las variables que se detallan en la Tabla 1.

Tabla 1: Descripción y tipo de datos de las variables agronómicas

Variable	Tipo	Descripción
FECHA	Ordinal	Fecha del registro
DIA	Entero	Días después de sembrado (DDS)
DISTANCIA_SIEMBRA	Categórica	Distancia de siembra
PARCELA_GRANDE	Categórica	Identificador de la distancia de siembra
PARCELA_PEQUENIA	Categórica	Variedad de semilla de maíz
REPETICIONES	Categórica	Repetición del conjunto de factores entre distancia de siembra y variedad de semilla de maíz
ETAPA_VEGETATIVA_EV	Ordinal	Registros de la etapa vegetativa de las plantas ejemplo (v1, V2, V3, ..., V12, ..., VN).
DIAMETRO_TALLO_EV	Decimal	Diámetro del tallo las plantas en milímetros (mm).
ALTURA_PLANTA_EV	Decimal	Alturas de la planta en centímetros (cm).
INFLORESCENCIA_MASCULINA_ER	Binario	Registros de la presencia de inflorescencia masculina de la etapa reproductiva de las plantas.
INFLORESCENCIA_FEMENINA_ER	Binario	Registros de la presencia de inflorescencia femenina de la etapa reproductiva de las plantas.
MAZORCAS_PLANTA_ER	Entero	Registros del número de mazorcas en las plantas.
ETAPA_REPRODUCTIVA_ER	Ordinal	Registros de la etapa reproductiva de las plantas ejemplo (R1, R2, R3, R4).
INSERCIÓN_MAZORCA_ER	Decimal	Registros de los centímetros de la inserción de mazorca en la planta.

FUENTE: AUTOR

Adicionalmente también se tiene registros del rendimiento del maíz los cuales se calcularon en base a los pesos frescos que se tomaron de las hojas, tallos y mazorcas de las 10 plantas de cada repetición que se indican en la Tabla 2.

Tabla 2: Descripción y tipo de datos de las variables de rendimiento

Variable	Tipo	Descripción
FECHA	Ordinal	Fecha del registro
DIA	Entero	Días después de sembrado (DDS)
DISTANCIA_SIEMBRA	Categórica	Distancia de siembra
PARCELA_GRANDE	Categórica	Identificador de la distancia de siembra
PARCELA_PEQUENIA	Categórica	Variedad de semilla de maíz
REPETICIONES	Categórica	Repetición del conjunto de factores entre distancia de siembra y variedad de semilla de maíz
PESO_TOTAL_KG	Decimal	Peso fresco total de diez plantas en kilogramos (kg)
PESO_HOJAS_KG	Decimal	Peso fresco de las hojas en kilogramos (kg) de diez plantas
PESO_TALLO_KG	Decimal	Peso fresco de los tallos en kilogramos (kg) de diez plantas
PESO_MAZORCA_KG	Decimal	Peso fresco de las mazorcas en kilogramos (kg) de diez plantas
PESO_TOTAL_TON_H	Decimal	Peso total estimado en toneladas (t) producidas por una hectárea (ha)
PESO_HOJAS_TON_H	Decimal	Peso de hojas en toneladas (t) producidas por una hectárea (ha)
PESO_TALLO_TON_H	Decimal	Peso de tallo en toneladas (t) producidas por una hectárea (ha)
PESO_MAZORCA_TON_H	Decimal	Peso de mazorca en toneladas (t) producidas por una hectárea (ha)

FUENTE: AUTOR

Los registros de las diferentes variables tienen diferentes periodos de recolección tal como lo indica la Tabla 3.

Tabla 3: Periodos de recolección de datos de las variables agronómicas y de las variables de rendimiento

Variables	Fecha Inicio	Fecha Final
ETAPA_VEGETATIVA_EV	29/06/2023	21/08/2023
DIAMETRO_TALLO_EV	29/06/2023	21/08/2023
ALTURA_PLANTA_EV	29/06/2023	21/08/2023
INFLORESCENCIA_MASCULINA_ER	07/08/2023	21/08/2023
INFLORESCENCIA_FEMENINA_ER	10/08/2023	21/08/2023
MAZORCAS_PLANTA_ER	14/08/2023	24/08/2023
ETAPA_REPRODUCTIVA_ER	21/08/2023	24/08/2023
INSERCIÓN_MAZORCA_ER	21/08/2023	21/08/2023
RENDIMIENTO	28/08/2023	28/08/2023
PESO_TOTAL_KG	28/08/2023	28/08/2023
PESO_HOJAS_KG	28/08/2023	28/08/2023
PESO_TALLO_KG	28/08/2023	28/08/2023
PESO_MAZORCA_KG	28/08/2023	28/08/2023
PESO_TOTAL_TON_H	28/08/2023	28/08/2023
PESO_HOJAS_TON_H	28/08/2023	28/08/2023
PESO_TALLO_TON_H	28/08/2023	28/08/2023
PESO_MAZORCA_TON_H	28/08/2023	28/08/2023

FUENTE: AUTOR

3.6. PROCESAMIENTO Y ANÁLISIS

La metodología CRISP-DM se utilizó para el desarrollo de la investigación, esta metodología es altamente iterativa y puede requerir múltiples ciclos de ajuste del modelo a medida que se recopilan más datos del comportamiento de las predicciones y se obtiene retroalimentación del desempeño. La colaboración con el investigador del INIAP y la comunicación efectiva de los resultados fueron fundamentales en este proceso.

La información se analizó a través de bibliotecas de Python que se especializan en tratamiento de datos y de aprendizaje automático de acuerdo con las siguientes etapas de la metodología CRISP-DM.

3.6.1. Comprensión del Negocio

De acuerdo con los objetivos planteados, la estructura, las necesidades del investigador del INIAP y la disponibilidad de los datos se definió realizar el pronóstico de rendimiento del maíz en base a las variables agronómicas las cuales fueron resultado de un estudio realizado por el INIAP para la evaluación del rendimiento del maíz forrajero la cual constituyo la fuente primaria de los datos de esta investigación y se la detallo en el anterior apartado.

3.6.2. Comprensión de los Datos

Las hojas de cálculo con todos los datos se cargaron y procesaron para su correcta lectura en las bibliotecas de Python. Se realizó un análisis exploratorio para entender mejor las características de los datos y las relaciones entre las variables. Para ello se realizó un análisis estadístico descriptivo, gráficas de frecuencias y análisis de correlación de las variables (Anexo 4).

3.6.3. Preparación de los Datos

De acuerdo con el análisis exploratorio realizado en la etapa anterior se realizaron los siguientes procesos:

Se utilizó la técnica de normalización min-max, como lo indica Kotsiantis et al., (2006).

La fórmula aplicada se presenta en la ecuación 8:

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

Ecuación 8: Técnica de normalización mim-max

Donde v representa el valor original de las características $\min_A$ y $\max_A$, que son el mínimo y el máximo del conjunto de datos, respectivamente new_min_A y new_max_A , son los nuevos valores definidos para la normalización. Esta técnica aseguró que todos los valores numéricos sean transformados y uniformes.

Se realizó una transformación de las variables categóricas mediante la codificación One-Hot (dummy). Este es uno de los métodos de codificación más populares. Para una variable categórica de cardinalidad C genera $C - 1$ nuevas características binarias, con un valor de 1 para el valor real y cero en caso contrario. Aunque este método funciona bien en modelos lineales, no es adecuado para variables con C grande. Este enfoque se adoptó siguiendo las recomendaciones de Gutiérrez et al., (2020) para optimizar el tratamiento de las variables categóricas en el análisis.

Para la selección de características inicialmente se aplicó un método de filtrado, en específico se aplicó el filtro de colinealidad al conjunto de datos para eliminar la multicolinealidad entre variables, utilizando el coeficiente de correlación de Spearman. Adicionalmente también se aplicó la Eliminación Recursiva de Características (RFE) una técnica perteneciente a los métodos de envoltorios. RFE descartó las variables menos

importantes de manera recursiva, para ello se utilizó como estimador el algoritmo de Random Forest (Clarke et al., 2024).

Los datos disponibles se dividieron en dos grupos, uno para entrenamiento y otro grupo para prueba en proporción 70 % y 30 % respectivamente.

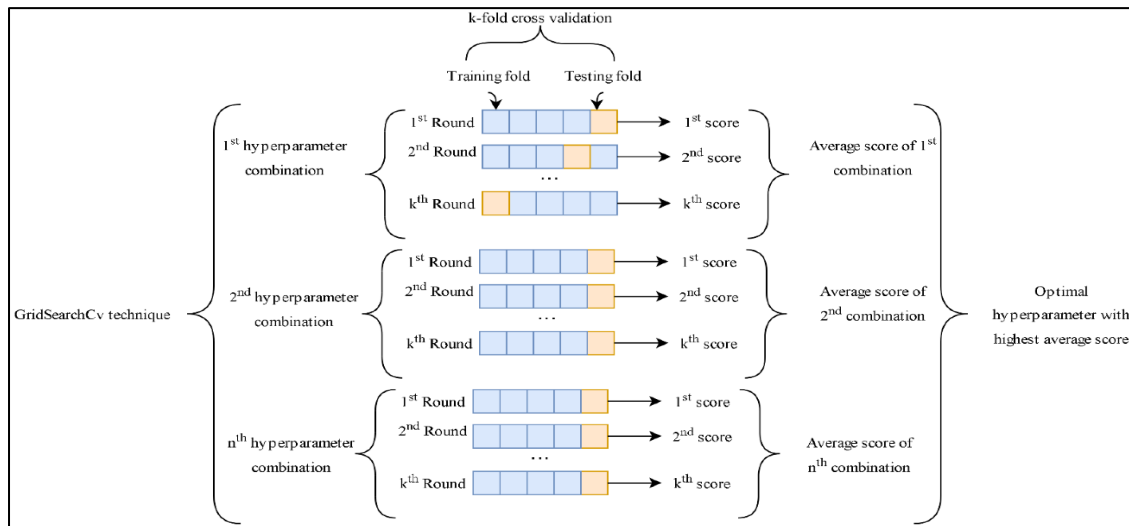
3.6.4. Modelado

De acuerdo con las características de los datos y el marco referencial consultado se seleccionaron los algoritmos de aprendizaje automático aplicables realizando búsquedas en Google Scholar y en las bases de datos de Scopus y SciELO, utilizando las palabras claves de “crop”, “regression”, “supervised learning”, “crop yield prediction”, “agricultural”.

Se realizó el ajuste de hiperparámetros de los algoritmos seleccionados mediante la función GridSearchCV que esta implementada en la biblioteca de Python scikit-learn (Anexo 5). Esta función es la implementación una técnica de validación cruzada que ayuda a obtener una estimación del rendimiento de los modelos y para determinar los valores óptimos de los hiperparámetros (Gutierrez et al., 2020), como estrategia de evaluación del desempeño del modelo con validación cruzada quíntuple (5-fold CV) se utilizaron las métricas de Raíz Cuadrada del Error Cuadrático Medio (RMSE) y el Coeficiente de Determinación (R^2) y para ajustar el estimador con los mejores hiperparámetros encontrados se utilizó la Raíz Cuadrada del Error Cuadrático Medio

(RMSE). En la Ilustración 3 se presenta un resumen del proceso de la técnica implementada en la función GridSearchCV.

Ilustración 3: Ajustes de hiperparámetros mediante función GridSearchCV.



FUENTE: Alhakeem et al. (2022)

3.6.5. Evaluación

Se evaluó el rendimiento de los modelos utilizando las métricas de Raíz Cuadrada del Error Cuadrático Medio y el Coeficiente de Determinación (Anexo 6) para así poder seleccionar el modelo con mejores predicciones en base a las métricas de evaluación seleccionadas:

Raíz Cuadrada del Error Cuadrático Medio (RMSE): Su uso fue fundamental para evaluar la precisión del modelo predictivo. Esta métrica, calculada como la raíz cuadrada del promedio de los cuadrados de las diferencias entre valores observados, permitió cuantificar el error de predicción del modelo, esta métrica se utilizó para el ajuste de

hiperparámetros, con el objetivo de minimizar el RMSE, y por ende, aumentar la fiabilidad de la predicciones del rendimiento del cultivo (Abbas et al., 2020). Se calculó utilizando la ecuación 9:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y_i^{obs} - y_i^{pred})^2}{N}}$$

Ecuación 9: Formula para cálculo de la métrica de RMSE.

Coefficiente de Determinación (R²): Esta métrica se utilizó como indicador del ajuste del modelo predictivo, puesto que midió la proporción de la variabilidad en la variable dependiente (rendimiento del cultivo), predecible a partir de las variables independientes (variables agronómicas). Este indicador reflejó como las predicciones se ajustan a los datos reales. La interpretación de un modelo predictivo con un valor R² cercano a 1 indica un ajuste perfecto y muy fiable para las predicciones, mientras que un valor de 0 refleja que no modela los datos con precisión. Para su cálculo se empleó la ecuación 10:

$$R^2 = \sqrt{\frac{\sum_{i=1}^N (y_i - \bar{y})^2 - \sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}}$$

Ecuación 10: Formula para cálculo del R².

Se realizó la comparación estadística de las métricas de evaluación RMSE y R² de los modelos predictivos construidos mediante la prueba estadística de Friedman's y la prueba post-hoc Nemenyi. Este enfoque permitió determinar si existían diferencias significativas en el desempeño de los modelos (Gutiérrez et al., 2020).

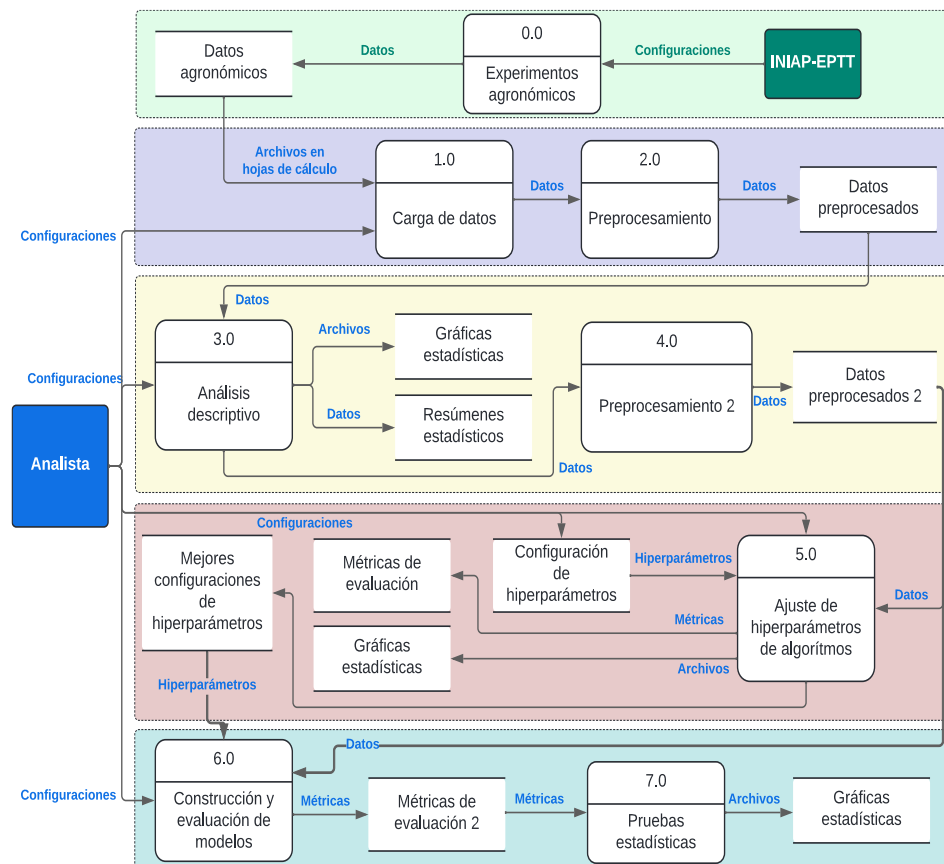
3.6.6. Despliegue

Esta etapa no se encuentra en el alcance de la presente investigación sin embargo se puede recomendar realizarla mediante la creación de una API REST que implemente el modelo de pronóstico en un entorno operativo y sea accesible para los usuarios finales, como agricultores o investigadores de áreas agronómicas.

3.6.7. Diagrama de flujo de datos

En la Ilustración 4 se presenta el diagrama de flujo de datos de la metodología de procesamiento y análisis aplicada.

Ilustración 4: Diagrama de flujo de datos de las etapas de la metodología.



FUENTE: AUTOR

CAPÍTULO IV
RESULTADOS Y DISCUSIÓN

A continuación, se describen los resultados que se obtuvieron por cada objetivo específico planteado en la investigación:

4.1. ANÁLISIS DESCRIPTIVO DE LOS DATOS AGRONÓMICOS

Los resultados del estudio agronómico de maíz forrajero se encontraban distribuidos en diferentes hojas de cálculo (Anexo 3), los cuales contenían los datos de las variables agronómica y de rendimiento. Para poder tener los datos en formato tabular se realizó un primer preprocesamiento y así poder realizar el análisis descriptivo.

El conjunto de variables agronómicas en formato tabular presentó 9 variables del tipo categórico y 4 variables del tipo numérico, adicionalmente se registró una variable de tipo fecha y una entera, tal y como lo indica la Tabla 4.

Tabla 4: Variables agronómicas del maíz forrajero

Variable	Tipo	Válidos	Nulls	% Nulls
FECHA	date	8160	0	0.00
DDS	int32	8160	0	0.00
DISTANCIA_SIEMBRA	category	8160	0	0.00
VARIEDAD_SEMILLA	category	8160	0	0.00
TRATAMIENTO	category	8160	0	0.00
REPETICIONES	category	8160	0	0.00
UNIDAD_EXPERIMENTAL	category	8160	0	0.00
ETAPA_VEGETATIVA_EV	category	7680	480	5.88
DIAMETRO_TALLO_EV	float64	7680	480	5.88
ALTURA_PLANTA_EV	float64	7680	480	5.88

INFLORESCENCIA_MASCULINA_ER	category	1947	6213	76.14
INFLORESCENCIA_FEMENINA_ER	category	1201	6959	85.28
MAZORCAS_PLANTA_ER	float64	1219	6941	85.06
ETAPA_REPRODUCTIVA_ER	category	960	7200	88.24
INSERCIÓN_MAZORCA_ER	float64	438	7722	94.63

FUENTE: AUTOR

Las variables FECHA y DDS indican la fecha y los días después de la siembra en que se registró la medición, la variable DISTANCIA_SIEMBRA indica la distancia de sembrado al que corresponde la medición, VARIEDAD_SEMILLA indica la variedad semilla de la planta de maíz forrajero al cual se registró la medición, la variable TRATAMIENTO indica la combinación de la distancia de siembra y de variedad de semilla, la variable REPETICIONES indica la parcela de repetición que pertenece la planta a la cual se registró las mediciones y la variable UNIDAD_EXPERIMENTAL refleja que planta dentro de la parcela de repetición se registró las mediciones.

A continuación, se analizarán a detalle cada una de las variables de DIAMETRO_TALLO_EV que indica el diámetro del tallo de la planta en centímetros, ALTURA_PLANTA_EV que indica la altura de la planta en centímetros, INFLORESCENCIA_MASCULINA_ER es la variable que indica si la planta tiene la presencia inflorescencia masculina (espiga), INFLORESCENCIA_FEMENINA_ER es la variable que indica si la planta tiene la presencia inflorescencia femenina (mazorca), MAZORCAS_PLANTA_ER indica la cantidad de mazorcas que presenta la planta de maíz, ETAPA_REPRODUCTIVA_ER indica la etapa reproductiva en que se encuentra la planta y INSERCIÓN_MAZORCA_ER indica la altura de inserción en centímetros de la distancia de la mazorca principal al suelo.

4.1.1 Análisis Descriptivo

A continuación, se presentan los estadísticos descriptivos de las variables agronómicas para lo cual se ha realizado un análisis exhaustivo comenzando con los días después de la siembra, que variaron entre 8 y 64 días. Para las variables diámetro del tallo, altura de la planta y etapa vegetativa se han analizado detalladamente de forma individual las cuales fueron representados mediante gráficos de líneas. En cambio, para las variables categóricas se han analizado mediante gráficos de barras las cuales se analizan las frecuencias de cada categoría.

En la Tabla 5 se presentan los estadísticos de las variables de tipo numérico, las columnas incluyen el número de observaciones, la media, desviación estándar, valores mínimo y máximo, percentiles, mediana, la cantidad y porcentajes de valores nulos.

Tabla 5: Estadística descriptiva de las variables numéricas

Variable	#	Media	Std	Mín	25 %	50 %	75 %	Max	Nulls	Nulls %
Días después de la siembra	8160	36.24	17.15	8.00	22.00	36.00	50.00	64.00	0	0.00
Diámetro del tallo (cm)	7680	18.45	6.90	2.87	14.00	20.92	23.67	30.95	480	5.88
Altura de la planta (cm)	7680	111.83	84.47	4.50	29.00	100.00	195.25	340.00	480	5.88
Número de mazorcas	1219	1.07	0.26	1.00	1.00	1.00	1.00	3.00	6941	85.06
Altura de inserción de mazorca (cm)	438	134.31	20.19	51.00	123.00	133.50	146.00	204.00	7722	94.63

FUENTE: AUTOR

4.1.1.1 Días Después de la Siembra

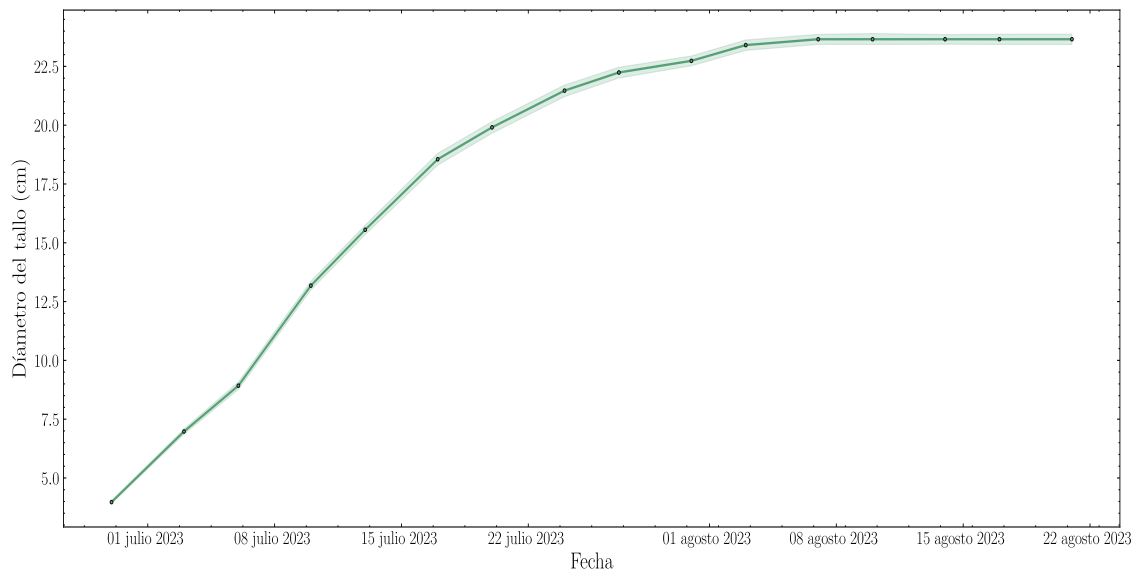
Se registraron 8160 observaciones sin datos nulos como lo reflejó la Tabla 5. El promedio de días transcurridos desde la siembra en que se registraron los datos fue de 36.24, con una desviación estándar de 17.15 indicando variabilidad moderada en el tiempo en que tomaron los datos agronómicos. El rango se dio desde un mínimo de 8 días después de siembra hasta un máximo de 64. Los percentiles 25 % y 75 % indicaron que la mitad de las observaciones se registraron entre los 22 y los 50 días después de la siembra.

4.1.1.2 Diámetro del Tallo

El diámetro del tallo de las plantas de maíz presentó un crecimiento considerable a lo largo del tiempo, como indican los datos recogidos entre el 29/06/2023 y el 21/08/2023. Se observó que el diámetro más pequeño registrado fue de 2.87 cm. A medida que las plantas maduraron, se observó un incremento en el diámetro del tallo, alcanzando valores superiores a los 20 cm después de 26 días, con un máximo de 30.95 cm perteneciente a una planta a los 50 días.

El análisis del conjunto de datos completo, compuesto por 7680 registros válidos, reflejó un diámetro medio del tallo de 18.45 cm, con una desviación estándar de 6.9 cm. Esto indica una dispersión moderada en torno a la media. Se observó que el 5.88 % de las entradas de datos estaban vacías, correspondientes a 480 valores nulos. En la Ilustración 5 se muestra la media de los valores del diámetro del tallo en cada fecha de medición y adicionalmente las bandas muestra el intervalo de confianza al 95 %.

Ilustración 5: Diámetro del tallo

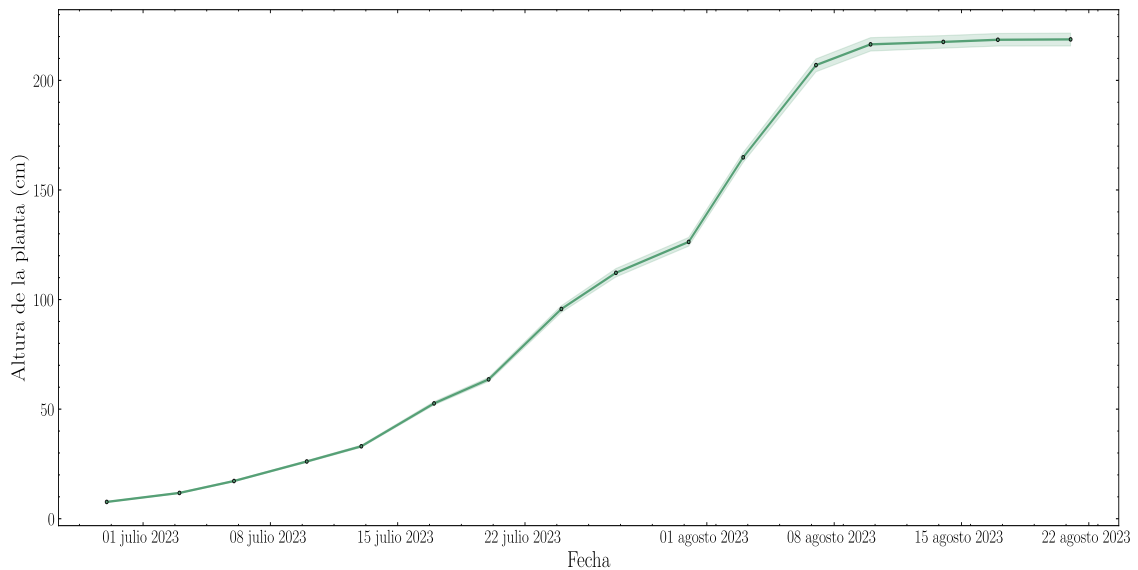


FUENTE: AUTOR

4.1.1.3 Altura de la Planta

Durante el periodo de estudio, se registraron alturas de plantas que variaron significativamente a través del tiempo. En el conjunto de datos de 7680 registros válidos muestran una altura media de 111.83 cm con una desviación estándar de 84.47 cm, lo cual implicó una amplia variación en la altura de las plantas. El rango de crecimiento respecto a la altura de la planta es extenso, con la planta más baja de 4.5 cm y la más alta de 340 cm. Alrededor del 25 % de las plantas alcanzaron 29 cm o menos, mientras que el 75 % no superó los 195.25 cm de altura. Es importante destacar que el 5.88 % de los registros estaban ausentes del conjunto de datos. En la Ilustración 6 se presenta un incremento en la altura de las plantas entre los 29 y 50 días, seguida de una fase de crecimiento más lento hasta el día 61, esta gráfica de líneas muestra la media de los valores de altura de la planta en cada fecha de medición y adicionalmente las bandas muestra el intervalo de confianza al 95 %.

Ilustración 6: Altura de la planta



FUENTE: AUTOR

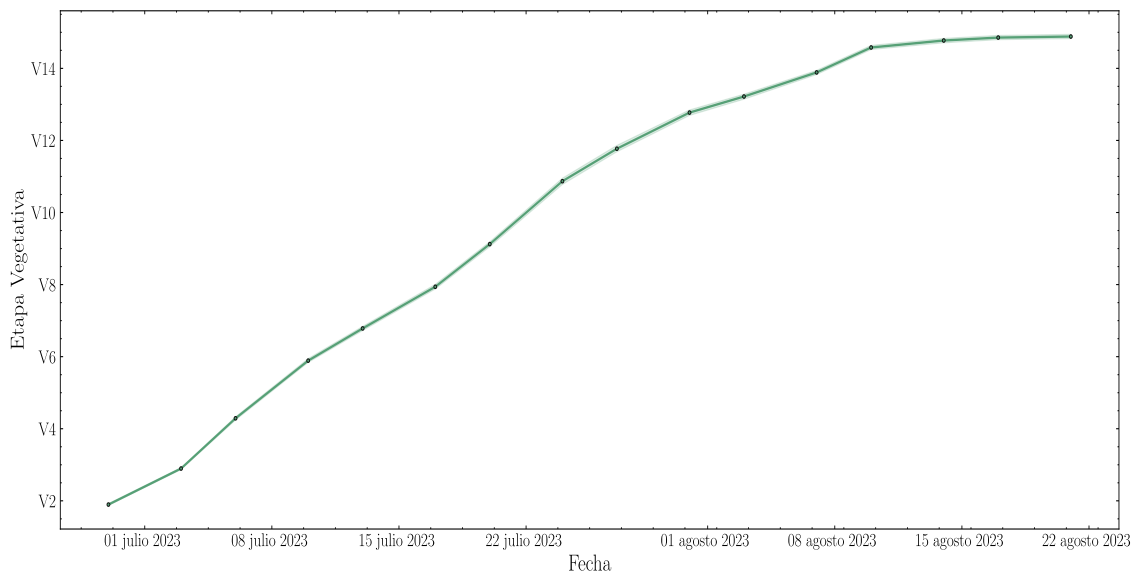
4.1.1.4 Etapa Vegetativa

En las etapas fenológicas el seguimiento comenzó el 29/06/2023, con plantas apenas en V1 y en V2 a los 8 días después de la siembra, lo que representa el establecimiento inicial de las plantas jóvenes. Al día 29, se observó plantas que alcanzaron la etapa vegetativa V9 e incluso V10, claro indicador del crecimiento vegetativo. Para el 27/07/2023, día 36 las plantas alcanzaron la etapa V13.

La fase de crecimiento más intensa se dio entre los días 50 y 57, con plantas alcanzado desde V13 hasta V16, característico de una aceleración hacia la madurez vegetativa. Adicionalmente se transformó los datos de categóricos a numéricos por los valores numéricos respectivos con lo cual se elaboró la Ilustración 7 en donde se presenta la media

de los valores de la etapa vegetativa en cada fecha de medición y adicionalmente las bandas muestra el intervalo de confianza al 95 %.

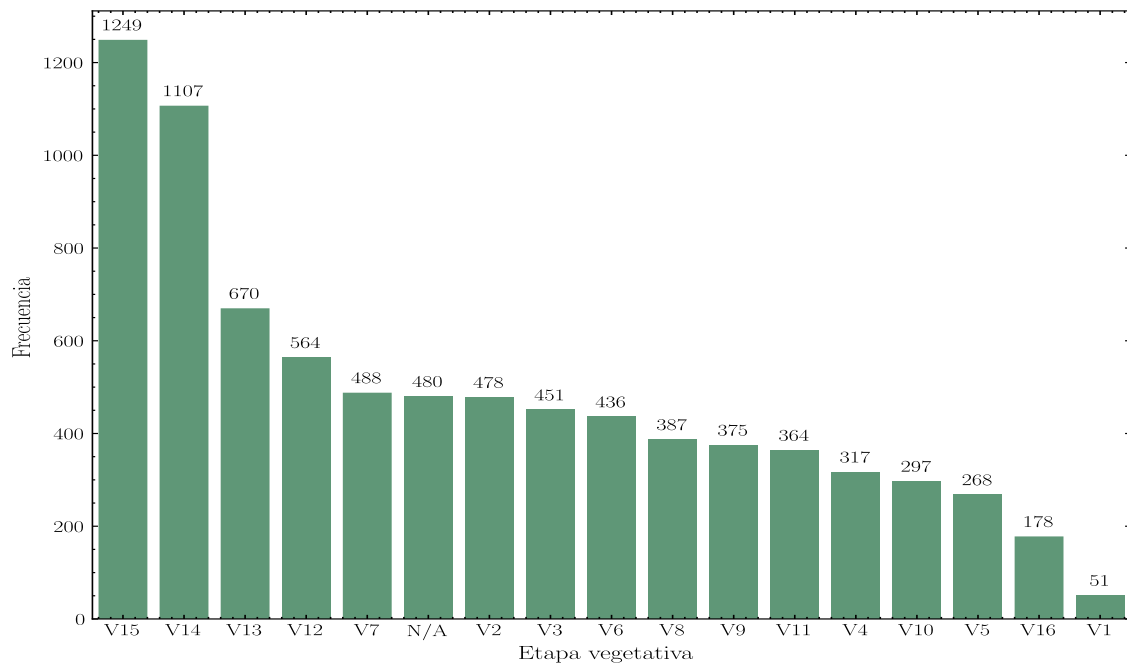
Ilustración 7: Etapa vegetativa



FUENTE: AUTOR

Se tienen 7680 observaciones del ciclo vegetativo, clasificadas en 16 etapas diferentes. Se reflejó que la etapa V15 fue la más prevalente, con 1249 registros, indicando mayor duración en comparación con otras etapas. También se identificó que el 5.88 % de los datos fueron ausentes, lo que equivale a 480 registros. La Ilustración 8 indica el diagrama de barras con la frecuencia de los datos de la variable etapa vegetativa.

Ilustración 8: Frecuencia de la etapa vegetativa

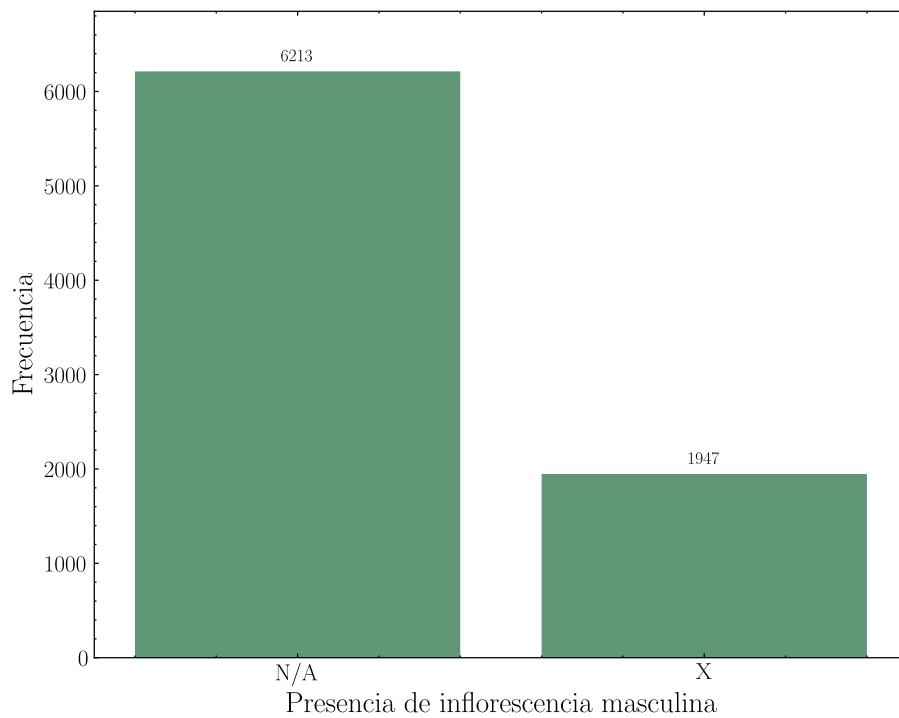


FUENTE: AUTOR

4.1.1.5 Presencia de Inflorescencia Masculina

Se registraron 1947 elementos, como se refleja en Ilustración 9. La variable estudiada presentó una única categoría marcada como X, que se registró en todas las observaciones validas. La alta cantidad de datos faltantes alcanzaron un 76.14 % equivalente a 6123 registros.

Ilustración 9: Frecuencia de la presencia de inflorescencia masculina

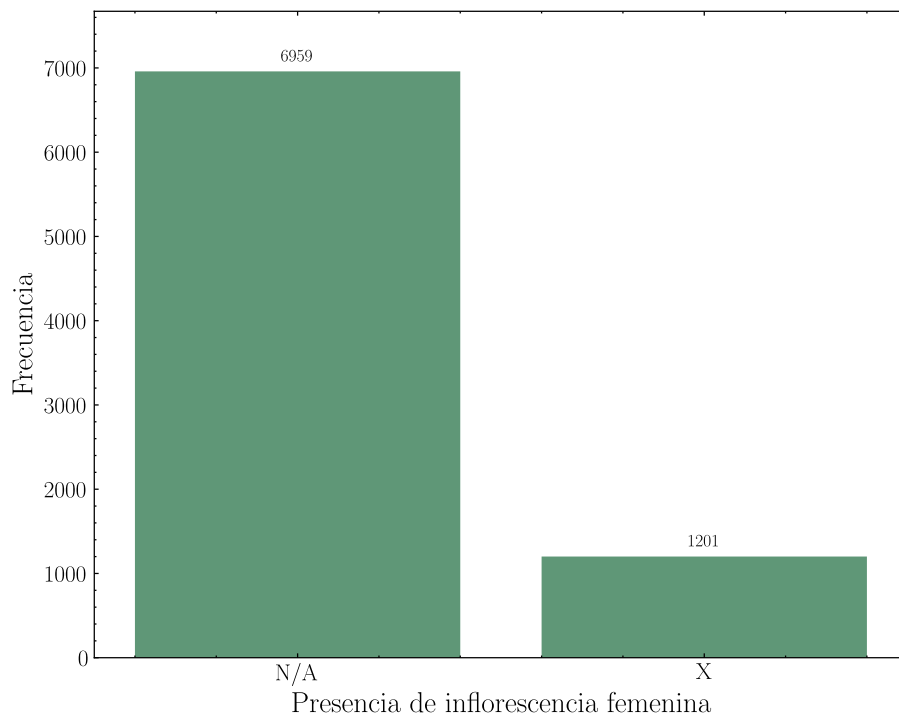


FUENTE: AUTOR

4.1.1.6 Presencia de Inflorescencia Femenina

Se registraron 1201 elementos, los cuales fueron clasificados en una categoría identificadas como X. No obstante, el 85.28 % de los datos, equivalente a 6959 observaciones, fueron clasificados como datos faltantes, donde la mayoría no reportaron la presencia de inflorescencia femenina. Conforme a lo observado en la Ilustración 10.

Ilustración 10: Frecuencia de la presencia de inflorescencia femenina

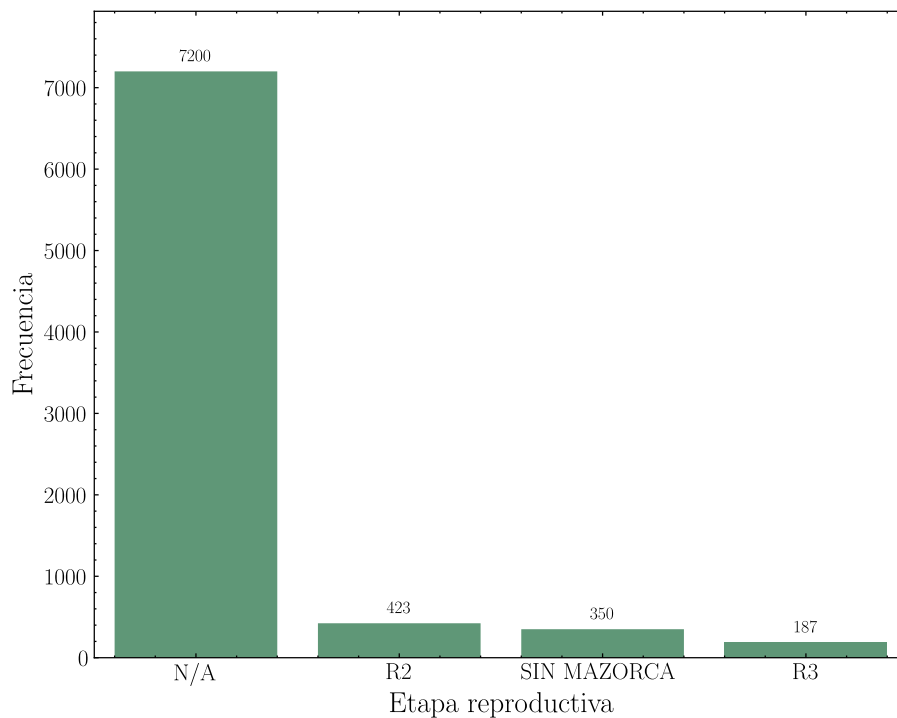


FUENTE: AUTOR

4.1.1.7 Frecuencia de la Etapa Reproductiva

Se evaluaron 610 registros que fueron calificados en dos categorías distintas de acuerdo con lo observado en la Ilustración 11. De estos, la categoría R2 fue la más frecuente con 423 registros. Sin embargo, se reflejó un porcentaje de datos 92.52 % que equivalen a 7750 registros con datos faltantes, lo que indicó una marcada ausencia de datos en la etapa reproductiva.

Ilustración 11: Frecuencia de la etapa reproductiva



FUENTE: AUTOR

4.1.2 Análisis Descriptivo de Registros del 21/08/2023

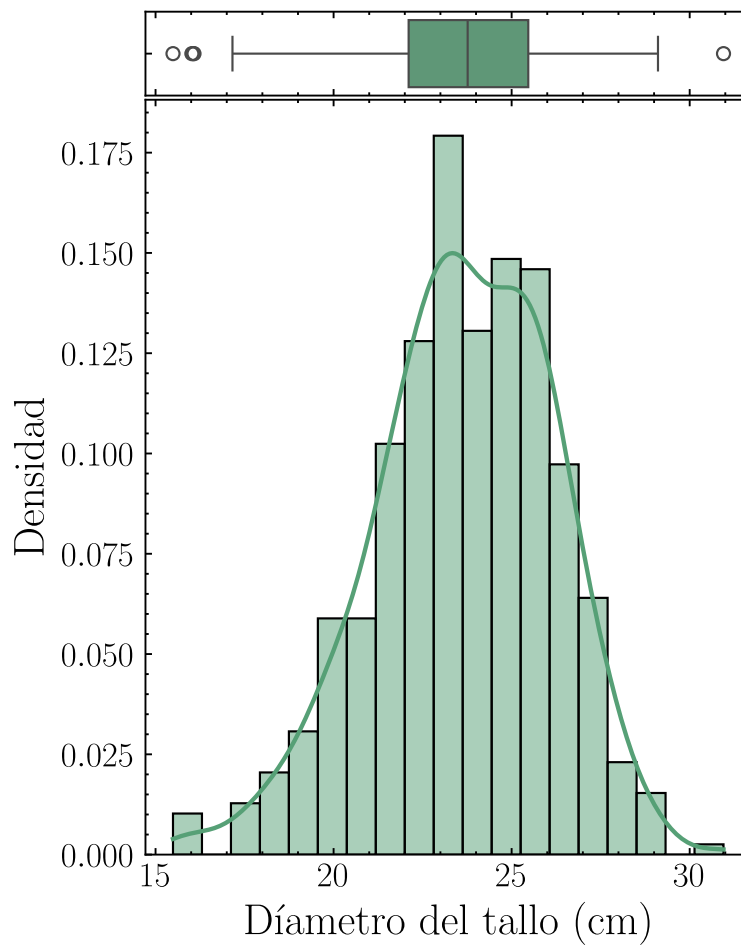
El siguiente análisis descriptivo se centró en los datos agronómicos recogidos el 21 de agosto de 2023, correspondientes a la última fecha de muestreo del cultivo de maíz forrajero, esto debido a la alta cantidad de valores faltantes en las variables antes descritas. Este análisis descriptivo reflejó las mediciones finales de las variables agronómicas, que se presentan a continuación:

4.1.2.1 Diámetro del Tallo

Con base a las 480 mediciones registradas. La media del diámetro fue de 23.65 cm reflejando la dimensión promedio de los tallos. La desviación estándar de 2.49 cm indicó

una variabilidad moderada. La distribución varió desde un mínimo de 15.49 cm hasta un máximo de 30.95 cm. Tal como se presenta en la Ilustración 12 la mediana de 23.76 cm se alineó con la media reflejando una distribución simétrica.

Ilustración 12: Diámetro del tallo para el 21/08/2023



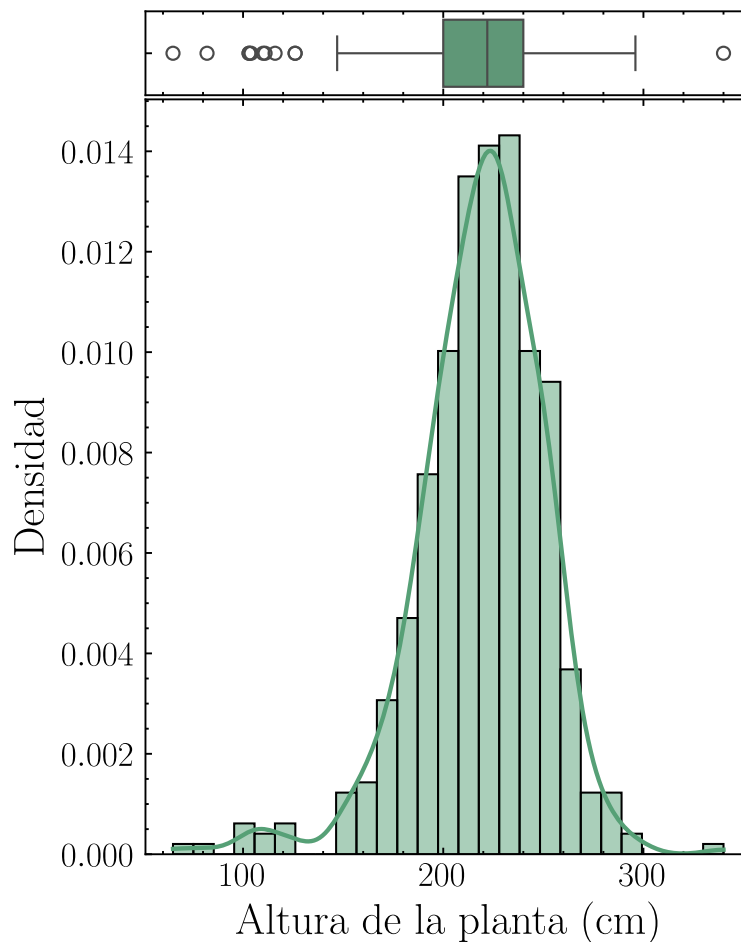
FUENTE: AUTOR

4.1.2.2 Altura de la Planta

En altura de la planta se reflejó una media de 218.70 cm a partir de 480 mediciones. La desviación estándar de 31.81 cm indica una variabilidad significativa. El rango observado fue amplio, con la planta más baja a 65 cm y la más alta a 340 cm. Los cuartiles indicaron

que el 50 % de los datos presentó una altura entre 200 cm y 240 cm. La distribución de los datos, sin valores nulos y una mediana de 222 cm cercana a la media sugiere una distribución simétrica con un leve sesgo hacia las plantas más altas, esto se puede evidenciar en la Ilustración 13.

Ilustración 13: Altura de la planta para el 21/08/2023



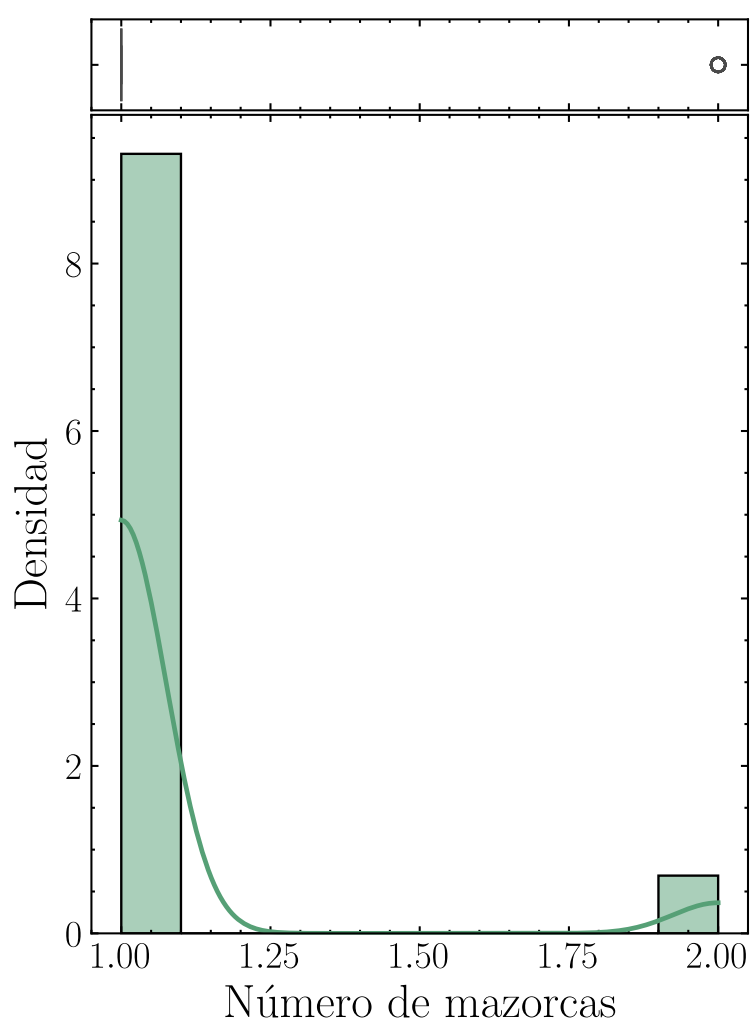
FUENTE: AUTOR

4.1.2.3 Número de Mazorcas

La variable analizada tuvo una media de 1.07 mazorcas por plantas basadas en 435 elementos registrados. Con una desviación estándar de 0.25, los datos sugieren una

variabilidad limitada en torno a la mayoría de las plantas que obtuvieron una sola mazorca. El valor mínimo registrado fue de 1 mazorca, con el primer, segundo (mediana) y tercer cuartil, todos en 1, y un valor máximo de 2 mazorcas, como se refleja en la Ilustración 14. Se registró un 9.38 % correspondiente a 45 observaciones faltantes donde no se reportó el número de mazorcas.

Ilustración 14: Número de mazorcas para el 21/08/2023

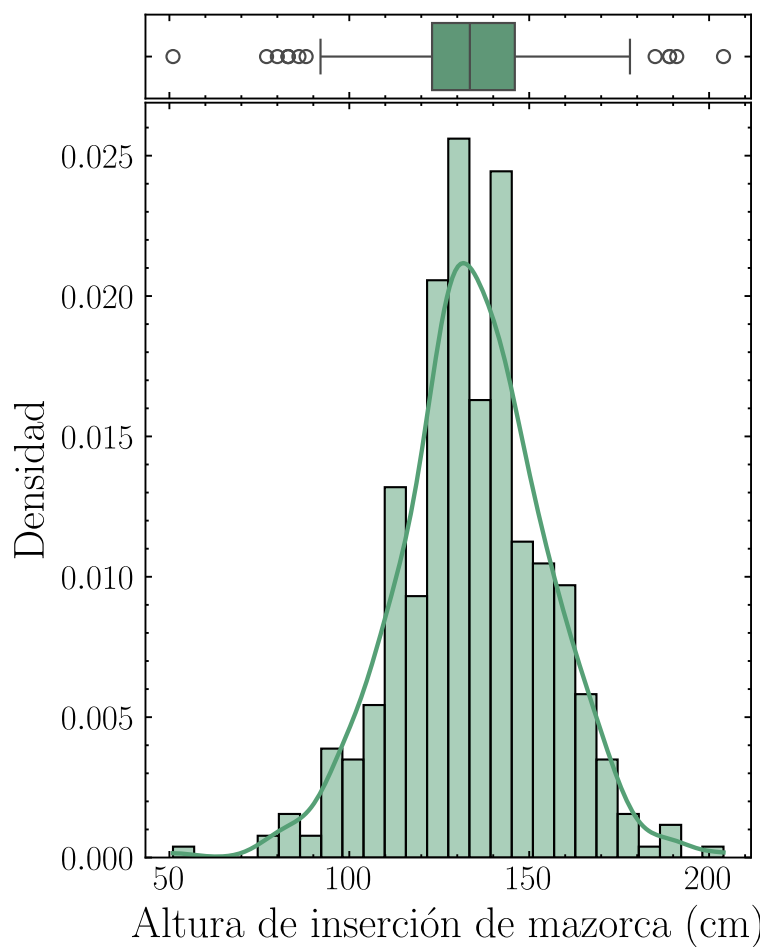


FUENTE: AUTOR

4.1.2.4 Altura de la Inserción de la Mazorca

La altura de inserción de mazorca en el maíz forrajero fue evaluada con 438 elementos, con una media de 134.31 cm. La desviación estándar de 20.19 cm indicó una variabilidad en la altura moderada en la altura de inserción de las mazorcas, conforme a la Ilustración 15. Con una altura mínima y máxima de 51 cm y 204 cm respectivamente, los datos mostraron un rango considerable. Los cuartiles reflejaron que el 50 % de los datos se ubicó entre 123 cm y 146 cm de altura, con una mediana de 133.5 cm, alineada estrechamente con la media, indicando una distribución simétrica.

Ilustración 15: Altura de la inserción de la mazorca para el 21/08/2023

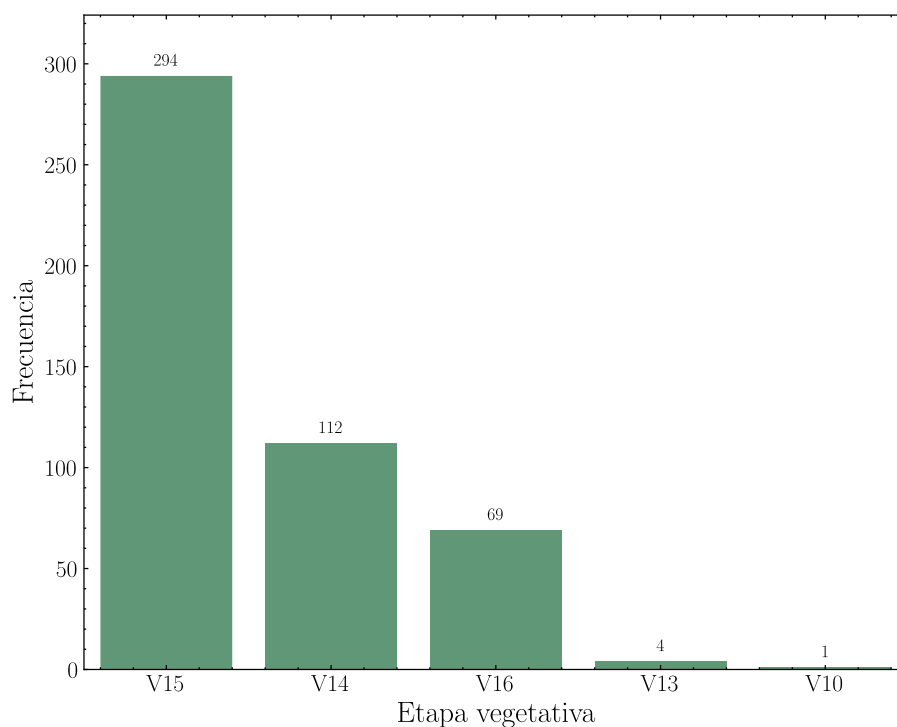


FUENTE: AUTOR

4.1.2.5 Etapa Vegetativa

Se evaluaron 480 registros donde se encontraron cinco categorías, la categoría V15 fue la más prevalente con 294 registros, lo que representa aproximadamente el 61.25 % de las muestras analizadas en esta fecha. Como se evidencia en la Ilustración 16.

Ilustración 16: Etapa vegetativa para el 21/08/2023

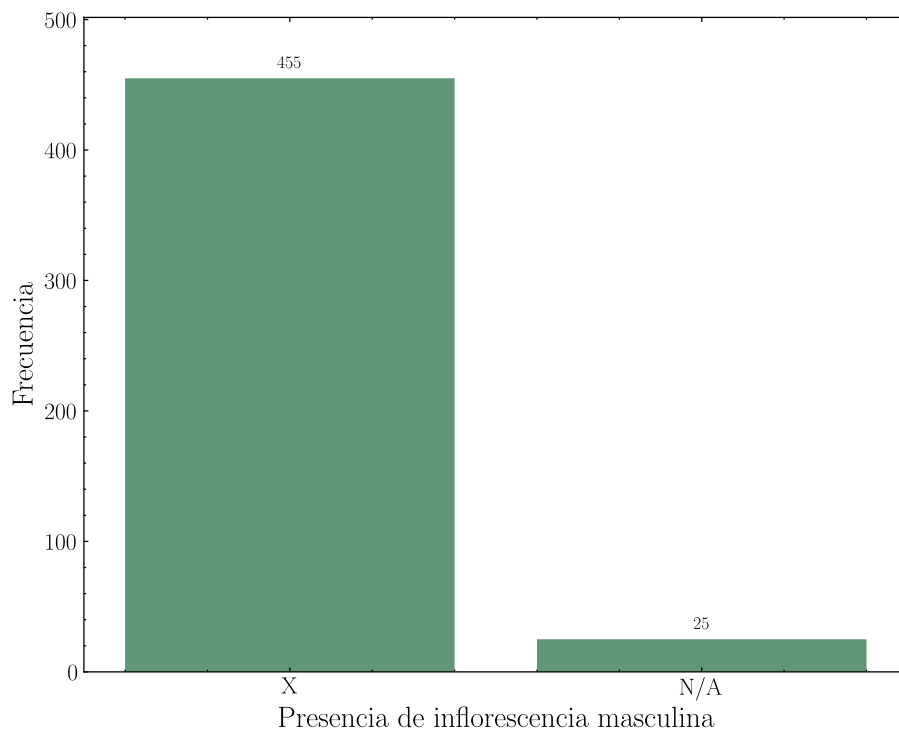


FUENTE: AUTOR

4.1.2.6 Presencia de Inflorescencia Masculina

Se registraron 455 elementos dentro de la única categoría existente (X). Lo cual representa la totalidad de las muestras validas recogidas. No obstante, un 5.21 % correspondiente a 25 elementos fueron clasificados como datos faltantes, como indica la Ilustración 17.

Ilustración 17: Presencia de inflorescencia masculina para el 21/08/2023

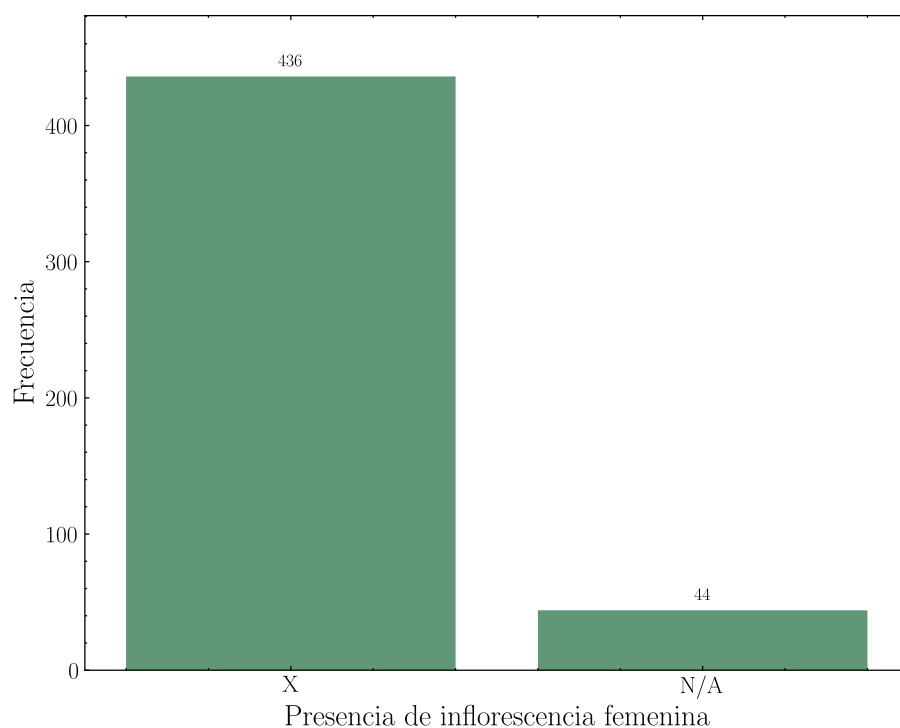


FUENTE: AUTOR

4.1.2.7 Presencia de Inflorescencia Femenina

Se observaron 436 elementos dentro de la categoría X, que indicó la presencia de inflorescencia femenina en el maíz forrajero. Además, se registraron 44 casos de datos faltantes, lo que corresponde al 9.17 % del total de las observaciones (Ilustración 18).

Ilustración 18: Presencia de inflorescencia femenina para el 21/08/2023

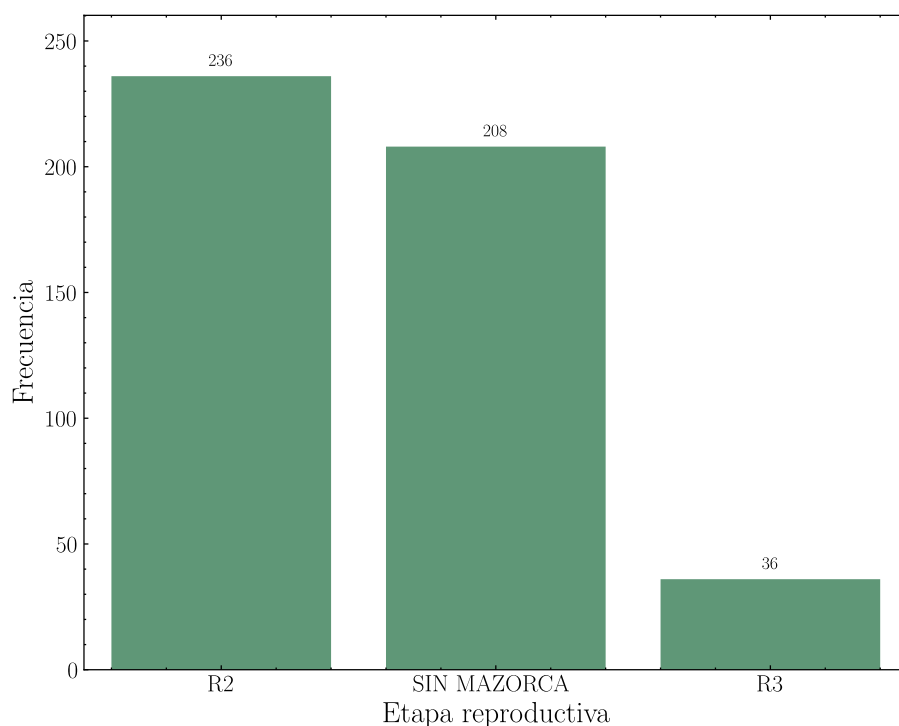


FUENTE: AUTOR

4.1.2.8 Etapa Reproductiva

Se registraron 480 elementos, en 3 categorías distintas. La categoría R2 presentó la mayoría con 236 casos. No obstante, se identificó 43.33 % de plantas sin mazorcas que corresponde a 208 observaciones. Tal y como se evidencia en la Ilustración 19.

Ilustración 19: Etapa reproductiva para el 21/08/2023



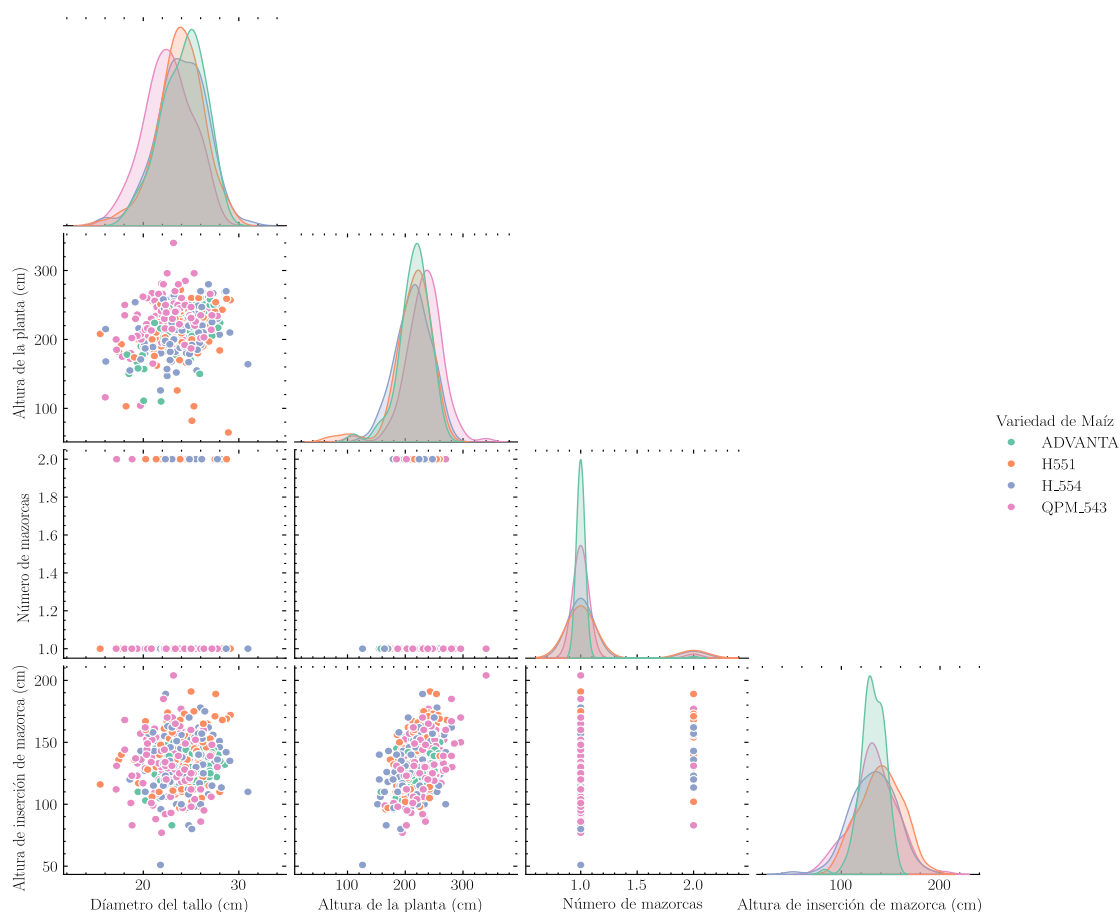
FUENTE: AUTOR

4.1.3 Análisis Multivariante de Variables Agronómicas Numéricas

4.1.3.1 Análisis Multivariante de Variables Agronómicas Numéricas por Variedad de Maíz

De acuerdo con lo presentado en la Ilustración 20 se puede establecer las relaciones entre las variables de altura de la planta y diámetro de tallo siendo que la variedad de semilla QPM 543 es la que más altos valores presenta. Este mismo comportamiento se repite en el gráfico de dispersión entre altura de inserción de mazorca y altura de la panta.

Ilustración 20: Gráfica por pares de variables agronómicas numéricas por variedad de maíz

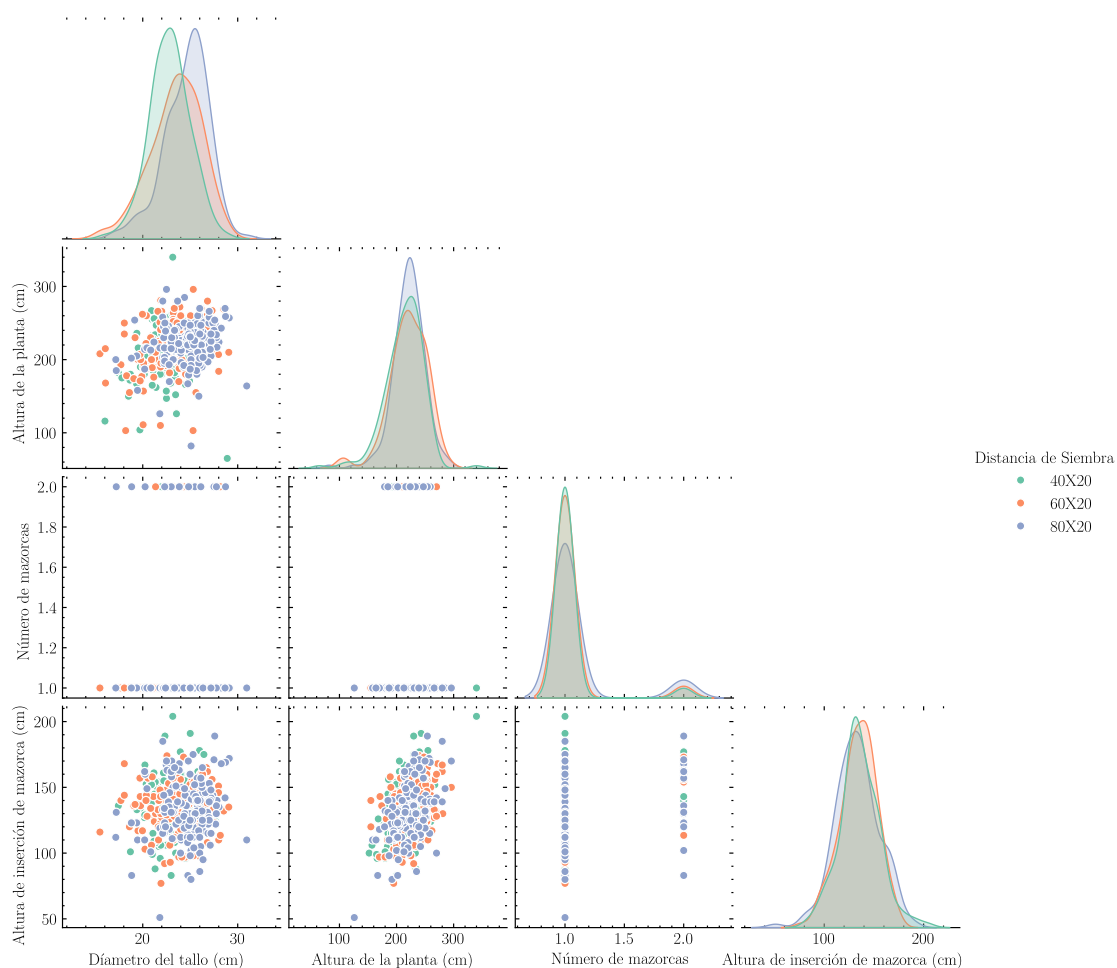


FUENTE: AUTOR

4.1.3.2 Análisis Multivariante de Variables Agronómicas Numéricas por Distancia de Siembra

De acuerdo con lo presentado en la Ilustración 21 se estableció que la distancia de siembra causa que las plantas tengan igual altura, pero presentan mayor diámetro de tallo a una distancia de siembra de 80x20 respecto a la de 60x20. Pero en cambio no se evidencia diferencia entre las plantas sembradas respecto a la altura de inserción de mazorca y altura de la panta.

Ilustración 21: Gráfica por pares de variables agronómicas numéricas agronómicas numéricas por distancia de siembra



FUENTE: AUTOR

4.1.4 Análisis Descriptivo de Datos de Rendimiento

Se contaron con datos de 480 elementos para cada variable de rendimiento analizada, que incluyeron peso total, peso de hojas, peso de tallo, y peso de mazorca, tanto en kilogramos como toneladas por hectárea. A continuación, se presenta la Tabla 6 donde se encuentran las estadísticas descriptivas de cada una de las variables:

Tabla 6: Estadística descriptiva de las variables de rendimiento

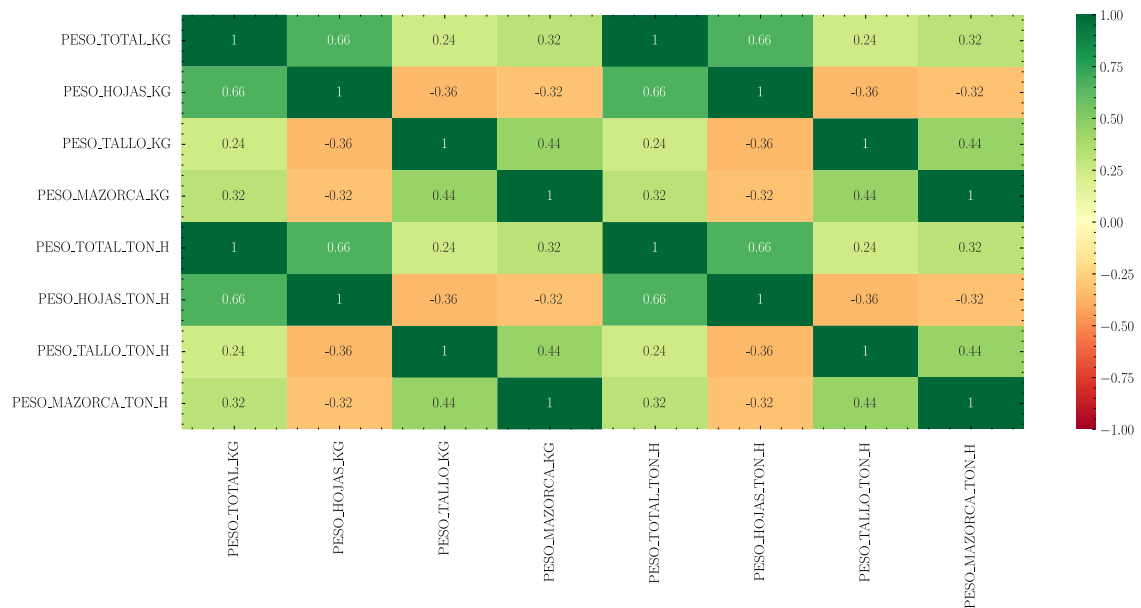
Variable	Elementos	Media	Std	Mín	25 %	50 %	75 %	Max
Peso Total (Kg)	480	0.66	0.13	0.45	0.55	0.65	0.75	0.95
Peso Hojas (Kg)	480	0.18	0.06	0.10	0.15	0.18	0.20	0.35
Peso Tallo (Kg)	480	0.31	0.08	0.03	0.27	0.30	0.35	0.45
Peso Mazorca (Kg)	480	1.73	0.79	0.70	1.00	1.60	2.00	5.00
Peso Total (t/ha)	480	57.21	11.22	38.75	48.59	54.17	62.50	81.25
Peso Hojas (t/ha)	480	15.49	4.81	6.25	12.50	15.42	18.75	25.00
Peso Tallo (t/ha)	480	26.51	6.84	3.75	21.88	25.00	31.25	43.75
Peso Mazorca (t/ha)	480	15.21	9.28	8.33	12.50	12.50	16.35	62.50

FUENTE: AUTOR

4.1.4.1 Matriz de Correlación de Spearman de Variables de Rendimiento del Maíz

Conforme a lo indicado en la Ilustración 22 se establece una correlación entre las variables de pesos frescos (peso total, peso de hojas, peso de tallos y peso de mazorcas) con las respectivas variables de rendimiento de toneladas por hectárea. Se observa correlaciones perfectas positivas, esto se origina por que el rendimiento de toneladas se basa a una formula agronómica la cual se multiplica el peso fresco de las hojas, tallos y mazorcas por un valor dependiendo de la distancia de siembra. Por ello se descartan las variables de rendimientos de toneladas por hectárea en el primer preprocesamiento de datos y solo se realiza un análisis descriptivo de las variables de peso fresco total, de hojas, tallo y mazorca.

Ilustración 22: Correlación de variables de rendimiento del maíz



FUENTE: AUTOR

4.1.4.2 Peso Total

Se registraron 480 elementos, el peso total promedio del maíz forrajero de 0.66 kg. La variación en los pesos totales fue relativamente baja, como lo indicó la desviación estándar de 0.13 kg. Los datos variaron en un mínimo de 0.45 kg hasta un máximo de 0.95 kg, el 50 % de las plantas pesaron 0.65 kg, reflejando una concentración de la mayoría de los datos a la media.

4.1.4.3 Peso de Hojas

Se registraron 480 elementos con un peso promedio de 0.18 kg. Con una desviación estándar de 0.06 kg, los pesos de las hojas mostraron poca dispersión. El rango se extendió de un mínimo de 0.10 kg a un máximo de 0.35 kg, y la mediana fue de 0.18 kg, lo cual reflejó una distribución simétrica de los pesos de las hojas alrededor de la media.

4.1.4.4 Peso de Tallo

Se registraron 480 elementos, con una media de 0.31 kg por planta, con una desviación estándar de 0.08 kg, lo que indicó una dispersión moderada de los datos. El mínimo registrado fue de 0.03 kg, mientras que el máximo se reflejó en 0.45 kg. El 50 % de los tallos tuvo un peso entre 0.27 kg y 0.35 kg.

4.1.4.5 Peso de Mazorca

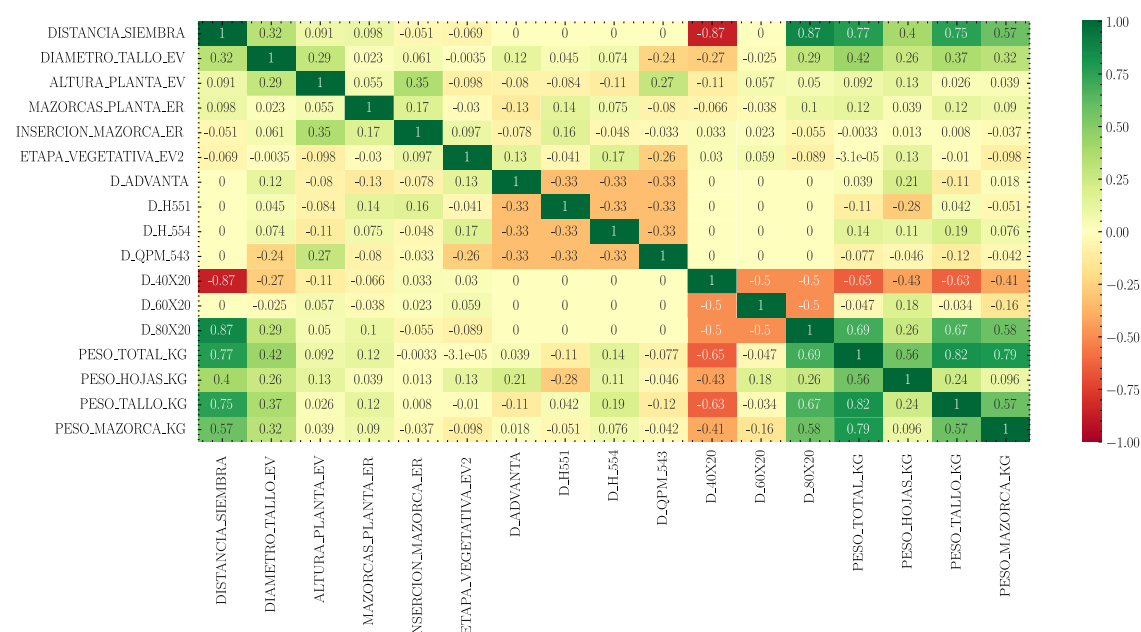
Se registraron 480 elementos los cuales presentaron el peso promedio más alto entre las variables analizadas, con 1.73 kg. La desviación estándar de 0.79 kg evidencio una variabilidad significativa en el peso de las mazorcas entre las plantas. El rango de peso fue de 0.70 kg hasta 5.00 kg. La mediana de 1.60 kg, más baja que la media, lo que indicó una distribución sesgada hacia pesos menores.

4.1.5 Matriz de Correlación de Spearman de Datos Agronómicos y de Rendimiento del Maíz

Para analizar la correlación de las variables agronómicas y de rendimiento se realizó una transformación de las variables distancia siembra y variedad semilla de tipo categóricas a variables binarias mediante la codificación One-Hot, adicionalmente también se realizó la transformación de la variable distancia siembra de categórica a numérica llevando los valores de distancias a su equivalente de 40 cm, 60 cm y 80 cm. La matriz de correlación de Spearman, como se muestra en la Ilustración 23, permitió observar las relaciones entre

variables agronómicas y de rendimiento de las plantas de maíz forrajero. La variable diámetro de tallo y la variable distancia de siembra presentan los más altos coeficientes de correlación. Las variables transformadas de distancia de siembra “D 40X20” y “D 80X20” también presentan altas correlaciones con las variables de rendimiento.

Ilustración 23: Correlación de datos Agronómicos y de rendimiento del maíz



FUENTE: AUTOR

4.1.6 Discusión

Se realizó un análisis descriptivo de las variables agronómicas y de rendimiento del maíz forrajero de un estudio realizado en la ciudad de Quevedo, con el propósito de seleccionar un modelo de pronóstico del rendimiento del cultivo. La base de datos incluyó un total de 8160 observaciones válidas para la mayoría de las variables (Anexo 3); como días después de la siembra, variedad de semilla, diámetro del tallo y altura de la planta etc., lo que

refleja solidez y completitud del conjunto de datos, puesto que permiten comprender las dinámicas de crecimiento y desarrollo.

En el análisis descriptivo de los datos agronómicos del maíz forrajero se observó una distribución variada en términos de la etapa vegetativa de las plantas comenzando con V1 y registrándose hasta la etapa V15, lo que refleja las diferentes etapas de desarrollo del cultivo desde la germinación hasta la madurez. Este rango es consistente con lo reportado en (Ransom, 2020) que indica que las fases fenológicas del maíz pueden variar significativamente dependiendo de las condiciones ambientales y de manejo.

Los datos muestran que el diámetro del tallo y la altura de la planta presentan variaciones significativas, lo cual es indicativo de diferencias en el crecimiento vegetativo influidas por las variedades de semilla y las condiciones de cultivo. La media del diámetro del tallo (18.45 cm) con una variabilidad moderada de 6.90, y de la altura de la planta (111.83 cm mostrando una alta variabilidad de 84.47 cm) son indicadores robustos de la vitalidad y el potencial de crecimiento del cultivo, con desviaciones estándar que reflejan la variabilidad inherente al entorno agrícola y a las prácticas de manejo.

Esta variabilidad es crucial, ya que características físicas robustas en plantas de maíz están relacionadas con la capacidad de las plantas para soportar mayores rendimientos de mazorca. Estos resultados son consistentes con los hallazgos de (Guamán Guamán et al., 2020) quienes destacaron la relación significativa entre altura de la planta y el rendimiento del cultivo, sugiriendo una mayor biomasa aérea está relacionada con mayores rendimientos, además evaluaron las mismas variables agronómicas, y características

como desarrollo y rendimiento. La correlación entre altura de la planta, diámetro del tallo y el rendimiento observado en este estudio sugiere que son indicadores útiles del vigor y la capacidad productiva. Estos resultados están en concordancia con lo reportado por (Sánchez-Hernández et al., 2011) quienes utilizaron análisis multivariante para identificar patrones de asociación entre diferentes características agronómicas y su impacto en el rendimiento.

(Cadet-Díaz & Guerrero-Escobar, 2018) utilizaron análisis multivariante para desarrollar modelos predictivos de rendimiento de maíz, similar al presente estudio, puesto que la integración de diversas variables agronómicas en modelos predictivos puede ayudar a anticipar el rendimiento. El análisis multivariante de las variables numéricas permitió identificar correlaciones significativas entre el diámetro del tallo, la altura de la planta y el rendimiento total, coincidiendo con los hallazgos de estudios previos que vinculan estas características morfológicas con la adaptabilidad y productividad (Guamán et al., 2020). Estas correlaciones son vitales para la aplicación en modelos de aprendizaje automático supervisado, ya que proporcionan una base sólida para predecir el rendimiento basado en indicadores de crecimiento fácilmente medibles.

4.2. AJUSTE DE HIPERPARÁMETROS DE LOS ALGORITMOS

Tras realizar el análisis descriptivo de las variables agronómicas y de rendimiento del maíz forrajero, se estableció que para cada una de las variables objetivo (variables de rendimiento), es decir Peso Total, Peso Hojas, Peso Tallo y de Peso Mazorca se entrenarían modelos independientes.

En este sentido se procedió a la selección de algoritmos, de las métricas de evaluación, de sus hiperparámetros y de los rangos de valores de cada hiperparámetro seleccionado. Luego se realizó un preprocesamiento de datos para poder llevar a cabo el ajuste de hiperparámetros. Durante el proceso de ajuste se registraron los tiempos de ajuste y los valores de las métricas RMSE y R^2 . A continuación, se describe a detalle cada uno de los procesos realizados y los resultados obtenidos.

4.2.1 Selección de Algoritmos

Para la selección de algoritmos se realizó búsqueda bibliográfica especializada la cual se analizó y se presenta en la Tabla 7 donde se cita el estudio, los algoritmos utilizados, las métricas de evaluación y el algoritmo que mejor evaluación obtuvo.

Tabla 7: Estudios relacionados con lista de algoritmos y métricas de evaluación utilizadas

Estudio	Algoritmos utilizados	Métricas de Evaluación	Mejor Algoritmo
(Dhaliwal & Williams, 2024)	Principal components regression, Partial least squares regression, Multiple linear regression, Regularized regression, MARS, Random Forest	RMSE	Random Forests
(Gowda & Reddy, 2020)	Polynomial regression, Decision Tree, Random Forests	Accuracy	Random Forests

(Hu et al., 2023)	Bayesian ensemble model (BM), ElasticNet, Neural Network, MARS, SVM, Random Forests, and XGBoost	RMSE	Bayesian ensemble model (BM)
(Choudhary et al., 2020)	k-NN, SVM, XGBoost, Gradient Boosting, LightGBM, Random Forest, Decision Trees, LR	RMSE, R^2	XGBoost
(Çatal, Çelik, & Bakoglu, 2023)	MARS, Bagging MARS	R^2 , RMSE, RRMSE	Bagging MARS
(Khan et al., 2022)	Random Forests	MSE, R^2	Random Forests
(Abbas et al., 2020)	LR, EN, k-NN, and SVM	RMSE	SVM
(Khaki & Wang, 2019)	DNN, Lasso, SNN, RT	RMSE	DNN

FUENTE: AUTOR

De acuerdo con los estudios presentados en la Tabla 7 para este proyecto se seleccionaron los algoritmos más comunes y que mejores resultados obtuvieron: Linear Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Multi-layer Perceptron (MLP), Deep Neural Network (DNN), Gradient Boosting Tree (GBT) y XGBoost (XGB). Las métricas de evaluación más usadas de acuerdo con los estudios analizados son Raíz Cuadrada del Error Cuadrático Medio (RMSE), Coeficiente de Determinación (R^2), Error Cuadrado Medio (MSE) y Error Absoluto Medio (MAE). Para el presente estudio se tomó

como métricas de evaluación a la Raíz Cuadrada del Error Cuadrático Medio (RMSE) y el Coeficiente de Determinación (R^2).

4.2.2 Selección de Rango de Valores para Hiperparámetros de Algoritmos

Dentro del proceso de la construcción de los modelos con los algoritmos seleccionados se tienen una amplia variedad de parámetros de configuración (hiperparámetros) los cuales fueron seleccionados y se establecieron sus rangos de valores en base a estudios previos y también algunos de los parámetros y rangos de valores fueron propuestos en el presente estudio tal como se puede evidenciar en la Tabla 8.

Tabla 8: Algoritmos rango de valores de hiperparámetros

Algoritmo	Parámetro	Valores	Fuente
LR	fit_intercept	False, True	(Abbas et al., 2020)
	n_jobs	None	
	n_estimators	100 - 300	(Guimarães et al., 2024)
	max_depth	1 - 10	
RF	criterion	'absolute_error', 'squared_error', 'poisson', 'friedman_mse'	Propuesta
	max_features	'sqrt', 'log2', 1.0	
SVM	kernel	'linear', 'rbf'	(Guimarães et al., 2024)
	C	0.01 - 100	(Abbas et al., 2020)
	gamma	0.01 - 100	
MLP	hidden_layer_sizes	(50,50,50), (50,100,50), (100,50)	(Oikonomidis, Catal, & Kassahun, 2022)
	learning_rate	'constant','adaptive'	
	max_iter	100 - 500	
	activation	'relu','tanh'	Propuesta
	alpha	0.0001 - 0.05	
	learning_rate_init	0.001 - 0.05	
	solver	'adam', 'sgd'	

DNN	optimizer	'adam', 'SGD', 'rmsprop'	Propuesta
	activations	'relu', 'tanh', 'sigmoid'	
	neurons	[50,50,50], [50,100,50], [100,50]	(Oikonomidis et al., 2022)
	epochs	100 - 500	
GBT	n_estimators	10 - 1000	(Mustapha et al., 2024)
	learning_rate	0.01 - 0.5	
	max_features	0.5 - 1	
	subsample	0.7 - 1	
	max_depth	1 - 10	
	criterion	'friedman_mse', 'squared_error'	Propuesta
	loss	'squared_error', 'absolute_error', 'huber', 'quantile'	
XGB	n_estimators	10 - 1000	(Mustapha et al., 2024)
	learning_rate	0.01 - 0.5	
	subsample	0.7 - 1	
	max_depth	1 - 10	

FUENTE: AUTOR

4.2.3 Preprocesamiento de Datos

Adicional a las transformaciones que se realizaron para obtener la matriz de correlación, los algoritmos seleccionados requirieron que los datos estén normalizados por lo cual se aplicó la normalización min-max.

Para la reducción de la dimensionalidad de los datos se realizó un análisis de correlación de Spearman y la Eliminación Recursiva de Características (RFE), el cual permitió ver la relevancia de las variables predictoras en cada una de las variables objetivo de pesos frescos, resultado de ello se estableció como variables independientes o predictoras DISTANCIA_SIEMBRA, DIAMETRO_TALLO_EV, ALTURA_PLANTA_EV, INSERCIÓN_MAZORCA_ER, D_ADVANTA, D_H551, D_H_554 y D_QPM_543.

Finalmente se dividió los datos previamente preprocesados en una proporción de 70 % para entrenamiento de los modelos y 30 % para la evaluación. A continuación, se presenta los tiempos de ajuste de hiperparámetros y las métricas de evaluación de RMSE y R^2 .

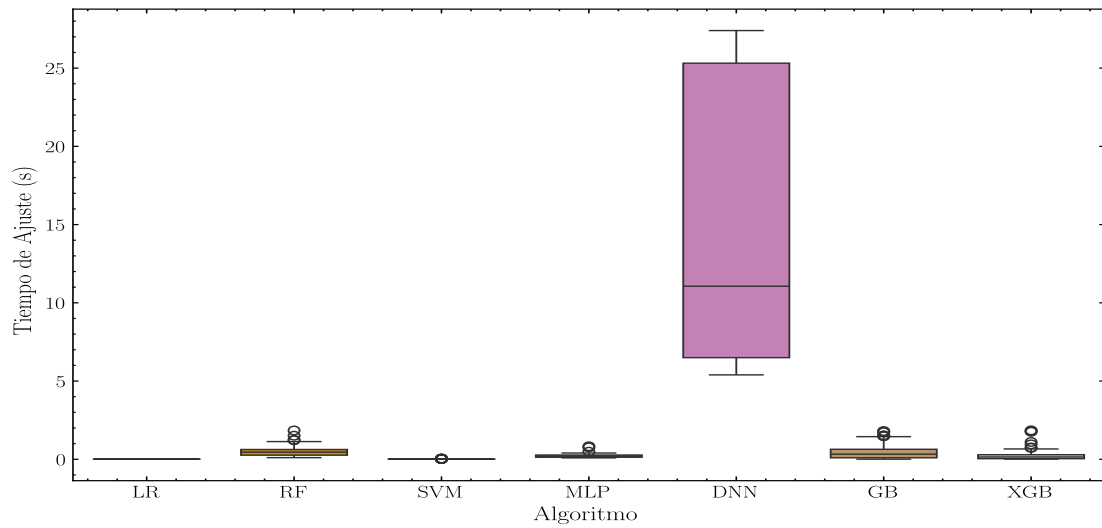
4.2.4 Tiempo de Ajuste de Hiperparámetros

Los tiempos de ajuste de hiperparámetros para los modelos de Peso Total, Peso Hojas, Peso Tallo y Peso Mazorca se registraron milisegundos (ms). El algoritmo de DNN fue el que mayor tiempo de ajuste de hiperparámetros en promedio obtuvo por lo que se optó por mostrar dos gráficas, una en segundos para mostrar los resultados generales y otra gráfica mostrando los tiempos de ajuste en milisegundos para detallar las diferencias entre el resto de los algoritmos.

4.2.4.1 Tiempo de Ajuste para Peso Total

En el análisis del tiempo de ajuste de hiperparámetros para el modelo de peso de total, medido en segundos, se observó una variabilidad considerable en los tiempos registrados. El algoritmo de DNN para Peso Total tiene el mayor tiempo de ajuste de hiperparámetros siendo el valor máximo de 27 segundos y una mediana de 11 segundos tal como lo presenta la Ilustración 24.

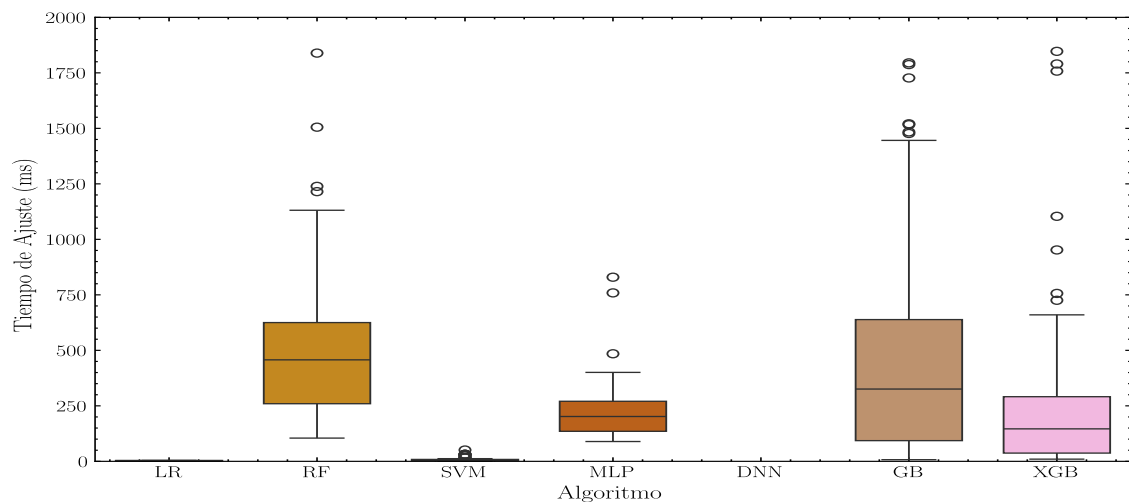
Ilustración 24: Tiempos de ajuste de hiperparámetros para modelos de peso total



FUENTE: AUTOR

El comportamiento del resto de algoritmos para Peso Total fue marcado por un menor tiempo de ajuste, siendo el algoritmo de LR el que menor tiempo obtuvo con una media de 3 ms seguido por el algoritmo SVM con una media de 9 ms. La Ilustración 25 muestra la distribución de los tiempos de ajuste de los diferentes algoritmos en milisegundos donde se evidencia que después de DNN las medianas en orden descendente de los tiempos de ajuste son de RF, GB, MLP y XGB.

Ilustración 25: Tiempos de ajuste de hiperparámetros en ms para modelos de peso total

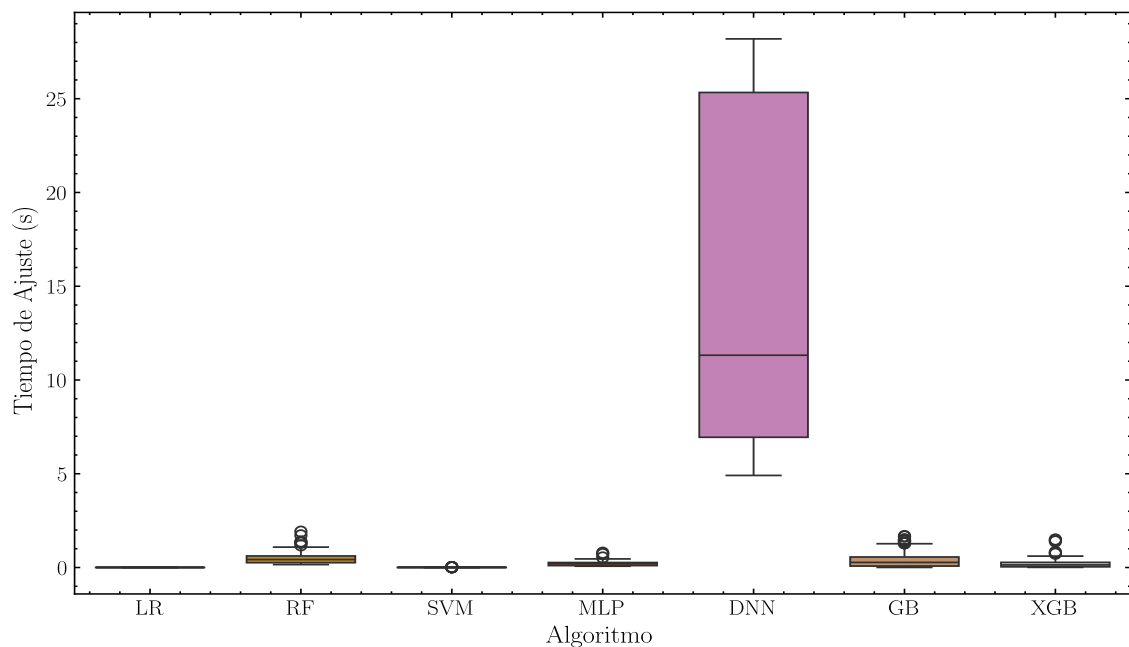


FUENTE: AUTOR

4.2.4.2 Tiempo de Ajuste para Peso Hojas

Los tiempos de ajuste para Peso Hojas del algoritmo DNN reflejaron tiempos con una mediana de 11 segundos y un máximo de 28 segundos, estos resultados son similares a los obtenidos para tal el modelo de Peso Total, los resultados se visualizan en la Ilustración 26.

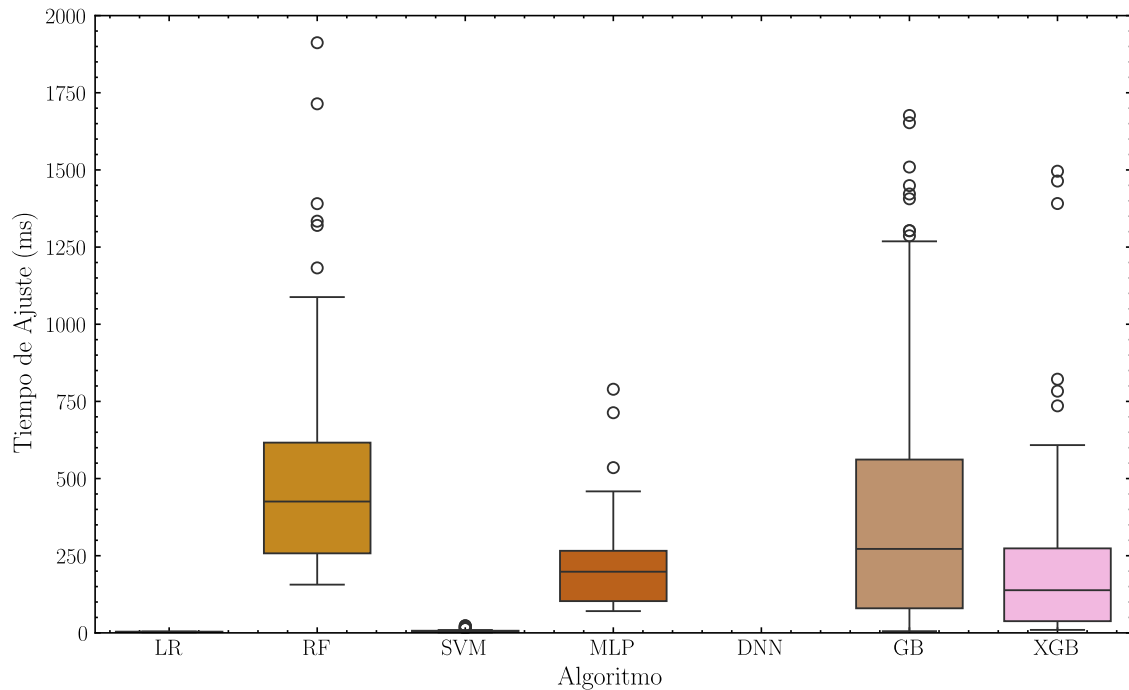
Ilustración 26: Tiempos de ajuste de hiperparámetros para modelos de peso hojas



FUENTE: AUTOR

Los algoritmos de LR y SVM demostraron tiempos de ajuste más rápidos, con medianas de 3 ms y 5 ms, respectivamente. Por su parte XGB y MLP registraron medianas de tiempos de 138 ms y 198 ms. La variabilidad en estos tiempos se refleja en los rangos intercuartílicos (IQR), indicando diferencias notables en la eficiencia de ajuste entre los algoritmos, como se observa en la Ilustración 27.

Ilustración 27: Tiempos de ajuste de hiperparámetros en ms para modelos de peso hojas

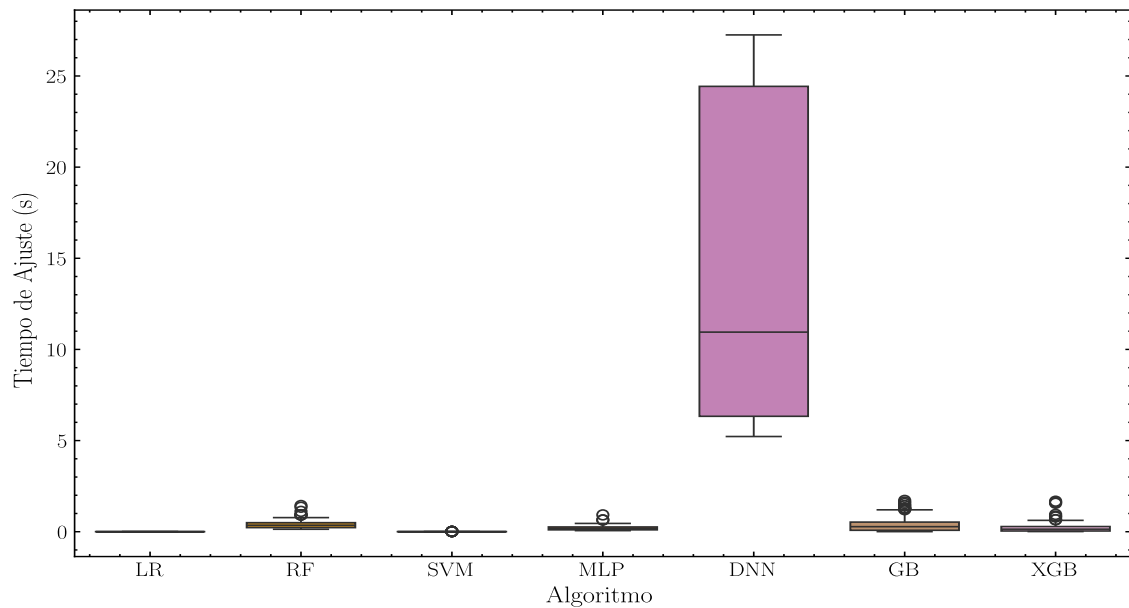


FUENTE: AUTOR

4.2.4.3 Tiempo de Ajuste para Peso Tallo

Los tiempos de ajuste de hiperparámetros para Peso Tallo de DNN muestra una desviación estándar relativamente alta lo que apoya la idea de que hay una variabilidad considerable en los tiempos de ajuste. DNN tuvo una mediana de ajuste de 11 segundos y un tiempo máximo de 27 segundos, lo que refleja un tiempo considerable comparado con otros algoritmos, como indica la Ilustración 28.

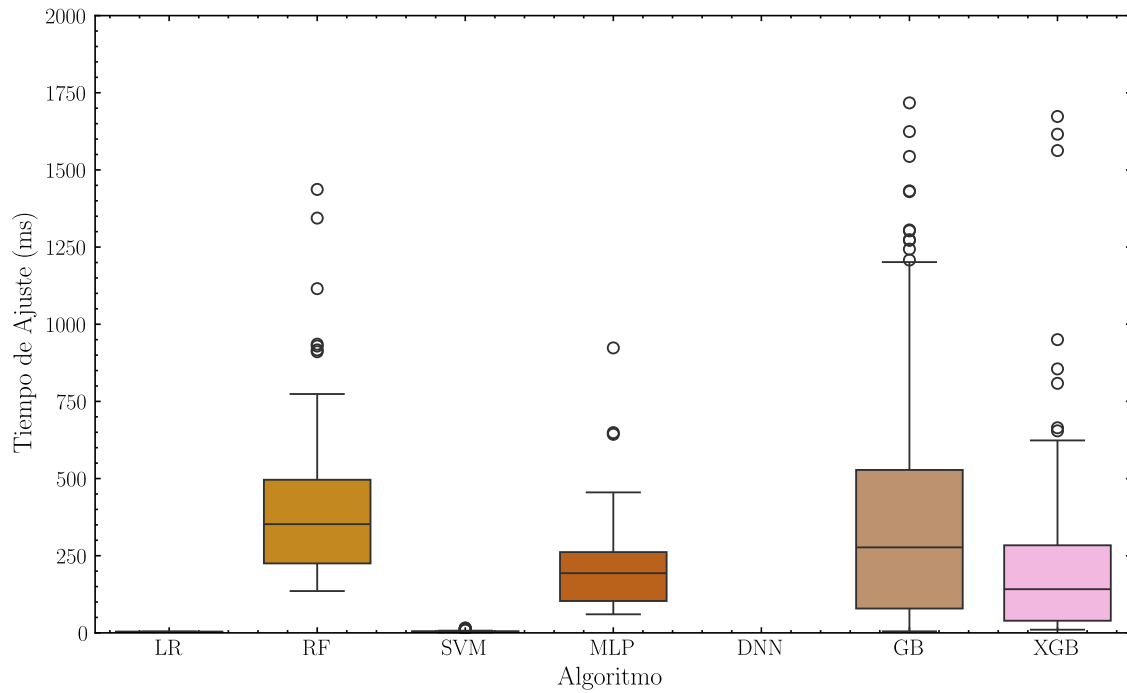
Ilustración 28: Tiempos de ajuste de hiperparámetros para modelos de peso tallo



FUENTE: AUTOR

En el análisis del tiempo de ajuste de hiperparámetros para el modelo de peso de tallo, medido en milisegundos (ms), se observaron variaciones significativas entre los diferentes algoritmos. LR tuvo el tiempo de ajuste más rápido, con una mediana de solo 3 ms, indicando una alta eficiencia en el ajuste de hiperparámetros. En contraste, RF mostró una mediana de 475 ms. GB presentó un tiempo de ajuste intermedio, con una mediana de 272 ms, debido a su naturaleza iterativa y compleja. Estos resultados, representados en la Ilustración 29, destacan las diferencias en los tiempos de ajuste de hiperparámetros entre los algoritmos.

Ilustración 29: Tiempos de ajuste de hiperparámetros en ms para modelos de peso tallo

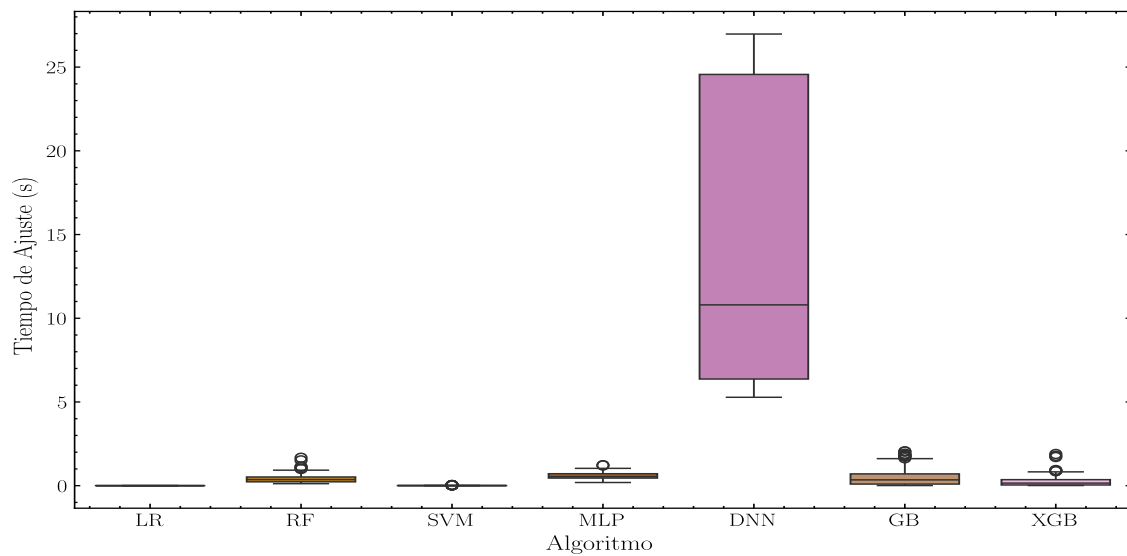


FUENTE: AUTOR

4.2.4.4 Tiempo de Ajuste para Peso Mazorca

El algoritmo DNN presentó un tiempo mínimo de 5.28 segundos y un tiempo máximo de 26.97 segundos. La mediana del tiempo de ajuste fue de 10.80 segundos, indicando que la mitad de los tiempos fueron menores a este valor. El primer cuartil fue de 6.37 segundos y el tercer cuartil de 24.56 segundos, resultando en un rango intercuartílico (IQR) de 18.19 segundos, lo que refleja una dispersión significativa entre los tiempos de ajuste, estos resultados se presentan en la Ilustración 30.

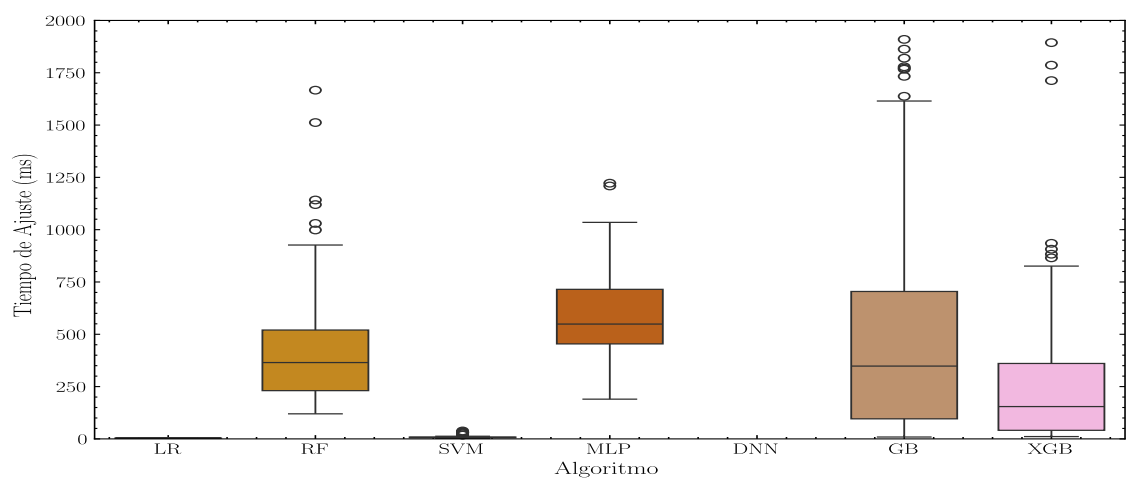
Ilustración 30: Tiempos de ajuste de hiperparámetros en para modelos de peso mazorca



FUENTE: AUTOR

LR presento los menores tiempos de ajuste de hiperparámetros para Peso Mazorca con una mediana de 3 ms. GB tuvo una mediana de 348 ms, reflejando un tiempo de ajuste considerablemente mayor comprado con LR pero menor que DNN. MLP y RF mostraron medianas en los tiempos de ajuste de 549 ms y 365 ms, como lo indica la Ilustración 31.

Ilustración 31: Tiempos de ajuste de hiperparámetros en ms para modelos de peso mazorca



FUENTE: AUTOR

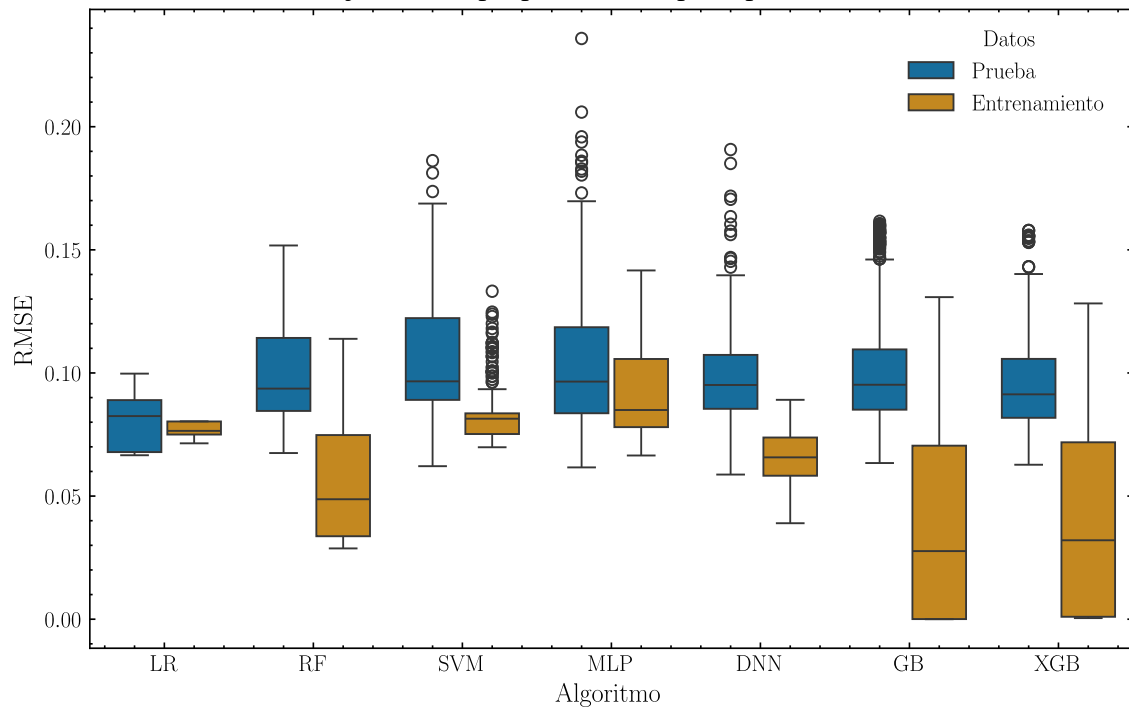
4.2.5 Evaluación de Métricas en Ajuste de Hiperparámetros

Adicionalmente se realizó la evaluación mediante las métricas RMSE y R^2 tanto para datos de entrenamiento como de pruebas para los diferentes algoritmos seleccionados los cuales se muestran a continuación organizados por variable objetivo de Peso Total, Peso Hojas, Peso Tallo y Peso Mazorca. Cabe destacar que mayores valores de R^2 indican un mejor ajuste del modelo, mientras que valores altos de RMSE reflejan una menor precisión del modelo.

4.2.5.1 Métricas de Evaluación para Peso Total

En los valores de la métrica de evaluación de RMSE obtenidos por los modelos contruidos por los algoritmos de LR, SVM y MLP mostraron una estabilidad considerable entre los valores de RMSE en la fase de entrenamiento y prueba. Los modelos de Peso Total que mejores resultados obtuvieron en la fase de entrenamiento fueron los que se entrenaron con los algoritmos de GB y XGB. En resumen, si bien todos los modelos presentan incrementos en RMSE en el conjunto de prueba, algunos, como LR y MLP, muestran estabilidad entre los conjuntos de datos tal como lo indica la Ilustración 32.

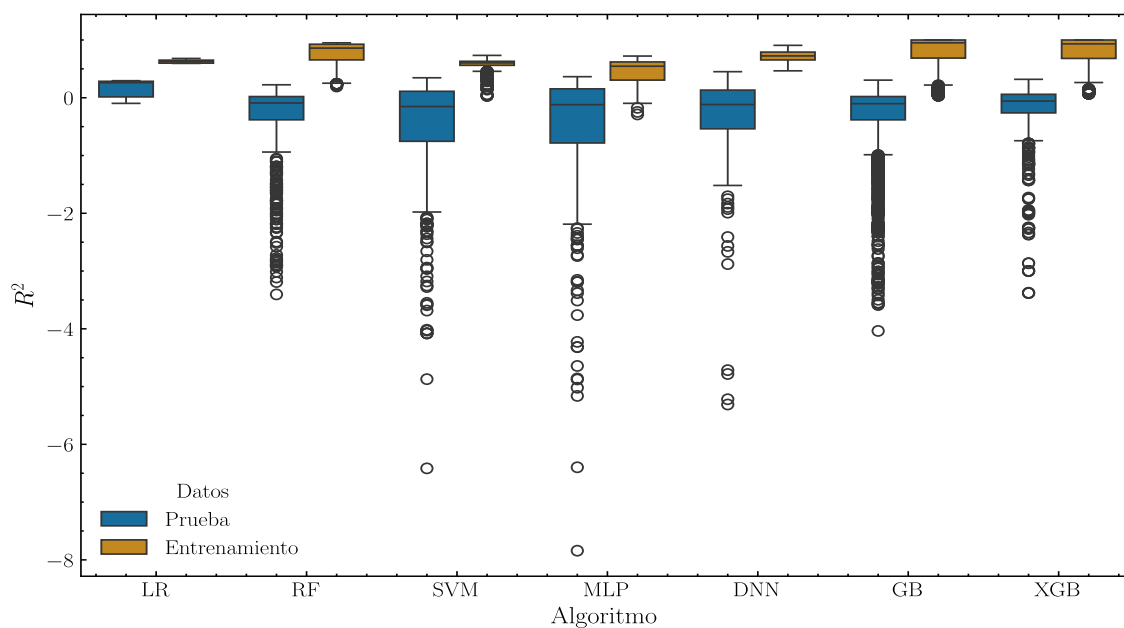
Ilustración 32: RMSE en ajuste de hiperparámetros para peso total



FUENTE: AUTOR

En relación con la métrica de evaluación de R^2 el análisis mostró que muchas de las combinaciones de hiperparámetros provocaba que los modelos presenten un rendimiento deficiente, sin embargo, la mediana de los modelos GB y XGB destacaron en la fase de entrenamiento, pero no en la de pruebas. Los modelos de LR en cambio mostraron un ajuste a los datos moderado en la fase de entrenamiento, pero un bajo ajuste en la fase de prueba, GB en la fase de entrenamiento alcanzó el valor máximo de 1.00, con una mediana de 0.95 y un rango intercuartílico de 0.31, indicando una alta consistencia con sus resultados, pero también un posible sobreajuste a los datos con ciertas configuraciones de hiperparámetros. Los resultados se los presenta en la Ilustración 33.

Ilustración 33: R^2 en ajuste de hiperparámetros para peso total

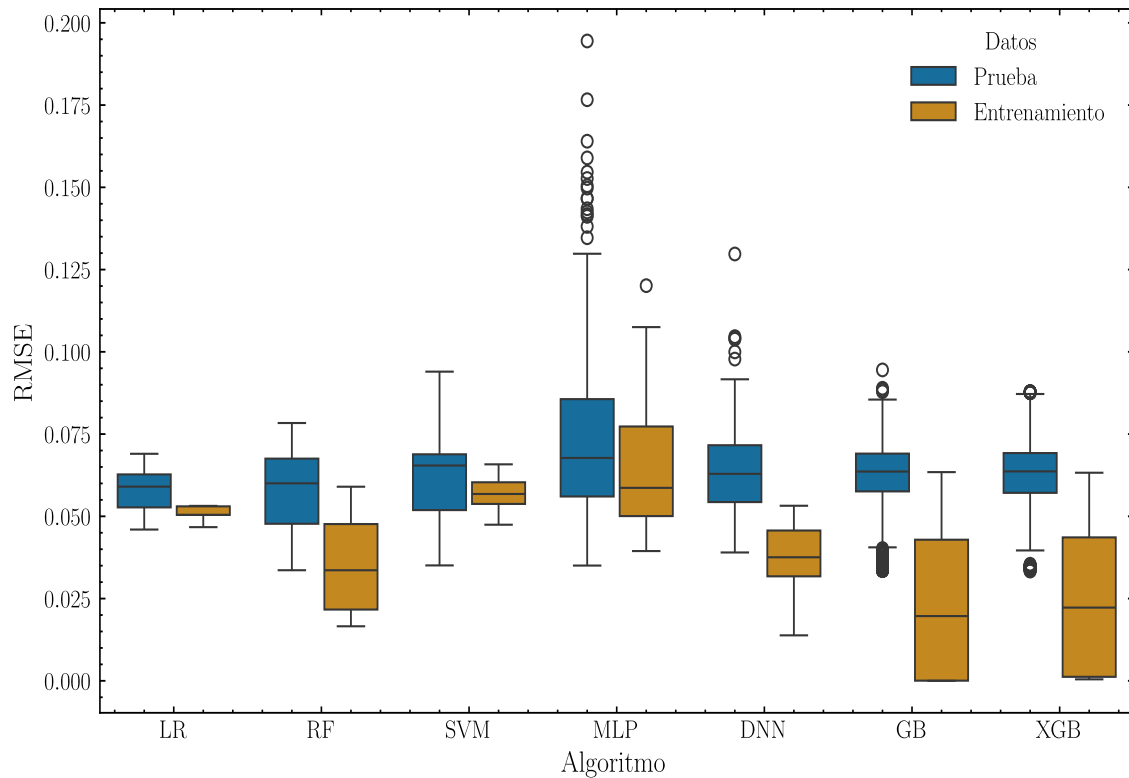


FUENTE: AUTOR

4.2.5.2 Métricas de Evaluación para Peso Hojas

Al igual que los modelos de Peso Total, los modelos de Peso Hojas que mejores resultados obtuvieron en la fase de entrenamiento fueron los que se entrenaron con los algoritmos de GB y XGB. MLP en el ajuste de hiperparámetros presentó mayor variabilidad en RMSE con los datos de prueba demostrando una precisión débil de los modelos construidos y destacando la importancia del ajuste de hiperparámetros para poder obtener mejores modelos a partir de un mismo algoritmo. En la Ilustración 34 se presentan los resultados.

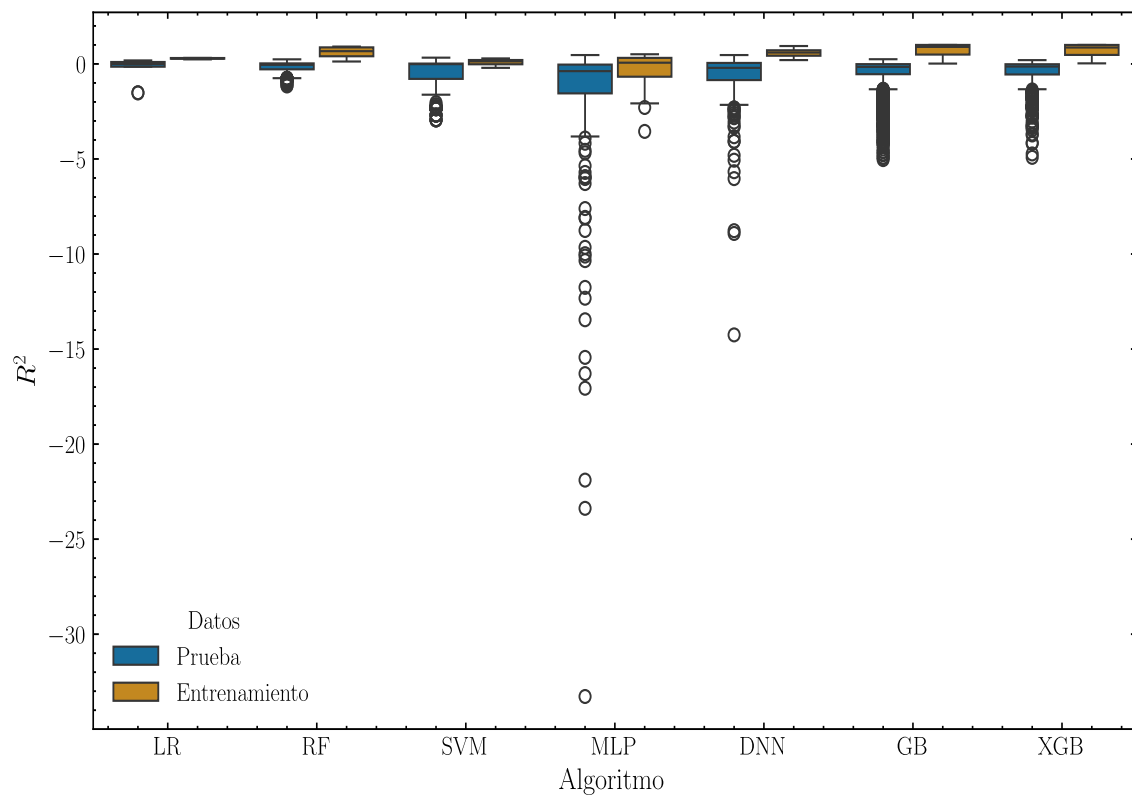
Ilustración 34: RMSE en ajuste de hiperparámetros para peso hojas



FUENTE: AUTOR

En el ajuste de hiperparámetros para la métrica R^2 del peso de hojas, los algoritmos de GB y DNN mostraron ser notables durante el entrenamiento, con GB alcanzando un máximo de 1.0 y una mediana significativamente alta de 0.896459. Sin embargo, ambos algoritmos experimentaron una caída en su desempeño en la fase prueba, donde DNN mostró una mediana negativa de -0.205996 y GB de -0.15544, lo que puede ser causado por un sobre ajuste a los datos de entrenamiento, tal como lo indica la Ilustración 35.

Ilustración 35: R^2 en ajuste de hiperparámetros para peso Hojas

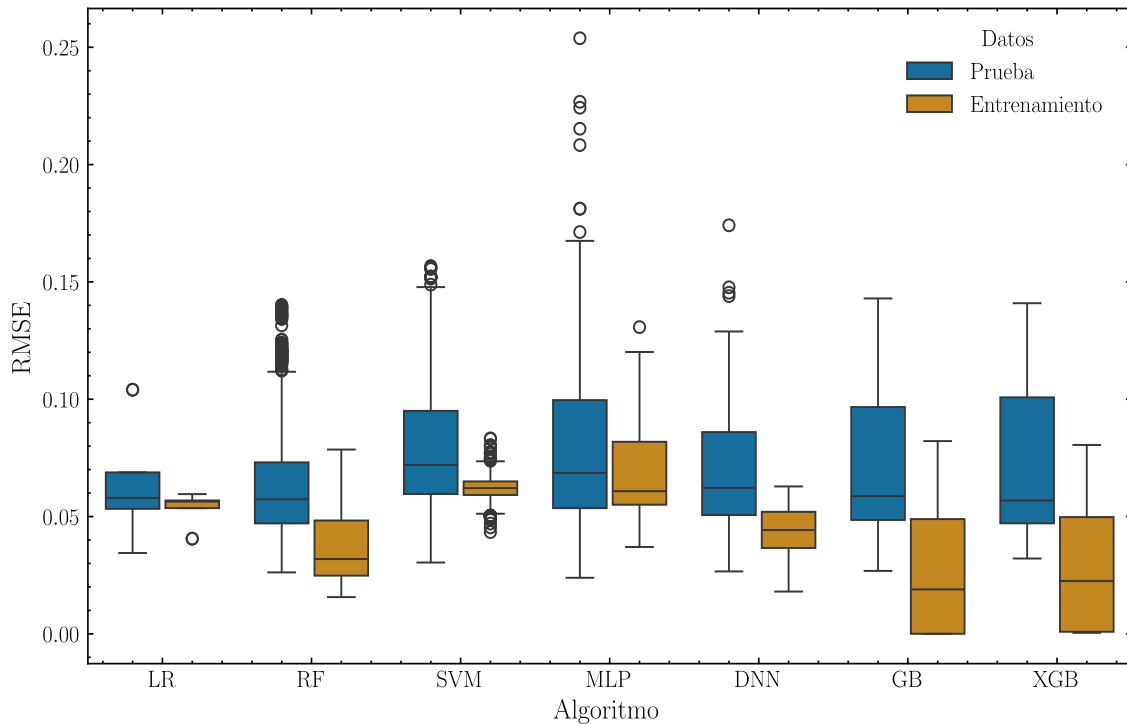


FUENTE: AUTOR

4.2.5.3 Métricas de Evaluación para Peso Tallo

En el ajuste de hiperparámetros para la métrica de RMSE del peso de tallo RF mostró una gran consistencia con medianas de 0.032 en entrenamiento y 0.057 en la prueba, presentando un desempeño robusto, demostrando una alta precisión y estabilidad entre entrenamiento y pruebas como lo muestra la Ilustración 36.

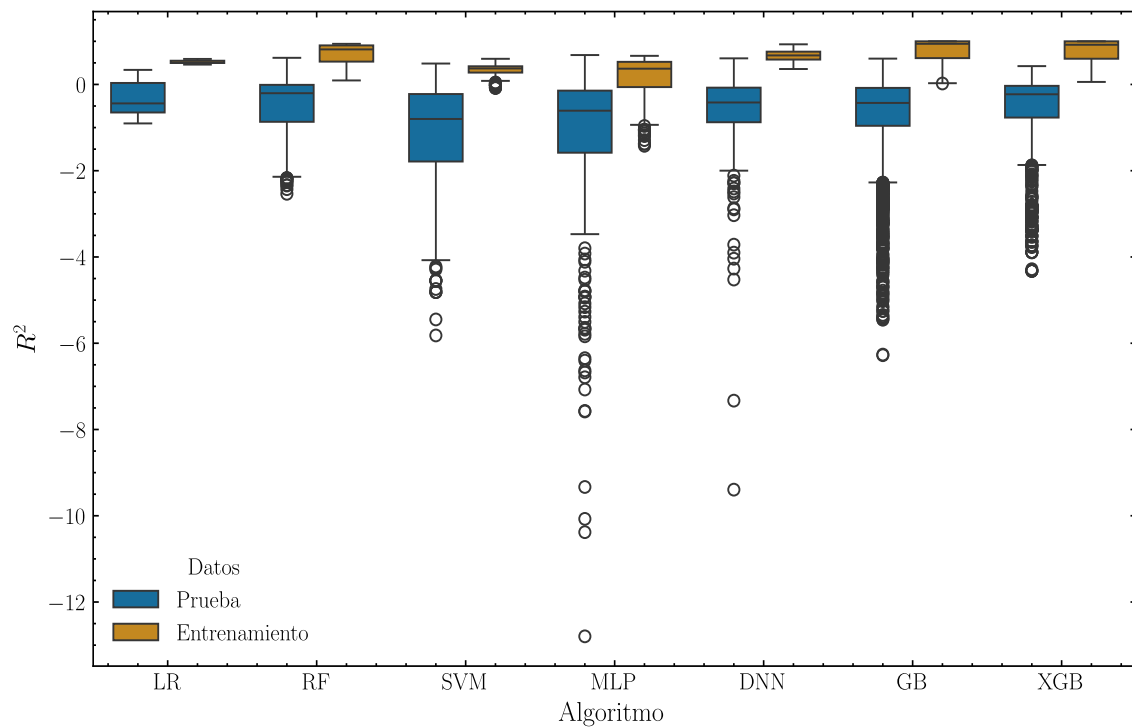
Ilustración 36: RMSE en ajuste de hiperparámetros para peso tallo



FUENTE: AUTOR

Para el ajuste de hiperparámetros con la métrica R^2 , en el peso del tallo, se observó una notable diferencia entre los resultados de entrenamiento y prueba en todos los algoritmos. GB y XGB presentaron altos valores de R^2 en el entrenamiento, alcanzando medias de 0.786767 y 0.77785, lo que demuestra un excelente ajuste a los datos de entrenamiento. No obstante, ambos experimentaron un declive significativo en la fase prueba con medias de -0.701331 y -0.660521, indicando un marcado sobreajuste a los datos de entrenamiento y un mal rendimiento predictivo del modelo. Por otro lado, LR mostró un comportamiento más estable entre el entrenamiento y la prueba, aunque con un rendimiento inferior comparado con GB y XGB como lo presenta la Ilustración 37.

Ilustración 37: R^2 en ajuste de hiperparámetros para peso tallo

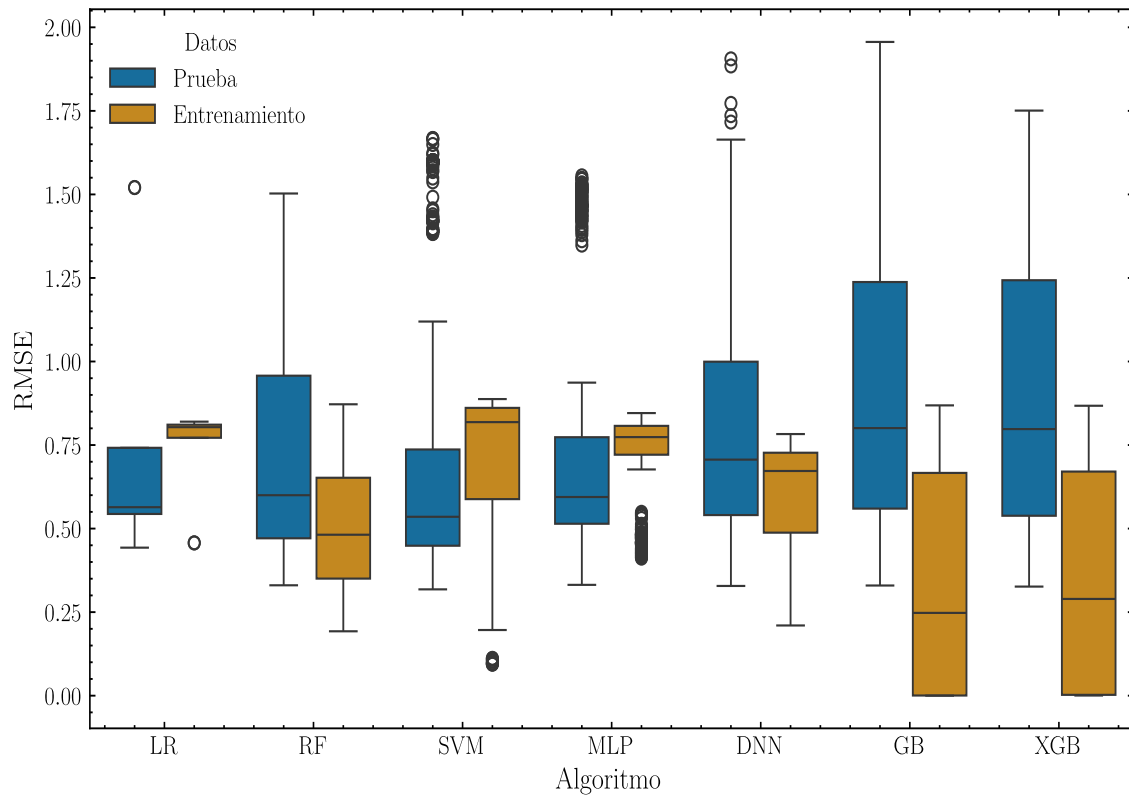


FUENTE: AUTOR

4.2.5.4 Métricas de Evaluación para Peso Mazorca

En la evaluación del RMSE para el peso de mazorca, los algoritmos mostraron una variabilidad entre el entrenamiento y las pruebas. GB y XGB se destacaron en el entrenamiento con medias de 0.321589 y 0.32589 y 0.337355, reflejando errores en la predicción de los datos de entrenamiento. Sin embargo, en la fase prueba, ambos experimentaron un incremento en la mediana y la media del RMSE, indicando una pérdida de precisión al generalizar nuevos datos tal como lo presenta la Ilustración 38.

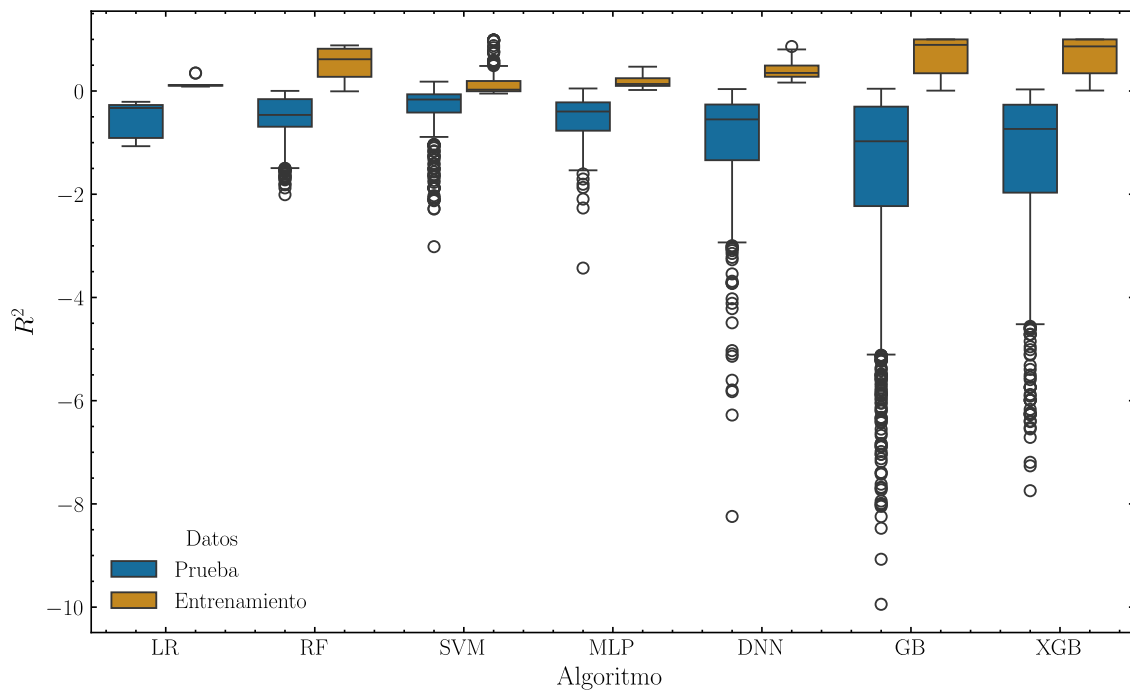
Ilustración 38: RMSE en ajuste de hiperparámetros para peso mazorca



FUENTE: AUTOR

DNN, GB, y XGB mostraron valores negativos de R^2 en los datos de prueba con medias de -1.077558, -1.499067 y -1.364395 respectivamente, indicando que estas configuraciones de hiperparámetros no generalizaron bien y que sus predicciones fueron peores que un modelo que simplemente predice la media. RF y LR presentaron resultados de R^2 negativos en los datos de prueba, pero con valores menos extremos, sugiriendo un rendimiento moderado y más consistente en comparación con DNN, GB y XGB, como lo indica la Ilustración 39.

Ilustración 39: R^2 en ajuste de hiperparámetros para peso mazorca



FUENTE: AUTOR

4.2.6 Valores de Hiperparámetros

De acuerdo con los resultados obtenidos por el proceso de ajuste de hiperparámetros se elabora la Tabla 9 donde se indican los valores a cada uno de los parámetros de los algoritmos y sus respectivos modelos de aplicación como lo es Peso Total, Peso Hojas, Peso Tallo y Peso Mazorca. Estos valores fueron seleccionados en base a la métrica de RMSE.

Tabla 9: Valores de hiperparámetros por modelo y algoritmo

Modelo	Algoritmo	Mejores Parametros de Configuración
PESO_TOTAL_KG	LR	{'fit_intercept': False, 'n_jobs': None}
PESO_TOTAL_KG	RF	{'criterion': 'friedman_mse', 'max_depth': 5, 'max_features': 1.0, 'n_estimators': 100}
PESO_TOTAL_KG	SVM	{'C': 1, 'gamma': 0.01, 'kernel': 'linear'}
PESO_TOTAL_KG	MLP	{'activation': 'tanh', 'alpha': 0.05, 'hidden_layer_sizes': (100, 50), 'max_iter': 500, 'solver': 'adam'}
PESO_TOTAL_KG	DNN	{'activations': 'relu', 'epochs': 200, 'neurons': [100, 50], 'optimizer': 'SGD'}
PESO_TOTAL_KG	GB	{'learning_rate': 0.1, 'max_depth': 1, 'n_estimators': 100, 'subsample': 0.7}
PESO_TOTAL_KG	XGB	{'learning_rate': 0.1, 'max_depth': 1, 'n_estimators': 100, 'subsample': 0.8}
PESO_HOJAS_KG	LR	{'fit_intercept': False, 'n_jobs': None}
PESO_HOJAS_KG	RF	{'criterion': 'absolute_error', 'max_depth': 5, 'max_features': 'log2', 'n_estimators': 300}
PESO_HOJAS_KG	SVM	{'C': 1, 'gamma': 0.01, 'kernel': 'rbf'}
PESO_HOJAS_KG	MLP	{'activation': 'relu', 'alpha': 0.05, 'hidden_layer_sizes': (50, 50, 50), 'max_iter': 200, 'solver': 'adam'}

PESO_HOJAS_KG	DNN	{'activations': 'relu', 'epochs': 200, 'neurons': [100, 50], 'optimizer': 'SGD'}
PESO_HOJAS_KG	GB	{'learning_rate': 0.01, 'max_depth': 5, 'max_features': 0.5, 'n_estimators': 100, 'subsample': 0.7}
PESO_HOJAS_KG	XGB	{'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 100, 'subsample': 0.7}
PESO_TALLO_KG	LR	{'fit_intercept': False, 'n_jobs': None}
PESO_TALLO_KG	RF	{'criterion': 'absolute_error', 'max_depth': 5, 'max_features': 'log2', 'n_estimators': 300}
PESO_TALLO_KG	SVM	{'C': 0.1, 'gamma': 0.01, 'kernel': 'linear'}
PESO_TALLO_KG	MLP	{'activation': 'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (50, 100, 50), 'max_iter': 200, 'solver': 'adam'}
PESO_TALLO_KG	DNN	{'activations': 'tanh', 'epochs': 500, 'neurons': [50, 50, 50], 'optimizer': 'SGD'}
PESO_TALLO_KG	GB	{'learning_rate': 0.01, 'max_depth': 5, 'max_features': 0.5, 'n_estimators': 300, 'subsample': 0.7}
PESO_TALLO_KG	XGB	{'learning_rate': 0.5, 'max_depth': 1, 'n_estimators': 10, 'subsample': 0.7}
PESO_MAZORCA_KG	LR	{'fit_intercept': False, 'n_jobs': None}
PESO_MAZORCA_KG	RF	{'criterion': 'absolute_error', 'max_depth': 1, 'max_features': 1.0, 'n_estimators': 300}

PESO_MAZORCA_KG	SVM	{'C': 1, 'gamma': 'auto', 'kernel': 'rbf'} {'activation': 'relu', 'alpha': 0.0001,
PESO_MAZORCA_KG	MLP	'hidden_layer_sizes': (50, 50, 50), 'max_iter': 100, 'solver': 'adam'}
PESO_MAZORCA_KG	DNN	{'activations': 'tanh', 'epochs': 100, 'neurons': [50, 100, 50], 'optimizer': 'SGD'} {'learning_rate': 0.1, 'max_depth': 1,
PESO_MAZORCA_KG	GB	'max_features': 0.7, 'n_estimators': 10, 'subsample': 0.7}
PESO_MAZORCA_KG	XGB	{'learning_rate': 0.05, 'max_depth': 1, 'n_estimators': 10, 'subsample': 1.0}

FUENTE: AUTOR

4.2.7 Discusión

La búsqueda de estudios relacionados los cuales permitieron seleccionar los algoritmos con mejor desempeño y los más utilizados concuerdan con el estudio de revisión de (Van, Kassahun, & Catal, 2020) donde se establece que los algoritmos más utilizados son Neural Networks, Linear Regression, Random Forest, SVM, Gradient Boosting. Igualmente, el ajuste de hiperparámetros puede mejorar significativamente el rendimiento de los modelos de regresión de aprendizaje automático esto se lo evidenció en los resultados obtenidos del proceso de ajuste (Isabona, Imoize, & Kim, 2022).

En el desarrollo de modelos predictivos para la agricultura, la selección adecuada de hiperparámetros es crucial para optimizar el rendimiento y la generalización de los

algoritmos. En este estudio, se evaluaron varios modelos de aprendizaje automático, incluidos Regresión Logística (LR), Random Forest (RF), Máquinas de Soporte Vectorial (SVM), Redes Neuronales (MLP y DNN), y métodos de ensamblaje como Gradient Boosting (GB) y XGBoost (XGB), cada uno con una configuración específica de hiperparámetros diseñada para maximizar su efectividad en la predicción del rendimiento de maíz forrajero. Similar al enfoque de Ramírez Gómez (2020) donde se priorizó la selección de hiperparámetros como 'max_depth', 'n_estimators' y 'learning_rate' para algoritmos de ensamblaje, y el ajuste de regularización en SVM y LR que maximizan la precisión y generalización de los modelos.

Comparando los resultados con estudios previos como el de Dey, Ferdous, & Ahmed (2024) que reportaron una precisión superior del modelo XGBoost, lo que es consistente con los hallazgos de la investigación, donde XGBoost también mostró un rendimiento destacado. Esto subraya la eficacia de XGBoost en manejar datasets complejos y su capacidad para adaptarse a diferentes características agronómicas y ambientales. Además, se identificaron configuraciones específicas para cada modelo que maximizaron la precisión predictiva, tales como: 'max_depth' y 'n_estimators' para RF y XGB, y 'C' y 'gamma' para SVM. Estas configuraciones son esenciales para manejar la variabilidad inherente en los datos de rendimiento de cultivos.

Comparativamente, el estudio referenciado de Dey et al. (2024) emplea SVM con un kernel de función de base radial (RBF) y parámetros de complejidad y gamma ajustados a $C=5$ y $\gamma=0.1$, respectivamente. Este conjunto de hiperparámetros, en especial el uso de un kernel RBF, se orienta a capturar relaciones no lineales complejas entre las

características, lo cual es adecuado para conjuntos de datos donde las interacciones entre variables son altamente no lineales y donde la dimensionalidad del espacio de características es significativa.

En la investigación de Srinivas Rao, Hema, Sai Priya, Vamsi Krishna, y Sakhavath Ali (2021) el regresor lineal implementado utilizó el hiperparámetro `fit_intercept`, el cual se estableció para calcular el término de intercepción del modelo, lo que es esencial cuando no se puede asumir que los datos están centrados en el origen. Este ajuste permite que el modelo adapte mejor la línea de regresión a los datos, proporcionando una precisión mejorada basada en sus características físicas. La opción de `n_jobs: None` indica que el entrenamiento del modelo no se paralelizó, lo que puede influir en el tiempo de procesamiento, pero es adecuado para conjuntos de datos de menor tamaño, como es el caso en este estudio, asegurando que los recursos computacionales se gestionen de manera efectiva sin necesidad de paralelización.

Estas configuraciones de hiperparámetros no solo reflejan las mejores prácticas en el modelado de aprendizaje automático, sino que también están alineadas con los hallazgos de este estudio, que sugieren una marcada mejora en el rendimiento de los modelos cuando se optimizan adecuadamente sus configuraciones. Así, esta investigación contribuye a un mejor entendimiento de cómo diversas configuraciones de hiperparámetros pueden ser explotadas eficazmente para predecir el rendimiento del maíz forrajero.

4.3 SELECCIÓN DEL MODELO PARA LA PREDICCIÓN DEL RENDIMIENTO DEL MAÍZ FORRAJERO

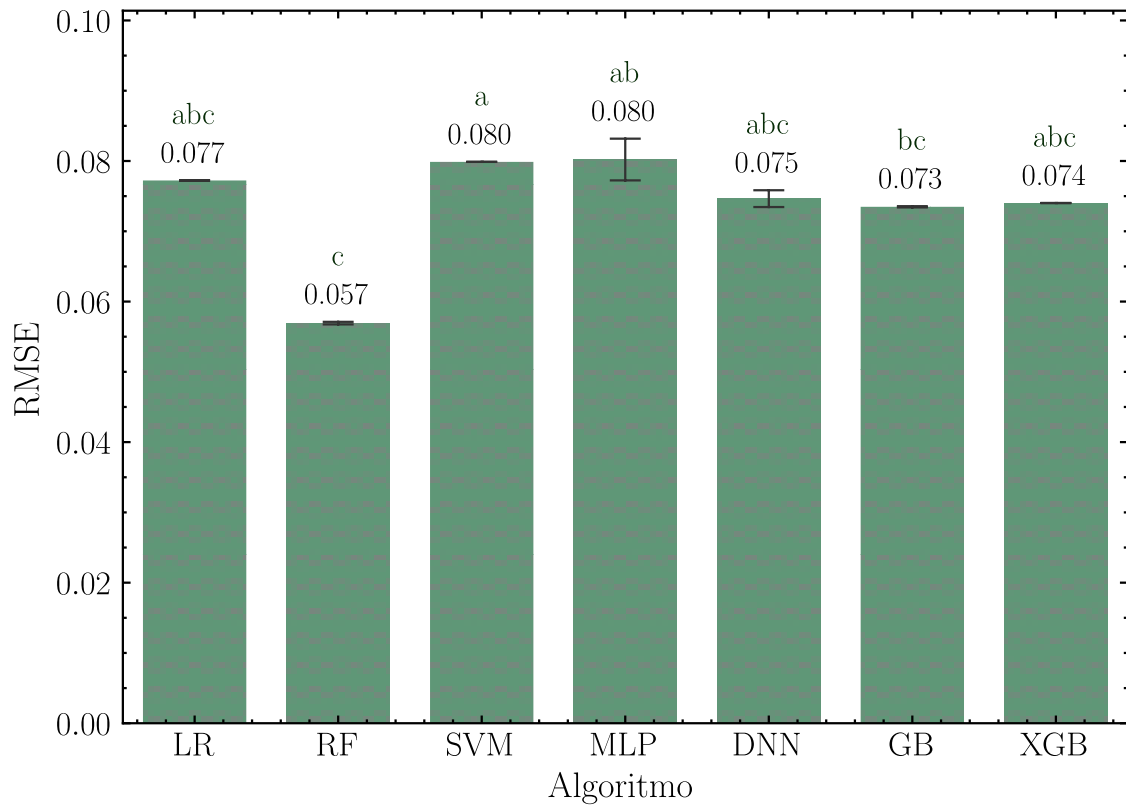
La selección del modelo que mejor desempeño obtuvo se lo realizó mediante la prueba de hipótesis de significancia estadística Friedman y la prueba post-hoc Nemenyi las cuales evaluaron las métricas de RMSE y R^2 . Adicionalmente se registraron los tiempos de entrenamiento y de prueba de los modelos construidos. Este proceso se lo realizó en 4 repeticiones con las mejores configuraciones de hiperparámetros encontrados en la fase anterior.

4.3.1 Evaluación de Modelos para Peso Total

4.3.1.1 Métricas de Evaluación con Datos de Entrenamiento

Los resultados mostraron que RF obtuvo el menor RMSE medio (0.057), indicando la mejor precisión de las predicciones en los datos de entrenamiento, seguido de GBT (0.073), XGB (0.074), y DNN (0.075). SVM y MLP presentaron los valores más altos de RMSE medio (0.080 y 0.080, respectivamente). La prueba de Friedman, seguida por la prueba post-hoc de Nemenyi reveló que RF (c) se diferencia significativamente de SVM (a) y MLP (ab), destacándose como el algoritmo con mejor rendimiento medio obtuvo en términos de RMSE, esto se lo presenta en la Ilustración 40.

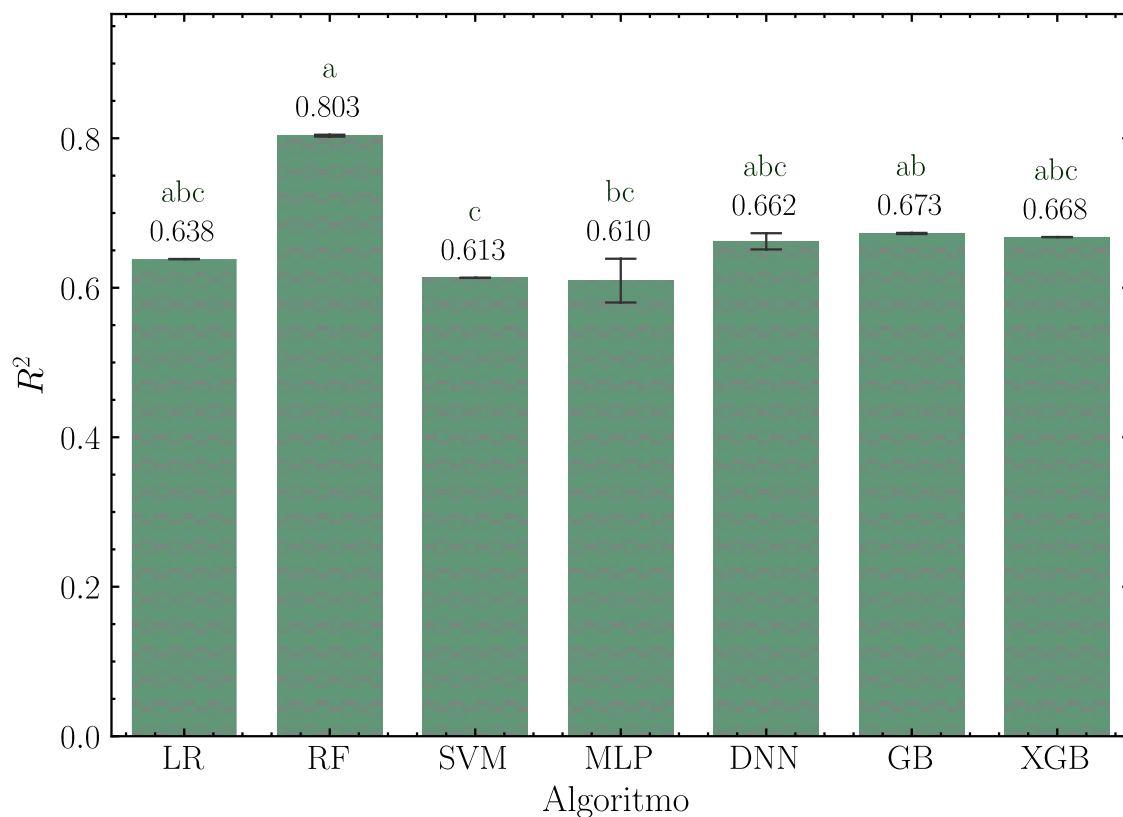
Ilustración 40: Métrica RMSE de datos de entrenamiento para Peso Total



FUENTE: AUTOR

El algoritmo RF obtuvo el valor medio más alto de R^2 (0.80), indicando que es el mejor modelo en términos de ajuste a los datos de entrenamiento. La prueba de Friedman, seguida por la prueba post-hoc de Nemenyi muestra que RF (a) es significativamente mejor que SVM (c) y MLP (bc). Sin embargo, RF no tiene diferencias significativas con los demás algoritmos, estos hallazgos se los presenta en Ilustración 41.

Ilustración 41: Métrica R^2 de datos de entrenamiento para Peso Total

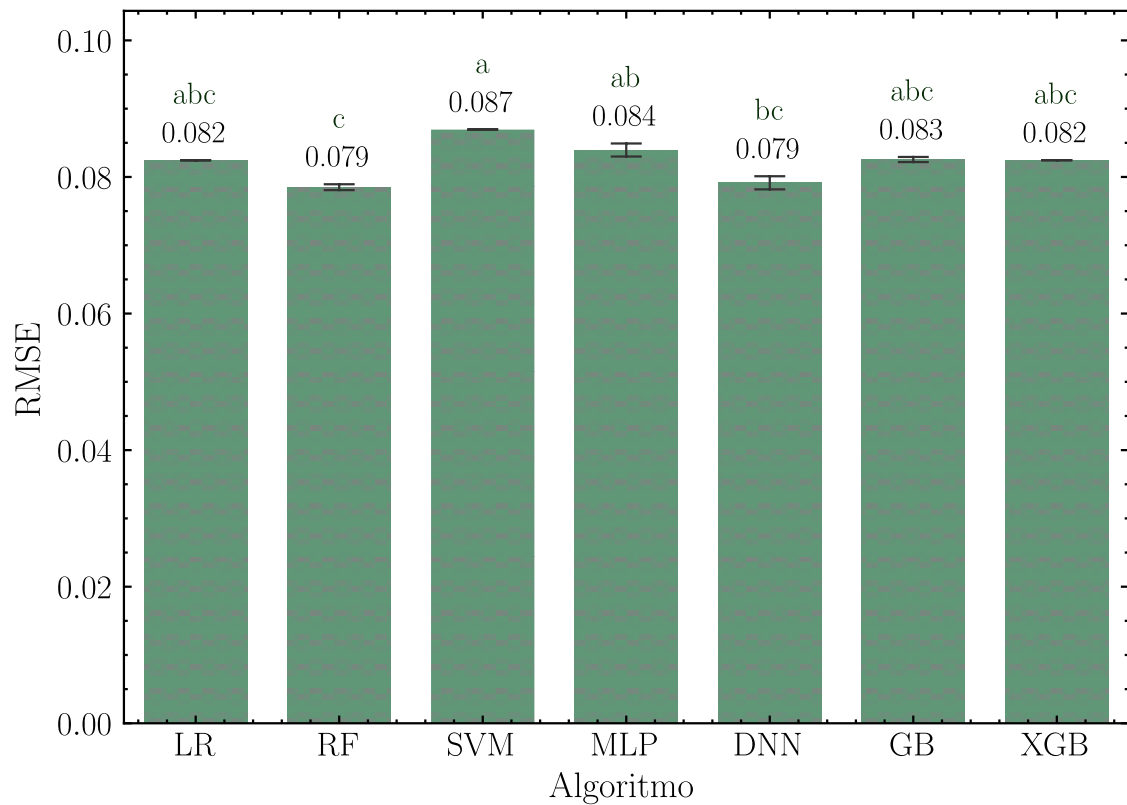


FUENTE: AUTOR

4.3.1.2 Métricas de Evaluación con Datos de Prueba

Los resultados mostraron que RF obtuvo el menor RMSE medio (0.079), indicando la mejor precisión de las predicciones en los datos de pruebas, seguido de DNN (0.079), LR (0.082), y XGB (0.082). SVM y MLP presentaron los valores más altos de RMSE medio (0.087 y 0.084, respectivamente). La prueba de Friedman, seguida por la prueba post-hoc de Nemenyi reveló que RF (c) se diferencia significativamente de SVM (a) y MLP (ab), destacándose como el algoritmo con mejor rendimiento medio obtuvo en términos de RMSE en la fase de prueba de los modelos, esto se lo presenta en la Ilustración 42.

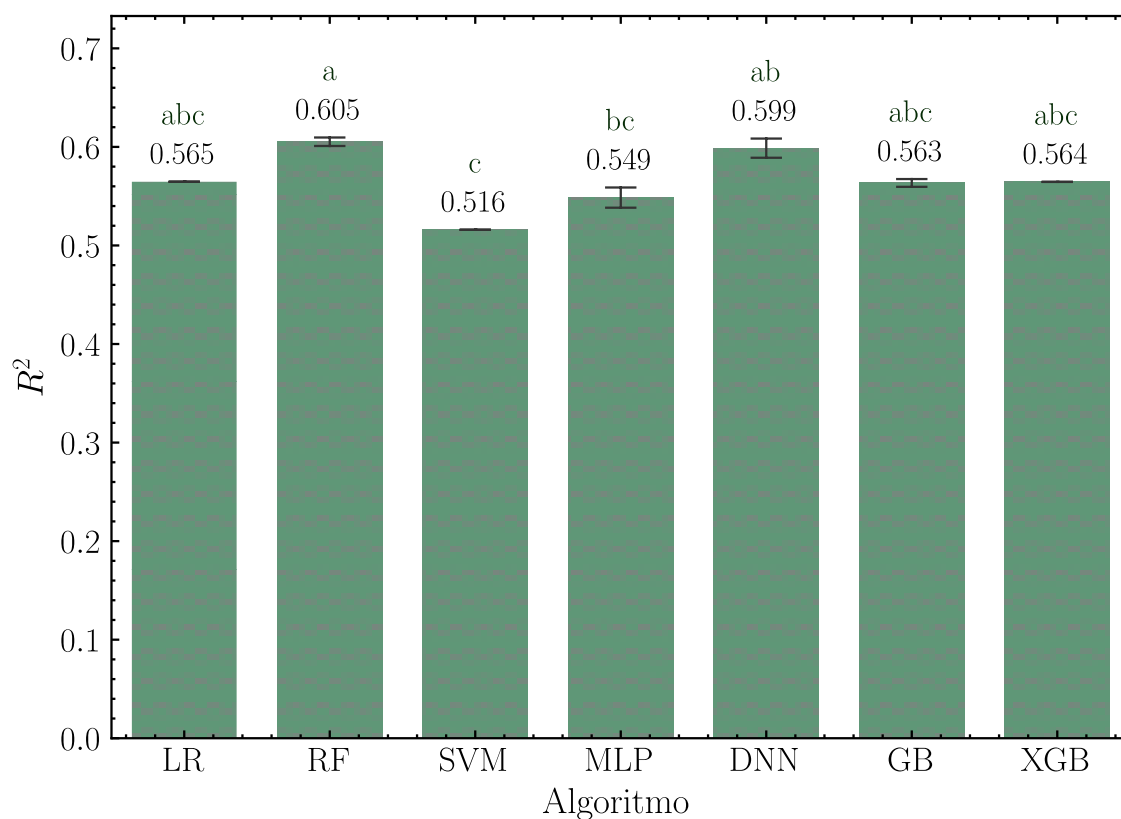
Ilustración 42: Métrica RMSE de datos de prueba para Peso Total



FUENTE: AUTOR

El algoritmo RF obtuvo el valor medio más alto de R^2 (0.605), indicando que es el mejor modelo en términos de ajuste a los datos de pruebas. La prueba de Friedman, seguida por la prueba post-hoc de Nemenyi muestra que RF (a) es significativamente mejor que SVM (c) y MLP (bc). Sin embargo, RF no tiene diferencias significativas con los demás algoritmos, estos hallazgos se los presenta en Ilustración 43.

Ilustración 43: Métrica R^2 de datos de prueba para Peso Total

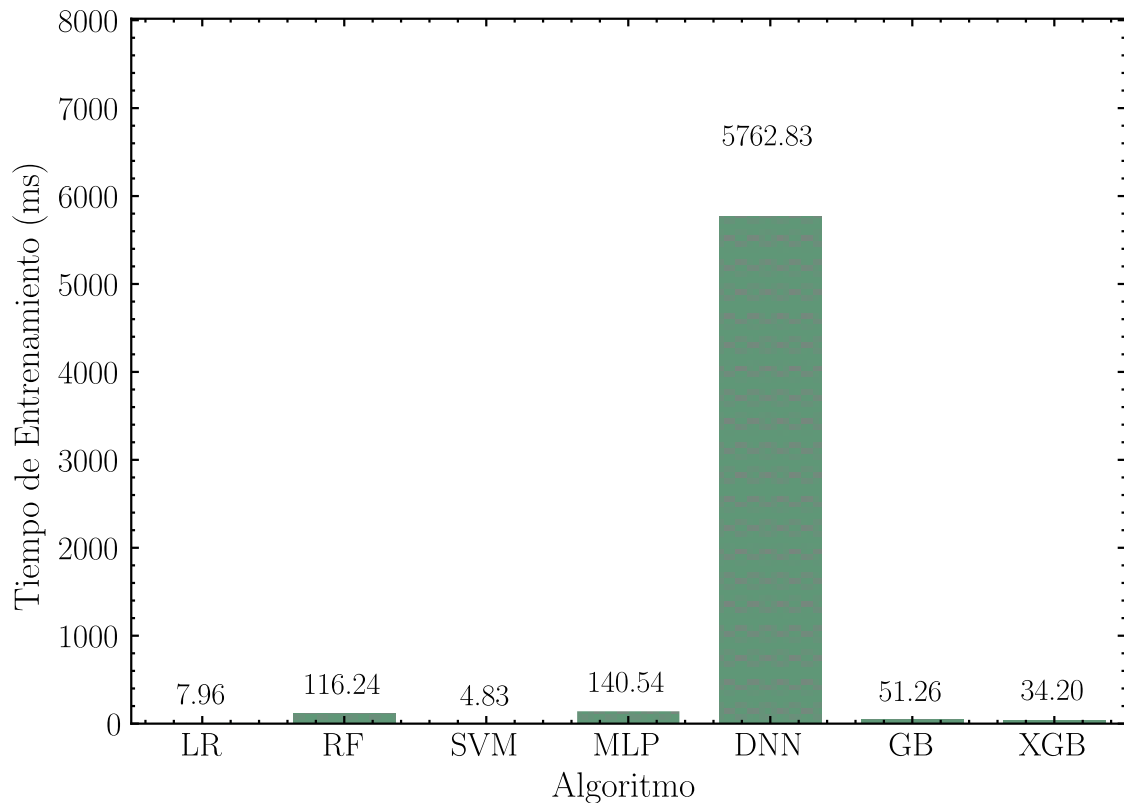


FUENTE: AUTOR

4.3.1.3 Tiempos de Entrenamiento

Los modelos entrenados por el algoritmo DNN presentaron el tiempo más elevado con 5762.83 ms, Por otro lado, los modelos de SVM y LR demostraron ser los más eficientes, con tiempo de 4.83 ms y 7.96 ms, siendo los más rápidos entre todos los modelos evaluados. XGB, GB y RF mostraron tiempos moderados entre 34.20 ms y 116.24 ms. SVM y LR son preferibles para aplicaciones que requieren rapidez de acuerdo con la ilustración 44.

Ilustración 44: Tiempo de entrenamiento para peso total

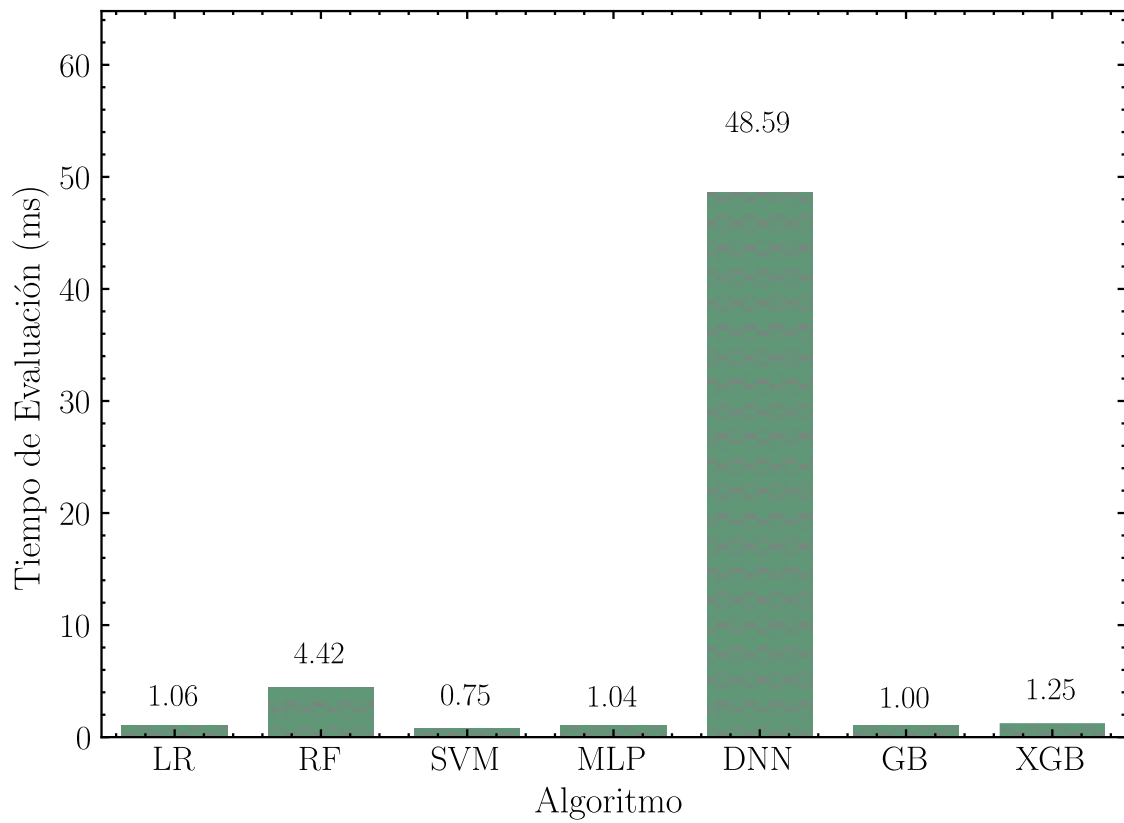


FUENTE: AUTOR

4.3.1.4 Tiempos de Evaluación

El modelo SVM demostró ser el más rápido con un tiempo de evaluación de 0.75 ms, seguido de GB con 1.00 ms y LR con 1.06 ms. XGB y MLP también mostraron tiempos de evaluación bajos, con 1.25 ms y 1.04 ms respectivamente. Por otro lado, DNN registró el tiempo más prolongado con 48.59 ms. Resaltando la eficiencia de SVM y GB en la evaluación rápida, haciendo del SVM el modelo optimo en velocidad, conforme se observa en la ilustración 45.

Ilustración 45: Tiempo de evaluación para peso total



FUENTE: AUTOR

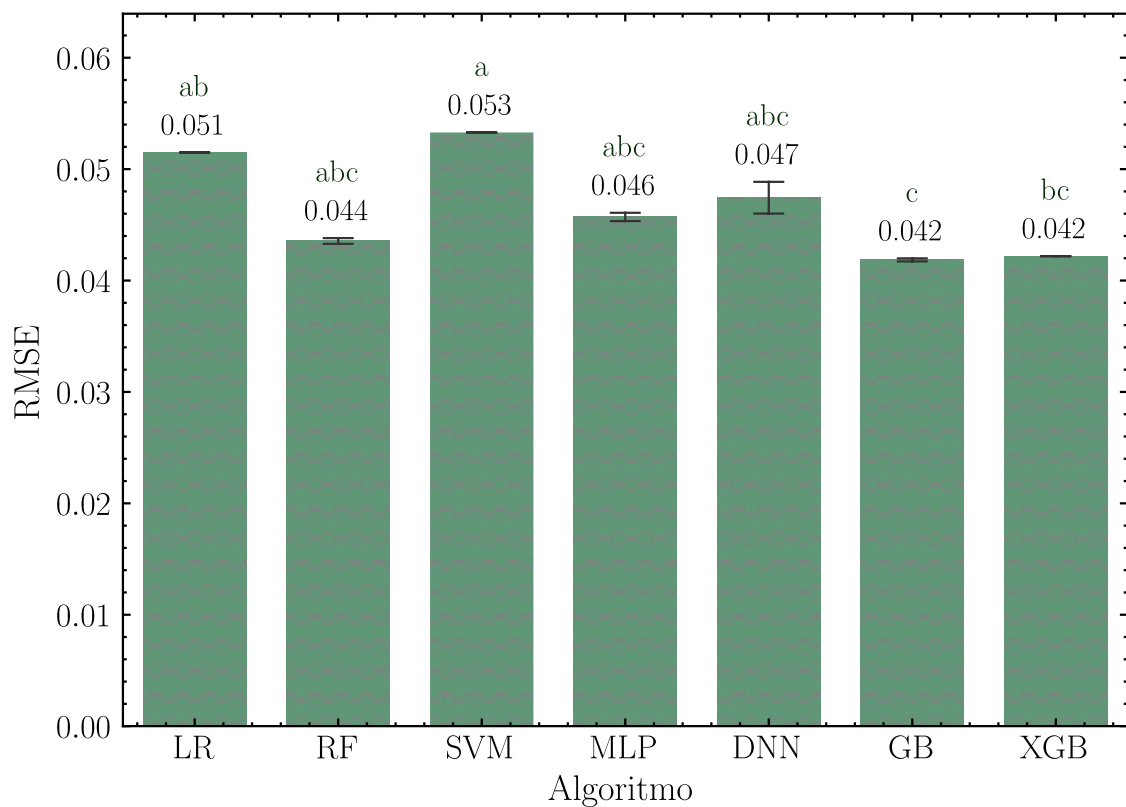
4.3.2 Evaluación de Modelos para Peso Hojas

4.3.2.1 Métricas de Evaluación con Datos de Entrenamiento

Los resultados mostraron que MLP obtuvo el menor RMSE medio (0.047), indicando la mejor precisión de las predicciones en los datos de entrenamiento, seguido de RF (0.048), GB (0.049), y XGB (0.049). LR y SVM presentaron los valores más altos de RMSE medio (0.051 y 0.054, respectivamente). La prueba de Friedman, seguida por la prueba post-hoc de Nemenyi reveló que MLP (b) se diferencia significativamente de SVM (a),

destacándose como el algoritmo con mejor rendimiento medio obtuvo en términos de RMSE, esto se lo presenta en la Ilustración 46.

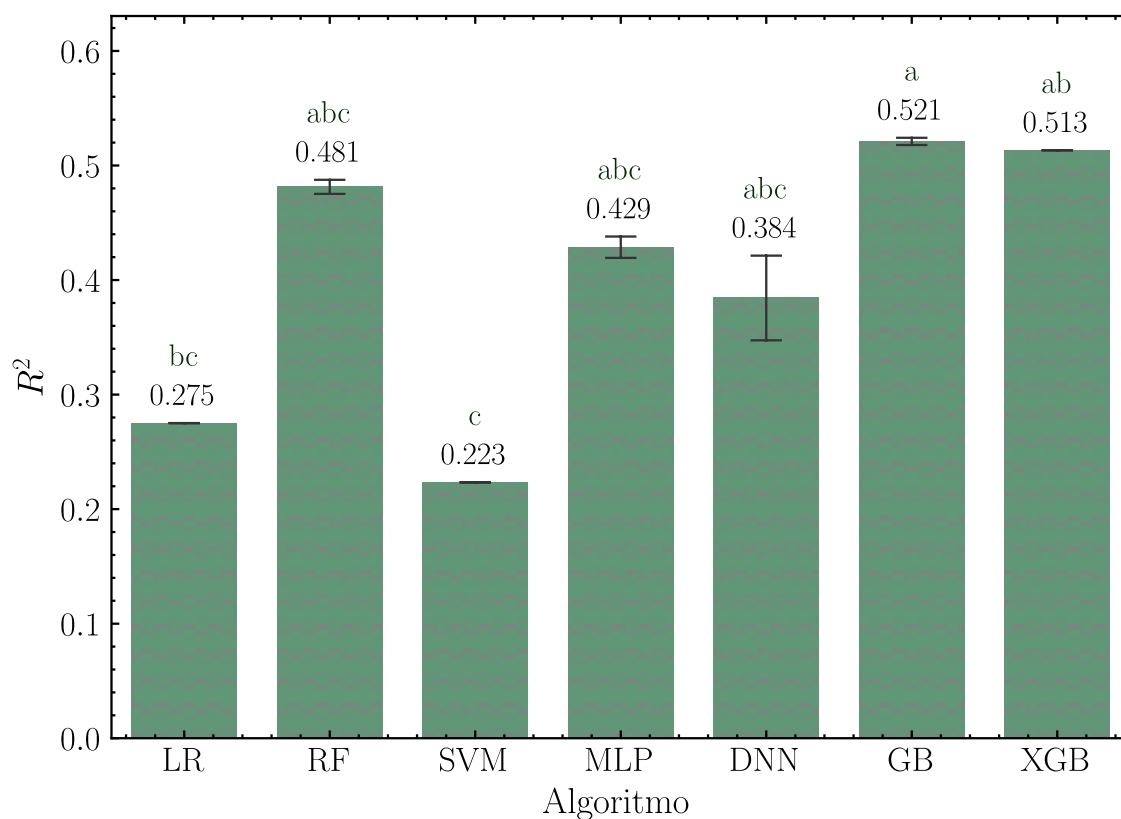
Ilustración 46: Métrica RMSE de datos de entrenamiento para Peso Hojas



FUENTE: AUTOR

El algoritmo GB obtuvo el valor medio más alto de R^2 (0.521), indicando que es el mejor en términos de ajuste a los datos de entrenamiento. La prueba de Friedman, seguida por la prueba post-hoc de Nemenyi muestra que GB (a) es significativamente mejor que SVM (c) y LR (bc). Sin embargo, GB no tiene diferencias significativas con los demás algoritmos, estos hallazgos se los presenta en Ilustración 47.

Ilustración 47: Métrica R^2 de datos de entrenamiento para Peso Hojas

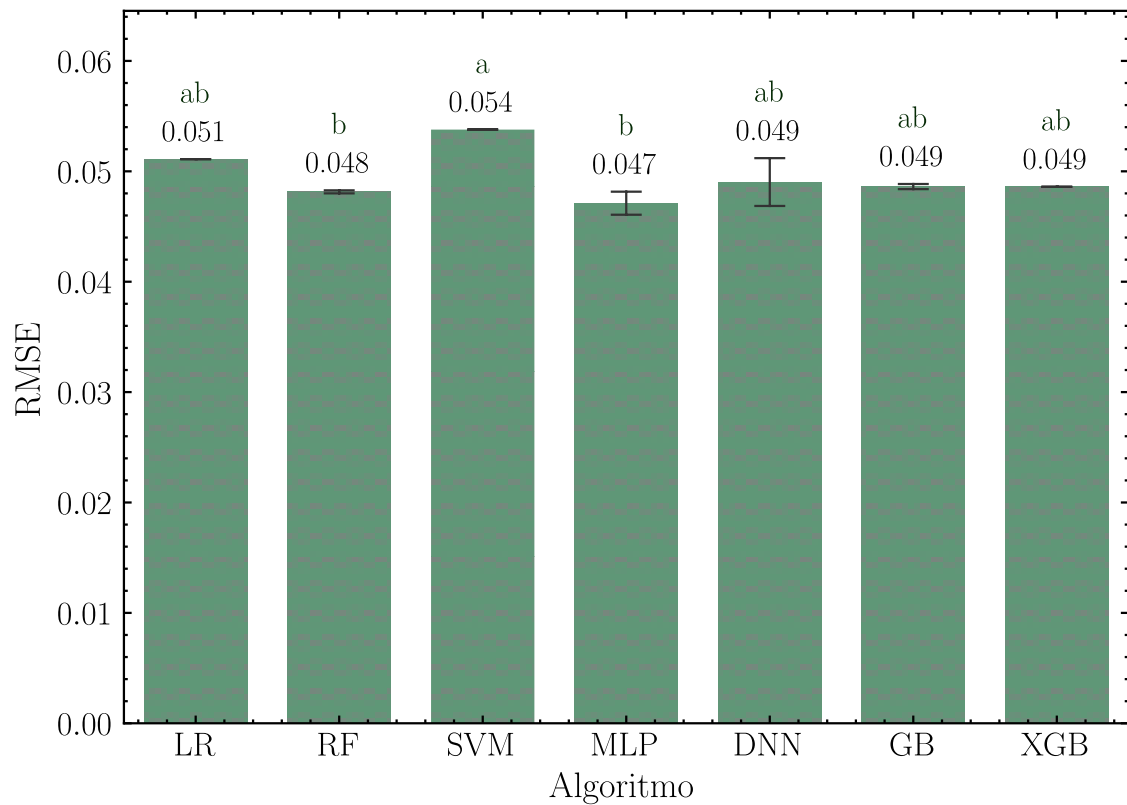


FUENTE: AUTOR

4.3.2.2 Métricas de Evaluación con Datos de Prueba

Los resultados mostraron que GB obtuvo el menor RMSE medio (0.042), indicando la mejor precisión de las predicciones en los datos de pruebas, seguido de XGB (0.042), RF (0.044), y MLP (0.046). SVM y LR presentaron los valores más altos de RMSE medio (0.053 y 0.051, respectivamente). La prueba de Friedman, seguida por la prueba post-hoc de Nemenyi reveló que GB (c) se diferencia significativamente de SVM (a) y LR (ab), destacándose como el algoritmo con mejor rendimiento medio obtuvo en términos de RMSE en la fase de prueba de los modelos, esto se lo presenta en la Ilustración 48.

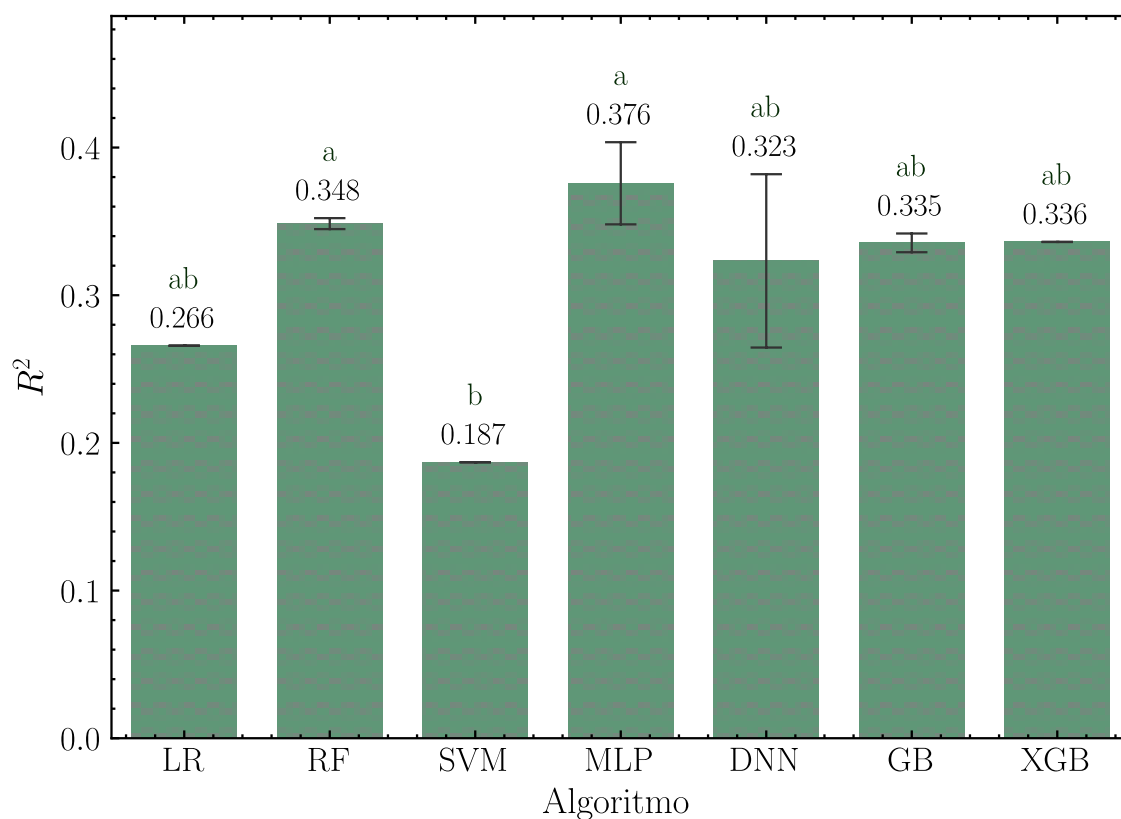
Ilustración 48: Métrica RMSE de datos de prueba para Peso Hojas



FUENTE: AUTOR

El algoritmo MLP obtuvo el valor medio más alto de R^2 (0.376), indicando que es el mejor modelo en términos de ajuste a los datos de pruebas. La prueba de Friedman, seguida por la prueba post-hoc de Nemenyi muestra que MLP (a) es significativamente mejor que SVM (b). Sin embargo, MLP no tiene diferencias significativas con los demás algoritmos, estos hallazgos se los presenta en Ilustración 49.

Ilustración 49: Métrica R^2 de datos de prueba para Peso Hojas

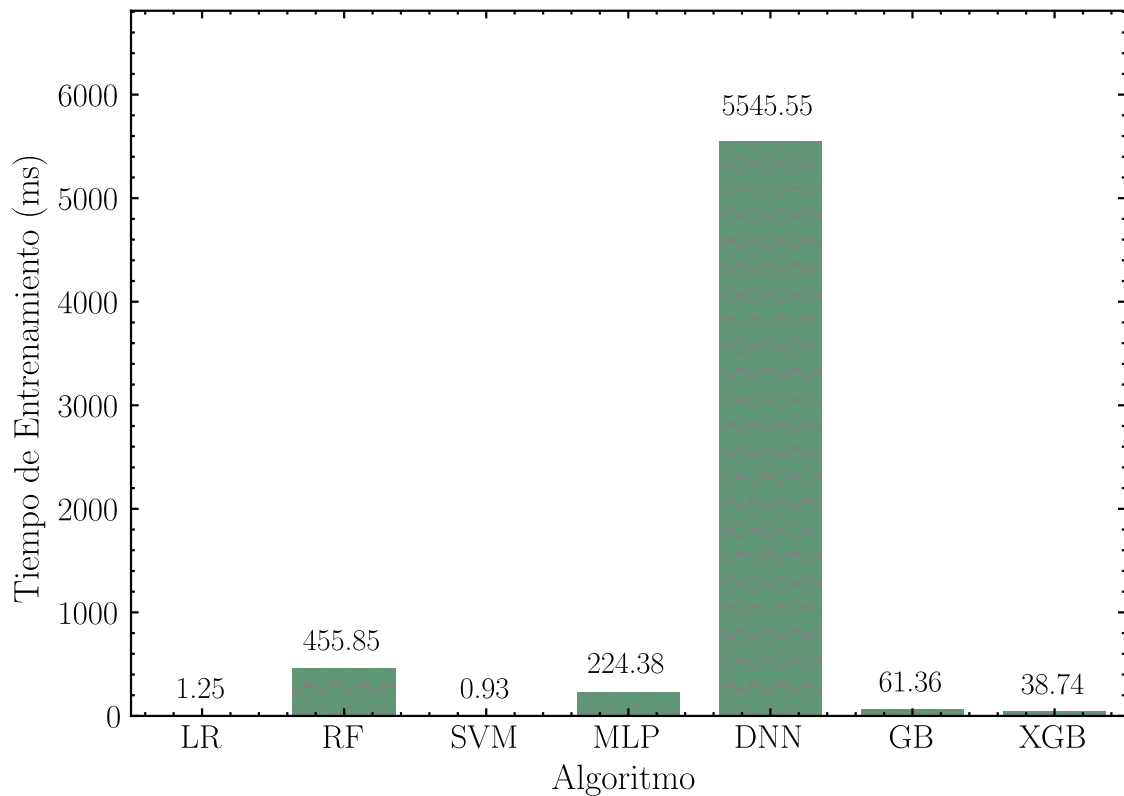


FUENTE: AUTOR

4.3.2.3 Tiempos de Entrenamiento

Para el peso de hojas de maíz forrajero, se observó que DNN registró el mayor promedio de tiempo 5545.55 ms, seguido por RF con 455.85 ms, y MLP con 224.38 ms. Los modelos GB y XGB mostraron tiempos inferiores con 61.36 ms y 38.74 ms, mientras que LR y SVM resultaron ser los más rápidos con 1.25 ms y 0.93. Los modelos GB y XGB fueron las opciones destacadas por su rapidez, como lo indica la Ilustración 50.

Ilustración 50: Tiempo de Entrenamiento para peso hojas

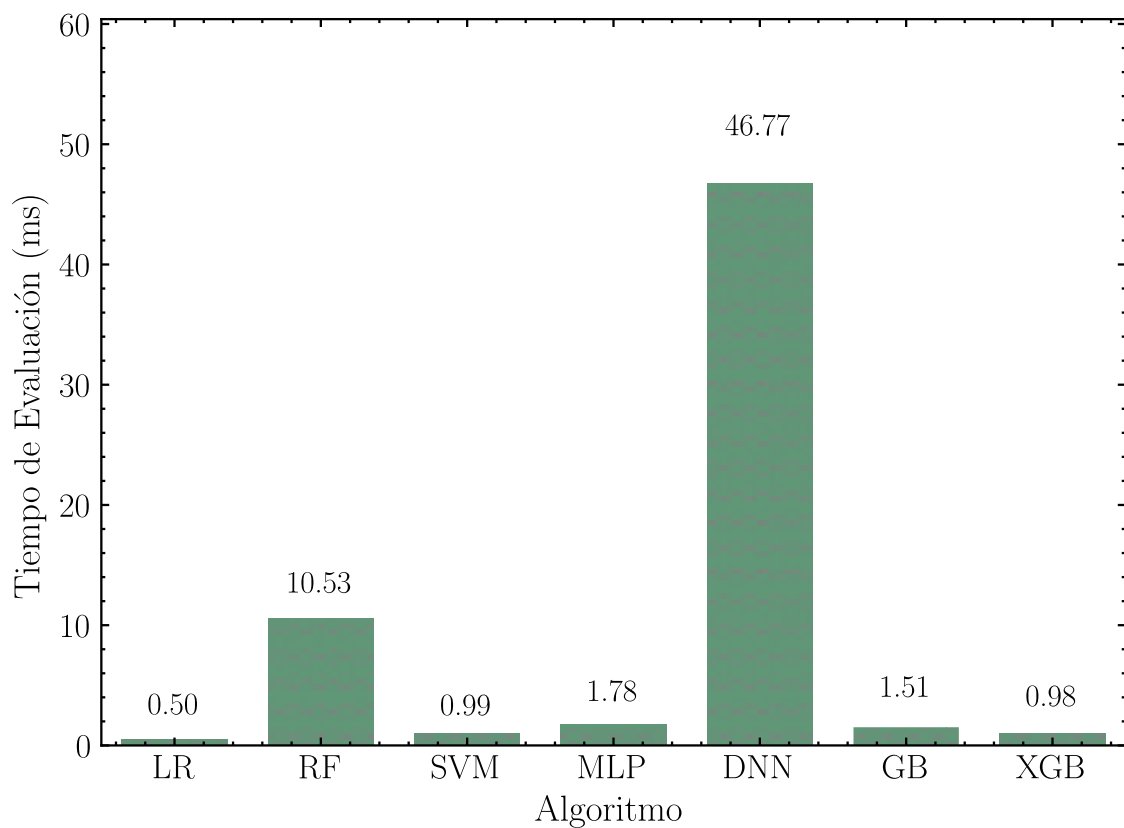


FUENTE: AUTOR

4.3.2.4 Tiempos de Evaluación

La predicción del peso de las hojas en milisegundos, DNN mostró el tiempo más extenso con 46.77 ms. En contraste RF y MLP mostraron de 10.53 y 1.78 ms indicando una rapidez mayor en la evaluación en comparación con DNN, GB, SVM, y XGB reflejaron tiempos más reducidos con 1.51 ms, 0.99 ms y 0.98 ms. El modelo de LR mostró ser el más rápido con un tiempo de 0.50 ms, como se muestra en la Ilustración 51.

Ilustración 51: Tiempo de evaluación para peso hojas



FUENTE: AUTOR

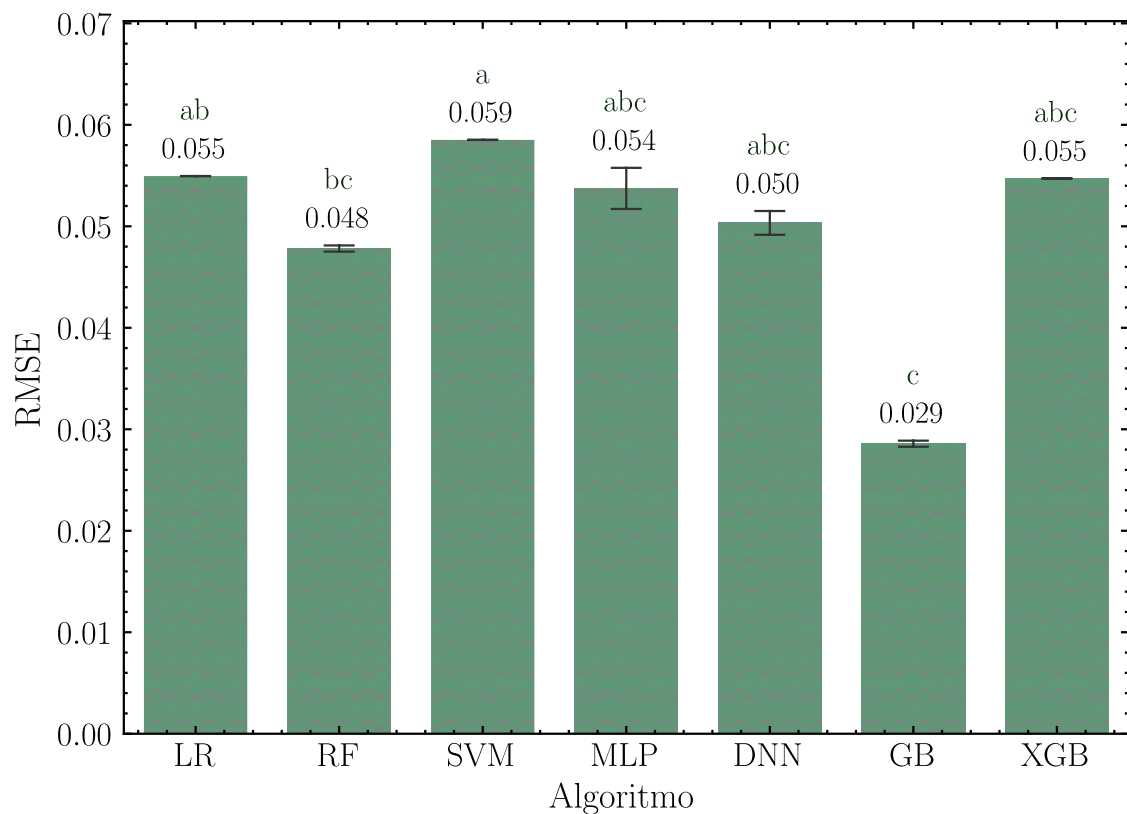
4.3.3 Evaluación de Modelos para Peso Tallo

4.3.3.1 Métricas de Evaluación con Datos de Entrenamiento

Los resultados mostraron que GB obtuvo el menor RMSE medio (0.029), indicando la mejor precisión de las predicciones en los datos de entrenamiento, seguido de RF (0.048), DNN (0.050), y MLP (0.054). SVM presentó los valores más altos de RMSE medio (0.059). La prueba de Friedman, seguida por la prueba post-hoc de Nemenyi reveló que GB (c) se diferencia significativamente de SVM (a), destacándose como el algoritmo con

mejor rendimiento medio obtuvo en términos de RMSE, esto se lo presenta en la Ilustración 52.

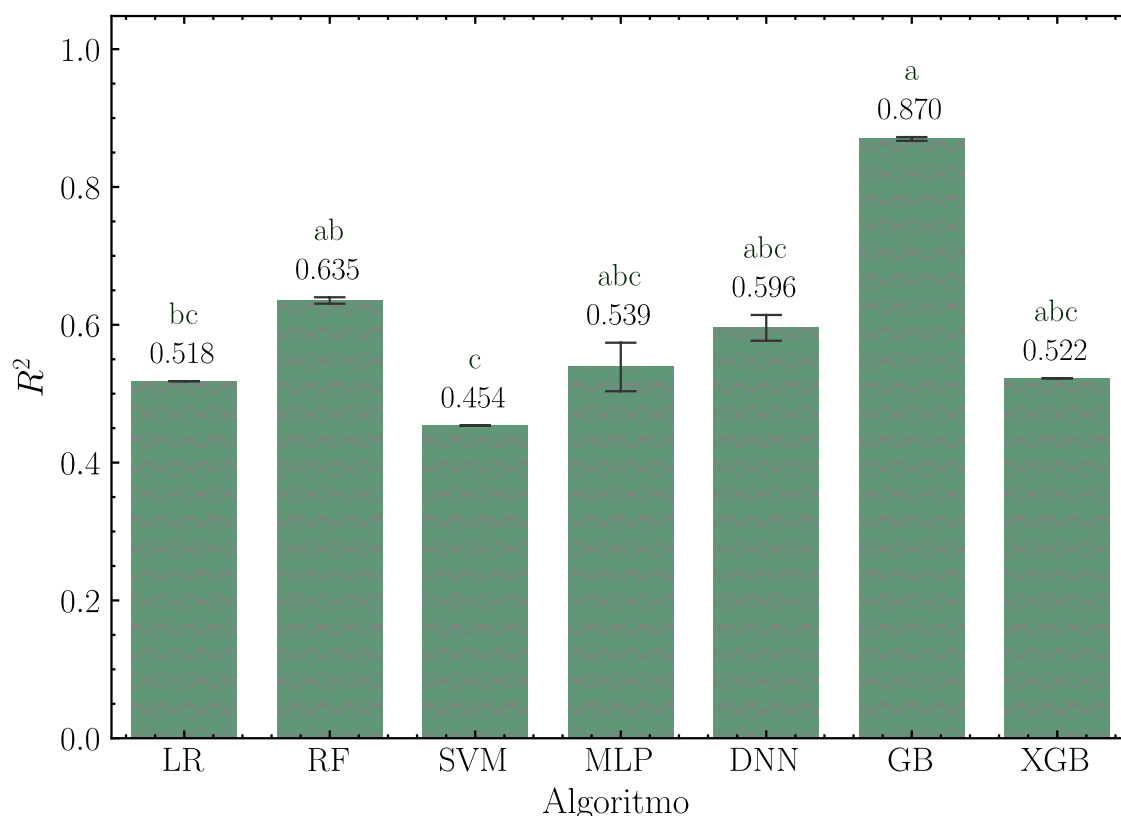
Ilustración 52: Métrica RMSE de datos de entrenamiento para Peso Tallo



FUENTE: AUTOR

El algoritmo GB obtuvo el valor medio más alto de R^2 (0.870), indicando que es el mejor en términos de ajuste a los datos de entrenamiento. La prueba de Friedman, seguida por la prueba post-hoc de Nemenyi muestra que GB (a) es significativamente mejor que SVM (c) y LR (bc). Sin embargo, GB no tiene diferencias significativas con los demás algoritmos, estos hallazgos se los presenta en Ilustración 53.

Ilustración 53: Métrica R^2 de datos de entrenamiento para Peso Tallo

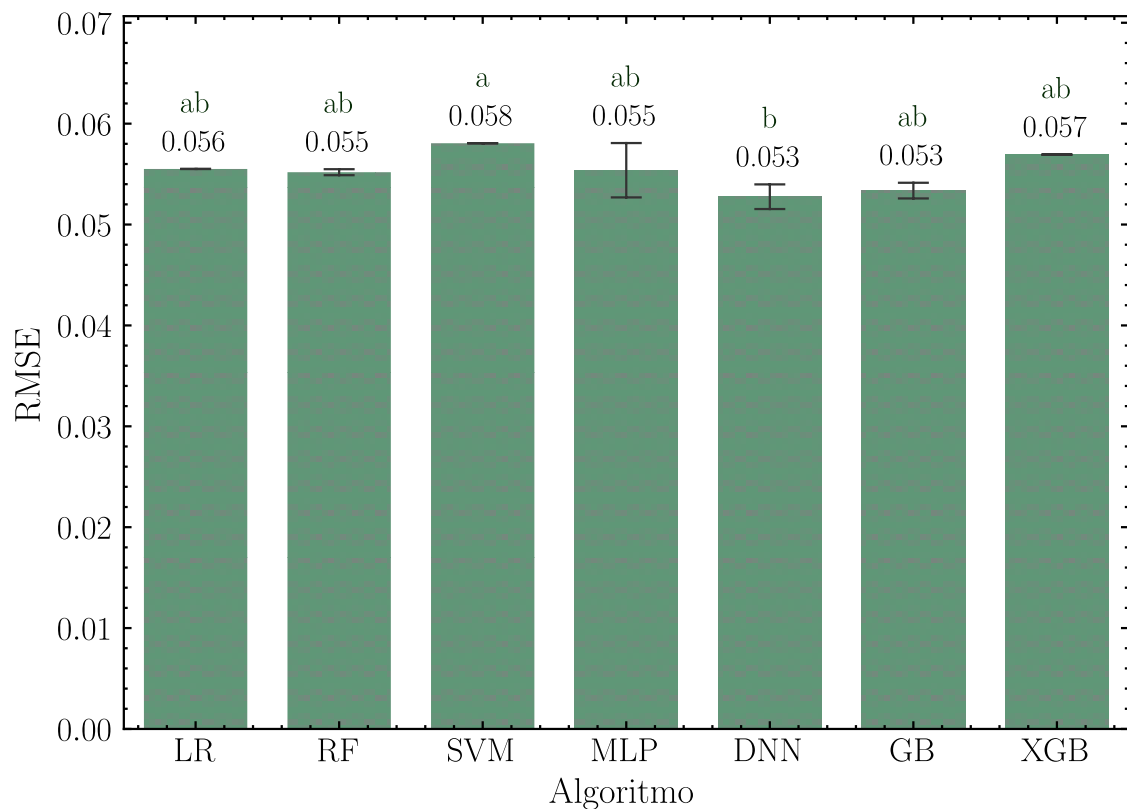


FUENTE: AUTOR

4.3.3.2 Métricas de Evaluación con Datos de Prueba

Los resultados mostraron que para Peso de Tallo los modelos entrenados por el algoritmo DNN obtuvo el menor RMSE medio (0.053), indicando mejor precisión de las predicciones en los datos de pruebas, seguido de GB (0.053), RF (0.055), y LR (0.056). SVM y XGB presentaron los valores más altos de RMSE medio (0.058 y 0.057, respectivamente). La prueba de Friedman, seguida por la prueba post-hoc de Nemenyi reveló que DNN (b) se diferencia significativamente de SVM (a), destacándose como el algoritmo con mejor rendimiento medio obtuvo en términos de RMSE en la fase de prueba de los modelos, esto se lo presenta en la Ilustración 54.

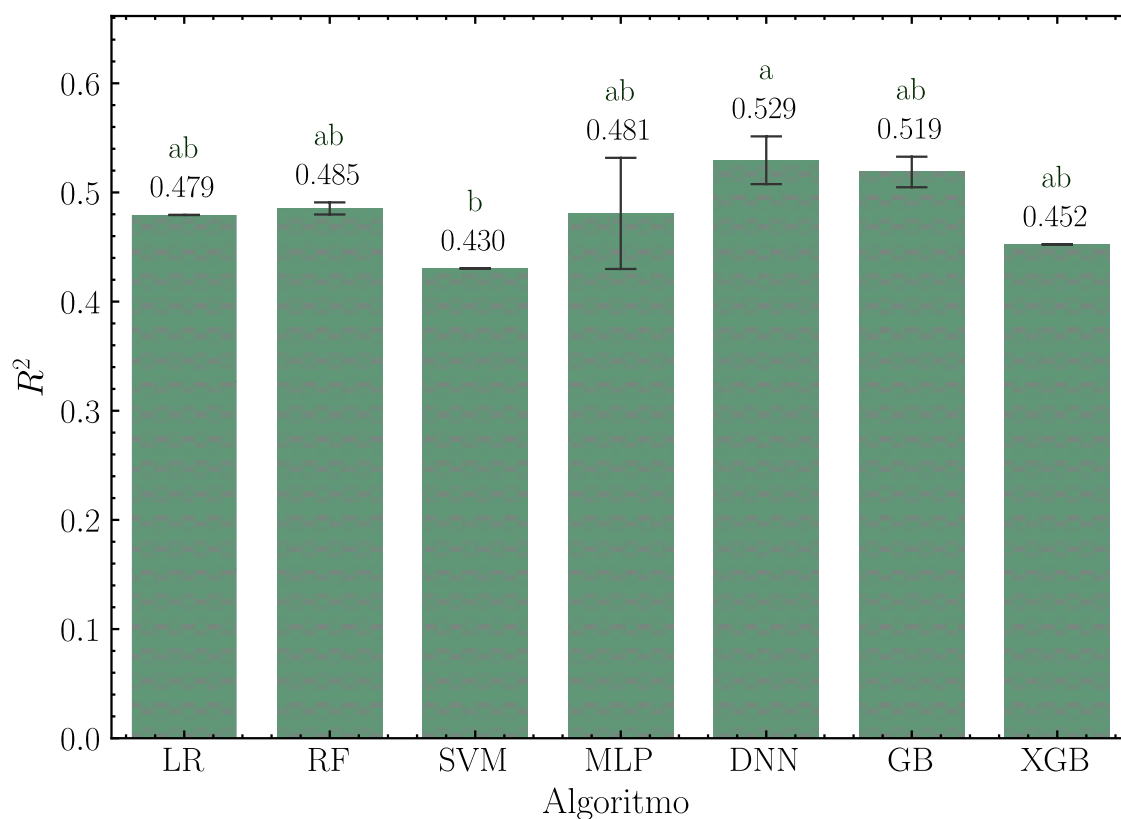
Ilustración 54: Métrica RMSE de datos de prueba para Peso Tallo



FUENTE: AUTOR

El algoritmo DNN obtuvo el valor medio más alto de R^2 (0.529), indicando que es el mejor modelo en términos de ajuste a los datos de pruebas. La prueba de Friedman, seguida por la prueba post-hoc de Nemenyi muestra que DNN (a) es significativamente mejor que SVM (b). Sin embargo, DNN no tiene diferencias significativas con los demás algoritmos, estos hallazgos se los presenta en Ilustración 55.

Ilustración 55: Métrica R^2 de datos de prueba para peso tallo

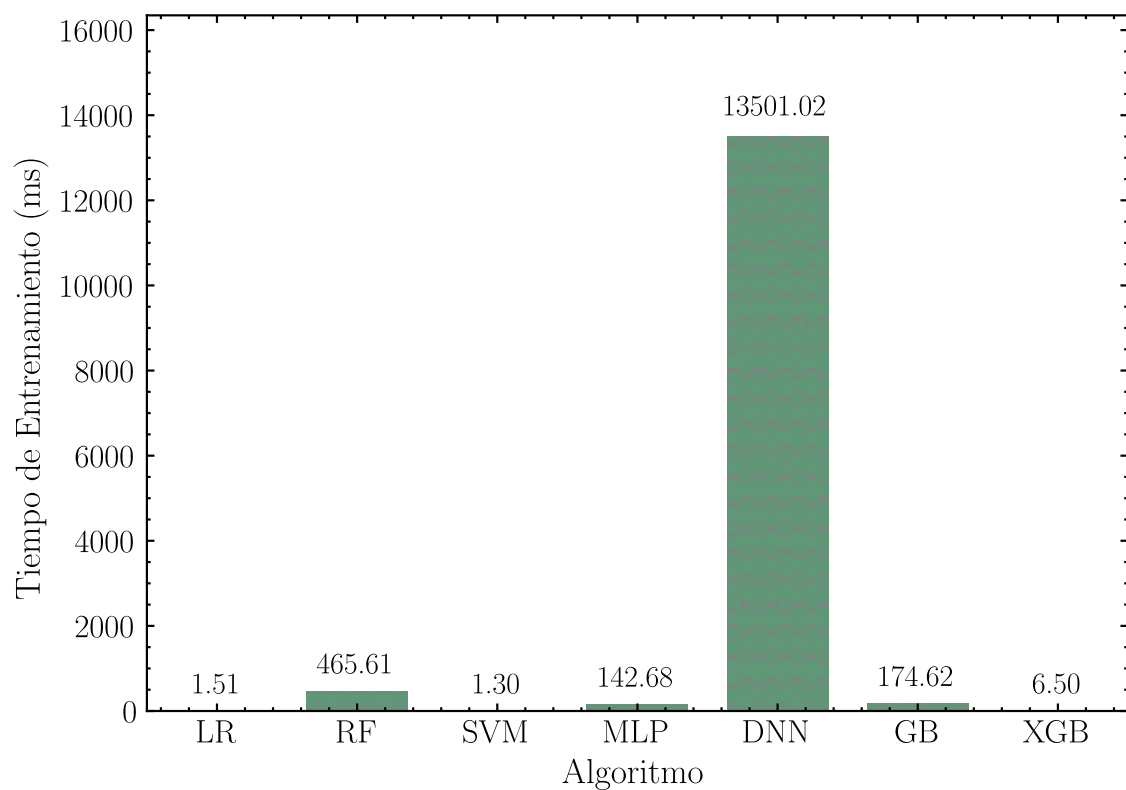


FUENTE: AUTOR

4.3.3.3 Tiempos de Entrenamiento

El modelo DNN mostró el tiempo más prolongado, con 13501.02 ms, indicando un consumo significativo de recursos durante su entrenamiento. En comparación, RF utilizó 465.61 ms y MLP 142.68 ms, demostrando ser menos demandante en tiempo, pero aun eficiente. XGB fue de 6.50 ms, y GB con 174.62 ms, resaltaron por su rapidez y eficiencia, como se refleja en la Ilustración 56. Por otro lado, SVM y LR presentaron los menores tiempos. 1.30 ms y 1.51 ms, destacando su eficiencia y rapidez.

Ilustración 56: Tiempo de entrenamiento para peso tallo

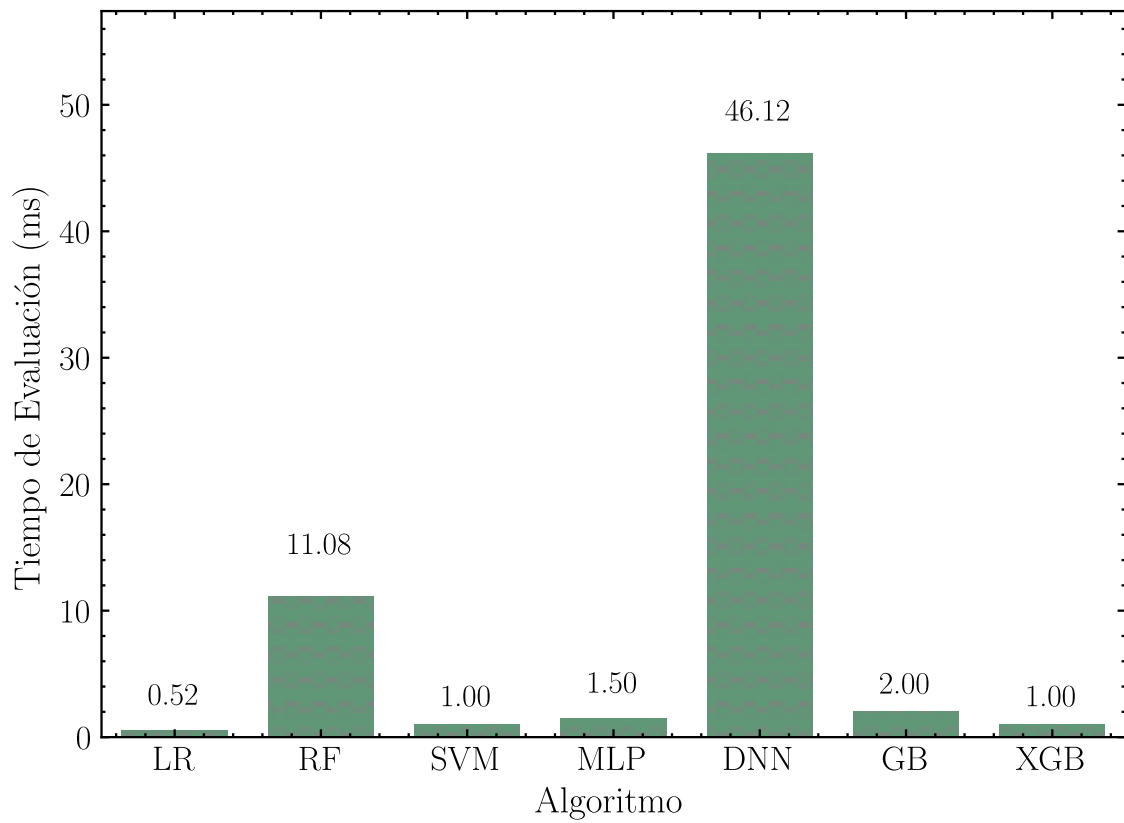


FUENTE: AUTOR

4.3.3.4 Tiempos de Evaluación

LR y, SVM muestran tiempos más bajos con 1.51 ms y 1.30 ms, lo que indica eficiencia en la evaluación rápida de los modelos. DNN registró el tiempo más largo con 13501.02 ms, lo cual indicó un procesamiento más lento durante la evaluación. Por otro lado, XGB también reflejó un desempeño de 6.50 ms con una velocidad y precisión balanceada, conforme se observa en la Ilustración 57.

Ilustración 57: Tiempo de evaluación para peso tallo



FUENTE: AUTOR

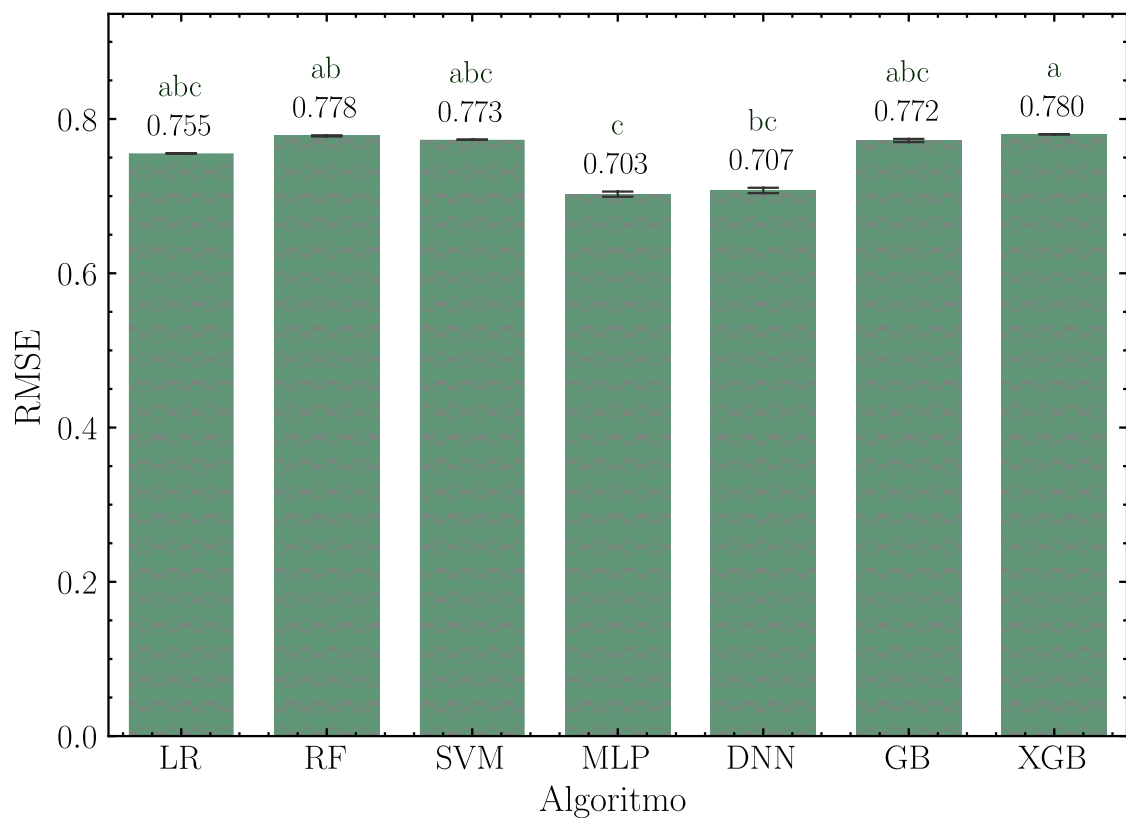
4.3.4 Evaluación de Modelos para Peso Mazorca

4.3.4.1 Métricas de Evaluación con Datos de Entrenamiento

Los resultados mostraron que MLP para el rendimiento de peso de mazorca obtuvo el menor RMSE medio (0.703), indicando la mejor precisión de las predicciones en los datos de entrenamiento, seguido de DNN (0.707) y LR (0.755). XGB y RF presentaron los valores más altos de RMSE medio (0.780 y 0.778, respectivamente). La prueba de Friedman, seguida por la prueba post-hoc de Nemenyi reveló que MLP (c) se diferencia

significativamente de XGB (a) y RF (ab), destacándose como el algoritmo con mejor rendimiento medio obtuvo en términos de RMSE, esto se lo presenta en la Ilustración 58.

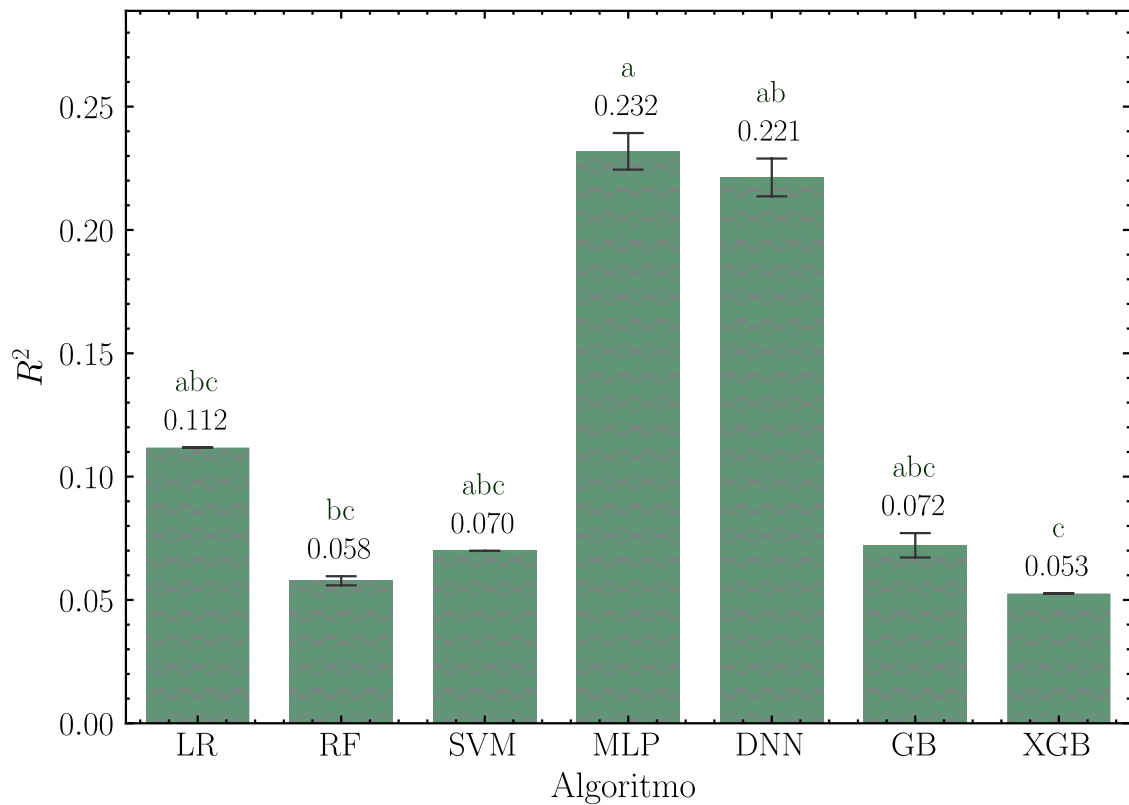
Ilustración 58: Métrica RMSE de datos de entrenamiento para Peso Mazorca



FUENTE: AUTOR

El algoritmo MLP obtuvo el valor medio más alto de R^2 (0.232), indicando que es el mejor en términos de ajuste a los datos de entrenamiento. La prueba de Friedman, seguida por la prueba post-hoc de Nemenyi muestra que MLP (a) es significativamente mejor que XGB (c) y RF (bc). Sin embargo, MLP no tiene diferencias significativas con los demás algoritmos, estos hallazgos se los presenta en Ilustración 59.

Ilustración 59: Métrica R^2 de datos de entrenamiento para Peso Mazorca

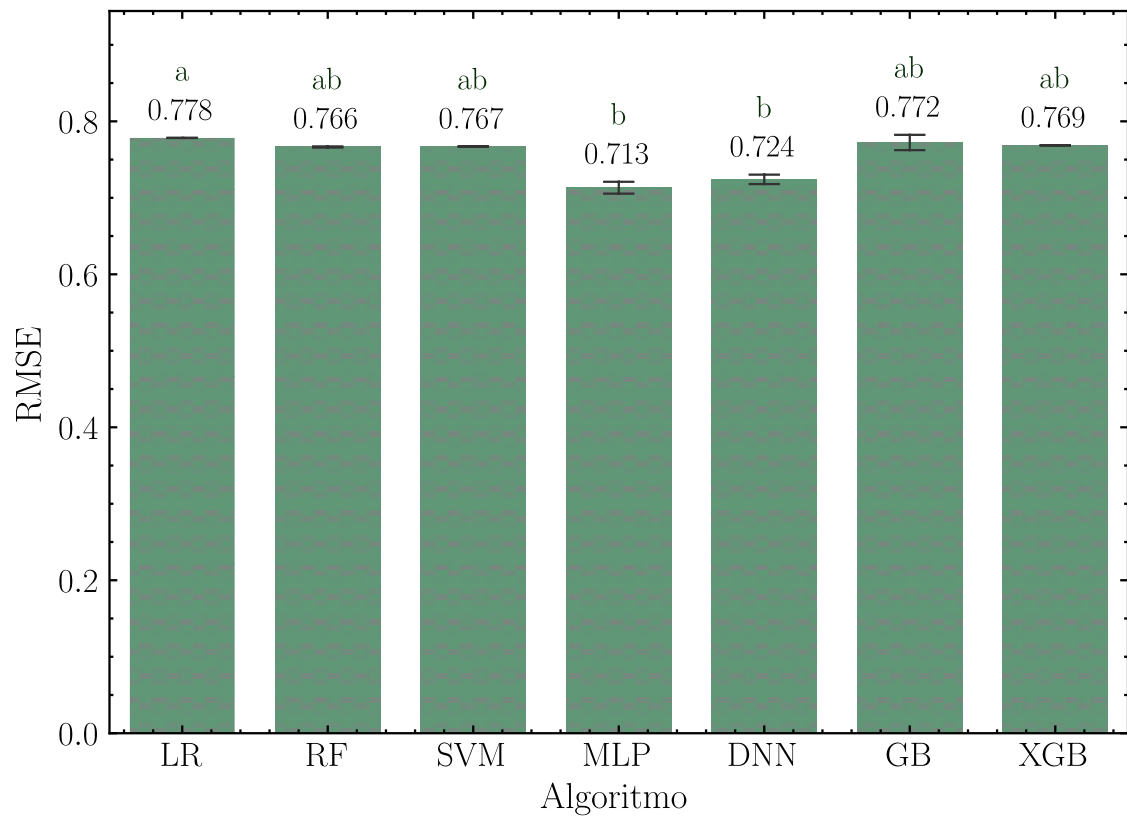


FUENTE: AUTOR

4.3.4.2 Métricas de Evaluación con Datos de Prueba

Los resultados mostraron que para Peso de Mazorca los modelos entrenados por el algoritmo MLP obtuvo el menor RMSE medio (0.713), indicando la mejor precisión de las predicciones en los datos de pruebas, seguido de DNN (0.724), RF (0.766), y XGB (0.769). LR y GB presentaron los valores más altos de RMSE medio (0.778 y 0.772, respectivamente). La prueba de Friedman, seguida por la prueba post-hoc de Nemenyi reveló que MLP (b) se diferencia significativamente de LR (a), destacándose como el algoritmo con mejor rendimiento medio obtuvo en términos de RMSE en la fase de prueba de los modelos, esto se lo presenta en la Ilustración 60.

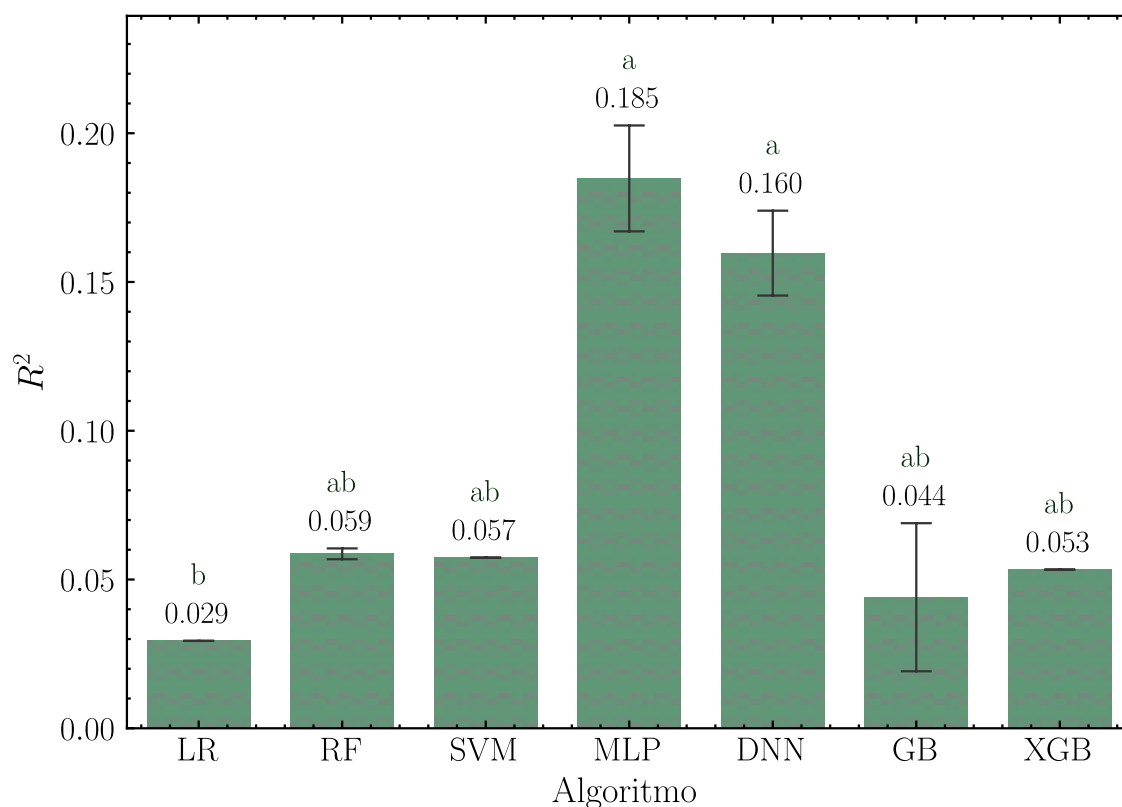
Ilustración 60: Métrica RMSE de datos de prueba para peso mazorca



FUENTE: AUTOR

El algoritmo MLP obtuvo el valor medio más alto de R^2 (0.185), indicando que es el mejor modelo en términos de ajuste a los datos de pruebas. La prueba de Friedman, seguida por la prueba post-hoc de Nemenyi muestra que MLP (a) es significativamente mejor que LR (b). Sin embargo, MLP no tiene diferencias significativas con los demás algoritmos, estos hallazgos se los presenta en Ilustración 61.

Ilustración 61: Métrica R^2 de datos de prueba para peso mazorca

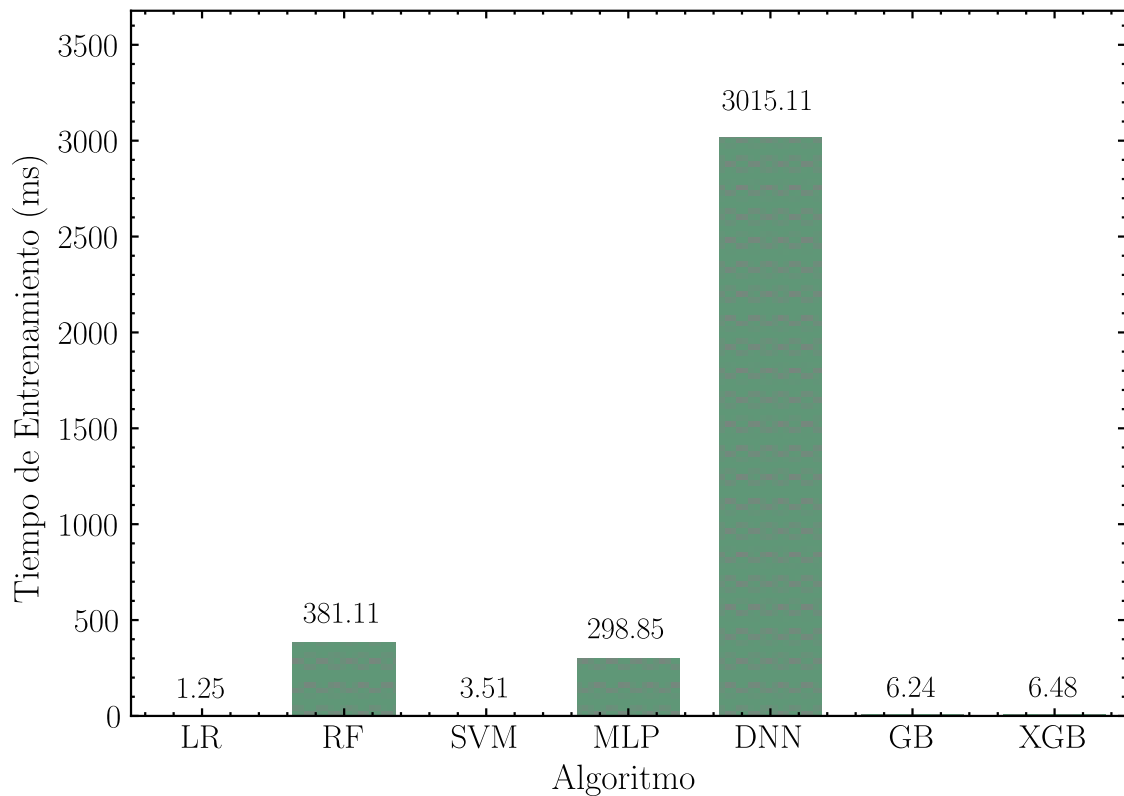


FUENTE: AUTOR

4.3.4.3 Tiempos de Entrenamiento

Para la predicción del rendimiento en el peso de mazorca, se observó que el tiempo medio de entrenamiento del modelo DNN fue significativamente mayor con 3015.11 ms, como lo indica la Ilustración 62. Para RF y MLP fueron de 381.11 ms y 298.85 ms, reflejando una moderada demanda de tiempo de cálculo. Por otro lado, Los modelos GB y XGB, SVM, LR presentaron los tiempos de entrenamiento más eficientes especialmente LR con solo 1.25.

Ilustración 62: Tiempo de entrenamiento para peso mazorca

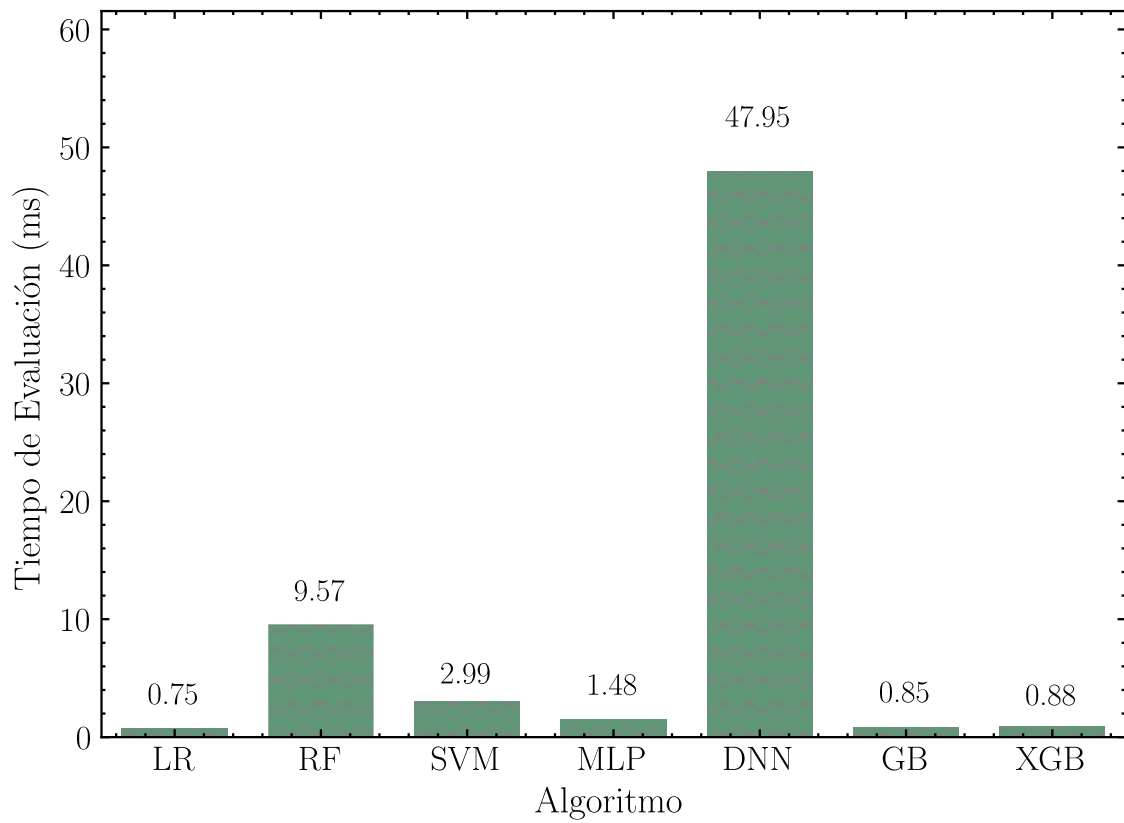


FUENTE: AUTOR

4.3.4.4 Tiempos de Evaluación

El modelo DNN registró un tiempo de evaluación de 47.95 ms, lo que indicó ofrece una capacidad predictiva considerable sin embargo demanda más tiempo, conforme se observa en la Ilustración 63. RF necesitó 9.57 ms y MLP 1.48 ms, mostrando que mientras RF ofrece un enfoque robusto, MLP promete una solución rápida. Los modelos GB y XGB se destacaron por su rapidez, con tiempos de 0.85 ms y 0.88, demostrando su eficiencia en el procesamiento. En contraste SVM requirió 2.99 ms la cual fue el más lento con comparación con los modelos Boosting, pero aun así eficiente. El modelo LR fue el más rápido con un tiempo de solo 0.75 ms.

Ilustración 63: Tiempo de evaluación para peso mazorca



FUENTE: AUTOR

4.3.5 Discusión

Las métricas de evaluación de la predicción del rendimiento de maíz concuerdan con el estudio de revisión Van, Kassahun, & Catal, (2020), donde se mencionan que las métricas de evaluación más usadas para la predicción del rendimiento de los cultivos son RMSE, R^2 , MAE y MSE.

En el estudio de Wu et al., (2023) se emplearon métricas como R^2 y RMSE para evaluar el rendimiento de los modelos de aprendizaje automático mediante el algoritmo SVM, ambos estudios comparten un enfoque en optimizar los resultados a través del aprendizaje automático, validando la efectividad de las técnicas avanzadas en campos de aplicación

similares. Un aspecto notable es que buscan mejorar la precisión y la eficiencia de las prácticas agrícolas.

Para el peso total, tanto el RF como DNN mostraron resultados similares en las pruebas, destacándose en la métrica R^2 con RF obteniendo el mayor valor de 0.605, lo que indica un ajuste adecuado de los datos. Aunque RF y DNN presentaron métricas de evaluación similares en términos de RMSE, RF sobresalió por su menor tiempo de evaluación, haciéndolo el algoritmo más adecuado para la predicción del peso total.

En cuanto al peso de las hojas, RF, GB y XGB mostraron resultados comparables en los datos de pruebas. GB se diferenció estadísticamente en los valores de RMSE respecto al SVM. En cuanto al coeficiente R^2 , los modelos MLP presentaron el mejor ajuste. Considerando los resultados en RMSE, XGB emerge como una buena alternativa para predecir el rendimiento en el peso de las hojas.

Para el peso del tallo, GB y DNN ofrecieron las mejores predicciones según RMSE, seguidos de cerca por RF. Aunque no hubo diferencias significativas con otros algoritmos, GB, con un R^2 de 0.519, se destaca como la mejor opción para predecir con precisión el rendimiento del peso del tallo.

En relación con el peso de la mazorca, los modelos MLP y DNN fueron los más efectivos según la métrica RMSE y también se destacaron en R^2 , mostrando los ajustes más altos. Considerando el tiempo de entrenamiento, el modelo MLP se perfila como una elección adecuada para la predicción del peso de la mazorca.

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

5.1 CONCLUSIONES

Para el análisis descriptivo de los datos agronómicos se llevó a cabo una evaluación detallada de las variables numéricas y categóricas, identificando un amplio rango de crecimiento y desarrollo en las plantas de maíz forrajero, evidenciando la influencia significativa de la distancia de siembra en el diámetro del tallo. A pesar de que la altura de las plantas se mantuvo relativamente constante, las plantas sembradas a mayores distancias exhibieron un aumento notable en el diámetro del tallo.

Además, la correlación de Spearman resaltó conexiones claras entre las variables agronómicas y el rendimiento del cultivo, especialmente en lo que respecta al peso de las mazorcas. Sin embargo, las etapas de inflorescencia y la inserción de la mazorca mostraron variabilidad y una frecuencia baja de datos, lo que subraya la necesidad de mejorar la recolección de información en estas etapas críticas para entender mejor su impacto en el rendimiento final y ajustar las prácticas agrícolas de acuerdo con estas observaciones.

Tras realizar un análisis descriptivo de los datos agronómicos se procedió a la selección de algoritmos y la configuración de sus hiperparámetros para la predicción de variables de rendimiento de maíz ("Peso Total", "Peso Hojas", "Peso Tallo" y "Peso Mazorca"). Se evaluaron siete algoritmos: Linear Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Multi-layer Perceptron (MLP), Deep Neural Network (DNN), Gradient Boosting Tree (GB) y XGBoost (XGB).

La optimización de hiperparámetros para los modelos predictivos de rendimiento del maíz forrajero mostró variaciones significativas entre los algoritmos en términos de eficiencia y precisión. Los tiempos de ajuste de hiperparámetros revelaron que DNN requería considerablemente más tiempo para ajustar comparado con algoritmos más rápidos como LR y SVM que mostraron tiempos de ajuste mucho más rápidos.

En términos de desempeño de los modelos, medido a través de RMSE y R^2 , se encontró que mientras algunos algoritmos como GB y XGB alcanzaron un ajuste casi perfecto en los datos de entrenamiento, su rendimiento se redujo significativamente en los datos de prueba, lo que indica un sobreajuste, la métrica R^2 registró valores negativos, indicando un ajuste predictivo deficiente fuera de la muestra de entrenamiento.

En contraste, algoritmos como RF mostraron un balance más consistente entre precisión en entrenamiento y generalización en pruebas, lo que lo convierte en el algoritmo más recomendable para aplicaciones prácticas donde la precisión y la generalización son críticas. Esta capacidad para equilibrar el ajuste y la generalización lo hace ideal para el uso en la predicción del rendimiento agrícola, especialmente en escenarios donde se necesitan decisiones rápidas y confiables basadas en datos.

La selección del modelo óptimo se basó en las pruebas de significancia estadística como las pruebas de Friedman y Nemenyi que permitió identificar los modelos con mejor desempeño en términos de RMSE y R^2 . RF demostró ser superiores, alcanzando los menores valores de RMSE y los más altos de R^2 en la mayoría de las pruebas, tanto en entrenamiento como en evaluación, especialmente para la variable de Peso Total. Esto

indica que RF no solo se ajusta bien los datos, sino que también mantiene una buena generalización en datos nuevos, lo que es crucial para aplicaciones prácticas en el campo por su velocidad precisión.

Los algoritmos de SVM y LR ofrecen los tiempos de respuesta más rápidos, sin embargo, su rendimiento en términos de RMSE y R^2 , fue inferior al de RF para las variables de rendimiento de peso de hojas y peso de mazorcas.

5.2 RECOMENDACIONES

Es crucial intensificar la recopilación de datos en las etapas tempranas y reproductivas del ciclo de cultivo para mejorar la comprensión de los factores que influyen en el desarrollo vegetativo y reproductivo del maíz forrajero. Implementar tecnologías de sensores y análisis en tiempo real puede ayudar a capturar datos más completos y precisos sobre inflorescencias y otros indicadores críticos, facilitando decisiones agronómicas más informadas y proactivas.

Para la determinación de la configuración óptima de hiperparámetros, se recomienda implementar estrategias de validación cruzada y técnicas de regularización durante el proceso de ajuste de hiperparámetros. Esto ayudará a prevenir el sobreajuste, especialmente en algoritmos complejos como GB y XGB, que demostraron altos niveles de ajuste en los datos de entrenamiento, pero rendimientos pobres en la generalización a datos nuevos. La validación cruzada no solo permite una evaluación más fiable de la capacidad predictiva del modelo, sino que también asegura que los hiperparámetros

seleccionados proporcionen una estimación robusta del rendimiento del modelo en diferentes subconjuntos de datos. Esto es crucial para garantizar que los modelos desarrollados sean aplicables y efectivos en entornos agrícolas reales.

Se recomienda implementar un enfoque de ensamble de modelos para combinar las fortalezas de diferentes algoritmos y mejorar la precisión y generalización de las predicciones. Considerando que modelos como Random Forest (RF) mostraron un desempeño destacado en términos de RMSE y R^2 , y teniendo en cuenta la eficiencia y la rapidez de otros algoritmos como SVM y LR en tiempos de evaluación y entrenamiento, un ensamble que integre estos modelos podría ofrecer un equilibrio entre precisión y velocidad operativa. Este enfoque permitiría aprovechar las ventajas individuales de cada modelo, asegurando así predicciones más robustas y confiables para la toma de decisiones en la agricultura de precisión.

Se recomienda implementar los modelos entrenados en un sistema informático mediante la creación de una API REST que implemente el modelo de pronóstico en un entorno operativo y sea accesible para los usuarios finales, como agricultores o investigadores de áreas agronómicas, esto ayudaría a establecer rendimientos de sus cultivos de maíz forrajero y anticipar falta de alimento para la demanda local o regional. El tener de antemano una estimación del rendimiento del cultivo de maíz forrajero ayudará a la seguridad alimentaria ya que este cultivo es crucial para la alimentación animales principalmente de ganado bovino, lo que garantizaría una correcta producción de carne y leche.

REFERENCIAS BIBLIOGRÁFICAS

- Abbas, F., Afzaal, H., Farooque, A., & Tang, S. (2020). Crop Yield Prediction through Proximal Sensing and Machine Learning Algorithms. *Agronomy*, 10, 1046. <https://doi.org/10.3390/agronomy10071046>
- Alhakeem, Z. M., Jebur, Y. M., Henedy, S. N., Imran, H., Bernardo, L. F. A., & Hussein, H. M. (2022). Prediction of Ecofriendly Concrete Compressive Strength Using Gradient Boosting Regression Tree Combined with GridSearchCV Hyperparameter-Optimization Techniques. *Materials*, 15(21), 7432. <https://doi.org/10.3390/ma15217432>
- Beserra, E. B., Dias, C. T. D. S., & Parra, J. R. P. (2002). DISTRIBUTION AND NATURAL PARASITISM OF SPODOPTERA FRUGIPERDA (LEPIDOPTERA: NOCTUIDAE) EGGS AT DIFFERENT PHENOLOGICAL STAGES OF CORN. *Florida Entomologist*, 85(4), 588-593. [https://doi.org/10.1653/0015-4040\(2002\)085\[0588:DANPOS\]2.0.CO;2](https://doi.org/10.1653/0015-4040(2002)085[0588:DANPOS]2.0.CO;2)
- Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. New York, NY, USA: Oxford University Press, Inc.
- Cadet-Díaz, S., & Guerrero-Escobar, S. (2018). Factores que determinan los rendimientos de la producción de maíz en México: Evidencia del censo agropecuario 2007. *Agricultura, sociedad y desarrollo*, 15(3), 311-337. Recuperado de http://www.scielo.org.mx/scielo.php?script=sci_abstract&pid=S1870-54722018000300311&lng=es&nrm=iso&tlng=es
- Çatal, M., Çelik, Ş., & Bakoglu, A. (2023). Practicability of MARS and bagging MARS algorithms in prediction of plant length of grass pea (*Lathyrus sativus* L.) in

- Turkey. *Acta Physiologiae Plantarum*, 45. <https://doi.org/10.1007/s11738-023-03587-8>
- Caviedes, M., Carvajal-Larenas, F. E., & Zambrano, J. L. (2020). Tecnologías para el cultivo de maíz (*Zea mays*. L) en el Ecuador. *ACI Avances en Ciencias e Ingenierías, especial*(1). Recuperado de <https://revistas.usfq.edu.ec/index.php/avances/article/download/2588/3111?inline=1>
- Cedeño, A., & Javier, C. (2022). *Análisis del comportamiento agronómico del cultivo de maíz, considerando fenología y densidad de siembra, con el uso de tecnología aeroespacial y espectral* (Pregrado, UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO). UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO, Quevedo. Recuperado de <https://repositorio.uteq.edu.ec/handle/43000/6959>
- Choudhary, N. K., Sree Laya Chukkapalli, S., Mittal, S., Gupta, M., Abdelsalam, M., & Joshi, A. (2020). YieldPredict: A Crop Yield Prediction Framework for Smart Farms. *2020 IEEE International Conference on Big Data (Big Data)*, 2340-2349. <https://doi.org/10.1109/BigData50022.2020.9377832>
- Cisternas, I., Velásquez, I., Caro, A., & Rodríguez, A. (2020). Systematic literature review of implementations of precision agriculture. *Computers and Electronics in Agriculture*, 176, 105626. <https://doi.org/10.1016/j.compag.2020.105626>
- Clarke, A., Yates, D., Blanchard, C., Islam, M. Z., Ford, R., Rehman, S., & Walsh, R. (2024). The effect of dataset construction and data pre-processing on the eXtreme Gradient Boosting algorithm applied to head rice yield prediction in Australia. *Computers and Electronics in Agriculture*, 219, 108716. <https://doi.org/10.1016/j.compag.2024.108716>

- Constitución de la República del Ecuador. (2021). *Constitución de la República del Ecuador*. Recuperado de https://www.defensa.gob.ec/wp-content/uploads/downloads/2021/02/Constitucion-de-la-Republica-del-Ecuador_act_ene-2021.pdf
- Dey, B., Ferdous, J., & Ahmed, R. (2024). Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables. *Heliyon*, 10(3), e25112. <https://doi.org/10.1016/j.heliyon.2024.e25112>
- Dhaliwal, D. S., & Williams, M. M. (2024). Sweet corn yield prediction using machine learning models and field-level data. *Precision Agriculture*, 25(1), 51-64. <https://doi.org/10.1007/s11119-023-10057-1>
- FAO. (2018). FAOSTAT [Organización de las Naciones Unidas para la Alimentación y la Agricultura]. Recuperado 10 de abril de 2024, de FAO website: <https://www.fao.org/faostat/es/#data/QCL>
- FAO. (2023, abril 4). Ley Orgánica del Régimen de la Soberanía Alimentaria. | FAOLEX. Recuperado 9 de octubre de 2023, de Organización de las Naciones Unidas para la Alimentación y la Agricultura website: <https://www.fao.org/faolex/results/details/es/c/LEX-FAOC088076/>
- Feng, S., Hao, Z., Zhang, X., & Hao, F. (2021). Changes in climate-crop yield relationships affect risks of crop yield reduction. *Agricultural and Forest Meteorology*, 304-305, 108401. <https://doi.org/10.1016/j.agrformet.2021.108401>
- Furukawa, F., Maruyama, K., Saito, Y. K., & Kaneko, M. (2020). Corn Height Estimation Using UAV for Yield Prediction and Crop Monitoring. En R. Avtar & T. Watanabe (Eds.), *Unmanned Aerial Vehicle: Applications in Agriculture and*

- Environment* (pp. 51-69). Cham: Springer International Publishing.
https://doi.org/10.1007/978-3-030-27157-2_5
- García-Arteaga, J., Zambrano-Zambrano, J., Alcivar-Cevallos, R., & Romero, W. (2020). Predicción del rendimiento de cultivos agrícolas usando aprendizaje automático. *Revista Arbitrada Interdisciplinaria Koinonía*, 5, 144.
<https://doi.org/10.35381/r.k.v5i2.1013>
- Gardezi, M., & Bronson, K. (2020). Examining the social and biophysical determinants of U.S. Midwestern corn farmers' adoption of precision agriculture. *Precision Agriculture*, 21(3), 549-568. <https://doi.org/10.1007/s11119-019-09681-7>
- Gowda, S., & Reddy, S. (2020). Design And Implementation Of Crop Yield Prediction Model In Agriculture. *International Journal of Scientific & Technology Research*, VOLUME 8, 544.
- Granados Niño, J., Sánchez Duarte, J. I., Ochoa Martínez, E., Rodríguez Hernández, K., Reta Sánchez, D. G., & López Calderón, M. J. (2022). Efecto del ciclo de producción sobre el potencial de rendimiento y calidad nutricional del maíz forrajero en la Comarca Lagunera. *Revista mexicana de ciencias agrícolas*, (Extra 28), 207-217. Recuperado de <https://dialnet.unirioja.es/servlet/articulo?codigo=8624625>
- Guamán Guamán, R. N., Desiderio Vera, T. X., Villavicencio Abril, Á. F., Ulloa Cortázar, S. M., Romero Salguero, E. J., Guamán Guamán, R. N., ... Romero Salguero, E. J. (2020). Evaluación del desarrollo y rendimiento del cultivo de maíz (*Zea mays* L.) utilizando cuatro híbridos. *Siembra*, 7(2), 47-56.
<https://doi.org/10.29166/siembra.v7i2.2196>

- Guimarães, N., Fraga, H., Sousa, J. J., Pádua, L., Bento, A., & Couto, P. (2024). Comparative Evaluation of Remote Sensing Platforms for Almond Yield Prediction. *AgriEngineering*, 6(1), 240-258. <https://doi.org/10.3390/agriengineering6010015>
- Guo, Y., Xiao, Y., Li, M., Hao, F., Zhang, X., Sun, H., ... He, Y. (2022). Identifying crop phenology using maize height constructed from multi-sources images. *International Journal of Applied Earth Observation and Geoinformation*, 115(4), 103121. <https://doi.org/10.1016/j.jag.2022.103121>
- Gutierrez-Gomez, L., Petry, F., & Khadraoui, D. (2020). A Comparison Framework of Machine Learning Algorithms for Mixed-Type Variables Datasets: A Case Study on Tire-Performances Prediction. *IEEE Access*, 8, 214902-214914. <https://doi.org/10.1109/ACCESS.2020.3041367>
- Hasang-Moran, E. S., García-Bendezú, S. J., Carrillo-Zenteno, M. D., Durango-Cabanilla, W. D., Cobos-Mora, F. J., Hasang-Moran, E. S., ... Cobos-Mora, F. J. (2021). Sustentabilidad del sistema de producción del maíz, en la provincia de Los Ríos (Ecuador), bajo la metodología multicriterio de Sarandón. *Journal of the Selva Andina Biosphere*, 9(1), 26-40. <https://doi.org/10.36610/j.jsab.2021.090100026>
- Hu, T., Zhang, X., Bohrer, G., Liu, Y., Zhou, Y., Martin, J., ... Zhao, K. (2023). Crop yield prediction via explainable AI and interpretable machine learning: Dangers of black box models for evaluating climate change impacts on crop yield. *Agricultural and Forest Meteorology*, 336, 109458. <https://doi.org/10.1016/j.agrformet.2023.109458>

- INEC, I. N. de E. y C. (2020). Encuesta de Superficie y Producción Agropecuaria Continua-2020. Recuperado 10 de abril de 2024, de Instituto Nacional de Estadística y Censos website: <https://www.ecuadorencifras.gob.ec/encuesta-de-superficie-y-produccion-agropecuaria-continua-2020/>
- Isabona, J., Imoize, A. L., & Kim, Y. (2022). Machine Learning-Based Boosted Regression Ensemble Combined with Hyperparameter Tuning for Optimal Adaptive Learning. *Sensors*, 22(10), 3776. <https://doi.org/10.3390/s22103776>
- Khaki, S., & Wang, L. (2019). Crop Yield Prediction Using Deep Neural Networks. *Frontiers in Plant Science*, 10. <https://doi.org/10.3389/fpls.2019.00621>
- Khan, N., Kamaruddin, M., Sheikh, U., Hafiz, M., Yusup, Y., Bakht, M., & Mohamed Noor, N. (2022). Prediction of Oil Palm Yield Using Machine Learning in the Perspective of Fluctuating Weather and Soil Moisture Conditions: Evaluation of a Generic Workflow. *Plants*, 11, 1697. <https://doi.org/10.3390/plants11131697>
- Koca, Y. O., & Ereku, O. (2016). Changes of Dry Matter, Biomass and Relative Growth Rate with Different Phenological Stages of Corn. *Agriculture and Agricultural Science Procedia*, 10, 67-75. <https://doi.org/10.1016/j.aaspro.2016.09.015>
- Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Data Preprocessing for Supervised Learning. *International Journal of Computer Science*, 1, 111-117.
- Kumar, S. (2019). Machine Learning (Supervised). En B. Pochiraju & S. Seshadri (Eds.), *Essentials of Business Analytics: An Introduction to the Methodology and its Applications* (pp. 507-568). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-68837-4_16

- Kumari, A., & Bhavanishankar, K. (2023). Precision Agriculture. *International Journal of Computing Algorithm*, 12(1). Recuperado de <https://typeset.io/papers/precision-agriculture-1i4qkxdl>
- Lamos-Díaz, H., Puentes-Garzón, D.-E., & Zarate-Caicedo, D.-A. (2020). Comparison Between Machine Learning Models for Yield Forecast in Cocoa Crops in Santander, Colombia. *Facultad de Ingeniería*, 29(54). Recuperado de <https://www.redalyc.org/articulo.oa?id=413962511018>
- Lexis S.A. (2023, octubre 9). Constitución de la República del Ecuador | Descargar PDF Constitución de la República del Ecuador | Actualizado 2023. Recuperado 10 de octubre de 2023, de Lexis S.A. website: <https://www.lexis.com.ec/biblioteca/constitucion-republica-ecuador>
- LEY ORGANICA DEL REGIMEN DE LA SOBERANIA ALIMENTARIA. (2010). *LEY ORGANICA DEL REGIMEN DE LA SOBERANIA ALIMENTARIA*. Recuperado de <https://faolex.fao.org/docs/pdf/ecu88076.pdf>
- López Talavera, C. A., Ortigoza Guerreño, J., & González Villalba, J. D. (2019). *Guía técnica cultivo de maíz* | ISBN 978-99967-940-5-6—Libro (1.^a ed.). Facultad de Ciencias Agrarias - UNA. Recuperado de <https://isbn.cloud/9789996794056/guia-tecnica-cultivo-de-maiz/>
- Madhu, B., Rahman, M. A., Mukherjee, A., Islam, M. Z., Roy, R., & Ali, L. E. (2021). A Comparative Study of Support Vector Machine and Artificial Neural Network for Option Price Prediction. *Journal of Computer and Communications*, 9(5), 78-91. <https://doi.org/10.4236/jcc.2021.95006>

- Mahore, P. S., & Bardekar, Dr. A. A. (2021). Crop Yield Prediction. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 7(3), 561-569. <https://doi.org/10.32628/CSEIT2173168>
- Mustapha, I. B., Abdulkareem, M., Jassam, T. M., AlAteah, A. H., Al-Sodani, K. A. A., Al-Tholaia, M. M. H., ... Ganiyu, A. (2024). Comparative Analysis of Gradient-Boosting Ensembles for Estimation of Compressive Strength of Quaternary Blend Concrete. *International Journal of Concrete Structures and Materials*, 18(1), 20. <https://doi.org/10.1186/s40069-023-00653-w>
- Navarro-Racines, C. E., Álvarez, P., Ríos, D., Borja, R., Padilla, G., Montalvo, D., ... Arce, A. (2021). *Implementación de Servicios Integrados Participativos de Clima para la Agricultura (PICSA) en comunidades de Ecuador, Perú y Colombia, con enfoque agroecológico* [Working Paper]. CGIAR Research Program on Climate Change, Agriculture and Food Security. Recuperado de CGIAR Research Program on Climate Change, Agriculture and Food Security website: <https://cgspace.cgiar.org/handle/10568/116717>
- Nieto, L., Houborg, R., Zajdband, A., Jumpasut, A., Prasad, P., Olson, B. J., & Ciampitti, I. (2022). How Relevant is High-Cadence Earth Observation for Maize Crop Phenology Classification? *Kansas Agricultural Experiment Station Research Reports*, 8(4). <https://doi.org/10.4148/2378-5977.8298>
- Niu, Q., Li, X., Huang, J., Huang, H., Huang, X., Su, W., & Yuan, W. (2022). *A 30-m annual maize phenology dataset from 1985 to 2020 in China* [Preprint]. Biogeosciences and biodiversity. <https://doi.org/10.5194/essd-2021-343>
- Nokeri, T. C. (2022). Exploring Machine Learning. En T. C. Nokeri (Ed.), *Data Science Solutions with Python: Fast and Scalable Models Using Keras, PySpark MLlib*,

- H2O, XGBoost, and Scikit-Learn* (pp. 1-5). Berkeley, CA: Apress.
https://doi.org/10.1007/978-1-4842-7762-1_1
- Nyéki, A., & Neményi, M. (2022). Crop Yield Prediction in Precision Agriculture. *Agronomy*, 12(10), 2460. <https://doi.org/10.3390/agronomy12102460>
- Oikonomidis, A., Catal, C., & Kassahun, A. (2022). Hybrid Deep Learning-based Models for Crop Yield Prediction. *Applied Artificial Intelligence*, 36(1), 2031822. <https://doi.org/10.1080/08839514.2022.2031823>
- PDOT. (2019, 2023). PDYOT-10-SEP-2021.pdf. Recuperado 10 de octubre de 2023, de Google Docs website: https://drive.google.com/file/d/1QdixzOD2SD_I5wpXKB2RoPchQjUkpHGy/view?usp=sharing&usp=embed_facebook
- Ramírez Gómez, C. A. (2020, diciembre 31). *Aplicación del Machine Learning en agricultura de precisión | Revista CINTEX*. 25(2), 14-27. Recuperado de <https://revistas.pascualbravo.edu.co/index.php/cintex/article/view/356>
- Ransom, J. (2020, junio 25). Guía rápida de crecimiento y manejo del maíz | Agricultura NDSU. Recuperado 13 de mayo de 2024, de <https://www.ndsu.edu/agriculture/extension/publications/corn-growth-and-management-quick-guide>
- Raskar, S., Abhang, A., Pethe, N., & Attar, V. (2022). Crop Yield Prediction and Recommendation System. *2022 IEEE Pune Section International Conference (PuneCon)*, 1-6. <https://doi.org/10.1109/PuneCon55413.2022.10014957>
- Reyes González, F., Galvis Spinola, A., Almaraz Suárez, J. J., & Hernández-Mendoza, T. M. (2021). Modelo estadístico para la predicción del rendimiento de grano en

- maíz. *Revista Mexicana de Ciencias Agrícolas*, 12, 447-459.
<https://doi.org/10.29312/remexca.v12i3.2482>
- Salas, E. M., & Maldonado, E. S. (2020). Breve historia de la ciencia del cambio climático y la respuesta política global: Un análisis contextual/Brief history of climate change science and global political response: a context analysis. *KnE Engineering*. <https://doi.org/10.18502/keg.v5i2.6294>
- Saltz, J. S. (2021). CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps. *2021 IEEE International Conference on Big Data (Big Data)*, 2337-2344. <https://doi.org/10.1109/BigData52589.2021.9671634>
- Sánchez-Hernández, M. Á., Aguilar-Martínez, C. U., Valenzuela-Jiménez, N., Sánchez-Hernández, C., Jiménez-Rojas, M. C., & Villanueva-Verduzco, C. (2011). Densidad de siembra y crecimiento de maíces forrajeros. *Agronomía Mesoamericana*, 22(2), 281-295. Recuperado de http://www.scielo.sa.cr/scielo.php?script=sci_abstract&pid=S1659-13212011000200005&lng=en&nrm=iso&tlng=es
- Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, 181, 526-534. <https://doi.org/10.1016/j.procs.2021.01.199>
- Sierra-Forero, B.-L., Baron-Velandia, J., & Vanegas-Ayala, S.-C. (2024). Assessment of the relevance of features associated with corn crop yield prediction in Colombia, a country in the Neotropical zone. *International Journal of Information Technology*, 16(4), 2129-2138. <https://doi.org/10.1007/s41870-024-01762-9>
- Srinivas Rao, T., Hema, M., Sai Priya, K., Vamsi Krishna, K., & Sakhavath Ali, M. (2021, junio). *Clasificación de flores de iris mediante aprendizaje automático*. 9(6).

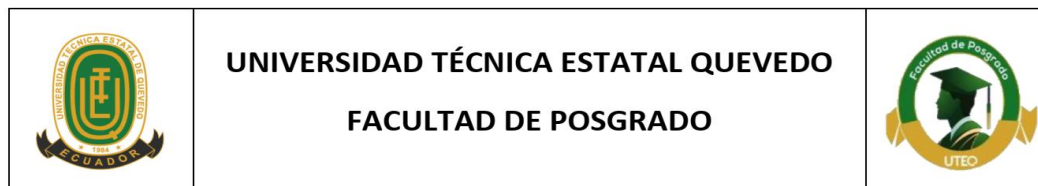
Recuperado de <https://www.ijaresm.com/iris-flower-classification-using-machine-learning>

- Steinberg, F., Burggaef, P., Wagner, J., & Heinbach, B. (2022). Impact of material data in assembly delay prediction—A machine learning-based case study in machinery industry. *The International Journal of Advanced Manufacturing Technology*, 120(1), 1333-1346. <https://doi.org/10.1007/s00170-022-08767-3>
- Stühler, H., Klau, D., Zöller, M.-A., Beiderwellen-Bedrikow, A., & Tutschku, C. (2024). End-to-End Implementation of Automated Price Forecasting Applications. *SN Comput. Sci.*, 5(4). <https://doi.org/10.1007/s42979-024-02735-2>
- Taeho, J. (2021). *Machine Learning Foundations: Supervised, Unsupervised, and Advanced Learning*. Springer Nature.
- Tarwidi, D., Pudjaprasetya, S. R., Adytia, D., & Apri, M. (2023). An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach. *MethodsX*, 10, 102119. <https://doi.org/10.1016/j.mex.2023.102119>
- Torgo, L. (2017). *Data Mining with R: Learning with Case Studies, Second Edition*. Taylor & Francis;Chapman and Hall/CRC.
- Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177, 105709. <https://doi.org/10.1016/j.compag.2020.105709>
- Villagran Munar, E., Arias, L., Gómez-Latorre, D., Giraldo, Y., Garzón, I., Vergel, S., ... Bonilla, R. (2023). *Estrategias de adaptación y mitigación al cambio climático en sistemas de producción agrícola: Un enfoque desde la agricultura protegida y técnicas de biotecnología para el manejo del cultivo*. <https://doi.org/10.21930/agrosavia.analisis.7406412>

- Wu, M., Xiong, J., Li, R., Dong, A., Lv, C., Sun, D., ... Niu, W. (2023, junio 6). *Predicting fertilizer concentration for precision irrigation under mixed variable-rate fertigation using machine learning: A case study of combined fertilization with dipotassium hydrogen phosphate and potassium chloride*. <https://doi.org/10.21203/rs.3.rs-2996273/v1>
- Zeng, L., Wardlow, B. D., Wang, R., Shan, J., Tadesse, T., Hayes, M. J., & Li, D. (2016). A hybrid approach for detecting corn and soybean phenology with time-series MODIS data. *Remote Sensing of Environment*, 181, 237-250. <https://doi.org/10.1016/j.rse.2016.03.039>

ANEXOS

Anexo 1: Informe de COMPILATIO



PARA: Ing. Byron Oviedo, PhD
Decano de Posgrado UTEQ

DE: Ing. Cristian Zambrano, PhD.

ASUNTO: Informe Proyecto de Investigación

FECHA: Quevedo, 27 de julio de 2024

Adjunto al presente sírvase encontrar el documento final del proyecto de investigación titulado: **Pronóstico del rendimiento de cultivo de maíz forrajero basado en variables agronómicas en la ciudad de Quevedo**, elaborado por el **Ing. Ricardo Rafael Villarroel Molina** posgradista de la **Maestría en Ciencia de Datos**. El Proyecto de investigación fue elaborado bajo mi dirección según lo asignado en el contrato PG-1238-2023-C-CIV-SERV-PROF de fecha 27 de noviembre de 2023, el mismo que cumple el informe de la herramienta COMPILATIO, el cual avala los niveles de originalidad, en un 96% del trabajo investigativo.



Atentamente,



Firmado electrónicamente por:
CRISTIAN GABRIEL ZAMBRANO VEGA

Ing. Cristian Zambrano, PhD.

DIRECTOR DE PROYECTO DE INVESTIGACIÓN

Anexo 2: Oficio de aceptación para colaboración en proyecto de investigación



Oficio Nro. INIAP-EETP_DIR-2024-0014-OF

Quevedo, 31 de enero de 2024

Asunto: Colaboración para desarrollo de proyecto de investigación.

Señor Ingeniero
Angel Ivan Torres Quijije
Coordinador Maestrías: Ciencia de Datos, Automatización y Control Industrial
UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO
En su Despacho

De mi consideración:

Reciba un cordial saludo del suscrito en calidad de Director de la Estación Experimental Tropical Pichilingue del INIAP.

Refiriéndome a su solicitud de respaldo en el proceso de coordinación y gestión para formalizar una colaboración entre la Universidad Técnica Estatal de Quevedo - UTEQ y la Estación Experimental Tropical Pichilingue - EETP del INIAP, en el marco del programa de Maestría en Ciencia de Datos, la cual busca desarrollar un proyecto de investigación que será elaborado por el estudiante Ricardo Rafael Villarroel Molina y dirigido por el Ing. Cristian Zambrano PhD, docente de la UTEQ, el cual tiene como objetivo principal la creación de un modelo de pronóstico del rendimiento de cultivo de maíz forrajero en la ciudad de Quevedo, haciendo uso de técnicas avanzadas de aprendizaje automático supervisado; así como también, la participación del Mgs. Javier Alejandro Maiguashca Guzmán - Investigador del INIAP, en la construcción del modelo predictivo.

Me permito informar que su pedido ha sido aceptado de manera favorable, para lo cual se requiere mantener reunión trabajo en la Estación Experimental Tropical Pichilingue el día martes 6 de febrero del 2024 a las 09h00, de acuerdo a su disponibilidad en Agenda.

Mucho agradeceré su confirmación al correo: pichilingue@iniap.gob.ec.

Con sentimientos de distinguida consideración.

Atentamente,

Documento firmado electrónicamente

Mgs. Carlos Alberto Molina Hidrovo
DIRECTOR DE ESTACIÓN EXPERIMENTAL TROPICAL PICHILINGUE, ENCARGADO

Referencias:
- INIAP-EETP_DIR-2024-0005-E

Copia:
Señor Ingeniero
Byron Wladimir Oviedo Bayas, PhD.
Decano de Postgrado
UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

Señor Magíster
Javier Alejandro Maiguashca Guzman
Investigador Auxiliar 1

Señor Médico Veterinario
Julio Rafael Gilces Lectong
Responsable de Ganadería y Pastos, Delegado

Instituto Nacional de Investigaciones Agropecuarias
Dirección: Km 5 Vía Quevedo - El Empalme, Provincia de los Ríos
Mocache - Ecuador. Teléfono: +593-5-2783128 / 2783044
www.iniap.gob.ec

* Documento firmado electrónicamente por Quijije



1/2



Oficio Nro. INIAP-EETP_DIR-2024-0014-OF

Quevedo, 31 de enero de 2024

Señor Ingeniero
Carlos Julio Macías Looz
Responsable de Talento Humano de Estación Experimental

Señora Ingeniera
Ruth Marjorie Tenesaca Mera
Responsable de Planificación y Gestión Estratégica, delegada (Unidad Desconcentrada)

Anexo 3: Capturas de pantalla de pantalla de hojas de cálculo con datos agronómicos

A		B			C	D	DK	DL	DM	DN	DO	DP	DQ	DR	DS	DT	DU	DV	DW	DX	▲
03/08/2023																					
DIAMETRO DEL TALLO mm																					
DISTANCIA DE SIEMBRA		Parcela GRANDE	Parcela PEQUEÑA	REPETICIONES (BLOQUE)	PLANTA 1	PLANTA 2	PLANTA 3	PLANTA 4	PLANTA 5	PLANTA 6	PLANTA 7	PLANTA 8	PLANTA 9	PLANTA 10		PLANTA 1	PLANTA 2	PLANTA 3	PL		
58				H551	1	29,11	27,13	24,08	25,17	23,55	26,87	25,10	25,36	23,07	22,47		29,11	27,13	24,08		
59				H_554	1	25,84	23,50	23,98	25,17	24,35	25,24	22,83	24,06	24,80	28,67		26,04	23,50	24,28		
60				QPM_543	1	22,11	25,43	24,40	23,50	23,67	22,33	18,83	17,16	26,01	23,84		22,11	27,00	24,40		
61				ADVANTA	1	22,61	27,08	24,37	26,37	25,12	25,41	26,08	25,91	24,70	24,8		22,61	27,68	24,47		
62				H551	2	26,15	23,85	27,58	28,17	23,39	25,80	25,55	24,59	23,34	28,72		26,15	23,85	27,58		
63				H_554	2	22,80	25,09	23,20	24,80	27,67	24,49	21,80	25,73	26,12	30,35		22,80	25,89	23,20		
64				QPM_543	2	27,11	23,18	24,37	19,36	22,33	19,28	25,23	24,15	22,21	27,21		27,11	23,78	24,37		
65				ADVANTA	2	26,26	25,40	25,60	26,97	26,16	25,50	26,06	24,86	24,49	26,90		26,26	26,34	26,03		
66				H551	3	26,60	22,81	26,03	25,70	25,39	25,49	23,30	24,59	23,26	25,08		26,60	22,93	26,26		
67				H_554	3	25,52	22,26	22,10	19,15	22,20	25,09	24,70	23,98	22,56	24,99		25,52	22,26	22,10		
68				QPM_543	3	22,11	18,80	24,33	20,20	22,51	25,36	25,51	20,14	27,45	25,38		23,00	18,80	24,33		
69				ADVANTA	3	27,98	21,81	26,38	25,07	26,03	21,18	25,11	26,79	27,12	22,70		28,03	21,81	27,00		
70				H551	4	22,29	26,24	23,10	26,05	23,26	27,35	20,25	24,42	26,28	24,05		22,29	26,24	23,10		
71				H_554	4	20,08	27,02	22,69	24,73	22,13	24,68	22,69	25,82	24,33	25,27		20,20	27,19	23,00		
72				QPM_543	4	23,17	24,28	22,30	26,78	24,25	24,20	20,70	24,21	22,32	25,85		23,35	25,00	22,50		
73				ADVANTA	4	24,91	24,3	24,67	24,36	26,72	26,23	22,00	19,45	23,09	23,80		25,31	25,09	25,87		
74				H551	1	24,80	21,82	22,72	21,69	22,55	22,09	23,11	23,30	22,96	22,65		25,30	21,82	22,72		
75				H_554	1	25,60	22,41	19,03	26,49	25,07	27,20	22,22	20,55	25,16	25,02		26,60	27,41	19,73		
76				QPM_543	1	18,06	22,30	21,92	26,49	25,32	21,04	22,45	21,90	22,52	23,21		18,06	22,30	21,92		
77				ADVANTA	1	27,60	25,51	25,53	28,53	25,06	26,53	26,07	26,67	24,65	23,82		27,60	25,51	25,53		
78				H551	2	25,97	24,30	15,49	24,53	20,26	18,20	23,32	21,18	20,12	21,35		25,97	24,30	15,49		
79				H_554	2	23,95	24,18	23,14	24,86	16,00	25,04	20,54	24,31	21,13	19,29		23,95	24,58	23,84		
80				QPM_543	2	18,06	23,46	23,27	20,38	23,38	21,88	23,40	22,85	24,66	24,12		18,06	24,00	23,27		
81				ADVANTA	2	22,36	21,20	20,23	23,52	18,30	21,22	23,75	21,90	22,46	24,34		23,16	21,20	20,23		
82				H551	3	24,74	23,10	17,70	26,70	21,66	25,72	23,95	25,12	25,63	25,03		24,74	23,10	17,70		
83				H_554	3	27,76	25,72	22,89	25,40	23,75	23,21	28,26	21,69	16,06	18,60		28,17	26,00	22,99		
84				QPM_543	3	19,97	23,59	23,20	21,60	25,74	25,95	25,90	22,70	23,61	21,64		19,97	23,59	23,20		
85				ADVANTA	3	22,19	23,10	22,52	24,20	25,20	21,86	19,90	24,19	25,36	20,14		22,19	23,10	23,10		
▶ ◀ ☰ DATOS AGRONÓMICOS MANEJO DEL CULTIVO ETAPA VEGETATIVA EVIDENCIA FOTO ETAPA REPRO: ☰ + - ☰																					

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
482	03/08/2023	43	80x20	D1	H551	1	29,11	27,13	24,08	25,17	23,55	26,87	25,10	25,36	23,07	22,47
483		H_554			1	25,84	23,50	23,98	25,17	24,35	25,24	22,83	24,06	24,80	28,67	
484		QPM_543			1	22,11	25,43	24,40	23,50	23,67	22,33	18,83	17,16	26,01	23,84	
485		ADVANTA			1	22,61	27,08	24,37	26,37	25,12	25,41	26,08	25,91	24,70	24,8	
486		H551			2	26,15	23,85	27,58	28,17	23,39	25,80	25,55	24,59	23,34	28,72	
487		H_554			2	22,80	25,09	23,20	24,80	27,67	24,49	21,80	25,73	26,12	30,35	
488		QPM_543			2	27,11	23,18	24,37	19,36	22,33	19,28	25,23	24,15	22,21	17,21	
489		ADVANTA			2	26,26	25,40	25,60	26,97	26,16	25,50	26,06	24,86	24,49	26,82	
490		H551			3	26,60	22,83	26,03	25,70	25,29	25,30	22,93	24,59	23,26	25,08	
491		H_554			3	25,52	22,26	22,10	19,15	22,20	25,09	24,70	23,98	22,56	24,99	
492		QPM_543			3	22,11	18,80	24,33	20,20	22,51	25,36	25,51	20,14	27,45	25,38	
493		ADVANTA			3	27,98	21,81	26,38	25,07	26,03	21,18	25,11	26,79	27,12	22,70	
494		H551			4	22,29	26,24	23,10	26,05	23,26	27,35	20,25	24,42	26,28	24,05	
495		H_554			4	20,08	27,02	22,69	24,73	22,13	24,68	22,69	25,82	24,33	25,27	
496		QPM_543			4	23,17	24,28	22,30	26,78	24,25	24,20	20,70	24,21	22,32	25,85	
497		ADVANTA			4	24,91	24,3	24,67	24,36	26,72	26,23	22,00	19,45	23,09	23,90	
498		60x20	D2	H551	1	24,80	21,82	22,72	21,69	22,55	22,09	23,11	23,30	22,96	22,65	
499				H_554	1	25,60	27,41	19,03	26,82	25,07	27,30	22,22	20,55	25,16	25,02	
500				QPM_543	1	18,06	22,30	21,92	26,49	25,32	21,04	22,45	21,90	22,52	23,21	
501				ADVANTA	1	27,60	25,51	25,53	28,53	25,06	26,53	26,07	26,67	24,65	23,82	
502				H551	2	25,97	24,30	15,49	24,53	20,36	18,20	23,32	21,18	20,12	21,35	
503				H_554	2	23,95	24,18	23,14	24,86	16,00	25,04	20,54	24,31	21,13	19,29	
504				QPM_543	2	18,06	23,46	23,27	20,38	23,38	21,88	23,40	22,85	24,66	24,12	
505				ADVANTA	2	22,36	21,20	20,23	23,52	18,30	21,22	23,75	21,90	22,46	24,34	
506				H551	3	24,74	23,10	17,70	26,70	21,66	25,72	23,95	25,12	25,63	25,03	
507				H_554	3	27,76	25,72	22,89	25,40	23,75	23,21	28,26	21,69	16,06	18,60	
508				QPM_543	3	19,97	23,59	23,20	21,60	25,74	25,95	25,90	22,70	23,61	21,64	
509				ADVANTA	3	22,19	23,10	22,52	24,20	25,20	21,86	19,90	24,19	25,36	20,04	
510				H551	4	23,70	27,02	24,90	18,96	24,00	23,73	22,70	18,56	24,37	20,18	
511				H_554	4	25,89	25,19	26,07	27,03	24,39	22,62	23,97	23,71	25,91	27,36	
512				QPM_543	4	19,77	20,65	19,61	18,92	25,70	23,32	25,46	24,35	20,34	25,76	
513				ADVANTA	4	23,18	24,56	20,07	24,47	22,43	24,25	23,90	21,90	25,36	26,99	
<	>	---	ETAPA_VEGETATIVA_EV		DIAMETRO_TALLO_EV		ALTURA_PLANTA_EV		INFLORESCENCIA_MASCULINA_ER		INF_MASCULINA_D_ER		INFLORESCENCIA_FEMENINA_ER		INF_FEME? --- +	

Anexo 4: Capturas de pantalla de código utilizado para obtención de resultados del objetivo 1

Matizipynb

```
[7]: # Procesamiento de tratamiento o asignación de tratamiento dependiendo de los factores existentes.
for index_sheet in range(len(sheet_names)):
    if len(grouped_cols)>0:
        ffill_df(sheets_agro, index_sheet, grouped_cols)

        if not var_tratamiento:
            for i_factor in range(len(var_factors)):
                sheets_agro[index_sheet].insert(sheets_agro[index_sheet].columns.get_loc(var_factors[i_factor])+1, 'N_'+var_factors[i_factor], sheets_agro[index_sheet][var_factors[i_factor]].replace(
                    sheets_agro[index_sheet].insert(sheets_agro[index_sheet].columns.get_loc('N_'+ var_factors[-1])+1, 'TRATAMIENTO', sheets_agro[index_sheet][var_factors[0]].str.cat(sheets_agro[index_she
                    sheets_agro[index_sheet].rename(columns={'DIA':'DOS', 'DISTANCIA_SIEMBRA ':'DISTANCIA_SIEMBRA', 'PARCELA_PEQUENIA':'VARIEDAD_SEMILLA'}, inplace=True)
                    sheets_agro[index_sheet].drop(['PARCELA_GRANDE', 'N_PARCELA_GRANDE', 'N_PARCELA_PEQUENIA'], axis=1, inplace=True)
                var_tratamiento = 'TRATAMIENTO'

p_df(sheets_agro[1].sample(n=10), title='Conjunto de los datos despues del proceso', tag_titles='H1')
```

Conjunto de los datos despues del proceso

FECHA	DOS	DISTANCIA_SIEMBRA	VARIEDAD_SEMILLA	TRATAMIENTO	REPETICIONES	PLANTA 1	PLANTA 2	PLANTA 3	PLANTA 4	PLANTA 5	PLANTA 6	PLANTA 7	PLANTA 8	PLANTA 9	PLANTA 10
2023-07-13	22.0	60x20	QPM_543	D2_S3	2	11.49	16.23	15.26	14.65	13.90	14.00	16.96	16.23	15.76	16.00
2023-07-03	12.0	80x20	H_554	D1_S2	4	5.05	7.83	6.34	7.35	7.84	5.30	6.24	5.91	5.50	7.63
2023-07-13	22.0	40x20	ADVANTA	D3_S4	1	14.02	14.87	18.68	14.18	17.80	14.97	13.55	11.79	17.16	15.82
2023-08-03	43.0	80x20	QPM_543	D1_S3	1	22.11	25.43	24.40	23.50	23.67	22.33	18.83	17.16	26.01	23.84
2023-07-03	12.0	80x20	H551	D1_S1	1	7.10	8.23	6.64	7.17	6.40	7.28	8.18	4.51	3.79	6.19
2023-08-10	50.0	40x20	H551	D3_S1	3	28.90	23.55	24.40	24.46	26.15	25.60	23.58	21.75	24.50	23.99
2023-08-21	61.0	40x20	H551	D3_S1	2	22.23	20.06	22.86	23.00	22.18	23.08	22.56	22.15	20.18	23.19
2023-07-31	40.0	60x20	QPM_543	D2_S3	2	17.82	21.36	23.12	19.81	23.20	21.50	23.34	22.79	24.00	23.87
2023-07-20	29.0	60x20	H551	D2_S1	1	20.69	18.36	18.63	19.53	18.00	18.35	21.29	20.25	18.69	17.98
2023-07-03	12.0	60x20	H551	D2_S1	3	7.50	7.62	5.91	10.47	5.35	9.68	6.68	6.37	5.74	4.33

Matizipynb

Filtrado de Datos agronomicos del 21 de agosto de 2023

```
[16]: fecha_principal = '21-08-2023'
df_agro_reduce = df_agro[df_agro['FECHA']==fecha_principal]
#df_agro_reduce = df_agro_reduce.sort_values(by=['N_PARCELA_GRANDE', 'N_PARCELA_PEQUENIA'])
df_agro_reduce['ETAPA_VEGETATIVA_EV'] = df_agro_reduce['ETAPA_VEGETATIVA_EV'].cat.remove_unused_categories()

labels = ['DIAMETRO_TALLO_EV', 'ALTURA_PLANTA_EV', 'MAZORCAS_PLANTA_ER', 'INSERCIÓN_MAZORCA_ER']
labels_new = ['Diámetro del tallo (cm)', 'Altura de la planta (cm)', 'Número de mazorcas', 'Altura de inserción de mazorca (cm)']
df_agro_reduce.drop(['FECHA', 'DOS'], axis=1, inplace=True)
df_agro_reduce['ETAPA_VEGETATIVA_EV2'] = df_agro_reduce['ETAPA_VEGETATIVA_EV'].str.replace('V', '')

df = df_agro_reduce.copy()

describe(dfndf.select_dtypes(include=['float']), columns=titles_vars_no_categorias[1:], title='Estadísticos descriptivos de datos Agronomicos Números del 21 de agosto de 2023')
df_agro_reduce['ETAPA_VEGETATIVA_EV2'] = df_agro_reduce['ETAPA_VEGETATIVA_EV2'].astype(float)
```

Estadísticos descriptivos de datos Agronomicos Números del 21 de agosto de 2023

Variable	Elementos	Media	Std	Min	25%	50%	75%	Max	Nulls	Nulls %
Diámetro del tallo (cm)	480.0	23.65	2.49	15.49	22.11	23.76	25.47	30.95	0	0.00
Altura de la planta (cm)	480.0	218.70	31.81	65.00	200.00	222.00	240.00	340.00	0	0.00
Número de mazorcas	435.0	1.07	0.25	1.00	1.00	1.00	1.00	2.00	45	9.38
Altura de inserción de mazorca (cm)	438.0	134.31	20.19	51.00	123.00	133.50	146.00	204.00	42	8.75

Anexo 5: Capturas de pantalla de código utilizado para obtención de resultados del objetivo 2

OBJ 2: Realizar experimentos para obtener la configuración optima de hiperparámetros de los algoritmos para la predicción del rendimiento de cultivo de maíz forrajero.

Libraries and functions

+ 4 cells hidden

Hyperparameters Tuning

```
[32]: ## READ DATA
scaler = MinMaxScaler().set_output(transform='pandas')
df_crop_yield = pd.read_csv('df_crop_yield.csv', sep=';', encoding='UTF-8')

[33]: ## CONFIGURE FEATURES AND TARGETS
numeric_features = ['DISTANCIA_SIEMBRA', 'DIAMETRO_TALLO_EV', 'ALTURA_PLANTA_EV', 'MAZORCAS_PLANTA_ER', 'INSERCIÓN_MAZORCA_ER', 'ETAPA_VEGETATIVA_EV2']
binary_features = ['D_ADVANTA', 'D_H551', 'D_H554', 'D_QPM_S43']
features_dummies = ['D_40X20', 'D_60X20', 'D_80X20']

features = numeric_features + binary_features + features_dummies
targets = ['PESO_TOTAL_KG', 'PESO_HOJAS_KG', 'PESO_TALLO_KG', 'PESO_MAZORCA_KG']

[34]: # FEATURE SELECTION BY TARGET
for target in targets:
    X_train, X_test, y_train, y_test = train_test_split(df_crop_yield, features, target, numeric_features)
    p_df(make_rfe(X_train, y_train), "Recursive Feature Elimination for " + target)
```

Recursive Feature Elimination for PESO_TOTAL_KG

feature	support	ranking
DISTANCIA_SIEMBRA	True	1
DIAMETRO_TALLO_EV	True	1
ALTURA_PLANTA_EV	True	1
MAZORCAS_PLANTA_ER	False	7

```
[35]: ## CONFIGURE FEATURES AND TARGETS
# Se eliminó MAZORCAS_PLANTA_ER y ETAPA_VEGETATIVA_EV2, además se optó por descartar las variables dummies de DISTANCIA_SIEMBRA
numeric_features = ['DIAMETRO_TALLO_EV', 'ALTURA_PLANTA_EV', 'INSERCIÓN_MAZORCA_ER']
binary_features = ['D_ADVANTA', 'D_H551', 'D_H554', 'D_QPM_S43']
features = numeric_features + binary_features
targets = ['PESO_TOTAL_KG', 'PESO_HOJAS_KG', 'PESO_TALLO_KG', 'PESO_MAZORCA_KG']

[ ]: df_tuning_results = []
df_best_params_results = []

for target in targets:
    X_train, X_test, y_train, y_test = train_test_split(df_crop_yield, features, target, numeric_features)
    p_title('Hyperparameters Tuning for ' + target)

    algorithms, tuning_results = hyperparameters_tuning(configure_estimators, X_train, y_train,
                                                         scorings=['neg_root_mean_squared_error', 'r2'],
                                                         refits='neg_root_mean_squared_error',
                                                         return_train_score=True, test_params=False, verbose=1)

    df_best_params = get_best_params(algorithms, tuning_results)
    df_best_params.insert(0, 'target', target)

    df_tuning = get_data(algorithms, tuning_results)
    df_tuning.insert(0, 'target', target)

    df_tuning_results = concat_dfs(df_tuning_results, df_tuning)
    df_best_params_results = concat_dfs(df_best_params_results, df_best_params)

df_tuning_results.to_csv('df_tuning_results.csv', index=False, sep=';', encoding='UTF-8')
df_best_params_results.to_csv('df_best_params_results.csv', index=False, sep=';', encoding='UTF-8')
```

Plot Results Hyperparameters Tuning

+ 3 cells hidden

Anexo 6: Capturas de pantalla de código utilizado para obtención de resultados del objetivo 3

OBJ 3: Seleccionar el modelo que mejores pronósticos proporcione sobre el rendimiento del maíz forrajero utilizando la métrica del error cuadrático medio.

Libraries and functions

+ 3 cells hidden

Model Evaluation

```
[ ]: df_metrics_final = []

for target in targets:
    tuning_results = df_best_params_results[df_best_params_results['target']==target]
    algorithms, tuning_results = make_tuning_results(tuning_results)

    X_train, X_test, y_train, y_test = train_test_split(df_crop_yield, features, target, numeric_features)
    p_title('Evaluating Models for ' + target)

    df_pr, df_metrics = models_evaluation(algorithms, tuning_results, X_train, X_test, y_train, y_test, repetitions=4, verbose=1)
    df_metrics.insert(0, 'TARGET', target)

    df_metrics_final = concat_dfs(df_metrics_final, df_metrics)
df_metrics_final.to_csv('df_models_evaluation_results.csv', index=False, sep=';', encoding='UTF-8')
```

Plot Results Model Evaluation

```
[14]: # Run libraries and functions of Configuration
df_metrics_final = pd.read_csv('df_models_evaluation_results.csv', sep=';', encoding='UTF-8')

[15]: df_metrics_final.replace(targets_new, inplace=True)
df_metrics_final.replace(algorithms_new, inplace=True)
df_metrics_final.columns = ['Target', 'Iteración', 'Algoritmo', 'Parametros', 'TRAIN_RMSE', 'TRAIN_R2', 'TRAIN_MSE',
                           'TRAIN_MAE', 'Tiempo de Entrenamiento (ms)', 'TEST_RMSE', 'TEST_R2', 'TEST_MSE', 'TEST_MAE', 'Tiempo de Evaluación (ms)']

[16]: df_metrics_final2 = df_metrics_final.copy()
```

```
[18]: for i_target, target in enumerate(targets_new):
    p_title('Métricas con ' + target, tag='h2')
    df = df_metrics_final[df_metrics_final['Target']==target]
    cdlc_temp = cdlc[i_target]
    for index in range(len(metrics)):

        metric = metrics[index]
        metric_label = metrics_label[index]
        p_title('Datos de ' + metric, tag='h3')
        f_result, n_result = get_statistical_test(df, cols_index=['Algoritmo'], cols_column=['Iteración'], metrics=metric, groups_names=algorithms_new)
        p_title('Friedman test statistic: {:.2f} pvalue: {:.5f}'.format(f_result.statistic, f_result.pvalue))
        p_title(n_result, title = 'Post-hoc Nemenyi test')
        pair_df = get_groups_nemenyi(nemenyi_result = n_result, groups = algorithms_new, alpha = 0.05)
        p_df(pair_df[pair_df['reject'] == True], title = 'Post-hoc Nemenyi test groups')
        results = df[['Algoritmo', metric]].groupby('Algoritmo', sort=False).mean().reset_index()
        results['cd'] = cdlc_temp[index]
        p_df(results, title = 'Mean by ' + metric)
        significant_pairs = pair_df[pair_df['reject'] == True][['group1', 'group2']].values.tolist()
        cld = compactletterdisplay(compact_letter_display(significant_pairs, algorithms_new)
        print(cld)
        cldcdls_temp[index]
        make_barplot(x='Algoritmo', y = metric, df=df, ylabel=metric_label, title = 'Métricas de ' + target + ' para ' + metric, errorbars='sd', format_text = '.3f', cdl_letters=cld)
        p_salto(2)
```

Métricas con Peso Total Kg

Datos de TRAIN_RMSE

Friedman test statistic: 23.25 pvalue: 0.00072

Post-hoc Nemenyi test

Index	LR	RF	SVM	MLP	DNN	GB	XGB
LR	1.000000	0.120596	0.900000	0.900000	0.900000	0.440184	0.900000
RF	0.120596	1.000000	0.003171	0.010545	0.637136	0.900000	0.637136
SVM	0.900000	0.003171	1.000000	0.900000	0.336257	0.030924	0.336257
MLP	0.900000	0.010545	0.900000	1.000000	0.540899	0.079497	0.540899
DNN	0.900000	0.637136	0.336257	0.540899	1.000000	0.900000	0.900000
GB	0.440184	0.900000	0.030924	0.079497	0.900000	1.000000	0.900000
XGB	0.900000	0.637136	0.336257	0.540899	0.900000	0.900000	1.000000