agronomy

MDPI

Article

Prediction of Grain Yield in Wheat by CHAID and MARS Algorithms Analyses

Fatih Demirel [1], Baris Eren [1], Abdurrahim Yilmaz [2], Aras Türkoğlu [3], Kamil Haliloğlu [4], Gniewko Niedbała [5], Henryk Bujak [6,7], Bita Jamshidi [8], Alireza Pour-Aboughadareh [9,*], Jan Bocianowski [10] and Kamila Nowosad [6,*]

1 Department of Agricultural Biotechnology, Faculty of Agriculture, Igdır University, Igdir 76000, Türkiye; fatih.demirel@igdir.edu.tr (F.D.); baris.eren@igdir.edu.tr (B.E.)
2 Department of Field Crops, Faculty of Agriculture, Bolu Abant Izzet Baysal University, Bolu 14030, Türkiye; abdurrahimyilmaz@ibu.edu.tr
3 Department of Field Crops, Faculty of Agriculture, Necmettin Erbakan University, Konya 42310, Türkiye; aras.turkoglu@erbakan.edu.tr
4 Department of Field Crops, Faculty of Agriculture, Ataturk University, Erzurum 25240, Türkiye; kamilh@atauni.edu.tr
5 Department of Biosystems Engineering, Faculty of Environmental and Mechanical Engineering, Poznań University of Life Sciences, Wojska Polskiego 50, 60-627 Poznań, Poland; gniewko.niedbala@up.poznan.pl
6 Department of Genetics, Plant Breeding and Seed Production, Wrocław University of Environmental and Life Sciences, Plac Grunwaldzki 24A, 53-535 Wrocław, Poland; henryk.bujak@upwr.edu.pl or h.bujak@coboru.gov.pl
7 Research Centre for Cultivar Testing, Słupia Wielka 34, 63-022 Słupia Wielka, Poland
8 Department of Food Security and Public Health, Khabat Technical Institute, Erbil Polytechnic University, Erbil 44001, Iraq; bita.alimer@epu.edu.iq
9 Seed and Plant Improvement Institute, Agricultural Research, Education and Extension Organization (AREEO), Karaj P.O. Box 3158854119, Iran
10 Department of Mathematical and Statistical Methods, Poznan University of Life Sciences, Wojska Polskiego 28, 60-637 Poznań, Poland; jan.bocianowski@up.poznan.pl
* Correspondence: a.poraboghadareh@edu.ikiu.ac.ir (A.P.-A.); kamila.nowosad@upwr.edu.pl (K.N.)

check for updates

Citation: Demirel, F.; Eren, B.; Yilmaz, A.; Türkoğlu, A.; Haliloğlu, K.; Niedbała, G.; Bujak, H.; Jamshidi, B.; Pour-Aboughadareh, A.; Bocianowski, J.; et al. Prediction of Grain Yield in Wheat by CHAID and MARS Algorithms Analyses. *Agronomy* **2023**, *13*, 1438. https://doi.org/10.3390/agronomy13061438

Academic Editor: Louis Kouadio

Received: 27 April 2023
Revised: 16 May 2023
Accepted: 19 May 2023
Published: 23 May 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract: Genetic information obtained from ancestral species of wheat and other registered wheat has brought about critical research, especially in wheat breeding, and shown great potential for the development of advanced breeding techniques. The purpose of this study was to determine correlations between some morphological traits of various wheat (*Triticum* spp.) species and to demonstrate the application of MARS and CHAID algorithms to wheat-derived data sets. Relationships among several morphological traits of wheat were investigated using a total of 26 different wheat genotypes. MARS and CHAID data mining methods were compared for grain yield prediction from different traits using cross-validation. In addition, an optimal CHAID tree structure with minimum RMSE was obtained and cross-validated with nine terminal nodes. Based on the smallest RMSE of the cross-validation, the eight-element MARS model was found to be the best model for grain yield prediction. The MARS algorithm proved superior to CHAID in grain yield prediction and accounted for 95.7% of the variation in grain yield among wheats. CHAID and MARS analyses on wheat grain yield were performed for the first time in this research. In this context, we showed how MARS and CHAID algorithms can help wheat breeders describe complex interaction effects more precisely. With the data mining methodology demonstrated in this study, breeders can predict which wheat traits are beneficial for increasing grain yield. The adaption of MARS and CHAID algorithms should benefit breeding research.

Keywords: morphological characterization; plant breeding; prediction; selection

1. Introduction

The origin of wheat is Southwest Asia, Türkiye, Syria, Iraq and Iran [1,2]. Small farmers in Türkiye, a prominent wheat gene center, still cultivate einkorn (*Triticum monococcum* L.) and emmer (*Triticum dicoccum* L.) wheat varieties, which are considered precursors to modern wheat [3,4]. The einkorn and emmer wheat varieties are considered significant genetic resources for increasing both the yield and quality of wheat as part of wheat breeding initiatives [3].

Cultivars have grown more genetically uniform and contain less genetic variability than their local varieties, transitional forms and wild relatives [5]. When local types are improved using the correct selection method, they have the potential to be used in plant breeding studies to increase genetic diversity [6,7]. It is essential to recognize the relationship between improved agronomic traits and their mutual influence. Knowing the means by which wild-grown genotypes increase seed yield will contribute to proper selection [8]. For this purpose, the use of correlation analyses is common [9]. However, in order to properly determine the relationship between yield and agronomic traits for plant breeding studies, more sophisticated analysis techniques are needed than are found in the literature, i.e., multivariate adaptive regression splines (MARSs) algorithms and chi-square automatic interaction detectors (CHAIDs) [10].

MARS is a non-parametric regression technique that detects the complex relationship between predictors and response variables [11]. It has been found that MARS modeling can enable plant breeders to establish agronomic traits that positively influence yield traits vis a criteria indirect selection framework in plant breeding studies [12]. In this regard, several previous studies on path analysis for different plant species can be found in the literature [13]. For example, through correlation and path analyses, Janmohammadi et al. [14] positively identified agronomic traits associated with grain yield in bread wheat. However, the application of a predictive model built by the MARS algorithm, without the need for distributional assumptions and functional variables, to different plant species is still rare in plant breeding research [12].

A regression tree is a method of analysis that detects the relationship and interaction between independent and dependent variables [15]. The hierarchical tree structure of decision tree algorithms makes them easy to understand and use for classification and regression [16]. Additionally, regression trees and the CHAID algorithm are used to determine the relationship between independent and dependent variables [17]. CHAID is a technique that performs Bonferroni correction to calculate adjusted *p*-values at the split points of a regression tree created for a continuous response using an *F*-significance test with ten-fold cross-validation [11]. CHAID has been successfully used to predict crop fields based on sunflower plant traits [10]. It has also been used to estimate the water requirement of maize at different growth stages [16].

To the best of our knowledge, the description of the relationship between grain yield (GY) and agronomic traits of current wheat genotypes under microclimate conditions has not yet been documented using MARS or CHAID analysis. Therefore, the present study attempts to determine the agronomic traits positively affecting the GY of these 26 wheat genotypes (which have different levels of ploidy) under the microclimate conditions of Igdir province in the eastern part of Türkiye, by means of correlation analysis (CHAID) techniques, and in particular used MARS analysis as a powerful non-parametric regression technique that allows for describing multivariate relationships between sets of dependent factors and predictors.

2. Material and Method

2.1. Material

Nine populations of emmer (*Triticum dicoccum* L.) and twelve populations of einkorn wheat (*Triticum monococcum* L.) as well as five registered varieties (Dogankent, Kızıltan91, C.1252, Sarıcanak98, and Demir-2000) were used in this study (Table 1). Dogankent is a registered bread wheat, and the others are registered durum wheats. Wheat varieties

belonging to different ploidy levels were used to maintain wide variability. The materials were sown in Igdir province during the summer seasons of 2017–2018 and 2018–2019. The experiment field is located in the Agricultural Application and Research field of Igdir University in Türkiye (39°55′42.9″ N latitude, 44°5′37.7″ E longitude at 852 m above sea level) at a distance of 8 km from the province's administrative center city of Igdir (Figure 1).

Table 1. Identity information of wheat genotypes evaluated in the study.

No	Wheat Genotypes	Province	Location
G1	*T. dicoccum* L.	Kayseri	38°17′04.1″ N–35°34′59.3″ E
G2	*T. dicoccum* L.	Kayseri	38°17′13.4″ N–35°35′43.3″ E
G3	*T. dicoccum* L.	Kayseri	38°15′47.9″ N–35°37′19.3″ E
G4	*T. dicoccum* L.	Kayseri	38°15′15.5″ N–35°37′40.1″ E
G5	*T. dicoccum* L.	Kayseri	38°13′17.6″ N–35°37′32.4″ E
G6	*T. dicoccum* L.	Kayseri	38°13′05.1″ N–35°37′32.7″ E
G7	*T. dicoccum* L.	Kayseri	38°11′23.1″ N–35°40′18.0″ E
G8	*T. dicoccum* L.	Kayseri	38°11′14.1″ N–35°40′25.0″ E
G9	*T. dicoccum* L.	Kayseri	38°11′09.9″ N–35°40′19.3″ E
G10	*T. monococcum* L.	Kastamonu	41°12′06.2″ N–33°33′25.5″ E
G11	*T. monococcum* L.	Kastamonu	41°11′56.9″ N–33°33′52.1″ E
G12	*T. monococcum* L.	Kastamonu	41°13′28.7″ N–33°31′58.6″ E
G13	*T. monococcum* L.	Kastamonu	41°13′08.4″ N–33°30′50.6″ E
G14	*T. monococcum* L.	Kastamonu	41°12′44.5″ N–33°30′27.1″ E
G15	*T. monococcum* L.	Kastamonu	41°13′11.9″ N–33°28′58.0″ E
G16	*T. monococcum* L.	Igdir	Igdir University—*Breeding line-1*
G17	*T. monococcum* L.	Igdir	Igdir University—*Breeding line-2*
G18	*T. monococcum* L.	Igdir	Igdir University—*Breeding line-3*
G19	*T. monococcum* L.	Igdir	Igdir University—*Breeding line-4*
G20	*T. monococcum* L.	Igdir	Igdir University—*Breeding line-5*
G21	*T. monococcum* L.	Igdir	Igdir University—*Breeding line-6*
G22	*T. aestivum* L. (Dogankent)	Igdir	Registered variety
G23	*T. durum* L. (Kiziltan-91)	Igdir	Registered variety
G24	*T. durum* L. (C. 1252)	Igdir	Registered variety
G25	*T. durum* L. (Sarıcanak98)	Igdir	Registered variety
G26	*T. durum* L. (Demir-2000)	Igdir	Registered variety

2.2. Climate Properties of the Research Area

The total precipitation for the months of March–August in 2018–2019 was 147.5–89.3 mm. During the period in which the experiment was conducted, precipitation was lower than the multi-year average (165.9 mm). While the multi-year average temperature was 18.66 °C, the growing seasons in which the experiment was conducted had average temperatures of 20.65 °C and 19.8 °C. The average relative humidity index by year was similar to that of the growing seasons (47.15–49.2%) and the multi-year average (48.88%) (Table 2).

2.3. Soil Properties of Research Area

The soil was taken from a depth of 0–20 cm at various points in the sample plot, and had a clay-loam structure. It had a pH of 8.6 (strongly alkaline), 1.20% organic matter and 22.25% lime ($CaCO_3$). It was determined that the amount of nitrogen (N) was 0.06%, the amount of potassium (K_2O) was 851.5 ppm, the amount of phosphorus was 51.5 ppm, and the value of electrical conductivity was 1.37 dS m^{-1} (Table 3).

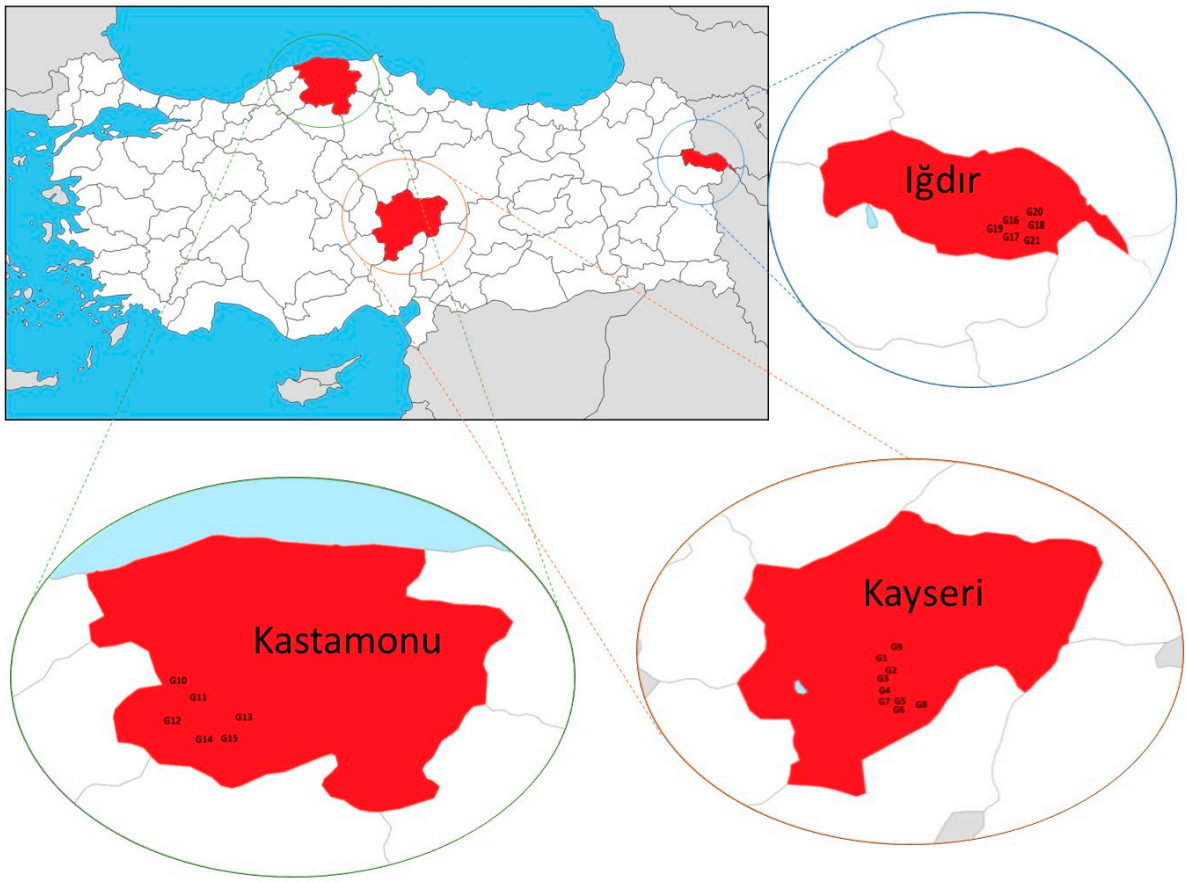


Figure 1. Locations where the wheat used in the study was delivered.

Table 2. Some of the climate data of 2018 and 2019 and the long years average (LYA) *.

Climate Factors	Years	Months						
		March	April	May	June	July	August	Total/Mean
Precipitation (mm)	2017–2018	16.5	18.2	69.3	31.8	5.9	5.8	147.5
	2018–2019	23.5	25.1	25.9	13.6	0.6	0.6	89.3
	LYA	21.9	37.4	49.4	33.1	14.5	9.6	165.9
Average Temperature (°C)	2017–2018	12.3	14.2	18.4	23.4	29.2	26.4	20.65
	2018–2019	6.8	12.1	19.9	25.6	27.3	27.0	19.8
	LYA	6.99	13.4	17.5	22.3	26.2	25.6	18.66
Relative Humidity (%)	2017–2018	45.7	47.7	45.7	46.1	45.1	52.6	47.15
	2018–2019	59.7	56.9	51.2	45.8	40.1	41.2	49.2
	LYA	52.2	49.9	51.5	47.3	45.3	47.1	48.88

* The Igdır region's climate data were sourced from the Turkish State Meteorological Service.

Table 3. Some soil data of the experimental area.

Soil Features	Values
pH	8.6
EC (dS m^{-1})	1.37
$CaCO_3$ (%)	22.25
Total N (%)	0.06
Organic matter (%)	1.2
P_2O_5 (ppm)	51.5
K_2O (ppm)	851.5

2.4. Measured Characteristics of Dependent and Independent Variables

Field trials were conducted in the Agricultural Application and Research field of Igdir University. According to the random block sampling scheme, research materials were planted in plots of 5 m^2 in three replicates in March. Totals of 3 kg da^{-1} nitrogen (N) and 6 kg da^{-1} pure phosphorus (P_2O_5) were applied, and 3 kg da^{-1} nitrogen (N) was applied as a top fertilizer during the tillering period. In obtaining data on the number of grains per spike (GNS), grain yield (GY) (g $plant^{-1}$), biological yield per plant (BY) (g $plant^{-1}$), the number of days to physiological maturity (NMD), the number of days to ripening (NRD), protein rate (PR), 1000 grains weight (1000–GW), plant height (PH), and spike length per plant (SL), the methods defined by Demirel et al. [18], Ahmad et al. [19], and Coşkun et al. [6] were used. The methods specified by Horwitz and Latimer [20] were also used for protein analysis. The mean values of years obtained for all analyzed parameters are given in Table 4. Pearson pair correlations between wheat traits were calculated in R Studio [21] using the "corrplot" package [22]. The obtained data were analyzed using SPSS package software for CHAID analysis. Cluster analysis was performed for wheat genotypes using XLSTAT software [23]. Seventy-eight measurements belonging to 26 genotypes were used as input data. Some influential predictors, i.e., GNS, BY, NMD, NRD, PR, 1000-GW, PH, and SL, were entered into MARS and CHAID to predict GY.

Table 4. Descriptive plant measurement statistics.

Variable	Minimum	Maximum	Mean	Standard Error	*F*-Statistic
GNS [1]	12.8	34.3	22.49	0.596	75.387 *
GY (g $plant^{-1}$)	0.29	0.95	0.57	0.022	53.992 *
BY (g $plant^{-1}$)	1.28	3.54	2.16	0.083	4177.015 *
1000-GW (g)	22.6	57.3	38.7	0.75	12.838 *
NRD	62	86	73.39	0.614	7.975 *
NMD	84	117	105.04	0.809	15.373 *
PR (%)	8.96	21.27	16.27	0.345	113.973 *
PH (cm $plant^{-1}$)	65	81	71.15	0.338	4.747 *
SL (cm $plant^{-1}$)	5.95	7.5	6.57	0.035	5.314 *

*: $p < 0.0001$ significance level. [1] GNS, the number of grains per spike; GY, grain yield (g $plant^{-1}$); BY, biological yield (g $plant^{-1}$), NMD, the number of days to physiological maturity; NRD, the number of days to ripening; PR, protein rate; 1000-GW, 1000 grains weight; PH, plant height; and SL, spike length.

2.5. CHAID and MARS Analysis

The CHAID technique can only by applied to categorized independent variables with nominal or ordinal levels. Therefore, continuous predictors are converted into ordinal predictors. For a known set of breakpoints a_1,a_2, . . . , a_{K-1}, x is mapped to the category $C(x)$.

$$C(x) = \begin{cases} 1 & x \le a_1 \\ k+1 \; a_k < x \le a_{k+1} & k = 1, \ldots, K-2 \\ K & a_{K-1} < x \end{cases} \quad (1)$$

where K is the desired number of bins, and x_i is the frequency weights combined to approximate the breakpoints when computing rankings. In case of a tie, the average rank is used. The list of ranks and their respective values in ascending order are as follows:

$$\left\{ r_{(i)}, x_{(i)} \right\}_{i=1}^{n} \quad (2)$$

For k = 0 to $(K - 1)$, we set:

$$I_k = \left\{ i : \left[r_{(i)} \frac{K}{N_f + 1} \right] = k \right\} \quad (3)$$

MARS is a method capable of generating a robust predictive equation for a response variable. MARS finds applications in estimating yield or other traits using agronomic traits.

Simple and multiple linear regression was used to estimate plant traits. Violating distributional assumptions can have a detrimental effect on the dependence of the underlying approaches [24].

In this study, the MARS method was used to forecast grain production from important independent factors, and its statistical notation is as follows:

$$\hat{y} = \beta_0 + \sum_{m=1}^{M} \beta_m \prod_{k=1}^{K_m} h_{\mathrm{km}}\left(X_{v(k,m)}\right) \tag{4}$$

where $\hat{y}$—the predicted value of grain yield (dependent variable), β_0—the constant (intercept), β_m—the coefficient of the basic functions, h_{km} $(X_{v(k,m)})$—the basis function in which $v(k,m)$ is an index of the independent variable in the mth component of the kth product, and K_m—the parameter that limits the order of interaction. After generating the most complex MARS model in the forward pass, basis functions that reduce model performance were removed via a pruning procedure (backward pass) based on generalized cross-validation error (GCV) [25,26]:

$$GCV(\breve{}) = \frac{\sum_{i=1}^{n}\left(y_i - y_{ip}\right)^2}{\left[1 - \frac{M(\breve{})}{n}\right]^2} \tag{5}$$

where n—sample size, y_i—observed grain yield value, y_{ip}—predicted grain yield value, and $M(\lambda)$—model complexity penalty function involving λ expressions.

The train function in the caret package was used to implement the processes of model construction, evaluation and optimization in MARS predictive modeling [27]. The dataset was split into a training set (70%) and a test set (30%). The goal of the training set was to select the best candidate MARS prediction model developed by combining degrees = 1:3 and nprune = 2:50 as the numbers of selected terms. For this purpose, v-tenfold cross-validation was defined among the resampling strategies. The optimal MARS model from the training dataset was validated using the test dataset. The predictive performance of the ideal MARS model was tested using the following model evaluation criteria [25,26,28,29]:

1. Pearson correlation coefficient (r) between the observed and predicted values;
2. Akaike information criterion (AIC),

$$\begin{cases} AIC = n.ln\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_{ip}\right)^2\right] + 2k, \text{ if } n/k > 40 \\ AIC_c = AIC + \frac{2k(k+1)}{n-k-1} \qquad\qquad otherwise \end{cases} \tag{6}$$

3. Root mean square error ($RMSE$),

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_{ip}\right)^2} \tag{7}$$

4. Mean error (ME),

$$ME = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_{ip}\right) \tag{8}$$

5. Mean absolute deviation (MAD),

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_{ip}\right| \tag{9}$$

6. Standard deviation ratio (SD_{ratio}),

$$SD_{ratio} = \frac{S_m}{S_d} \tag{10}$$

7. Global relative approximation error (*RAE*),

$$RAE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_{ip})^2}{\sum_{i=1}^{n} {y_i}^2}} \quad (11)$$

8. Mean absolute percentage error (*MAPE*),

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y_{ip}}{y_i}\right| * 100 \quad (12)$$

where n—the training sample size in the dataset, k—the number of model parameters, y_i—the actual GY value, y_{ip}—the predicted GY value, s_m—the standard deviation of model errors, and s_d—the standard deviation of GY.

Descriptive statistics of quantitative variables were calculated using the *psych* package in R. The statistical evaluation of the MARS data mining method for wheat GY prediction used the *caret* and *earth* packages in the R environment [30,31]. To evaluate the prediction performance of the optimal MARS model in R, the *ehaGoF* package (version 0.1.1) developed by Eyduran [25] was used.

3. Results

The results of descriptive statistics for the wheat variables are shown in Table 4. Significant coefficients ($p < 0.0001$) were observed for each trait. Based on the data obtained from the study of factors associated with 26 different wheat genotypes, correlation, cluster, MARS, and CHAID analyses were conducted. Depending on the purpose of the study, correlations between grain yield and other variables were examined.

3.1. Determination of Correlations between Variables

Pearson's correlation coefficients between pairs of agronomic traits tested in wheat are shown in Figure 2. Based on the correlation results, there was a strong positive relationship between BY and GY, while NMD, PR, and PH were significantly negatively correlated with GY.

3.2. Determination of the Relationship between Genotypes

In this study, cluster analysis was performed for wheat genotypes using variables (Figure 3). The analysis showed that two main groups (A and B) were formed in the dendrogram. Genotype number G26 (Demir-2000) was in group A. Group B was divided into two subgroups (D and C). Some species of *T. monococcum* L. (G19, G12, G16, G13, G20, and G18) formed clusters in group C. In group D, two subgroups (E and F) were distinguished, while subgroup E included the registered variety (G25—Sarıcanak98, G23—Kiziltan-91, G24——C. 1252) Dogankent (registered variety), and *T. monococcum* L. species (G11 and G10) were included in subgroup F.

Using cluster analysis, we were able to visually combine the variance among wheat genotypes based on observations.

3.3. Comparison of the MARS and CHAID Algorithms

Table 5 presents a summary of the goodness-of-fit criteria that were used in the process of testing and comparing the predictive capabilities of the MARS and CHAID data mining algorithms. It was found that the MARS algorithm has higher predictive accuracy compared to the evaluation criteria of the CART algorithm. Both methods estimated highly significant correlations between the variables they used and the predicted GY values ($p < 0.001$).



Figure 2. Correlation coefficients of the studied properties of wheat. *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$. [1] GNS, the number of grains per spike; GY, grain yield (g plant^{-1}); BY, biological yield (g plant^{-1}), NMD, the number of days to physiological maturity; NRD, the number of days to ripening; PR, protein rate; 1000-GW, 1000 grains weight; PH, plant height; and SL, spike length.

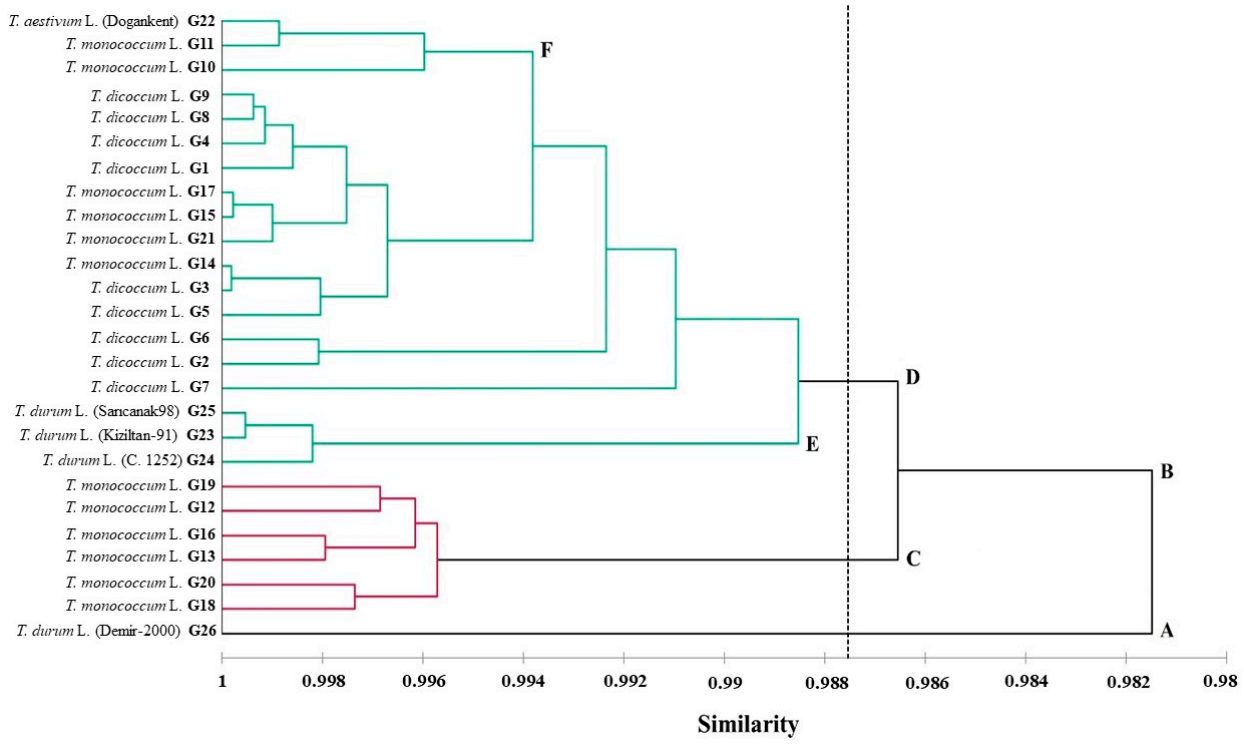


Figure 3. The dendrogram rendered by cluster analysis for wheat genotypes; A and B including main groups, C, D, E, and F comprising subgroups.

Table 5. The goodness of fit criteria for MARS and CHAID algorithms.

Criteria	MARS		CHAID	
	Training SP	Test SP	Training SP	Test SP
r [1]	0.992 ***	0.979 ***	0.968 ***	0.954 ***
R^2	0.985	0.957	0.939	0.89
Adj. R^2	0.985	0.957	0.937	0.88
AIC	−404.773	−155.449	−325.273	−131.099
RMSE (%)	0.024	0.039	0.047	0.06
ME (%)	0.00	0.009	0.001	−0.025
MAD (%)	0.018	0.031	0.038	0.048
SD_{ratio}	0.124	0.203	0.248	0.301
RAE	0.002	0.004	0.006	0.011
MAPE (%)	3.539	6.521	7.258	10.532

*** $p < 0.001$ shows the significance level of the Pearson coefficient between the actual and predicted GY values. (SP: set performance). [1] r: Pearson correlation coefficient. R^2: Coefficient of determination. Adj. R^2: Adjusted coefficient of determination. AIC: Akaike's information criterion. RMSE: Relative root mean square error. ME: Mean error. MAD: Mean absolute deviation. SDratio: Standard deviation ratio. RAE: Relative approximation error. MAPE: Mean absolute percentage error.

The optimum predictive MARS model was found to be the MARS model with degrees equal to 1 and nprune equal to 8 for the training set in this study. The optimal model in the analysis was found to have eight terms, including seven intercept terms with no interaction effect. The best predictive performance of the MARS model applied to the training and test datasets is shown in Table 5.

When the second term and coefficient were examined, genotypes with a BY lower than 2.86 g were considered, and GY was expected to decrease (Table 6). For example, for genotypes with BY = 2 g, when the second term and coefficient (−0.3612226) were considered, the calculation was as follows: $-0.36122260 \times \max(0, 2.86 - 2) = -0.36122260 \times 0.86 = -0.310651436$. When the third term and coefficient (+0.00933546) were examined, considering genotypes with NRD values greater than 77 g, an increase in GL could be expected. When fourth term and coefficient were considered, a decrease in GY could be expected for genotypes with 1000-GW greater than 41.8 g. An increase in GY was expected when genotypes with a PH greater than 70.5 cm were considered (sixth term). A decrease in GY would be expected when genotypes whose SL was higher than 6.65 (eighth term) were considered (Table 6).

Table 6. Results of the MARS algorithm for grain yield trait.

Terms	Model	Coefficients
1	Intercept	+0.841
2	max (0, 2.86-BY [1])	−0.361
3	max (0, NRD-77)	+0.009
4	max (0, 1000-GW-41.8)	−0.006
5	max (0, 70.5-PH)	−0.032
6	max (0, PH-70.5)	+0.032
7	max (0, 6.65-SL)	+0.137
8	max (0, SL-6.65)	−0.291

[1] BY, biological yield (g plant^{-1}); NRD, the number of days to ripening; 1000-GW, 1000 grains weight; PH, plant height; and SL, spike length.

Supplementary Figure S1 displays the tree diagram produced by the CHAID algorithm. The minimum ratio of parent to child size was determined to be 8:4. The SD coefficient value (0.248) for the CHAID algorithm was less than 0.40 and represented a good fit. The CHAID algorithm in wheat GY prediction showed that GY scores were significantly correlated ($r = 0.968$, $p < 0.001$). The CHAID regression tree shows that BY was the most influential predictor of GY (adjusted $p = 0.000$, $F = 95.891$, df1 = 5, and df2 = 48). At the top of the diagram, the root node (Node 0) containing all the wheat in the study generated a

mean GY of 0.583 g. The root node was divided into six smaller subgroups (Nodes 1–6) by BY as a good predictor. The mean GY from Node 1 to Node 6 increased as a result of increasing BY (Supplementary Figure S1). Node 6 was a subgroup of wheat with a BY of 2.54 g or heavier, and the sixth group was predicted to have an average GY of 0.805 g. Node 6 was examined, and the effect of NMD on BY in wheat was the most influential predictor (adjusted $p = 0.004$, $F = 16.431$, df1 = 2, and df2 = 13). The CHAID regression tree diagram shows that when the NMD was 104 days or more for wheat, the weight of BY increased (Supplementary Figure S1). As a result, the optimal CHAID structure showed that wheat grain yield at NMD > 104 days and BY > 2.54 g plant^{-1} increased.

4. Discussion

Landraces are considered important genetic sources for the development of new cultivars due to their exceptional resistance to conditions featuring both biotic and abiotic stress factors [32,33]. In studies, landraces have been shown to be very useful sources for breeding due to their considerable variation among populations [34,35]. In the present study, the range of variability remained high. This is evidenced by the large difference between the lowest and highest values of the observed variables for the wheat plant studied (Table 4). Significant coefficients ($p < 0.0001$) were found for all variables.

Correlation analysis is used to measure the relationship between pairs of different agronomic traits, and the information obtained can be used to determine agronomic traits positively related to crop yield as part of the intermediate selection criteria to improve crop yield for breeding purposes [36,37]. The results show that BY has a significant positive correlation with grain yield, but NMD, PR and PH are significantly negatively correlated with grain yield (Figure 2). Aydogan and Soylu [38] reported significant negative correlations between GY and PR. Polat et al. [8] reported a significant positive relationship between PH and SL. Çığ and Karaman [39] found a significant positive correlation between BY and GY. İpeksever and Özberk [40] showed a significant positive correlation between NMD and NRD. Yağmur and Kaydan [41] obtained a significant positive correlation between PH, SL, and GNS and SL. Kara and Akman [42] found a significant negative correlation between 1000-GW and GNS. In addition, they noted a significant positive correlation between GNS and SL. Güngör and Dumlupinar [43] showed a significant positive correlation between SL and GNS. The correlation results in the present study are consistent with these findings.

Wheat yield prediction is based on a number of factors, including crop area, production, rainfall, genotype and climate conditions, among others [44]. Understanding how one or more of these components work can help produce more precise estimates. In this context, the correct determination of grain yield estimate is strongly linked to the use of effective statistical methods, particularly data mining algorithms such as MARS and CHAID [44,45]. This is because these approaches are closely related to the use of effective statistical techniques. It is also important that the statistical techniques used to predict GY based on specific traits, which can vary even between species, have a high level of accuracy.

Using MARS and CHAID algorithms, various GY prediction models can be developed for this study. Within this context, the performance of the models is evaluated according to goodness-of-fit criteria [39]. Table 6 shows the presentation of the model that resulted from the MARS analysis. The SD coefficient value (0.203) for the MARS algorithm was less than 0.40 and represented a good fit. Grzesiak and Zaborski [28] stated that a good model fit was satisfactory if the SD coefficient was less than 0.40, and Eyduran et al. [24] confirmed this. The MARS model explained 95.7% of the total variance (Table 5). Several previous studies [9–11,24] are available on the application of the MARS algorithm in agricultural science. The better performance of MARS modeling studies has been reported in describing the relationship between different agronomic traits in soybean varieties for plant breeding purposes [12]. However, there is a lack of information on the prediction of GY based on several agronomic traits in wheat. Therefore, further studies on GY prediction may become more importance. These results will allow plant breeders to gain valuable insights related to wheat breeding programs.

Regression trees and CHAID analysis methods are easy to understand visually because their assumptions are smaller than those of other statistical methods [17]. The SD coefficient value (0.301) for the CHAID algorithm was less than 0.40 and represented a good fit. Eyduran et al. [24] reported that a good model fit was considered sufficient if the SD coefficient was less than 0.40. The CHAID model was able to account for 89% of the overall variability in the data. No literature data found significant predictors affecting the GY of wheat plantings that were included in the CHAID algorithm. In contrast, Celik et al. [45] successfully determined algorithms to analyze the influence of yield factors, oil production rate and plant height in sunflowers using various models, including CHAID.

A high correlation was found between biological yield and grain yield (Figure 2). Therefore, similar to the correlation result, biological yield was included in the model as the most important factor in the CHAID algorithm (Supplementary Figure S1). In addition, the second term in the MARS algorithm was also included in the model (Table 6). Ahmad et al. [19] also reported that the correlation coefficient and direct effect of biological yield on grain yield are the highest. It was reported that the effect of biological yield on grain yield is significant even in drought-stressed wheat [31,46]. In addition, the results of the study are consistent with studies on durum wheat [47] and bread wheat [48,49].

5. Conclusions

In this study, CHAID and MARS analyses were conducted on wheat grain yield for the first time. A significant positive correlation was found between grain yield and biological yield. In favor of both models, based on the relevant criteria, is the very high performance of the optimal models in the training harvest, which turned out to be almost the same as the performance found in the test harvest. This means that there is no problem of overfitting, and the best MARS and CHAID models could be used to predict grain yield (GY) in wheat, since they can generalize well. In addition, the MARS algorithm performed significantly better than CHAID in predicting grain yield, and was able to explain 95.7% of the variance in grain production across wheat varieties. Although this study was conducted under microclimatic conditions, the results indicate that it was consistent with studies conducted on other wheat species and under varied climatic conditions. It has been shown that the MARS and CHAID algorithms can allow a more precise description of complex interaction effects for wheat breeders. We suggest that the adoption of MARS and CHAID algorithms will benefit breeding research.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/agronomy13061438/s1, Figure S1. The decision tree diagram obtained by CHAID algorithm for grain yield (GY).

Author Contributions: Conceptualization, F.D., B.E. and A.Y. methodology, F.D. and B.E.; software, F.D.; validation, F.D., B.E. and A.Y.; formal analysis, F.D. and B.E.; investigation, F.D. and B.E.; resources, F.D. and B.E.; data curation, F.D., B.E. and A.Y.; writing—original draft preparation, F.D. and B.E.; writing—review and editing, A.T., K.H., G.N., B.J., A.P.-A., H.B. and J.B.; visualization, F.D., B.E., A.Y., A.T. and K.H.; funding acquisition, A.P.-A., J.B. and K.H., supervision, A.T., K.H., G.N., H.B., B.J., A.P.-A., J.B. and K.N.; project administration, G.N., H.B., A.P.-A., J.B. and K.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Demİrel, F.; Yildirim, B. Investigation of Some Gernik Wheat (*Triticum dicoccum* L.) Genotypes by Agromorphological Characteristics and Biplot Analysis Method. *J. Agric.* **2020**, *3*, 49–56.
2. Özkan, H.; Willcox, G.; Graner, A.; Salamini, F.; Kilian, B. Geographic distribution and domestication of wild emmer wheat (Triticum dicoccoides). *Genet. Resour. Crop Evol.* **2011**, *58*, 11–53. [CrossRef]
3. Demİrel, F.; Barış, E. Production Projection of Einkorn and Emmer Wheat Cultivated in Turkey. *J. Agric.* **2020**, *3*, 1–5. [CrossRef]

4. Lev-Yadun, S.; Gopher, A.; Abbo, S. The cradle of agriculture. *Science* **2000**, *288*, 1602–1603. [CrossRef] [PubMed]
5. Rauf, S.; da Silva, J.T.; Khan, A.A.; Naveed, A. Consequences of plant breeding on genetic diversity. *Int. J. Plant Breed.* **2010**, *4*, 1–21.
6. Coşkun, İ.; Tekin, M.; Akar, T. Characterization of Turkish diploid and tetraploid hulled wheat lines for some agromorphological traits. *Int. J. Agri. Wildlife Sci.* **2019**, *5*, 322–334.
7. Gollin, D.; Smale, M.; Skovmand, B. Searching an ex situ collection of wheat genetic resources. *Am. J. Agric. Econ.* **2000**, *82*, 812–827. [CrossRef]
8. Polat, P.; Çİfcİ, E.; Yağdı, K. Determination of relationships between grain yield and some yield components in bread wheat (*Triticum aestivum* L.). *Tarim Bilim. Derg.* **2015**, *21*, 355–362.
9. Srivastava, A.K.; Safaei, N.; Khaki, S.; Lopez, G.; Zeng, W.; Ewert, F.; Gaiser, T.; Rahimi, J. Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Sci. Rep.* **2022**, *12*, 3215. [CrossRef] [PubMed]
10. Celik, S.; Eyduran, E.; Tatliyer, A.; Karadas, K.; Kara, M.K.; Waheed, A. Comparing predictive performances of some nonlinear functions and Multivariate Adaptive Regression Splines (MARS) for describing the growth of Daera Din Panah (DDP) goat in Pakistan. *Pak. J. Zool.* **2018**, *50*, 1–4. [CrossRef]
11. Aytekin, İ.; Eyduran, E.; Karadas, K.; Akşahan, R.; Keskin, İ. Prediction of Fattening Final Live Weight from some Body Measurements and Fattening Period in Young Bulls of Crossbred and Exotic Breeds using MARS Data Mining Algorithm. *Pak. J. Zool.* **2018**, *50*, 189–195. [CrossRef]
12. Celik, S.; Boydak, E. Description of the relationships between different plant characteristics in soybean using multivariate adaptive regression splines (MARS) algorithm. *J. Anim. Plant Sci.* **2020**, *30*, 431–441.
13. Akçura, M. The relationships of some traits in Turkish winter bread wheat landraces. *Turk. J. Agric. For.* **2011**, *35*, 115–125. [CrossRef]
14. Janmohammadi, M.; Sabaghnia, N.; Nouraein, M. Path analysis of grain yield and yield components and some agronomic traits in bread wheat. *Acta Univ. Agric. Silvic. Mendel. Brun.* **2014**, *62*, 945–952. [CrossRef]
15. Kayri, M.; Boysan, M. Assesment of relation between cognitive vulnerability and depression's level by using classification and regression tree analysis. *Hacettepe Uni. J. Educ.* **2008**, *34*, 168–177.
16. Mirhashemi, S.H.; Panahi, M. Investigation and prediction of maize water requirements in four growth stages under the influence of natural factors (Case study: Qazvin plain, Iran). *Environ. Technol. Innov.* **2021**, *24*, 102062. [CrossRef]
17. Olfaz, M.; Tirink, C.; Önder, H. Use of CART and CHAID algorithms in Karayaka sheep breeding. *J. Facul. Vet. Med. Kafkas Uni.* **2019**, *25*, 105–110.
18. Demirel, F.; Gurcan, K.; Akar, T. Clustering analysis of morphological and phenological data in einkorn and emmer wheats collected from Kastamonu region. *Int. J. Sci. Technol. Res.* **2019**, *5*, 25–36.
19. Ahmad, T.; Kumar, A.; Pandey, D.; Prasad, B. Correlation and path coefficient analysis for yield and its attributing traits in bread wheat (*Triticum aestivum* L. em Thell). *J. Appl. Nat. Sci.* **2018**, *10*, 1078–1084. [CrossRef]
20. Horwitz, W. *Official Methods of Analysis of AOAC International. Volume I, Agricultural Chemicals, Contaminants, Drugs/Edited by William Horwitz*; AOAC International: Gaithersburg, MD, USA, 2010.
21. Allaire, J. RStudio: Integrated development environment for R. *Boston MA* **2012**, *770*, 165–171.
22. Wei, T.; Simko, V.; Levy, M.; Xie, Y.; Jin, Y.; Zemla, J. Package 'corrplot'. *Statistician* **2017**, *56*, e24.
23. Addinsoft, X. *Data Analysis and Statistics Software for Microsoft Excel*; Addinsoft: Paris, France, 2015.
24. Eyduran, E.; Akin, M.; Eyduran, S. *Application of Multivariate Adaptive Regression Splines through R Software*; Nobel Academic Publishing: Ankara, Turkey, 2019.
25. Eyduran, E. ehaGoF: Calculates Goodness of Fit Statistics, R Package Version 0.1.1; 2020. Available online: https://CRAN.R-project.org/package=ehaGoF (accessed on 10 February 2023).
26. Zaborski, D.; Ali, M.; Eyduran, E.; Grzesiak, W.; Tariq, M.M.; Abbas, F.; Waheed, A.; Tirink, C. Prediction of selected reproductive traits of indigenous Harnai sheep under the farm management system via various data mining algorithms. *Pak. J. Zool.* **2019**, *51*, 421. [CrossRef]
27. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 26.
28. Grzesiak, W.; Zaborski, D. Examples of the use of data mining methods in animal breeding. In *Data Mining Applications in Engineering and Medicine*; IntechOpen: London, UK, 2012; pp. 303–324.
29. Tyasi, T.L.; Eyduran, E.; Celik, S. Comparison of tree-based regression tree methods for predicting live body weight from morphological traits in Hy-line silver brown commercial layer and indigenous Potchefstroom Koekoek breeds raised in South Africa. *Trop. Anim. Health Prod.* **2021**, *53*, 1–8. [CrossRef] [PubMed]
30. Kuhn, M.; Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T.; Mayer, Z.; Kenkel, B.; Benesty, M. Caret: Classification and Regression Training: R Package, R package version 6.0-77; 2019. Available online: https://CRAN.R-project.org/package=caret (accessed on 10 February 2023).
31. Milborrow, S. Derived from mda: Mars by Trevor Hastie and Rob Tibshirani. *Uses Alan Mill. Fortran*. 2019. Available online: https://CRAN.R-project.org/package=earth (accessed on 10 February 2023).
32. Maccaferri, M.; Harris, N.S.; Twardziok, S.O.; Pasam, R.K.; Gundlach, H.; Spannagl, M.; Ormanbekova, D.; Lux, T.; Prade, V.M.; Milner, S.G. Durum wheat genome highlights past domestication signatures and future improvement targets. *Nat. Genet.* **2019**, *51*, 885–895. [CrossRef]

33. Soriano, J.M.; Villegas, D.; Sorrells, M.E.; Royo, C. Durum wheat landraces from east and west regions of the mediterranean basin are genetically distinct for yield components and phenology. *Front. Plant Sci.* **2018**, *9*, 80. [CrossRef]
34. Aoun, M.; Kolmer, J.A.; Rouse, M.N.; Elias, E.M.; Breiland, M.; Bulbula, W.D.; Chao, S.; Acevedo, M. Mapping of novel leaf rust and stem rust resistance genes in the Portuguese durum wheat landrace PI 192051. *G3 Genes Genomes Genet.* **2019**, *9*, 2535–2547. [CrossRef]
35. Chacón, E.A.; Vázquez, F.J.; Giraldo, P.; Carrillo, J.M.; Benavente, E.; Rodríguez-Quijano, M. Allelic variation for prolamins in Spanish durum wheat landraces and its relationship with quality traits. *Agronomy* **2020**, *10*, 136. [CrossRef]
36. Karaköy, T.; Baloch, F.S.; Toklu, F.; Özkan, H. Variation for selected morphological and quality-related traits among 178 faba bean landraces collected from Turkey. *Plant Genet. Resour.* **2014**, *12*, 5–13. [CrossRef]
37. Kumar, R.; Bhushan, B.; Pal, R.; Gaurav, S. Correlation and path coefficient analysis for quantitative traits in wheat (*Triticum aestivum* L.) under normal condition. *Ann. Agri-Bio Res.* **2014**, *19*, 447–450.
38. Aydoğan, S.; Soylu, S. Determination of yield, yield components and some quality properties of bread wheat varieties. *J. Cent. Res. Inst. Field Crops* **2017**, *26*, 24–30.
39. Çığ, F.; Karaman, M. Evaluation of Durum Wheat (*Triticum durum* Desf.) Genotypes originated from southeast Anatolia region for some agricultural character. *Turk. J. Agric. Res.* **2019**, *6*, 10–19.
40. İpeksever, F.; Özberk, İ. Investigation of yield and quality characteristics and economic returns of some plain and mixed bread wheat (*Triticum aestivum* L.) cultivars. *J. Cent. Res. Inst. Field Crops* **2019**, *28*, 80–91.
41. Yağmur, M.K.D. Relationship between grain yield, yield components and phenological periods in winter wheat. *Harran J. Agric. Food Sci.* **2008**, *12*, 9–18.
42. Kara, B.; Akman, Z. Trait relationships and path analysis in local wheat ecotypes. *SDU J. Nat. Appl. Sci.* **2007**, *11*, 219–224.
43. Güngör, H.; Dumlupinar, Z. Correlation and path analysis in terms of some agricultural characteristics in bread wheat (*Triticum aestivum* L.) cultivars. *J. Agric. Nat.* **2019**, *22*, 851–858. [CrossRef]
44. Nayana, B.; Kumar, K.R.; Chesneau, C. Wheat Yield Prediction in India Using Principal Component Analysis-Multivariate Adaptive Regression Splines (PCA-MARS). *AgriEngineering* **2022**, *4*, 461–474. [CrossRef]
45. Celik, S.; Boydak, E.; Firat, R. An analysis of factors affecting yield, oil production rate and plant height in sunflowers using selected data mining algorithms. *JAPS J. Anim. Plant Sci.* **2018**, *28*, 1085–1093.
46. Cyplik, A.; Czyczyło-Mysza, I.M.; Jankowicz-Cieslak, J.; Bocianowski, J. QTL×QTL×QTL Interaction Effects for Total Phenolic Content of Wheat Mapping Population of CSDH Lines under Drought Stress by Weighted Multiple Linear Regression. *Agriculture* **2023**, *13*, 850. [CrossRef]
47. Tsegaye, D.; Dessalegn, T.; Dessalegn, Y.; Share, G. Genetic variability, correlation and path analysis in durum wheat germplasm (*Triticum durum* Desf). *Agric. Res. Rev.* **2012**, *1*, 107–112.
48. Sabit, Z.; Yadav, B.; Rai, P. Genetic variability, correlation and path analysis for yield and its components in f5 generation of bread wheat (*Triticum aestivum* L.). *J. Pharmacogn. Phytochem.* **2017**, *6*, 680–687.
49. Jędzura, S.; Bocianowski, J.; Matysik, P. The AMMI model application to analyze the genotype–environmental interaction of spring wheat grain yield for the breeding program purposes. *Cereal Res. Commun.* **2023**, *51*, 197–205. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.