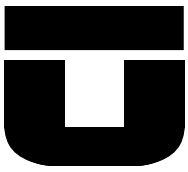


UNIVERSIDAD DE TALCA
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA CIVIL EN COMPUTACIÓN

Modelo Predictivo de los Rasgos Fenotípicos del Trigo: Una Metodología Sistemática

MARIANELA ANDREA ITURRIAGA JIMÉNEZ

Profesor Guía: CÉSAR ASTUDILLO

Memoria para optar al título de
Ingeniero Civil en Computación

Curicó – Chile
Agosto, 2017

CONSTANCIA

La Dirección del Sistema de Bibliotecas a través de su encargado Biblioteca Campus Curicó certifica que el autor del siguiente trabajo de titulación ha firmado su autorización para la reproducción en forma total o parcial e ilimitada del mismo.

Curicó, 2019

Dedicado a mi familia

AGRADECIMIENTOS

Quiero agradecer a mi familia por el apoyo entregado durante todo mi paso por la universidad, pero en particular por el apoyo y paciencia durante los meses que trabaje en este proyecto. Destacar la paciencia que mi abuelita tuvo conmigo y mis compañeros en nuestras interminables noches de estudio. A mi madre y hermana por estar ahí y hacer la típica pregunta "y, ¿Cuando terminas?" que tanto molesta, pero que ayuda a dar termino a este bella etapa de mi vida, este título que en unos meses estará en nuestras manos es para ustedes.

También agradecer a mis amigos y compañeros, Maha, Anita y Karen, por soportarme y apoyarme pero por sobre todo por dejarme ser su amiga y compartir tantos momentos inolvidables y porque sé que seguiremos forjando y fortaleciendo nuestra amistad.

Finalmente agradecer a los profesores de la carrera que fueron parte fundamental de la ingeniera que hoy soy. Gracias a la profe Ruth por darme el último empujoncito para cambiarme de carrera y a la Marce por ayudarme con tanto trámite que tuve que hacer, porque por dios que fue complejo mi paso por esta carrera!, obvio no podía hacerla fácil. Y finalmente, agradecer a los profesores Rodolfo, Rapa y profe Meza quienes estuvieron molestando al final del camino, unos más que otros, para que entregará la memoria, o como diría un (cabeza) grande, "termina la weá!!".

Este trabajo ha sido parcialmente financiado a través del proyecto FONDECYT 11121350 "Tree-Based Pattern Recognition".

TABLA DE CONTENIDOS

página

Dedicatoria I

Agradecimientos II

Tabla de Contenidos III

Índice de Figuras VI

Índice de Tablas VIII

Resumen IX

Abstract X

1. Introducción 1

1.1. Descripción de la Propuesta 1

1.1.1. Contexto del Proyecto 2

1.1.2. Trabajo Relacionado 3

1.1.3. Definición del Problema 3

1.1.4. Propuesta de Solución 4

1.2. Hipótesis 6

1.3. Objetivos 6

1.4. Alcances 7

2. Marco Teórico 8

2.1. Mejoramiento Genético del Trigo 8

2.2. Reflectancia Espectral 9

2.3. Análisis de Regresión 11

2.3.1. ¿Qué es regresión? 12

2.3.2. Multiple Linear Regression 13

2.3.3. Partial Least Squares Regression 14

2.3.4. Ridge Regression 15

2.3.5. Support Vector Regression 16

2.4. Otros Algoritmos Utilizados 19
2.4.1. Principal Component Analysis (PCA) 19
2.4.2. Local Outlier Factor (LOF) 21
2.5. Medidas de Rendimiento 22

3. Descripción de la Librería (API) 25
3.1. Subproblemas en Minería de Datos 25
3.1.1. Valores Faltantes 25
3.1.2. Normalización 26
3.2. Detección de Ruido 27
3.3. Gráficos de Diagnóstico 31
3.4. Otros Gráficos 34
3.4.1. Gráfico del Codo: Elbow Plot 34
3.4.2. Gráfico de Dispersión 35
3.4.3. Gráfico de Dispersión Matricial 37
3.4.4. Gráfico 3D 38
3.4.5. Gráfico de Densidad 39
3.4.6. Ridge Plot 41

4. Metodología Experimental 44
4.1. Adquisición de Datos 45
4.2. Variables de Estudio 46
4.3. Pre-procesamiento 47
4.3.1. Datos Faltantes 48
4.3.2. Detección de Ruido 48
4.3.3. Detección de Datos Atípicos 50
4.3.4. Normalización de Datos 52
4.3.5. Reducción de la Dimensionalidad 53
4.4. Modelado Predictivo 53
4.5. Validación del Modelo 56

5. Resultados 58

6. Conclusiones 68

Bibliografía **71**

Anexos

A: Resultados 2011 y 2012 **76**

ÍNDICE DE FIGURAS

página

2.1. Ejemplo de reflectancia en una hoja 10
2.2. Rangos del espectro y características medibles en las plantas 11
2.3. Plano que se ajusta a los datos cuando $\mathcal{X} \in \mathbf{R^2}$ 13
2.4. Hiperplano que clasifica correctamente y maximiza el margen 16
2.5. Ejemplo de una transformación del espacio de entrada 17
2.6. SVR que muestra la holgura de la función de pérdida ϵ 19
2.7. Componentes principales para un distribución de ejemplo 20
2.8. Componentes principales para un distribución de ejemplo 21

3.1. Ejemplo de la firma espectral de un subconjunto de los datos 28
3.2. Gráficos de influencia, leverange y residuos para el conjunto de datos *iris* 32
3.3. Gráfico del codo de los componentes principales del conjunto de datos *cars* 35
3.4. Gráfico de dispersión de los 2 primeros componentes principales del conjunto de datos *cars* 36
3.5. Gráfico de dispersión matricial de los primeros 4 componentes principales del conjunto de datos *cars* 38
3.6. Gráfico de densidad para los puntajes resultantes de LOF para el conjunto de datos *cars* 40
3.7. Gráfico de ejemplo de optimización del parámetro λ 42

4.1. Arquitectura del modelado predictivo 45
4.2. Zonas comunes con ruido en los datos de reflectancia espectral 49
4.3. Gráfico de densidad con las puntuaciones asignadas por el algoritmo LOF 52
4.4. Gráfico del codo de los componentes principales de un conjunto de datos 54
4.5. Gráfico de optimización del parámetro λ 55
4.6. Ejemplificación del funcionamiento de 10 veces validación cruzada . . 56

5.1. Datos predichos vs observados después de la validación cruzada para rendimiento 65

5.2. Datos predichos vs observados después de la validación cruzada para IAF 66
5.3. Datos predichos vs observados después de la validación cruzada para $\Delta 13C$ 67

ÍNDICE DE TABLAS

página

4.1. Conjuntos de datos de estudio 46

5.1. Media y desviación estándar del estadístico R^2 para los cuatro modelos con datos del 2011 59
5.2. Media y desviación estándar del estadístico R^2 para los cuatro modelos con datos del 2012 60
5.3. Detalle de los resultados obtenidos para los 3 rasgos con mejores rendimiento durante el 2011 62
5.4. Detalle de los resultados obtenidos para los 3 rasgos con mejores rendimiento durante el 2012 63

A.1. Resultados estadísticos de la predicción de los rasgos evaluados en los 4 ambientes, para los 2 estados fenológicos medidos durante el 2011 para los 4 modelos de regresión evaluados 76
A.2. Resultados estadísticos de la predicción de los rasgos evaluados en los 4 ambientes, para los 2 estados fenológicos medidos durante el 2012 para los 4 modelos de regresión evaluados 81

RESUMEN

El cambio climático y la creciente población mundial plantean la necesidad de adaptar cultivos a diferentes condiciones. Hoy en día, los científicos y las empresas se enfrentan a un reto importante en el aumento de la productividad del suelo con el fin de obtener más alimentos. Uno de los cereales más utilizados es el trigo debido a sus propiedades nutricionales deseables. Científicos en Chile están llevando a cabo experimentos con diferentes variaciones del gen del trigo para encontrar uno que sea más rico y nutritivo. El proceso tradicional consiste en cultivar diferentes tipos de trigo y observar a través del tiempo su rendimiento. Sin embargo, esta tarea es costosa, ya que requiere mano de obra especializada y equipos costosos. El objetivo de esta investigación es estimar analíticamente el rendimiento del trigo sobre la base de información estadística. Esto se logra construyendo un modelo de regresión que sea capaz de predecir la capacidad de producción basada en mediciones de la reflectancia espectral de los individuos.

Diseñamos un modelo de regresión basado en datos de reflectancia espectral para predecir el rendimiento del trigo.

Como parte de la metodología, procesamos previamente los datos, eliminando las instancias con datos incompletos, encontrando valores atípicos, reduciendo la dimensionalidad de los datos y construyendo los regresores. Esta metodología se explica de la manera más genérica posible, con la esperanza de que sea un recurso útil para futuros investigadores que quieran reproducir nuestro método para estimar el rendimiento del trigo o cualquier otro tipo de planta.

ABSTRACT

Climate change and the increasing world population bring about the need of adapting crops to different conditions. Nowadays, scientists and companies face a major challenge in increasing the productivity of the soil so as to obtain more food. One of the most widely used cereals is wheat, due to its desirable nutrition properties. Scientist in Chile are carrying out experiments using different variation of the wheat gene, in order to find a strain that is wealthier and more nutritious. The traditional process consists in growing different types of wheat and observe its performance over a lengthy time span. This task is costly, since it necessitates specialized labor and expensive equipment. The goal of this research is to analytically estimate the performance of wheat based on statistical information. This is achieved by constructing a regression model that is able to predict the production capacity based on measurements of the spectral reflectance of the individuals.

We propose a regression model based on spectral reflectance data which allows to predict the performance of the wheat.

As part of the methodology, we pre-process the data, removing incomplete data, finding outliers, reducing the dimensionality of the data, and building the regressors. This methodology is explained as generic as possible, in hopes that it will be used as a useful resource for future researchers that want to reproduce our method for estimating the performance of wheat or any other type of plant.

1. Introducción

1.1. Descripción de la Propuesta

El cambio climático y la creciente población mundial plantean una demanda especial, que es adaptar los cultivos a diferentes condiciones [13], [31]. Hoy en día, científicos y empresas se enfrentan a un reto importante en el aumento de la productividad del suelo a fin de obtener más alimentos. Uno de los cereales más ampliamente utilizado es el trigo, debido a sus propiedades nutricionales deseables. Científicos en Chile están realizando experimentos con diferentes variaciones del gen con el fin de encontrar uno que sea más rico y nutritivo. El proceso tradicional consiste en el cultivo de diferentes tipos de trigo y observar a través del tiempo su rendimiento utilizando métodos destructivos. Esta tarea es costosa, ya que requiere mano de obra especializada y equipos de alto valor. Es parte de este proyecto estimar analíticamente el desempeño del trigo con base en información estadística. Esto se logra mediante la construcción de un modelo de regresión que es capaz de predecir el desempeño basado en las mediciones de la reflectancia espectral de las observaciones.

Como parte de la metodología, se realizará un pre-procesamiento del subconjunto de datos de entrada, removiendo los datos faltantes, eliminando el ruido, buscando observaciones inusuales, reduciendo la dimensionalidad de ellos y aplicando diferentes modelos de regresión con el afán de encontrar aquel que mejor estime cada una de las variables de estudio e identificando las ventajas y desventajas de cada uno de ellos.

La metodología aplicada se explica de forma genérica, y por lo tanto, no está limitado por el equipo específico utilizado para medir las características particulares de la cosecha. El espíritu de este proyecto, es presentar un método genérico que

puede ser adaptado por otros investigadores cuyo objetivo sea predecir una variable dependiente cualquiera, basado en la reflectancia espectral de las observaciones.

1.1.1. Contexto del Proyecto

Ante los cambios climáticos en el mundo, la adaptación de los cultivos es un factor clave en la producción de algunos alimentos [21], ya que podría generar impactos negativos para grandes poblaciones que estarían expuestas a una escases de alimentos. Por tal motivo, es de preocupación que aquellos alimentos mayormente consumidos por la población sean capaces de soportar y/o adaptarse a cambios climáticos extremos de tal forma que no se vea afectada drásticamente su producción.

Según Lobell [21], el trigo y otros alimentos constituyen una parte importante de la alimentación de muchas personas en el mundo, sobre todo en personas con escasez de recursos, donde pasa a contribuir una parte importante de las calorías que en la actualidad consumen.

En Chile y otros países, la producción del trigo ha aumentado durante los últimos siglos [6], [10] lo que ha llamado la atención de científicos y agrónomos particularmente a preguntarse qué tiene en particular el trigo en estas regiones que incluso ha sido capaz de soportar los cambios climáticos de sequías severas y precipitaciones intensas [24].

Ante este escenario, agrónomos de la Universidad de Talca en conjunto con el Instituto de Investigaciones Agropecuarias CRI-Quilamapu han estado trabajando en el mejoramiento genético del trigo con una mayor adaptabilidad a condiciones ambientales cambiantes y con una alta capacidad de producción. Este trabajo, se ha centrado en el sembrado de diversos genotipos que permitirán realizar el estudio para determinar aquel genotipo que destaque en cuanto a rendimiento y diversas variables fisiológicas y morfológicas, ante diversos eventos climáticos y condiciones ambientales extremas.

Actualmente, es posible calcular el desempeño de una parcela de trigo utilizando métodos destructivos que son costosos y requieren de bastante tiempo, personal y herramientas especializadas. Estos métodos generalmente necesitan esperar hasta el período de cosecha y destruir el cultivo para calcular el desempeño de éste.

Por lo anterior, utilizar un método alternativo y no destructivo que permita estimar el desempeño de una parcela de trigo comienza a ser necesario y relevante

para hacer factible el desarrollo de más y nuevas investigaciones en el área.

1.1.2. Trabajo Relacionado

Los autores de [11], [16] y [22] han propuesto utilizar métodos no destructivos para estimar la capacidad de producción mediante la reflectancia espectral de las observaciones. Este método además resulta fácil de implementar y permite obtener información del cultivo antes del período de cosecha.

Según el autor de [2], la reflectancia espectral proporciona información acerca de las propiedades ópticas de muchos materiales. Esto se debe a que los materiales absorben, reflejan o transmiten la energía o radiación incidente que llega a la superficie. Mediante un espectrorradiómetro es posible medir la distribución de potencia espectral de un material cualquiera el que puede ser graficado como la curva de la reflectancia en función de la longitud de onda.

En el área de investigación relacionada al trigo, existen varios trabajos similares utilizando este mecanismo [11], [16] y [22] en los que mayormente se utilizan algoritmos de la familia de modelos de regresión lineales como regresión de mínimos cuadrados parciales (PLS, del inglés *Partial Least Square Regression*) y *Regresión Ridge*, ocupando diferentes regiones de longitudes de onda para el análisis. Estos algoritmos pertenecen a la misma familia de modelos de regresión y por tanto, el trabajo en esta área específica aún no ha sido completamente explotada.

Este tipo de investigaciones, requieren de una metodología sistemática que nos permita reportar de forma clara los hallazgos científicos encontrados para que cualquier lector pueda replicar el trabajo realizado. En particular, en minería de datos, existen ciertos pasos de la metodología experimental que se suelen obviar por profesionales no expertos en el área. Esta metodología es un plan para detallar todo el proceso de la confección del experimento y evitar la omisión de pasos que puedan afectar los resultados de este.

1.1.3. Definición del Problema

Nuestro problema es la construcción de diferentes modelos capaces de predecir cada una de las variables de estudio de una parcela de trigo sobre la base de mediciones de la reflectancia espectral de las observaciones. Esto se obtendrá modificando la variable de respuesta a predecir y aplicando diversos modelos de regresión pertene-

cientes a diferentes familias con el objeto de abarcar nuevos modelos que actualmente no hayan sido utilizados en este problema.

Como parte del proyecto, se detallará una metodología sistemática con el espíritu de que pueda ser replicada o adaptada por otros investigadores cuyo objetivo sea predecir una variable dependiente cualquiera, basado en la reflectancia espectral de las observaciones.

Este proyecto contribuirá a las investigaciones futuras de agrónomos, permitiéndoles realizar análisis de datos en un menor tiempo y utilizando un modelo de regresión con alto nivel de certeza y validez estadística. Este mecanismo resulta interesante, pues es un método no destructivo para estimar el desempeño del trigo y otras variables morfo-fisiológicas de interés para el estudio.

1.1.4. Propuesta de Solución

El problema planteado lo resolveremos utilizando minería de datos, técnica con la cual descubriremos algún patrón en los datos que nos permita abordar el problema y generar una solución aproximada. Según el autor [17], la minería de datos es un conjunto de herramientas que son utilizadas para modelar y comprender conjuntos de datos complejos en tamaño o variables que para el razonamiento humano no tienen sentido lógico. En general utilizamos estas herramientas para aprender de los datos existentes y obtener información nueva y desconocida a partir de ellos. Otros autores [23] definen la minería de datos como un proceso de identificación de patrones y reglas significativas a partir de un conjunto de datos grandes y complejos y definen un conjunto de pasos para el proceso de descubrimiento de conocimiento.

Cuando hablamos de descubrimiento de nuevo conocimiento, los autores [17] y [23] concuerdan en que las herramientas utilizadas en el área se pueden clasificar en dos grupos: aprendizaje supervisado y aprendizaje no supervisado. El aprendizaje supervisado esta relacionado con la predicción de datos, mientras que el aprendizaje no supervisado tiene relación con la descripción y visualización de los datos. La gran diferencia entre ambos tiene relación con el objetivo del descubrimiento de conocimiento y con el tipo de datos con el que se cuenta. En un aprendizaje no supervisado se cuenta con un conjunto de medidas o variables no asociadas a una variable de respuesta, por lo que resulta imposible adaptar un modelo a los datos, por el contrario en un problema de predicción cada instancia tiene asociada una variable

de respuesta por lo que es posible ajustar un modelo que relacione la variable de respuesta con las observaciones. En particular nos interesa aprender, comprender y utilizar herramientas de aprendizaje supervisado ya que éstas nos permitirán predecir instancias futuras.

Los autores [17] y [23] clasifican el aprendizaje supervisado en dos tipos: clasificación y regresión. Estas se diferencian en la naturaleza de la variable de respuesta, también llamada variable dependiente. Por un lado, cuando la variable dependiente es del tipo cualitativa estamos en presencia de un problema de clasificación, mientras que nos referimos a problemas de regresión cuando la variable dependiente es del tipo cuantitativa.

Considerando los antecedentes, la situación planteada es consistente con un *modelo de regresión*, donde las variables independientes corresponden a la reflectancia espectral de cada parcela de trigo y el desempeño obtenido por parcela o cualquiera de las variables de estudio constituye la variable dependiente o de respuesta.

La regresión es una técnica utilizada para ajustar o adaptar un modelo matemático a un conjunto de datos. La forma más simple de regresión es la regresión lineal simple (*Simple Linear Regression*), la que utiliza la formula de la línea recta ($y = mx + b$) como modelo y estima los coeficientes m y b a partir del conjunto de datos formado por las variables independientes x y la variable de respuesta conocida y. Con esta información podemos ajustar el modelo para predecir la variable dependiente $\hat{y}$ de instancias futuras. Este modelo es ideal para aquellos casos en los que se cuenta con 1 variable independiente.

En la práctica, como es nuestro caso, se cuenta con más de una variable independiente por lo que el modelo anterior resulta inadecuado para abordar el problema. El autor de [17] menciona una alternativa basada en el modelo de regresión lineal simple que extiende la definición anterior para múltiples variables independientes. Este modelo se conoce como regresión lineal múltiple (MLR, del inglés *Multiple Lineal Regression*).

Los modelos mencionados corresponden a una familia de modelos de regresión, sin embargo, el estado del arte sugiere un conjunto más grande de técnicas que pueden ser aplicadas para este tipo de problemas y que eventualmente podrían mejorar los resultados de los modelos más sencillos.

Nuestra propuesta es aplicar diversos modelos de regresión y abarcar con ello un nuevo grupo de familia de regresores con el afán de encontrar aquel que se destaque

sobre el resto. Adicionalmente, queremos identificar las ventajas y desventajas de dichos modelos, con objeto de en un futuro realizar cambios a estos para mejorar los resultados obtenidos. Para cada uno de estos modelos se variará la variable dependiente de tal forma que el modelo sea capaz de predecir diferentes indicadores de desempeño de un cultivo y variables fisiológicas y morfológicas del trigo.

Durante este proyecto se trabajará utilizando una metodología sistemática que incluirá todos los aspectos relevantes del análisis de datos. Dicha metodología, detallará el proceso utilizado para dar respuesta a las conjeturas propuestas.

1.2. Hipótesis

El modelo de regresión es una técnica adecuada para resolver el problema de la predicción del desempeño de una parcela de trigo, con alto nivel de certeza y validez estadística, basado en datos de frecuencia de reflectancia espectral.

1.3. Objetivos

Objetivo General

- Evaluar y comparar diferentes modelos de regresión, en función de la predicción del conjunto de variables de estudio de una parcela de trigo basado en la reflectancia espectral de las observaciones, utilizando una metodología sistemática.

Objetivos Específicos

- Recopilar la literatura sobre el estado del arte utilizado para la predicción del desempeño de cultivos basado en la reflectancia espectral de las observaciones.
- Recopilar la literatura sobre la taxonomía de modelos de regresión en minería de datos.
- Seleccionar los modelos de regresión a evaluar.
- Adquirir el conjunto de datos junto con las diferentes variables de respuestas del estudio.
- Preparar el conjunto de datos usando técnicas de minería de datos.

- Aplicar los modelos de regresión seleccionados para la predicción de cada una de las variables de respuesta.
- Medir y validar la exactitud del ajuste de cada modelo de regresión utilizando estadísticos ampliamente usados para este tipo de problemas.
- Comparar los resultados estadísticos de los diferentes modelos de regresión.
- Construir un modelo de regresión capaz de predecir el desempeño del trigo basado en las medidas de la reflectancia espectral de las observaciones.
- Evaluar el comportamiento de los diferentes modelos de regresión, detallando sus ventajas y desventajas.

1.4. Alcances

- En este proyecto se evaluarán y compararán un máximo de 4 modelos de regresión, en los que se incluyen Regresión Lineal Múltiple (MLR), Regresión de Mínimos Cuadrados Parciales (PLS), Regresión Ridge y una variante de Máquinas de Vectores de Soporte para Regresión (SVR, del inglés *Support Vector Regression*).
- En este proyecto se trabajará con a lo menos 3 variables de respuestas diferentes para el conjunto de datos con un máximo de 15.
- En este proyecto se trabajaran con diferentes conjuntos de datos para dos años en particular (2011/2012), los que son subdivididos a su vez por estrés hídrico (medio, severo, riego completo) y etapa fenológica (antesis, llenado de grano).
- Este proyecto no considera la mejora de modelos existentes, por lo que sólo se limita a la evaluación de estos.

2. Marco Teórico

2.1. Mejoramiento Genético del Trigo

Dada la problemática existente, los programas de mejoramiento deben focalizar sus esfuerzos en: (1) incrementar el potencial de rendimiento de las cosechas, y (2) mejorar la resistencia a los estreses abióticos de las plantas, los que comprenden fenómenos físicos y químicos como lluvia, aire, suelo, minerales, salinidad, etc. [28]. En particular, es de interés para este estudio, la tolerancia de las cosechas al estrés hídrico. Para ello, los programas de mejoramiento requieren hacer más eficiente la liberación de genotipos de mayor rendimiento y mejor aclimatados a condiciones adversas. Afortunadamente, el amplio rango de ambientes en los cuales se cultiva el trigo, nos permite tener una amplia variabilidad genética para hacer frente a las diversas condiciones medioambientales.

El mejoramiento de una especie es conocido como fitomejoramiento. El concepto se refiere a la mejora de rasgos cuantitativos de la especie que permitan plantas más resistentes y productivas a diversos ambientes, sin embargo, el éxito de este proceso dependerá de un adecuado fenotipeo [5]. El desarrollo de técnicas de fenotipeo eficientes será esencial para acelerar el desarrollo de nuevas variedades de trigo, con mayor potencial de rendimiento, con mejor eficiencia en el uso del agua o la tolerancia a la sequía provocada por el cambio climático y la disminución del agua de riego [32]. Algunas características morfo-fisiológicas relacionadas al rendimiento del trigo se describen a continuación:

Discriminación de isótopos de carbono ($\Delta^{13}C$): Utilizada para seleccionar genotipos mediante la eficiencia de transpiración en ambientes [1], [7], [22].

Potencial hídrico de la hoja: Mide el estado hídrico de hoja, es decir, la cantidad de moléculas de agua en la hoja potenciales para realizar el trabajo. Es una de las mediciones más utilizadas para medir la condición hídrica de los cultivos y es un reconocido indicador de estrés por sequía en la planta [22].

Contenido relativo de agua: Mide la cantidad de agua actual en la hoja, por tanto indica su estado hídrico y puede indicar el grado de estrés expresado bajo sequía y calor.

Contenido de clorofila (SPAD): Es un indicador del complejo fotosintético entero. La pérdida de clorofila es un indicativo del estrés inducido por calor, sequía, salinidad, deficiencia nutricional, envejecimiento, etc., y refleja una pérdida del potencial fotosintético de la planta [26].

Índice de área foliar (IAF): Aporta información respecto al crecimiento vegetativo del cultivo, lo cual está fuertemente relacionado con el desarrollo, crecimiento y rendimiento final del grano en condiciones óptimas.

Fluorescencia de la clorofila-a (CHO): El pigmento de *clorofila-a* que existe en las hojas de las plantas, es el encargado de absorber la luz incidente para utilizarla en el proceso de fotosíntesis para convertir la energía radiante en energía química estable. Parte de esta energía se disipa como calor y otra como fluorescencia. La fluorescencia de la clorofila está directamente relacionada con la actividad fotosintética y se ha indicado que su uso permite explicar la variación en potenciales de rendimiento de plantas de trigo [1].

Si bien existe un sin número de características que podrían ser de interés, e incluso aunque diversos autores sugieren una evaluación fenotípica profunda, no todas son factibles para abarcar en un estudio de fitomejoramiento, pues requiere un gran esfuerzo de tiempo y costo lo que lo hace casi inviable [5], [22].

2.2. Reflectancia Espectral

En la literatura se han reportado diversos estudios que han utilizado reflectancia espectral para medir diversas características morfo-fisiológicas del trigo. Los autores

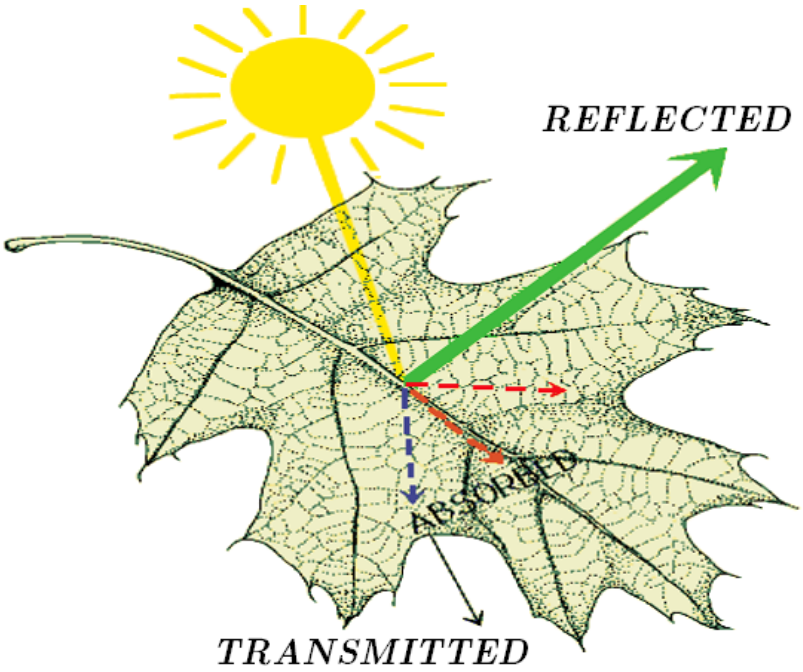


Figura 2.1: Ejemplo de reflectancia en una hoja

concuerdan en que este método no destructivo permite estimar de forma anticipada las características de estudio y han demostrado ser de gran ayuda en la caracterización del fenotipo orientado al mejoramiento genético [11], [16], [19], [22], [25].

Como se aprecia en la figura 2.1, de la energía incidente proveniente desde el sol, sólo una parte es reflejada desde las hojas y el resto es absorbida o transmitida [18]. La reflectancia espectral es la proporción entre la cantidad de fotones de luz que incide en un objeto respecto de la cantidad que refleja [3], [15].

La firma espectral (caracterización gráfica de la reflectancia) está estrechamente asociada a la absorción de determinadas longitudes de onda que son vinculadas a caracteres o condiciones específicas de las plantas. En la figura 2.2 se aprecian las características medibles en los distintos rangos de espectro en una planta. De este modo, la reflectancia dentro del espectro visible (400-700 nm) depende de la absorción de luz por parte de la clorofila de la hoja y otros pigmentos asociados; en este rango, la reflectancia es menor, debido a que hay una mayor absorción de luz por parte de dichos pigmentos [30]. Al contrario, en la banda del 700-1300 nm, conocido como infrarrojo cercano, existe una reflectancia máxima, determinada principalmente por características estructurales de la planta/hoja. Dentro del espectro

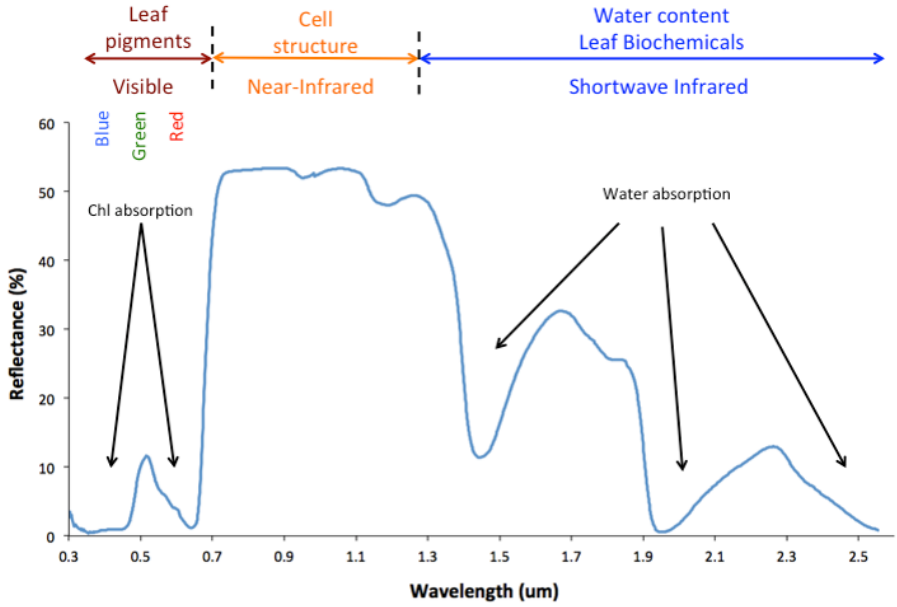


Figura 2.2: Rangos del espectro y características medibles en las plantas

del infrarrojo medio o de onda corta (>1300 nm), se presentan variaciones atribuidas a características de absorción del agua y otros componentes [18]. Sin embargo, los factores que influyen sobre la firma espectral son variados, tales como el ángulo de la medición y heterogeneidad de la superficie de la vegetación, factores ambientales, variabilidad atmosférica y lumínica, entre otros [18].

2.3. Análisis de Regresión

En el área de agronomía, existen diversos estudios que han utilizado técnicas de análisis de regresión para predecir o estimar algunas características morfo-fisiológicas del trigo a partir de datos de reflectancia espectral. Los autores de [29] han utilizado análisis multivariable como PLS y *Stepwise Multiple Linear Regression* (SMLR) para predecir el contenido de clorofila, producción de grano y contenido de proteínas en el trigo obteniendo resultados con un nivel de certeza superior al 80 %. Otros autores por ejemplo [19], han utilizado PLS para medir el contenido de nitrógeno del trigo de invierno, obteniendo resultados buenos cercanos al 75 % de certeza.

Algunas investigaciones más recientes sobre la producción de grano han utilizado técnicas diferentes a las anteriores. Los autores de [16], utilizaron *Ridge Regression*

para estimar el rendimiento del trigo obteniendo resultados satisfactorios y mejores con respecto a los utilizados con otros algoritmos.

Por otro lado, mediante el uso de redes neuronales, los autores de [20] y [25] utilizaron datos de reflectancia espectral para estudiar características morfo-fisiológicas en diferentes especies. El primero [20], con una certeza superior al 90 % pudo discriminar los niveles de infección fúngica en arroz. El segundo [25], mediante el uso de perceptrones multicapa detectó el óxido amarillo con una certeza superior al 95 %.

Si bien pareciera que los modelos basados en redes neuronales tienen mejores resultados, en algunos casos, modelos más sencillos obtienen resultados óptimos. Por tal razón, en esta investigación se analizarán diversos modelos para cada una de las variables de estudio seleccionadas.

2.3.1. ¿Qué es regresión?

Cuando se intenta aprender de los datos con el fin de predecir una variable continua y, el problema se conoce como análisis de regresión. Sea $\mathcal{X}$ el espacio de entrada e $\mathcal{Y}$ un subconjunto medible del dominio real. Sea D un conjunto de observaciones, donde existe una función subyacente $f : \mathcal{X} \to \mathcal{Y}$ que relaciona una observación $\mathbf{x} \in \mathcal{X}$ con una variable dependiente $y \in \mathcal{Y}$. En el problema de la regresión f se supone que es desconocida, y el objetivo es construir una función g que se aproxime a f tanto como sea posible. El análisis de regresión es un ejemplo de aprendizaje supervisado, porque el modelo de aprendizaje recibe n pares de la forma $(\mathbf{x}, y)$, donde $\mathbf{x} \in \mathbf{R^d}$ es la observación e $y = f(\mathbf{x})$ corresponde al valor real de la variable dependiente que queremos predecir. El modelo se llama supervisado porque, para cada instancia $\mathbf{x}$, estamos proporcionando al modelo con el ejemplo correcto de la variable de salida y.

Aquí, f es la función verdadera que explica la variable dependiente y en términos de las variables independientes $\mathbf{x}$. Si conociéramos f el problema estaría resuelto y sólo deberíamos aplicar $f(\mathbf{x})$ para obtener el correcto valor para y. Desafortunadamente, no sabemos la forma explícita de esta función. Sin embargo, los ejemplos proporcionados por los pares $(\mathbf{x}, y)$ nos dan una pista sobre cual es el patrón que relaciona $\mathbf{x}$ e y. Basado en esta información, somos capaces de construir un modelo matemático que mezcla las variables independientes con el fin de predecir la variable dependiente. Hay muchas maneras diferentes en las que estas variables se pueden combinar, y usualmente los científicos en esta área se enfrentan a la necesidad de

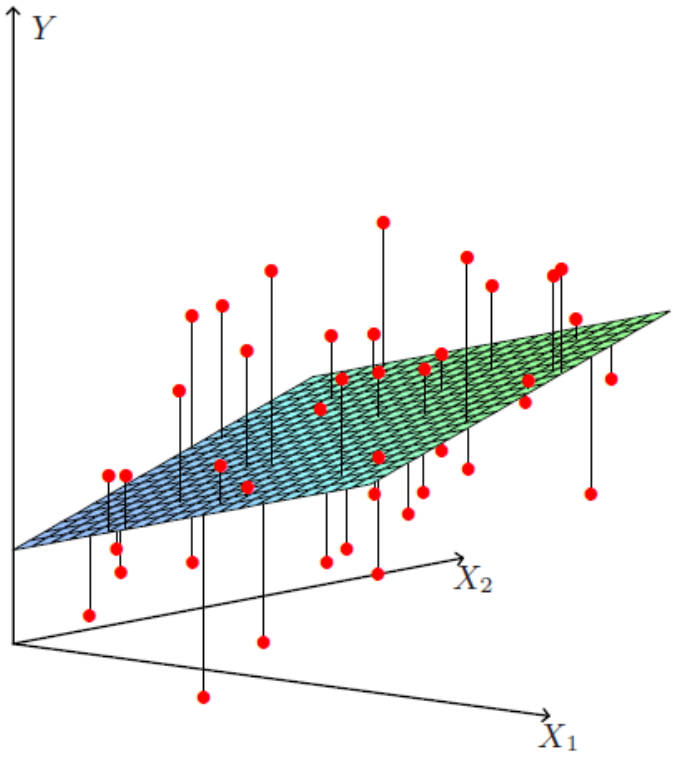


Figura 2.3: Plano que se ajusta a los datos cuando $\mathcal{X} \in \mathbf{R^2}$

elegir un esquema particular. Esta estrategia se llama algoritmo, y es denotado por $\mathcal{A}$. Un algoritmo $\mathcal{A}$ recibe como entrada un conjunto de datos $\mathcal{X}$ y el conjunto de salidas correspondientes $\mathcal{Y}$, generando una función $g(\cdot)$. El objetivo es que $g(\mathbf{x})$ y $f(\mathbf{x})$ sean lo más similar posible. La salida de g se llama predicción y se denota por $\hat{y} = g(\mathbf{x})$.

2.3.2. Multiple Linear Regression

La regresión linear múltiple parte del supuesto que la función $f(\mathbf{x})$ es lineal, es decir, que existe una correlación lineal entre el espacio de entrada $\mathcal{X}$ y el subconjunto medible del dominio real $\mathcal{Y}$ [14].

Cuando el espacio $\mathcal{X} \in \mathbf{R^1}$ nuestro problema se basa en ajustar los datos a una recta en el espacio unidimensional, es decir, la función que intentamos ajustar es de la forma $y_i = \beta_0 + \beta_1 x_i + \mathcal{E}_i$, donde β_0 y β_1 son los coeficientes del modelo que desconocemos y deseamos identificar a partir de la información estadística propor-

cionada por lo pares $(\mathbf{x}, y)$ e $\mathcal{E}_i$ corresponde al error asociado a la estimación de la i-ésima observación $\mathbf{x}$. Si bien existen muchas posibles rectas que se acercan a $f(\mathbf{x})$, solo aquella que genere el menor error para todas las observaciones del conjunto es de interés [14], [17].

Si ampliamos el espacio de tal forma que $\mathcal{X} \in \mathbf{R^2}$ el problema pasa de querer ajustar una recta a un plano lineal como el de la figura 2.3. En este caso, la lógica de la ecuación se mantiene, pero se complica levemente. La ecuación que queremos ajustar es de la forma, $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,1} + \mathcal{E}_i$.

A medida que el espacio comienza a aumentar, es decir, cuando $\mathcal{X} \in \mathbf{R^d}$, para el ser humano comienza a ser difícil de imaginar y por tanto no es posible de llevar a un gráfico que pueda ser comprendido. La ecuación por otro lado, mantiene la misma idea, sin embargo se debe extrapolar a las n dimensiones del problema. De este modo, la ecuación que relaciona una observación $\mathbf{x} \in \mathbf{X}$ con una variable dependiente $y \in \mathbf{Y}$, es de la forma $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,1} + \ldots + \beta_d x_{i,d} + \mathcal{E}_i$. La ecuación 2.1, representa una generalización del modelo lineal para un problema de n dimensiones.

$$\mathcal{Y} = \beta_0 + \beta\mathcal{X} + \mathcal{E} \tag{2.1}$$

donde β_0 corresponde al coeficiente de intercepción y β son todos los coeficientes del modelo.

2.3.3. Partial Least Squares Regression

Esta técnica combina y generaliza características de análisis de componentes principales y regresión lineal múltiple. PLS además de ser un modelo de regresión, se encarga de reducir la cantidad de dimensiones, por lo que es muy utilizado cuando la cantidad de variables independientes es alta.

Una de las consideraciones que se debe tener con PLS, es que la magnitud de los valores de las variables independientes influyen en la estimación del modelo, por lo que cada observación $\mathbf{x}$ debe ser estandarizada de tal modo que tenga media 0 y varianza 1.

Como mencionan los autores de [14] y [17], PLS identifica un nuevo subconjunto de variables independientes $Z_i, \ldots, Z_m$ que son combinaciones lineales construidas a partir del espacio de entrada $\mathcal{X}$. Este nuevo subconjunto no solo esta relacionado a $\mathcal{X}$, sino que también a la variable de respuesta del dominio real $\mathcal{Y}$. Esta característica

es lo que hace a este método pertenecer al grupo de algoritmos supervisados. Como menciona el autor de [17], PLS intenta encontrar combinaciones lineales que ayuden a explicar los predictores y que tengan una alta correlación con la respuesta.

Una vez que es identificado este nuevo subconjunto de variables independientes $Z_i, \ldots, Z_m$, se ajusta el modelo lineal a través de mínimos cuadrados utilizando estas nuevas características y se selecciona el mejor de ellos.

Al generar modelos vía PLS, se obtiene un vector amplio debido a la utilización del rango espectral completo, pero los resultados pueden estar alterados por la multicolinealidad, problema característico en estos métodos de calibración.

2.3.4. Ridge Regression

Como se menciona en la subsección anterior, un problema que tanto PLS como MLR no son capaces de identificar tiene relación con la multicolinealidad, es decir, no son capaces de determinar la correlación existente entre las variables independientes $\mathcal{X}$ del modelo. Este hecho afecta los resultados de las estimaciones y junto con ello amplifican los errores del modelo y producen medidas de rendimiento no realistas.

Regresión Ridge propone un mecanismo que permite subsanar el problema anterior atacando la causa de este, la multicolinealidad. En esencia Ridge funciona de forma similar a MLR pero incluye una contracción (conocido como "*shrinkage*" en inglés) de los coeficientes de la regresión del modelo, reduciéndolos en igual magnitud, mediante la incorporación de un factor de contracción de los estimadores [14], [17]. Los coeficientes de regresión estimados, son entonces aquellos valores que minimicen la ecuación 2.2.

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2.2}$$

donde $\lambda \geq 0$ y corresponde al parámetro de ajuste.

La primera parte de la ecuación 2.2 corresponde al ajuste del modelo en base a las observaciones proporcionadas, mientras que la segunda parte incluye el castigo de los coeficientes del modelo. Cuando $\lambda = 0$, estamos en presencia de una regresión lineal cualquiera, es decir, sin castigo. Por el contrario, mientras λ comienza a crecer, mayor será el factor de contracción de los coeficientes y por ende mayor es el castigo [17].

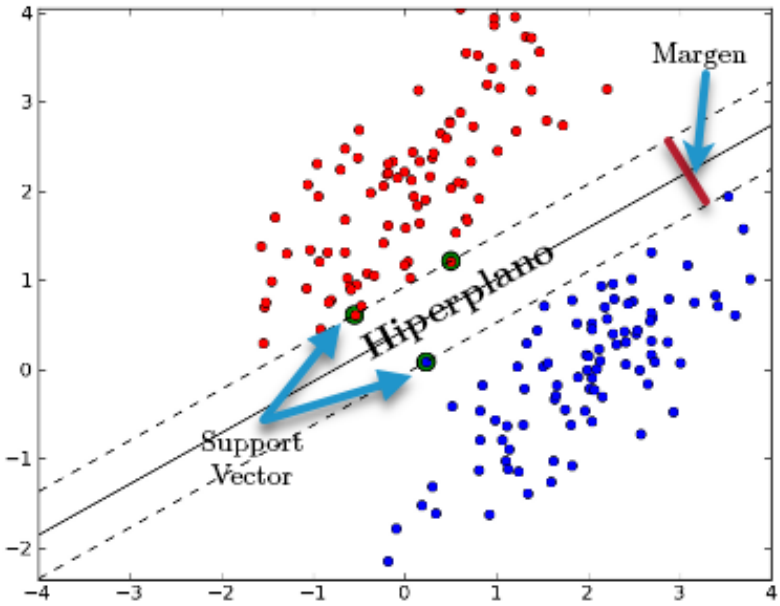


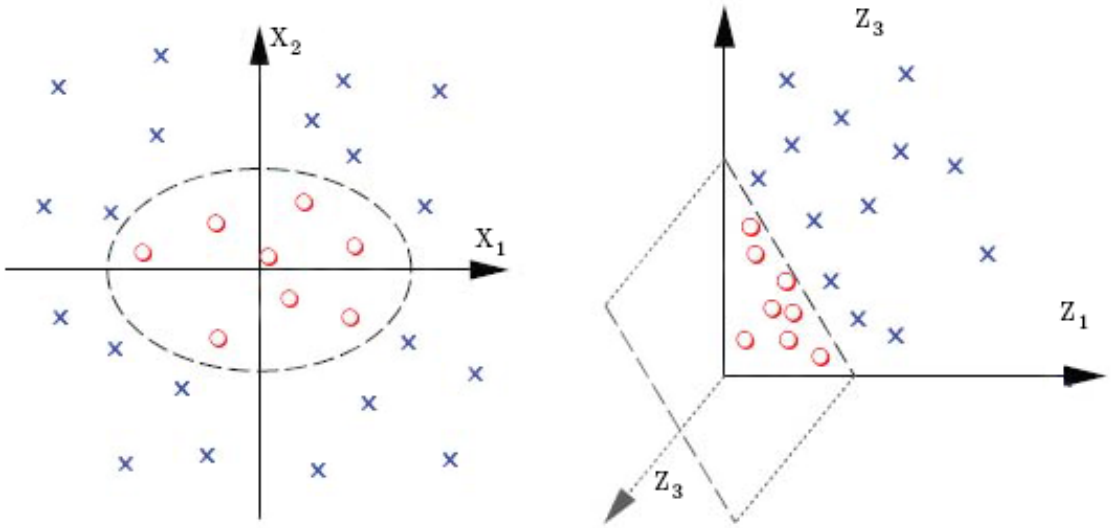
Figura 2.4: Hiperplano que clasifica correctamente y maximiza el margen

La optimización del parámetro λ es un problema en este modelo. El autor de [17] sugiere la utilización de validación cruzada para determinar el mejor valor para λ dentro un rango amplio de valores posibles. No queda claro cual es el rango que se debe utilizar, sin embargo dependiendo del problema y de los datos utilizados cada investigador deberá determinar el mejor rango de valores para su optimización.

2.3.5. Support Vector Regression

Como mencionamos en la sección 1.1.4, los autores [17] y [23] identifican dos tipos de aprendizajes supervisados, la clasificación y la regresión. La primera de ellas en particular, busca clasificar o separar en grupos un conjunto de datos. Las máquinas de vectores de soporte (SVM, del inglés *Support Vector Machines*) son un claro ejemplo de ello [14], [17].

Una primera aproximación a lo que SVM hace es encontrar una recta separadora, o más genéricamente llamado hiperplano, entre datos de dos clases. Por lo tanto, supongamos que tenemos algunos datos de dos clases diferentes como el presentado en la figura 2.4, Support Vector Machine toma estos datos como una entrada, y genera de salida una línea que separa esas clases, si es posible [8].



(a) Datos no linealmente separables

(b) Datos linealmente separables en un espacio transformado

Figura 2.5: Ejemplo de una transformación del espacio de entrada

En la figura 2.4 se muestra un ejemplo de un conjunto de datos de dos clases diferentes separadas por un hiperplano. Lo que hace esta línea, es maximizar la distancia hasta el punto más cercano con cualquier clase, esa distancia a menudo se llama "margen". Aquellas observaciones que caen en la frontera de dichos márgenes reciben el nombre de "vectores de soporte". En resumen, el primer objetivo de SVM es clasificar correctamente todas las observaciones y luego maximizar el margen, produciendo un clasificador robusto a los errores y por ende maximizando la robustez de los resultados.

Hasta aquí todo bien, sin embargo cuando los datos tienen un comportamiento diferente como el de la figura 2.5 (a), es muy difícil trazar una línea entre las dos clases. Si las observaciones no son separables linealmente en el espacio original, la búsqueda del hiperplano de separación en espacios transformados, normalmente de muy alta dimensión, como sugiere la figura 2.5 (b) parece ser la solución. Esto se realiza utilizando las llamadas funciones de "*kernel trick*". Hasta ahora hemos asumido que introducimos nuestros datos, $\mathbf{X}$ e $\mathbf{Y}$, en una caja negra pero mágica llamada SVM que nos entrega como resultado la etiqueta o clasificación de una observación. Ahora introduciremos a lo anterior una función de *kernel* cualquiera que toma un espacio de entrada de dimensiones bajas o un espacio de características y lo mapea

a un espacio de alta dimensión como en la figura 2.5. En esta nueva dimensión los datos son claramente separables por un hiperplano, entonces el truco esta en seleccionar sabiamente el *kernel trick* a utilizar. Existen diferentes tipos de kernel que se pueden utilizar, algunos de ellos son:

- Lineal
- Polinomial
- Sigmoidal
- Radial

SVM requiere la selección de algunos parámetros previos al ajuste del modelo. Uno de ellos es el tipo de kernel a utilizar, que corresponde a uno de los listados arriba. Otro parámetro, esta relacionado a controlar el equilibrio entre un límite de decisión suave y uno que clasifica correctamente todos los puntos de entrenamiento, este parámetro es conocido como C. El límite de decisión podría ser algo que es considerablemente más ondulado, pero se obtiene potencialmente todos los puntos de entrenamiento correctos. Este caso esta asociado a un gran valor de C y obtiene un modelo que es más complicado, pero también es probable que no va a generalizar muy bien el conjunto de pruebas. Así que algo que es un poco más recto y sencillo puede ser realmente mejor opción a la hora de buscar un modelo con mayor certeza sobre el conjunto de prueba.

Support Vector Machines puede aplicarse no solo a problemas de clasificación, sino también al caso de la regresión. Support Vector Regression (SVR) hereda las propiedades de SVM para ajustar un modelo no lineal para aprender de los datos. Dado que en regresión la salida debe ser un valor real es que se introduce una función de pérdida que ignora los errores que se sitúan dentro de cierta distancia del valor verdadero [9]. Esta función se llama épsilon (ϵ).

Como se aprecia en la figura 2.6, la función de pérdida ϵ permite que exista cierta dispersión en la solución de tal forma que todas las observaciones que se encuentran en la región que comprende, no sean considerados vectores de soporte. En resumen la idea sigue siendo la misma, buscar un hiperplano que maximice el margen, teniendo en cuenta que parte del error es tolerado.

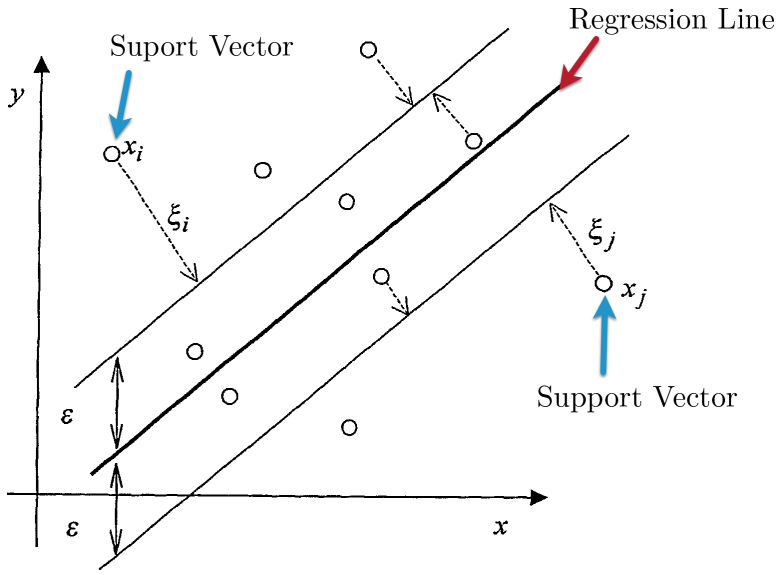


Figura 2.6: SVR que muestra la holgura de la función de pérdida ϵ

2.4. Otros Algoritmos Utilizados

2.4.1. Principal Component Analysis (PCA)

Análisis de componente principales (PCA, del inglés *Principal Component Analysis*) es un algoritmo no supervisado comúnmente utilizado para la reducción de la dimensionalidad, es decir para transformar un conjunto de datos de alta dimensión en un espacio mucho más pequeño.

Supongamos que contamos con un conjunto de puntos de alguna distribución como la que se muestra en la figura 2.7. Así que tenemos dos características en el plano real $\mathbf{R^2}$. PCA encuentra los vectores que maximicen la variación de los datos, es decir, encuentra aquellos vectores que explican los puntos y por tanto que están relacionados a ellos [17], [27].

Una vez que se encuentra un vector, se calcula la varianza de los puntos proyectados allí. Como se aprecia en la zona encerrada por un rectángulo en la figura 2.7, la varianza se calcula mediante la distancia entre cada punto y su proyección en el vector. La varianza de los puntos a medida que se proyectan en esta línea tendrá una varianza mucho mayor que en cualquiera de estas dos dimensiones.

Hasta ahora sabemos como encontrar el primer componente, que por cierto es quien explica gran parte de los datos. Pero como mencionamos anteriormente el

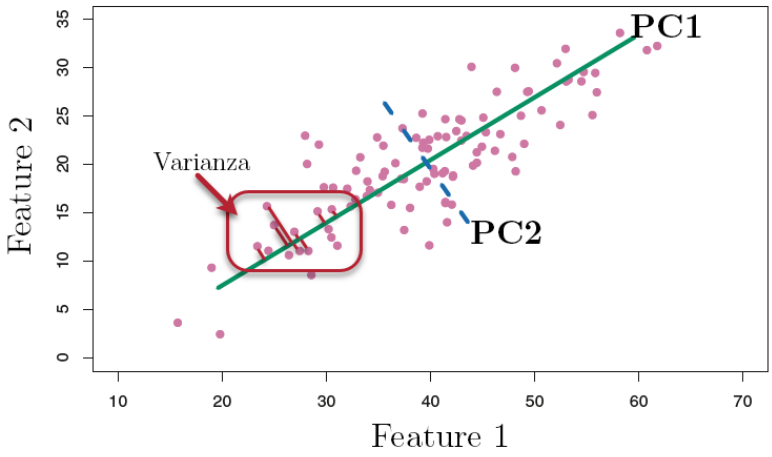


Figura 2.7: Componentes principales para un distribución de ejemplo

algoritmo no solo encuentra un vector, sino más de uno. Lo segundo a tener en consideración es que PCA encuentra direcciones que son mutuamente ortogonales, en el caso de un problema de dos dimensiones, buscaremos un vector que sea perpendicular al primer componente encontrado.

Una de las características más importantes de PCA es que no genera pérdida de información, es decir, produce la rotación lineal de las dimensiones originales. En otras palabras, es sólo una especie de re-etiquetado de las dimensiones.

Como menciona el autor de [17], el primer componente principal define la mejor línea que se encuentra más cercana a los datos. Lo que quiere decir, es que si proyecto un punto en el primer componente principal, y luego lo comparo respecto de donde estaba en el espacio original, la suma de todas las distancias entre esos puntos será realmente el mínimo que podría obtener para cualquier otra proyección.

Hasta ahora no hemos visto como esto puede permitirnos reducir la dimensionalidad de un problema. De hecho, cuando calculamos los componentes principales si comenzamos con N dimensiones, obtendremos otras N dimensiones. Por lo tanto, ahora nuestro objetivo se reduce a seleccionar un subconjunto M de ellos que sea mucho menor que las N dimensiones originales. Cada una de estas nuevas dimensiones que conseguimos tiene asociado un valor llamado "*eigenvalue*". Un *eigenvalue* garantiza que no es negativo, pero aún más importante, es que estos valores no aumentan monótonamente, es decir, terminan siendo más pequeños a medida que aumentan los componentes principales. En este sentido, el segundo componente principal tiene

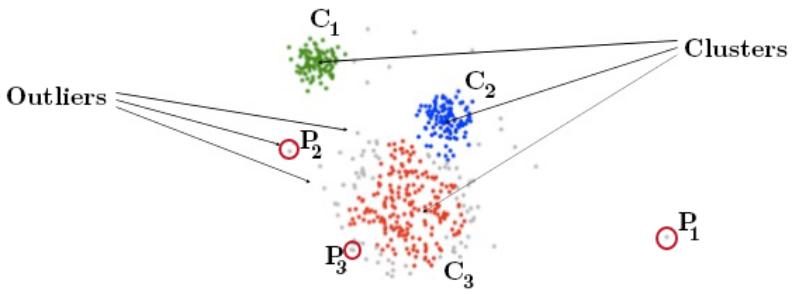


Figura 2.8: Componentes principales para un distribución de ejemplo

un *eigenvalue* más pequeño que el primero, y a su vez el tercer componente tiene un valor más pequeño que el segundo y así sucesivamente hasta el n-ésimo componente. Con esta característica, es que podemos seleccionar aquellos componentes que expliquen la mayor parte de los datos.

2.4.2. Local Outlier Factor (LOF)

Un subproblema tradicional en la minería de datos es identificar y tratar aquellos datos atípicos dentro de una distribución. En problemas con pocas dimensiones son fácilmente identificables mediante una inspección visual, en particular porque son puntos en el espacio que no siguen la tendencia de la distribución.

Como hemos visto hasta ahora, cuando el conjunto de datos tiene una gran cantidad de variables independientes las soluciones que para problemas de menor magnitud de dimensiones funcionan, no siempre resultan ser adecuadas para problemas de mayor complejidad. Por otro lado, algunas de estás técnicas requieren tener conocimiento del subconjunto medible del dominio real $\mathcal{Y}$ para detectar los datos atípicos presentes en la distribución.

El algoritmo propuesto por los autores de [4], permite encontrar datos atípicos en un conjunto de datos con alta dimensionalidad. El método llamado factor atípico local (LOF, del inglés *Local Outlier Factor*) introduce el concepto de localidad en el sentido de que solo se considera la vecindad cercana de un punto para determinar que tan atípico es o no. Supongamos que tenemos un conjunto de datos como el que se aprecia en la figura 2.8, con algunos puntos que difieren de la tendencia de los datos. Obviamente, no sabemos cuales son un dato atípico. En la figura 2.8, la observación que denotamos por P_1 es claramente un punto atípico porque se encuentra totalmente

aislado de la población, sin embargo los datos P_2 y P_3 que también son atípicos para este algoritmo, podrían no ser considerados así por otros métodos.

Local outlier factor, se basa en la densidad de la población, en la que se calcula la desviación local de una instancia respecto de sus k vecinos más cercanos. La distancia en este caso, es una función matemática que mide la disimilitud entre dos observaciones. Basado en esta información, LOF asigna una puntuación a cada instancia que indica el grado de ser un dato atípico o no. De este modo instancias solitarias en el espacio como P_1 en la figura 2.8 tienden a tener valores muy altos de LOF ($LOF \gg 1$). Por el contrario, observaciones que se encuentran en el centro de una población o que pertenecen a una región con una densidad homogénea alrededor de el y sus vecinos, son propensos a valores de $LOF \approx 1$.

La principal diferencia de este algoritmo con otros, es que no es una propiedad binaria, es decir, no es blanco o negro. En su lugar, como se menciona antes, a cada instancia se le asigna un factor, que es el grado de ser atípico. Por lo tanto, con este enfoque no solo podemos decir que una instancia es blanca o negra, sino que podría ser gris. La decisión de determinar cuales serán las instancias atípicas es traspasada a los científicos, quienes basados en la información de LOF de cada instancia tomaran la determinación que crean pertinente.

2.5. Medidas de Rendimiento

Una medida de rendimiento es una función cuantitativa para evaluar la bondad de una observación. Como en la regresión la salida es continua, no es realista esperar que el modelo de aprendizaje prediga exactamente la salida de una instancia particular. En lugar de ello, el rendimiento de un modelo de regresión se mide generalmente mediante el calculo de la diferencia entre el valor predicho ($\hat{y}$) y el real (y). La diferencia es llamada usualmente residuo.

En general, el rendimiento de un modelo de regresión se calcula con una función de pérdida $L : \mathcal{Y} \times \mathcal{Y} \to \mathbf{R}$. De acuerdo a los autores de [12], la función de pérdida más común es $L_2 = L(y_i, \hat{y}_i) = |y_i, \hat{y}_i|^2$. Con frecuencia, la función de pérdida se calcula para cada instancia de prueba y se calcula la suma total. Esta expresión se conoce como *Suma de cuadrados Residuales* (RSS) y se calcula utilizando la ecuación 2.3. Esta es una medida apropiada para la evaluación de la precisión de la estimación

de los coeficientes del modelo.

$$\mathsf{RSS} = \sum_{i=1}^{n} |y_i - \hat{y}_i|^2 \tag{2.3}$$

El valor de RSS esta relacionado al número de observaciones, por lo que a mayor cantidad de observaciones el valor de RSS aumenta. El objetivo es encontrar una función $g(\cdot)$ que minimice estos errores.

De acuerdo con los autores de [17], la exactitud del ajuste obtenido con una regresión lineal se evalúa típicamente usando el *Error Estándar Residual* (RSE) y el R^2 estadístico, los que son explicados en los siguientes párrafos.

El RSE representa la desviación estándar de los residuos y se calcula utilizando la ecuación 2.4, donde $n - p - 1$ corresponde a los grados de libertad. Esta medida de rendimiento es una estimación de la desviación entre el valor real de la variable dependiente y el valor estimado obtenido a partir de la línea de regresión. En general, si el modelo es cercano a la realidad, se podría esperar un valor pequeño de RSE. En caso contrario, es decir, si nuestro modelo no representa adecuadamente los datos, el valor de RSE debiese ser alto. Desafortunadamente, no es posible determinar que puede ser un buen (o mal) RSE *a priori*, sobre todo porque este valor se mide en las unidades de la variable dependiente, por lo que el rango de la variable debe ser tomado en cuenta para una correcta interpretación.

$$\mathsf{RSE} = \sqrt{\frac{1}{n-p-1} RSS} = \sqrt{\frac{1}{n-p-1} \sum_{i=1}^{n} |y_i - \hat{y}_i|^2} \tag{2.4}$$

En un escenario ideal, una sola medida de rendimiento debiese decirnos todo acerca de la exactitud del ajuste de un modelo predictivo. En la práctica, es todo lo contrario. Las medidas de rendimiento como el RSE son muy informativos, pero sólo ofrecen información parcial sobre el modelo. Una medida complementaría es el R^2 estadístico, el que viene dado por la siguiente ecuación:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{2.5}$$

donde $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$ es la suma total de los cuadrados.

El R^2 estadístico cuantifica la proporción de la varianza explicada, y cuyo valor varía entre 0 y 1. Normalmente un valor cercano a 1 está asociado a un buen modelo

de predicción, mientras que un R^2 cercano a 0 representa un predictor pobre. A diferencia del RSE, R^2 no depende de la escala de la variable dependiente y.

El autor de [33] advierte que R^2 es una medida insuficiente. El explica que el R^2 estadístico no es capaz de detectar el aditivo y las diferencias proporcionales en el valor real de la variable dependiente, así como las respectivas medias muéstrales y varianzas. Como una solución para este inconveniente, el autor de [33] propuso el Índice de Agreement (IA), y muestra su superioridad cuando se hacen comparaciones entre el valor real de la variable dependiente y el valor estimado a partir del modelo de regresión.

El Índice de Agreement desarrollado por Willmott (1981) es una medida estandarizada para estimar el error de predicción del modelo y varía entre 0 y 1. Un valor de IA igual a 1 indica una coincidencia perfecta, mientras que un valor de 0 se asocia a un modelo defectuoso [33]. IA se calcula según la ecuación 2.6.

$$\mathsf{IA} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (|\hat{y}_i - \bar{y}| + |y_i - \bar{y}|)^2} \tag{2.6}$$

3. Descripción de la Librería (API)

En este capítulo, se describen las diversas funciones creadas mediante una librería. Cada una de las subsecciones, aborda entre una o cuatro funciones para proveer una solución para un problema particular de la minería de datos, y en particular para el problema a solucionar mediante esta investigación.

En cada subsección se detalla el problema, se explica la solución, se muestran gráficos de ser necesarios y se presenta un ejemplo práctico utilizando la o las funciones de la librería.

3.1. Subproblemas en Minería de Datos

3.1.1. Valores Faltantes

Es común que el conjunto de datos se encuentre incompleto, es decir, con valores faltantes ya sea en el conjunto de variables dependientes $\mathcal{X}$, independientes $\mathcal{Y}$ o ambas. Por tanto, es uno de los primeros subproblema que los investigadores deben enfrentar.

La solución de este subproblema dependerá del tipo de algoritmo que se utilizará como solución. Por ejemplo, existen algoritmos un poco más complejos en su implementación que contienen mecanismos para identificar y tratar posibles valores faltantes en los datos, pero si este no es el caso, entonces debemos realizar pasos previos de limpieza de estos.

Los pasos necesarios son sencillos y se explican a continuación acompañado con un ejemplo:

1. Identificar cada uno de los valores faltantes del conjunto de datos `findMissing Values`.

2. Modificar el conjunto de datos.

   a) Reemplazar con información estadística

   b) Eliminar instancias u observaciones afectadas

Ejemplo:

```
# Instalación y carga de la librería
install.packages("devtools")
devtools::install_github("mariytu/RegressionLibs")
library(RegressionLibs)

# Conjunto de datos con valores faltantes
cars <- read.csv("https://dl.dropboxusercontent.com/u/12599702/
    autos.csv", sep = ";", dec = ",")

missingValues <- findMissingValues(cars) # Valores faltantes

if (any(missingValues)) {
  # Eliminando filas repetidas
  missingValues <- missingValues[!duplicated(missingValues[,1]),]
  print(missingValues)
  # Eliminando todas las filas que contienen datos faltantes
  cars <- cars[-missingValues[,1],]
}
```

3.1.2. Normalización

La normalización es un método de transformación de los datos utilizado como paso previo al ajuste de un modelo. Esta herramienta se utiliza debido a que un conjunto amplio de algoritmos de regresión son sensibles a los rangos de las variables, es decir, una variable independiente cuyo rango varía entre $[1, 50]$ tendrá mayor importancia que otra variable cuyo rango sea inferior que este.

No existe ninguna formula maestra que nos diga cual es la mejor normalización y/o transformación para nuestros datos, por lo que probar diferentes combinaciones es lo que usualmente se suele realizar.

Para cumplir con este propósito, la librería cuenta con 3 funciones enfocadas en la transformación de los datos:

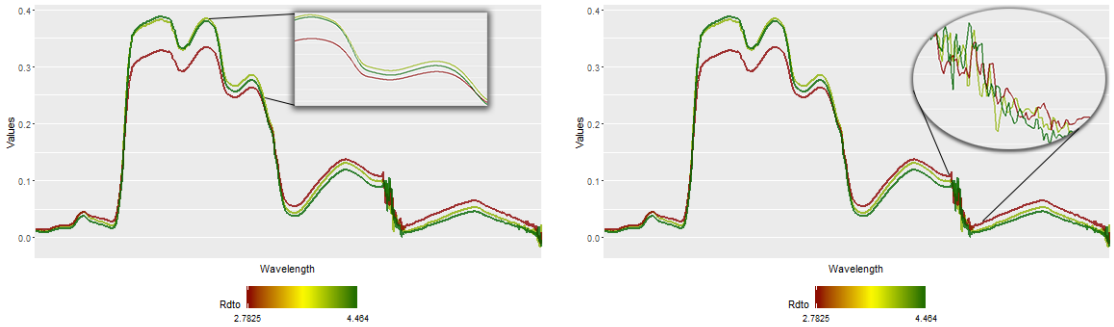
- **Normalize:** Esta función normaliza cada columna del conjunto de datos en un rango entre 0 y 1.
- **ScaleData:** Esta función permite escalar cada columna del conjunto de datos en un rango definido. Si se utiliza la función `normalize` primero y luego la función `scaleData`, se puede obtener un conjunto de datos normalizado en un rango $[min, max]$.
- **NormalizeData:** Esta función normaliza un conjunto de datos en un rango cualquiera $[min, max]$, pero si no se especifica cual será el rango por defecto se calcula para el rango $[0, 1]$.

Ejemplo:

```
# Instalación y carga de la librería
install.packages("devtools")
devtools::install_github("mariytu/RegressionLibs")
library(RegressionLibs)

# Conjunto de datos limpio (Sin datos faltantes)
cars <- read.csv("https://dl.dropboxusercontent.com/u/12599702/
    autosclean.csv", sep = ";", dec = ",")

# En un rango [0,1]
normedCars <- as.data.frame(lapply(cars, normalize))
# En un rango [1,10]
scaleCars <- as.data.frame(lapply(normedCars, scaleData, 1, 10))
#En un rango [1,5]
normed <- normalizeData(cars, 1, 5)
```

3.2. Detección de Ruido

El proceso de detección de ruido es uno de los subproblemas de la minería de datos más difíciles de resolver, porque cada problema requiere una solución particular. Para



(a) Zona del espectro sin ruido (b) Zona del espectro con ruido

Figura 3.1: Ejemplo de la firma espectral de un subconjunto de los datos

el problema con datos de reflectancia espectral, las mediciones que se realizan en el campo pueden verse afectadas principalmente por una mala utilización del equipo de medición o por un mal entrenamiento del personal encargado de tomar las mediciones y muestras correspondientes.

En un escenario ideal, los valores para cada longitud de onda varían levemente entre su antecesor y sucesor, incluso a nivel de números estos comienzan a variar recién en el cuarto o quinto decimal. Gráficamente podemos apreciar esta situación en la figura 3.1 (a), donde se ha destacado una zona del rango espectral cuya onda es suave y es lo que normalmente se debe esperar.

Por otro lado, hay zonas en este caso de estudio que generan ruido y es común para todas las mediciones realizadas. En la figura 3.1 (b) se han destacado las longitudes de onda entre los [1470-1580] que presentan ruido. En esta zona se producen cambios drásticos como picos entre una longitud de onda y su antecesora, incluso los valores varían en el segundo decimal.

Considerando todos estos antecedentes, la solución propuesta fue dividida en dos partes. El algoritmo 1 define el proceso para calcular la diferencia entre una longitud de onda y su antecesor. Mientras que el algoritmo 2 permite identificar aquellas longitudes de ondas que estén sobre cierto umbral (Θ) de tolerancia de ruido.

Algoritmo 1 calculateDiff

Entrada: Conjunto de datos $\mathcal{X}_{n,m}$, limites inf y sup
Salida: $\mathcal{T}_{n,m}$ un arreglo con la variación de cada longitud de onda con su vecino

$\mathcal{T} \leftarrow \phi$
para todo $i \in 1 : nrow(\mathcal{X})$ **hacer**
  $ant \leftarrow \mathcal{X}_{i,1}$
  para todo $j \in 1 : ncol(\mathcal{X})$ **hacer**
    si $j >= inf$ **y** $j <= sup$ **entonces**
      $\mathcal{T}_{i,j} \leftarrow |\mathcal{X}_{i,j} - ant|/ant$
      si $is.nan(\mathcal{T}_{i,j})$ **entonces**
        $\mathcal{T}_{i,j} \leftarrow 0$
      fin si
    fin si
    $ant \leftarrow \mathcal{X}_{i,j}$
  fin para
fin para
devolver $\mathcal{T}$

Algoritmo 2 getColumnsNoise

Entrada: Arreglo con la variación de cada longitud de onda $\mathcal{T}_{n,m}$, y los limites Θ, inf y sup
Salida: $cols$ un arreglo con las columnas con ruido

$cols \leftarrow \phi$
para todo $j \in 1 : ncol(\mathcal{T})$ **hacer**
  si $j >= inf$ **y** $j <= sup$ **entonces**
    $x \leftarrow \mathcal{T}_{,j}$
    $aux \leftarrow count(x >= \Theta)$
    para todo $i \in 1 : nrow(aux)$ **hacer**
      si $aux_{i,1} ==$ **cierto entonces**
        $cols \leftarrow cols + j$
      fin si
    fin para
  fin si
fin para
devolver $cols$

Ejemplo:

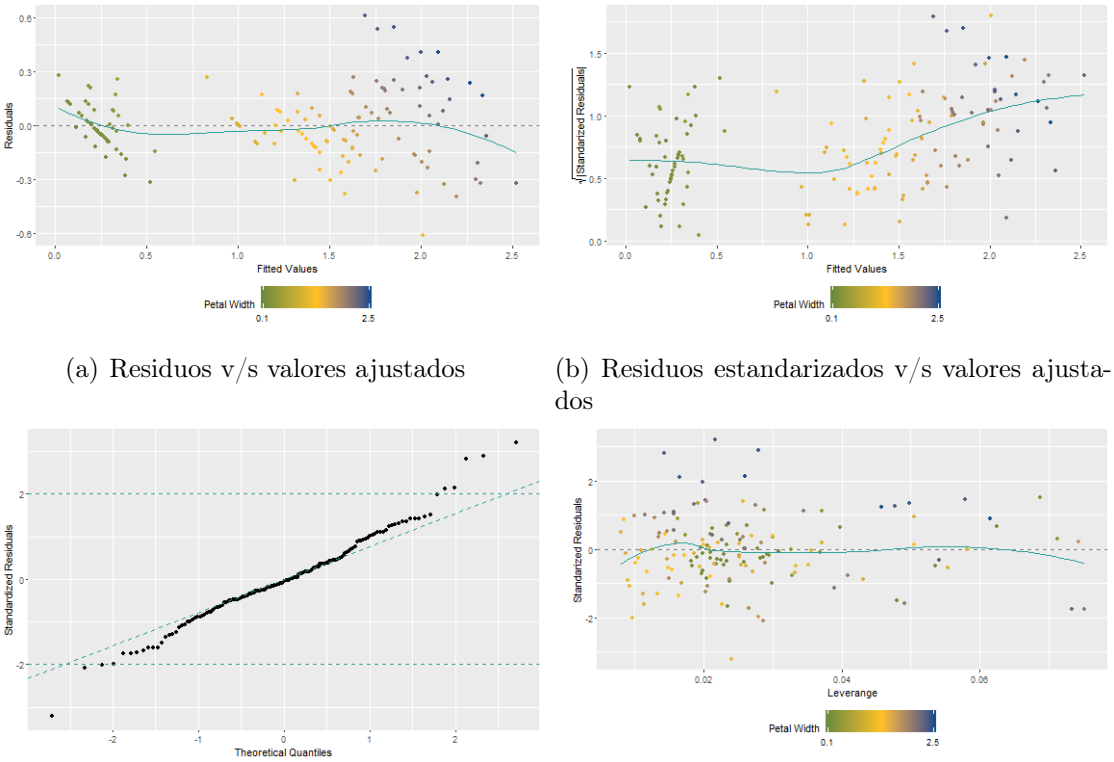
```
# Instalación y carga de la librería
install.packages("devtools")
devtools::install_github("mariytu/RegressionLibs")
library(RegressionLibs)

data <- read.csv("https://dl.dropboxusercontent.com/u/12599702/data.csv",
    sep = ";", dec = ",")
data.x <- data[,2:ncol(data)] #Variables independientes
data.y <- data[,1] #Variables dependientes

#Grafico con datos con ruido
ParallelPlot(data.x, seq(1,nrow(data.x),1), seq(1,ncol(data.x),1),
    data.y, "Rdto", lineSize = 1, alphaLine = 0.8, x_lab = FALSE)

#Proceso de DETECCION DE RUIDO
diffValues <- calculateDiff(data.x)
limit = 0.15
columnsNoise <- getColumnsNoise(diffValues, limit)

#Eliminacion de las columnas con ruido
if (nrow(columnsNoise) > 0) {
  print(paste("hay", nrow(columnsNoise), "columnas con ruido."))

  data.x <- data.x[,-columnsNoise[,1]]
  data <- data.frame(data[,1], data.x)

  #Grafico con datos sin ruido
  ParallelPlot(data.x, seq(1,nrow(data.x),1), seq(1,ncol(data.x),1),
    data.y, "Rdto", lineSize = 1, alphaLine = 0.8, x_lab = FALSE)
} else {
  print(paste("No hay columnas con ruido."))
}
```

3.3. Gráficos de Diagnóstico

Dentro del proceso de ajuste de un modelo de regresión, los investigadores suelen apoyarse con frecuencia de gráficos que les facilite la comprensión y obtención de información de los datos. En especial, con modelos de tipos lineales se utilizan un conjunto de gráficos llamados *gráficos de diagnóstico* que son generados a partir de los valores reales y estimados de la variable dependiente o de estudio. Este tipo de gráficos son usados frecuentemente para la detección de datos atípicos y para tener una idea preliminar del tipo de tendencia en los datos.

Esta sección de la librería dispone de 4 tipos de gráficos distintos, los que solo se pueden generar para modelos lineales de regresión. Previo a la generación de cualquier gráfico de diagnóstico, se requiere de la ejecución de la función `diagnosticData`, la que dado un objeto de clase `lm` que contiene el ajuste de un modelo lineal, calcula y retorna los valores observados, estimados, diferentes tipos de residuos y estadísticos para determinar datos influenciables. Específicamente se calculan los siguientes estadísticos:

- y: Valor real de una observación, también conocido como variable dependiente.
- $\hat{y}$: Valor estimado de una observación.
- **Residuo Ordinario**: Corresponde a la diferencia entre el valor observado de la variable dependiente (y) y el valor estimado ($\hat{y}$).
- **Residuo Estandarizado**: Los residuos dividido por sus desviaciones estándar.
- **Residuo Estudentizado**: Los residuos divido por sus desviaciones estándar, donde la n-ésima observación ha sido eliminada en el calculo de la desviación estándar para el residuo que sigue una distribución t.
- **Distancia de Cook**: Medida de impacto de cada observación en el grupo de coeficientes de regresión, así como con el grupo de valores ajustados. Los valores superiores a $4/n$ son considerados de gran influencia.
- **DFFITS**: Medida de cuanto una observación ha afectado su valor ajustado el modelo de regresión. Los valores mayores que $\left|2\sqrt{(k+1)/n}\right|$, son considerados altamente influyentes.



(a) Residuos v/s valores ajustados

(b) Residuos estandarizados v/s valores ajustados

(c) Normal QQ

(d) Leverange v/s residuos estandarizados

Figura 3.2: Gráficos de influencia, leverange y residuos para el conjunto de datos *iris*

- **DFBETAS**: Medida de cuanto una observación ha afectado la estimación de un coeficiente de regresión. Existe un *dfbeta* por cada coeficiente, incluyendo la intersección. Valores mayores a $|2/\sqrt{n}|$ son considerados de gran influencia.

- **Leverange**: Medida de que tan lejos una observación esta de los otros en términos de los niveles de las variables independientes. Observaciones con valores mayores a $2(k+1)/n$ son consideradas potencialmente de gran influencia, donde k es el número de predictores y n es el tamaño de la muestra.

- **Covratio**: Medida del impacto de cada observación en las varianzas de los coeficientes de regresión y sus covarianzas. Valores fuera del intervalo de $1 \pm 3(k+1)/n$ son consideradas de gran influencia.

Con la información anterior, se pueden generar los 4 tipos de gráficos. El primero de ellos 3.2 (a) muestra si los residuos tienen patrones no lineales. Por otro lado, en el

gráfico 3.2 (b) podemos identificar si los residuos se distribuyen de manera equitativa a lo largo del rango de los predictores.

El gráfico conocido como *Normal Q-Q* evidencia si los residuos se distribuyen normalmente. En este gráfico es esperable que las observaciones se distribuyan en una línea recta. Como se aprecia en la figura 3.2 (c) solo algunas observaciones se encuentran más desviadas de la recta, las que podríamos considerar potencialmente atípicos.

Finalmente el gráfico que se aprecia en la figura 3.2 (d) nos permite detectar aquellas observaciones potencialmente influyentes, es decir, observaciones que afectan fuertemente los coeficientes del modelo predictivo. No todos los valores atípicos son influyentes en el análisis de regresión.

Ejemplo:

```
# Instalación y carga de la librería
install.packages("devtools")
devtools::install_github("mariytu/RegressionLibs")
library(RegressionLibs)

iris.x <- iris[,1:3] # Estas son las variables independientes
Petal.Width <- iris[,4] # Esta es la variable dependiente

# Ejecución de los componentes principales
ir.pca <- prcomp(iris.x, center = TRUE, scale. = TRUE)

PCA <- as.data.frame(ir.pca$x)
PC1 <- PCA[,1] # Componente principal 1
PC2 <- PCA[,2] # Componente principal 2
PC3 <- PCA[,3] # Componente principal 3

# Ejecución de la regresión lineal
fit <- lm(Petal.Width ~ PC1 + PC2 + PC3, data = PCA)

# Generando los datos para los diferentes tipos de gráficos
diagnostic <- diagnosticData(fit)
```

```
# Gráfico: Residuos v/s valores ajustados
ResidualsFitted(diagnostic, "Petal Width")
# Gráfico: Residuos estandarizados v/s valores ajustados
StResidualsFitted(diagnostic, "Petal Width")
# Gráfico: Normal QQ
NormalQQ(diagnostic, "Petal Width")
# Gráfico: Leverange v/s residuos estandarizados
StResidualsLeverange(diagnostic, "Petal Width")

# Gráficos con una paleta de colores diferentes
myPalette <- c("darkolivegreen4", "goldenrod1", "dodgerblue4")
ResidualsFitted(diagnostic, "Petal Width", colours = myPalette)
# Gráfico: Residuos v/s valores ajustados
StResidualsFitted(diagnostic, "Petal Width", colours = myPalette)
# Gráfico: Residuos estandarizados v/s valores ajustados
StResidualsLeverange(diagnostic, "Petal Width",colours = myPalette)
# Gráfico: Leverange v/s residuos estandarizados
```

3.4. Otros Gráficos

3.4.1. Gráfico del Codo: Elbow Plot

El gráfico del codo recibe su particular nombre debido al punto de inflexión que se produce. Este tipo de gráfico es utilizado en el análisis de componentes principales para detectar la cantidad de componentes o factores que explican gran parte o la mayoría de los datos. La figura 3.3 muestra los valores propios asociados con un componente o factor en orden descendente. En el ejemplo, hasta el componente 3 o 4 se explica la mayor parte de la variabilidad de los datos.

Ejemplo:

```
# Instalación y carga de la librería
install.packages("devtools")
devtools::install_github("mariytu/RegressionLibs")
library(RegressionLibs)
```

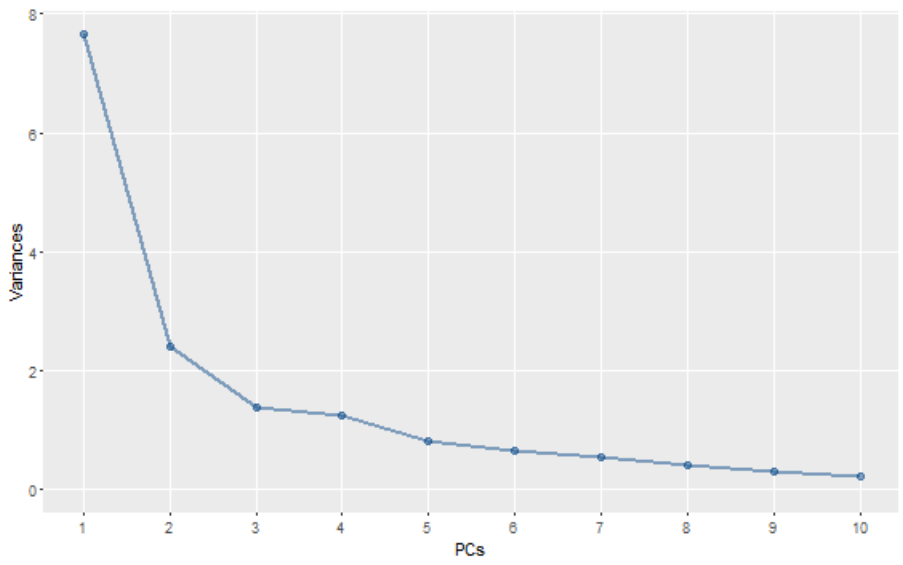


Figura 3.3: Gráfico del codo de los componentes principales del conjunto de datos *cars*

```
# Conjunto de datos limpio (Sin datos faltantes)
cars <- read.csv("https://dl.dropboxusercontent.com/u/12599702/
    autosclean.csv", sep = ";", dec = ",")
cars.x <- cars[,1:16] # Variables dependientes

# Ejecutando el análisis de componentes principales
cars.pca <- prcomp(cars.x, center = TRUE, scale. = TRUE)

# Gráfico del codo para detectar los componentes principales más
# importantes
elbowPlot(cars.pca)
```

3.4.2. Gráfico de Dispersión

El gráfico de dispersión permite identificar la existencia de correlación lineal entre dos variables. Esto se produce cuando los valores de una de ellas varía, la otra varía de forma sistemática. Si la correlación lineal es fuerte, entonces gráficamente deberíamos observar una nube de puntos formando una clara línea recta, en caso contrario la

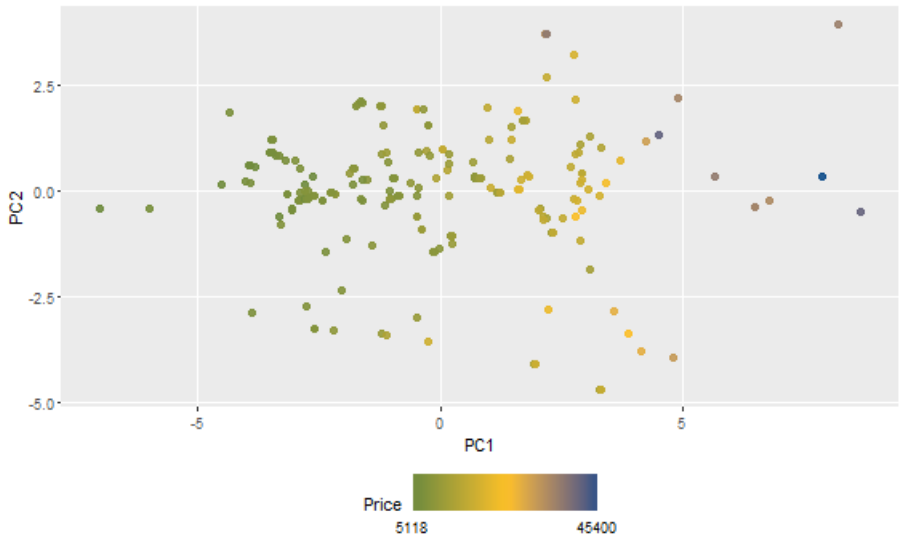


Figura 3.4: Gráfico de dispersión de los 2 primeros componentes principales del conjunto de datos *cars*

nube de puntos podría tomar una forma circular.

La figura 3.4 corresponde al gráfico de dispersión del ejemplo presentado a continuación en el que a partir de un conjunto de datos, estamos visualizando la correlación existente entre los 2 primeros componentes principales calculados para el conjunto de datos.

Ejemplo:

```
# Instalación y carga de la librería
install.packages("devtools")
devtools::install_github("mariytu/RegressionLibs")
library(RegressionLibs)

# Conjunto de datos limpio (Sin datos faltantes)
cars <- read.csv("https://dl.dropboxusercontent.com/u/12599702/
    autosclean.csv", sep = ";", dec = ",")
cars.x <- cars[,1:16] # Variables dependientes
cars.y <- cars[,17] # Variable independiente
```

```
# Ejecutando análisis de componentes principales
cars.pca <- prcomp(cars.x, center = TRUE, scale. = TRUE)

# Gráfico de los 2 primeros componentes principales
simplePlot(as.data.frame(cars.pca$x), cars.y, 1, 2, "Price", 2, 0.9)
# Gráfico con un paleta de colores diferente
myPalette <- c("darkolivegreen4", "goldenrod1", "dodgerblue4")
simplePlot(as.data.frame(cars.pca$x), cars.y, 1, 2, "Price", 2, 0.9,
    colours = myPalette)
```

3.4.3. Gráfico de Dispersión Matricial

El gráfico de dispersión matricial como el presentado en la figura 3.5 permite identificar la existencia de una correlación lineal entre múltiples variables. También son utilizados para diagnosticar el tipo de problema e identificar patrones en los datos.

Para problemas como el presentado en este proyecto, incluso este tipo de gráficos resultan inútiles debido a la cantidad de variables con las que se cuenta. Una forma de enfrentar esta situación es reduciendo la cantidad de variables dependientes utilizando técnicas como análisis de componentes principales (PCA) antes de crear el gráfico.

Ejemplo:

```
# Instalación y carga de la librería
install.packages("devtools")
devtools::install_github("mariytu/RegressionLibs")
library(RegressionLibs)

# Conjunto de datos limpio (Sin datos faltantes)
cars <- read.csv("https://dl.dropboxusercontent.com/u/12599702/
    autosclean.csv", sep = ";", dec = ",")
cars.x <- cars[,1:16] # Variables dependientes
cars.y <- cars[,17] # Variable independiente
```

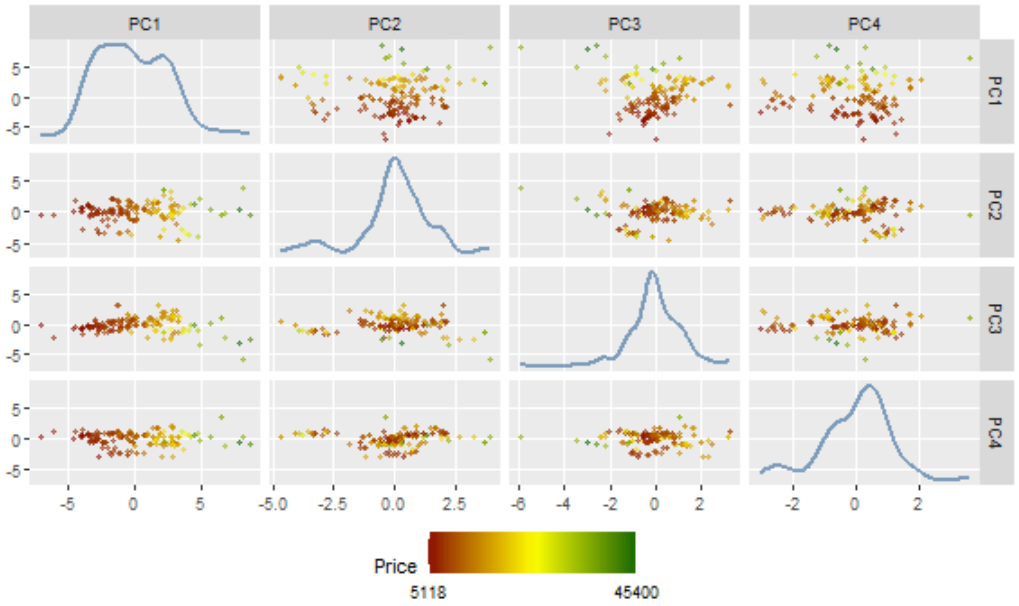


Figura 3.5: Gráfico de dispersión matricial de los primeros 4 componentes principales del conjunto de datos *cars*

```
# Gráfico de dispersión de algunas columnas
ScatterplotMatrix(cars.x, seq(2, 4, 1), cars.y, "Price", 2, 1)

# Ejecutando el análisis de componentes principales
cars.pca <- prcomp(cars.x, center = TRUE, scale. = TRUE)

# Gráfico de dispersión de los componentes principales más
# importantes
aux <- as.data.frame(cars.pca$x)
ScatterplotMatrix(aux, seq(1, 4, 1), cars.y, "Price")
```

3.4.4. Gráfico 3D

Los gráficos de dispersión son ampliamente utilizados como un estudio preliminar para enfrentar un problema. En este mismo sentido, los gráficos de dispersión matricial nos proporcionan un buen acercamiento para entender la correlación entre las variables.

Cuando se tiene pocas variables o cuando después de un proceso de reducción

de variables como análisis de componentes principales (PCA) se tienen pocas variables para el estudio, puede resultar oportuno utilizar un gráfico de dispersión 3D para obtener una idea más clara de la correlación existente entre todas o varias de variables.

Ejemplo:

```
# Instalación y carga de la librería
install.packages("devtools")
devtools::install_github("mariytu/RegressionLibs")
library(RegressionLibs)

# Conjunto de datos limpio (Sin datos faltantes)
iris.x <- iris[,1:4] # Variables dependientes
Species <- iris[,5] # Variable independiente

# Gráfico 3D de las 3 primeras variables
Plot3D(iris.x, c(1,2,3), Species)
# Ejecutando análisis de componentes principales
ir.pca <- prcomp(iris.x, center = TRUE, scale. = TRUE)
Plot3D(as.data.frame(ir.pca$x), c(1,2,3), Species)
```

3.4.5. Gráfico de Densidad

Un gráfico de densidad por definición describe la distribución de una única variable. En minería de datos, un gráfico de este tipo no sería muy útil si consideramos que en la gran mayoría de los problemas tendremos múltiples variables, pero puede ser utilizado para otros propósitos.

En este proyecto utilizamos este tipo de gráficos como el presentado en la figura 3.6 para mostrar la densidad del puntaje asignado a cada observación por el algoritmo de detección de datos atípicos. En este gráfico podemos apreciar como la gran mayoría de las observaciones (más del 90 %) tienen un puntaje cercano al 1.0 (entre 1 y 1.5), mientras que solo unos pocos, los datos atípicos, tienen puntajes mayores.

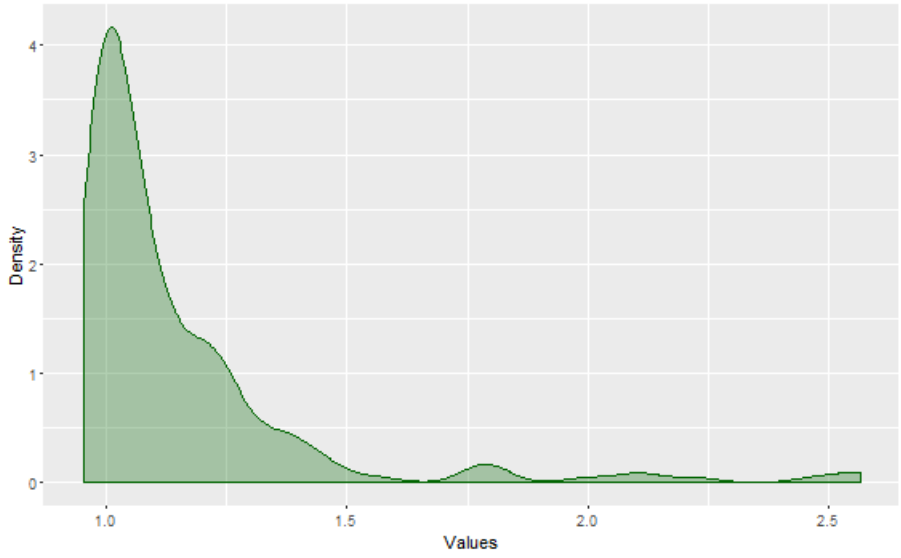


Figura 3.6: Gráfico de densidad para los puntajes resultantes de LOF para el conjunto de datos *cars*

Ejemplo:

```
# Instalación y carga de la librería
install.packages("devtools")
devtools::install_github("mariytu/RegressionLibs")
library(RegressionLibs)
install.packages("Rlof")
library(Rlof) #Librería de detección de datos atípicos
library(plyr)

# Conjunto de datos limpio (Sin datos faltantes)
cars <- read.csv("https://dl.dropboxusercontent.com/u/12599702/
    autosclean.csv", sep = ";", dec = ",")
cars.x <- cars[,1:16] # Variables dependientes
cars.y <- cars[,17] # Variable independiente

# Ejecutando detección de datos atípicos
outlier.scores <- lof(cars.x, k = c(5:10))
mean <- rowMeans(outlier.scores)
```

```
outlier.scores <- data.frame(outlier.scores, mean)

# Gráfico de densidad del puntaje de LOF
DensityPlot(outlier.scores, ncol(outlier.scores))

aux <- outlier.scores[,7]>1.7 #Umbral
count(aux)[2,2] #Total de datos atípicos
# Obteniendo datos que son atípicos
outliers <- order(outlier.scores[,7], decreasing=T)[1:count(aux)[2,2]]
#Obteniendo el puntaje de los datos atípicos
Score <- outlier.scores[outliers,7]
outliers <- data.frame(outliers,Score)
names(outliers) <- c("Position","Score")
View(outliers)
```

3.4.6. Ridge Plot

El algoritmo de regresión *Ridge* pertenece a la familia de regresores lineales. *Ridge* permite detectar la multicolinealidad dentro del modelo, es decir, es capaz de detectar aquellas columnas que presentan una alta correlación entre ellas y las penaliza. Este algoritmo esta pensado para trabajar con modelos que presentan sesgo, por lo mismo para la estimación se debe determinar con anterioridad el valor del sesgo que conocemos como λ.

En este proyecto se utilizo la metodología descrita por los autores de [17] para la optimización del parámetro (lambda), en la que mediante una grilla de 100 posibles valores de lambda en un rango de $[10^{-2}, 10^{10}]$ se realiza 10 veces validación cruzada para identificar el mejor lambda para ser utilizado en el modelo.

En la figura 3.7, se muestran todos los posibles valores de lambda junto con su error cuadrático medio (MSE). Como podemos apreciar en la figura 3.7 el mejor valor de lambda esta asociado a su vez con el MSE más pequeño.

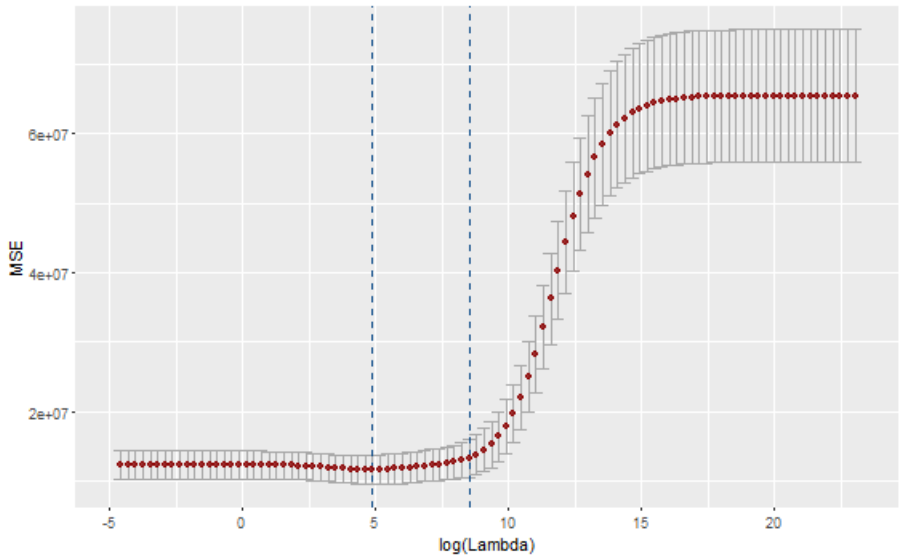


Figura 3.7: Gráfico de ejemplo de optimización del parámetro λ

Ejemplo:

```
# Instalación y carga de la librería
install.packages("devtools")
devtools::install_github("mariytu/RegressionLibs")
library(RegressionLibs)
install.packages("MASS")
library(MASS) #para la regresión ridge
install.packages("glmnet")
library(glmnet) #para la optimización de parámetros

# Conjunto de datos limpio (Sin datos faltantes)
cars <- read.csv("https://dl.dropboxusercontent.com/u/12599702/
    autosclean.csv", sep = ";", dec = ",")
cars.x <- cars[,1:16] # Variables dependientes
cars.y <- cars[,17] # Variable independiente

#Optimización de parámetros
#Se define una cuadrícula grande para los valores de lambda
grid <- 10^seq(10, -2, length = 100)
```

```
set.seed(2015)
#alpha = 0 para ridge, alpha = 1 para lasso
ridge <- cv.glmnet(as.matrix(cars.x), cars.y, alpha = 0, lambda = grid)
RidgePlot(ridge)
RidgePlot(ridge, errorMode = "ribbon")

bestLambda <- ridge$lambda.min #El lambda óptimo: 132.1941
ridge.final <- lm.ridge(cars.y ~ ., data = cars.x, lambda = bestLambda)
```

4. Metodología Experimental

En computación, y en particular en Ingeniería de Software, es frecuente el uso de una arquitectura para proporcionar una visión general del sistema como un todo. Esta arquitectura es un diagrama general que específica los módulos principales del sistema y su interacción. A pesar de que este trabajo no es explícitamente un desarrollo de software sino más bien se encuentra enfocado en una investigación científica, todo lo relacionado con la implementación y generación de código ha sido pensado y trabajado como un desarrollo tradicional en Ingeniería de Software.

En el campo de la minería de datos, muchos investigadores han reportado algunos módulos genéricos que son comúnmente utilizados en aplicaciones de minería de datos. La figura 4.1 muestra un diagrama de la arquitectura general para la predicción de rasgos fenotípicos a partir de datos de reflectancia espectral. La arquitectura que se presenta no es exhaustiva. La idea es proporcionar una serie de pasos que se realizan comúnmente en una aplicación de este tipo.

Desde el punto de vista científico, el uso de una plantilla, como la que se presenta en la figura 4.1, facilita la reproducción de los experimentos reportados por otros investigadores, un elemento esencial del método científico. Además, estos módulos permiten la descripción precisa de las diferentes técnicas utilizadas en el proceso de modelado y la secuencia de tiempo en que se aplican estas técnicas. Esta norma facilita a su vez, la comunicación de los resultados científicos en el campo de la minería de datos.

Vale la pena mencionar que no siempre se requieren de todos los componentes presentados en la figura 4.1, y en función de las necesidades de cada estudio de investigación, se puede agregar o quitar componentes, así como la modificación de los componentes existentes de acuerdo con el problema y modelos específicos utilizados.

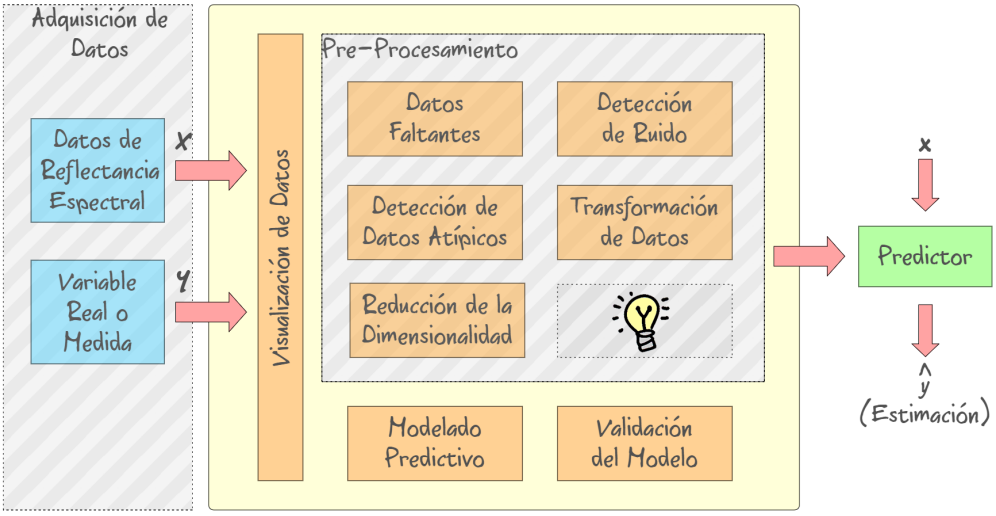


Figura 4.1: Arquitectura del modelado predictivo

A continuación se detalla cada uno de los módulos presentados en la figura 4.1 y se ejemplifica con el caso de estudio trabajado en esta investigación.

4.1. Adquisición de Datos

La fase de adquisición de datos, se refiere a la medición y/o obtención de los datos que queremos predecir. Esto se realiza normalmente con algún tipo de sensor o herramienta.

Por lo general, se incluye una descripción de los datos en su forma "en bruto", es decir, antes de realizar cualquier modificación. La idea es dar una visión general de la complejidad inicial de los datos. Con frecuencia, también se informa del lugar de obtención o descarga de estos, cuestiones de confidencialidad, cardinalidad y dimensionalidad original, tipos de datos (numéricos, ordinales, nominales) y en general, cualquier cosa que pueda ser de interés o para la comprensión de la situación.

En nuestro caso de estudio se evaluó un conjunto de 384 genotipos de trigo de pan en dos sitios mediterráneos de Chile: Cauquenes (35°58' S, 72°17' O; 177 m.s.n.m.) con estrés hídrico severo (SWS) y Santa Rosa (36°32' S, 71°55' O; 220 m.s.n.m.) bajo condiciones de riego completo (FI) y de estrés hídrico moderado (MWS).

El diseño experimental fue un α lattice con dos repeticiones, excepto en Cauquenes con una repetición. Los 384 genotipos fueron asignados al azar en 20 bloques

Cuadro 4.1: Conjuntos de datos de estudio

	2011		2012	
	AN	GF	AN	GF
FI	✓	✓	✓	✓
MWS	✓	✓	✓	✓
SWS	×	✓	✓	✓
Combinado	✓	✓	✓	✓

incompletos por repetición, cada bloque contiene 20 genotipos. Dos genotipos fueron incluidos ocho veces, generando en total 800 genotipos. Cada genotipo se sembró en parcelas de cinco hileras de 2 metros de largo y con una separación de 0,2 metros. Se considero una parcela como unidad experimental.

Las fechas de siembra fueron: 07 de septiembre en Cauquenes y 31 de agosto en Santa Rosa. Debido a que la fecha de siembra en Cauquenes fue mucho más tarde de lo típico para esa zona (finales de mayo), el estrés hídrico fue más severo de lo esperado.

Para cada una de estas parcelas, la reflectancia espectral se midió en un promedio de 3 disparos por parcela, que proporciona información relevante de la cosecha que son necesarios para el estudio [22]. Por ejemplo, los rangos del espectro 700-1.110nm entregar información sobre la absorción de luz por la clorofila y pigmentos asociados a la cosecha. En este trabajo, los rangos del espectro necesarios para el estudio corresponde a 350-2500nm (2150) atributos.

Cada conjunto de datos generado fueron separados de acuerdo a su ubicación (Cauquenes / Santa Rosa), el régimen hídrico (SWS / MWS / FI), etapa fenológica (Antesis (AN) / llenado de grano (GF)) y el año de la cosecha (2011/2012). Para cada conjunto de datos, se midieron 13 variables de estudio diferentes del tipo: fisiológica, morfológica y de rendimiento.

En total se generaron 15 conjuntos de datos que se resumen en el cuadro 4.1. Para el año 2011 en Cauquenes para un estrés hídrico severo, no se realizaron mediciones de espectro en la primera fecha de medición (antesis).

4.2. Variables de Estudio

Para cada conjunto de datos generado, junto con los valores de reflectancia espectral, se midieron 13 características morfo-fisiológicas que se detallan a continuación:

- Componentes de Rendimiento
  - El número de espigas por metro cuadrado (Espigas/m^2)
  - El número de granos por espiga (Granos/espiga)
  - El peso de 1000G, tomando 25 espigas al azar (Peso 100G)
  - El rendimiento del grano cosechando la parcela de $2m^2$ (Grain Yield t/ha)
- Contenido de Clorofila: Se midió el contenido de clorofila en 5 hojas por parcela (SPAD 1, SPAD 2, SPAD 3).
- Carbohidratos Hidrosolubles: Se determino la concentración de carbohidratos hidrosolubles en tallos durante antesis (CHOa mg/tallo y CHOa mg/g) y madures (CHOm mg/tallo y CHOm mg/g) sobre cinco tallos por parcela.
- Discriminación de Isótopos de Carbono ($\Delta^{13}C$)
- Índice de Área Foliar (IAF)

4.3. Pre-procesamiento

Esta fase consiste en realizar una serie de modificaciones a los datos originales con el fin de facilitar su procesamiento. Dependiendo de la situación, se deben aplicar una variedad de técnicas tales como normalización, discretización, binarización, rotación, reducción de la dimensionalidad, selección de atributos, detección de datos atípicos, detección de datos faltantes, detección de ruido, entre otros. Cada uno de estos sub-problemas se solucionan con una familia de algoritmos, respectivamente, y cada investigador debe indicar que algoritmo en particular ha seleccionado, justificando su decisión cuando sea necesario.

La plantilla de pre-procesamiento propuesta en la figura 4.1 es sólo una sugerencia, lo que significa que alguno de los componentes podrían ser obviados o incluso se podrían agregar componentes al sistema para dar solución a sub-problemas puntuales del caso de estudio, idea que esta representada con la caja vacía con una ampolleta.

Ahora vamos a describir los principales componentes de la fase de pre-procesamiento usando como ejemplo el caso de estudio desarrollado en este proyecto.

4.3.1. Datos Faltantes

El escenario ideal para cualquier investigador sería tener un conjunto de datos perfectos, es decir sin valores faltantes, sin embargo esto es casi difícil de imaginar que ocurra en la realidad. Lo más probable es que el conjunto de datos este incompleto, por lo que los científicos deben enfrentarse a la necesidad de decidir si modificar el conjunto de datos o no. Cuando se opta por la modificación de los valores faltantes, se pueden reemplazar aquellos datos con información estadística recopilada a partir de los datos disponibles o simplemente eliminar aquellas instancias u observaciones que contienen valores faltantes.

Por otro lado, existen casos en que a pesar de la existencia de datos faltantes ya sea en las variables independientes como dependientes, los algoritmos utilizados en los subsecuentes pasos o en el modelo de predicción no son influenciados por dichos problemas y contienen mecanismos para tratar con ellos.

En este caso de estudio, el dispositivo de reflectancia espectral que captura las mediciones de campo para un determinado rango de frecuencias proporciona la garantía de que todos los valores están presentes en las variables independientes. Sin embargo, no ocurre lo mismo con las variables dependientes, porque, entre otras razones, son propensos a errores humanos o requieren un tamaño de muestra que no esta disponible. Como estamos interesados en el análisis de regresión, las variables dependientes son cruciales para la construcción y ajuste de los modelos. Por esta razón, se decidió excluir todas las instancias que posean uno o más valor faltante en cualquiera de sus variables dependientes. Lo anterior, corresponde aproximadamente a un 5 % de los datos por conjunto de datos.

En este caso, somos afortunados de contar con una cantidad de instancias relativamente altas por conjunto de datos, y por ende podemos asumir los riesgos de eliminar algunas de ellas.

4.3.2. Detección de Ruido

Cuando se trabaja con datos de reflectancia espectral, es común obtener datos ruidosos en rangos específicos del espectro. Este ruido se produce principalmente por la humedad presente en la atmósfera produciendo alteraciones en la reflectancia. Para los datos trabajados en este proyecto, se pueden apreciar 3 zonas del espectro que contienen ruido, que son apreciables en la figura 4.2 como áreas donde las longitudes

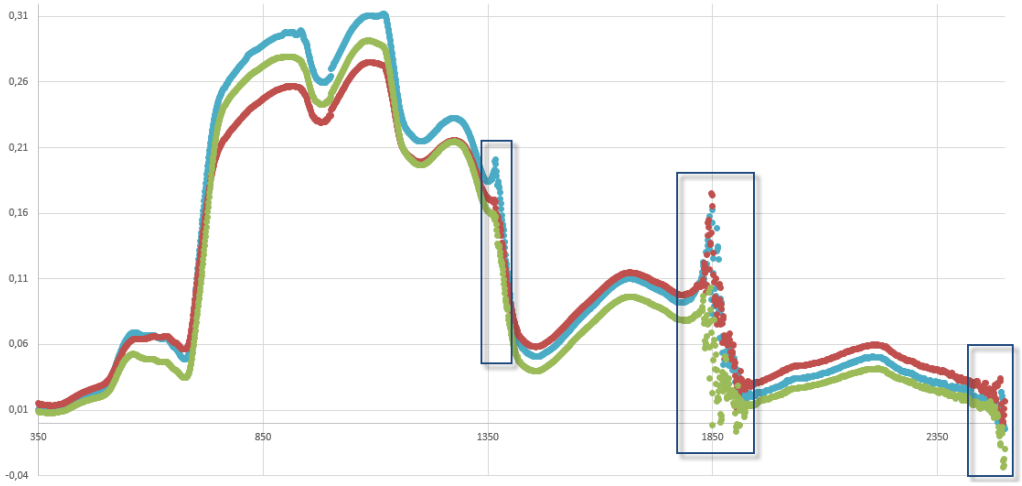


Figura 4.2: Zonas comunes con ruido en los datos de reflectancia espectral

de onda tienen cambios abruptos.

Aquellas zonas cercanas a las longitudes de onda 1350nm, 1850nm y 2500nm es común encontrar ruido, sin embargo podrían existir otras zonas de ruido particulares para una instancia que se producen debido a una mala manipulación o calibración del equipo de medición. En estos casos, no es posible determinar una zona específica de ruido, por lo que una solución un poco más general debe proponerse.

La solución propuesta ha consistido en el desarrollo de un algoritmo *ad-hoc* para la detección del ruido basado en un umbral diferencial, como se explica en los siguientes pasos:

1. Para cada longitud de onda, se calcula la diferencia discreta entre la longitud de onda y su antecesor usando la ecuación 4.1, para obtener una nueva matriz con las mismas dimensiones que la matriz original de datos.

$$T_{i,j} = \begin{cases} 0 & j = 0 \\ \frac{\mathcal{X}_{i,j} - \mathcal{X}_{i,j-1}}{\mathcal{X}_{i,j-1}} & j > 0 \end{cases} \tag{4.1}$$

   donde, $\mathcal{X}$ corresponde a la matriz de datos de reflectancia espectral, el índice i se refiere una una observación en particular y el índice j representa una longitud de onda específica.

2. Identificar y eliminar todas las longitudes de onda que sean mayores a un umbral (Θ). Dicho umbral indica el nivel de sensibilidad del algoritmo para

detectar ruido. Un valor de $\Theta = 0$ indica que el algoritmo no detectará ruido, mientras que a valores cercanos a 1 indica una alta susceptibilidad al ruido y podría considerar como ruido zonas que no lo son.

En este proyecto hemos seleccionado $\Theta = 0{,}15$ y encontramos dos fuentes de ruido muy marcadas cercanas a los 1830-1940nm y una segunda alrededor de la frecuencia 2500nm. Estas zonas de ruido fueron identificadas y eliminadas automáticamente para luego validar el proceso mediante una inspección visual.

El tiempo de ejecución depende del número de instancias que estemos procesando y también de la dimensionalidad del conjunto de datos. Para este proyecto, los experimentos fueron procesados en un ordenador portátil promedio (procesador i3, 8 GB de RAM y disco duro SSD), tomando aproximadamente 5 minutos para procesar un conjunto de datos de 1600 instancias y 2150 dimensiones. Si bien el proceso parece un poco lento, se podría optimizar la implementación para reducir el tiempo de ejecución y eventualmente realizar un mejor manejo del espacio en memoria, sin embargo es algo que no se abarcará en este proyecto.

4.3.3. Detección de Datos Atípicos

Los datos derivados de sensores pueden contener datos erróneos debido a la calibración de los equipos o por error en las mediciones. La mayoría de los estadísticos como el promedio, la desviación estándar y todos los estadísticos basados en ellos, son altamente sensibles a los datos atípicos. Debido a que las medidas de rendimiento utilizadas en este proyecto se basan en estos estadísticos, todos nuestros modelos se verían afectados por los valores atípicos y pueden realmente estropear nuestro análisis.

Afortunadamente, la cantidad de datos anormales en los conjuntos de datos son bajos, por lo que en la mayoría de los casos se sugiere extraerlos y con ello mejorar los resultados en las tareas de análisis, sin embargo se debe tener cierto cuidado con esto, porque podríamos estar eliminando una legitima observación y son a veces las más interesantes. Es importante investigar la naturaleza del valor atípico antes de tomar una decisión y por ende cada caso debería tratarse de forma única.

Para esta tarea, resulta más fácil y sencillo apoyarse de los gráficos de diagnostico presentados en la sección 3.3, los que basados en rango amplio de estadísticos permiten clarificar que tan atípico es un valor o no.

En resumen, el objetivo de esta fase es identificar aquellas instancias que son muy diferentes a los patrones esperados. Como se menciono anteriormente, muchos de los modelos de predicción son susceptibles a la presencia de datos atípicos y una práctica común es identificarlos y eliminarlos durante la fase de pre-procesamiento. Estos datos son quienes afectan los resultados del modelo, por lo que al eliminarlos evidentemente los resultados se verán levemente mejorados.

En nuestro caso de estudio eliminamos los datos atípicos utilizando un método bien conocido llamado factor de valor atípico local (LOF, del inglés *Local Outlier Factor*). El método se basa en la densidad de la población, en la que se calcula la desviación local de una instancia respecto de sus k vecinos más cercanos, por lo que si una instancia i se desvía considerablemente de la proximidad de sus vecinos, entonces se considera un valor atípico. En otras palabras, una instancia que se encuentra sola en el espacio o con muy pocos vecinos respecto de la población total se considera un valor atípico. La distancia en este caso es una función matemática que mide la disimilitud entre dos observaciones. Basado en esta distancia, LOF asigna una puntuación a cada instancia, indicando el grado de ser un dato atípico. Objetivamente, instancias con valores de $LOF \approx 1$ pertenecen a una región con una densidad homogénea alrededor de el y sus vecinos, mientras que puntos con valores de $LOF \gg 1$ son considerados atípicos (Véase sección 2.4.2).

Una de las ventajas de LOF es que corresponde a un método no supervisado. Esto significa que el método de detección de datos atípicos se basa únicamente en las variables independientes, y por tanto se puede aplicar antes de la construcción y ajuste del modelo de predicción.

La figura 4.3 muestra un gráfico de LOF para un conjunto de datos de trigo (Santa Rosa, estrés leve de agua, Antesis, 2011). El eje x representa los diferentes valores para el puntaje de LOF y el eje y muestra la densidad. Del gráfico 4.3 se puede observar que la mayoría de los casos poseen una puntuación de LOF cercana a 1.0 aproximadamente. La cola del gráfico muestra algunas instancias con un puntaje de LOF bastante alto, y que probablemente corresponden a los datos atípicos. En este caso, y basado en la inspección visual, se selecciono un umbral de Θ_{LOF} =1.5 y luego se eliminaron las instancias que poseen un valor de Θ por encima de esta cantidad. Como era de esperar, la cantidad de instancias eliminadas por efecto de este proceso resulto ser menor al 0.004 % de los datos, lo que significo unas 2 o 3 instancias por conjunto de datos.

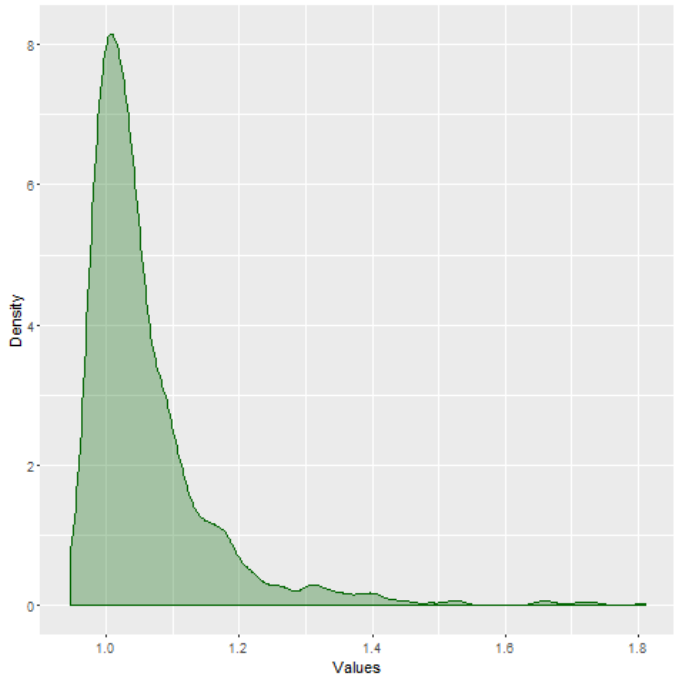


Figura 4.3: Gráfico de densidad con las puntuaciones asignadas por el algoritmo LOF

4.3.4. Normalización de Datos

La mayoría de los algoritmos de regresión, en particular los utilizados en esta investigación, son sensibles a los rangos de las variables. Por ejemplo, aquellos atributos que varían entre $[0, 100]$ tendrán mayor importancia en comparación con aquellas variables que varían en un intervalo $[0, 1]$. En principio, cuando no se cuenta con información adicional, se supone que todas las variables tienen igual relevancia. Por esta razón, es frecuente que los rangos de cada variable se ajusten a una forma canónica. En esta investigación se aplicó la llamada normalización unitaria. Sea $Z = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T$ la matriz de datos, la normalización en el rango $[0, 1]$ de una variable Z_{ij} puede ser calculado aplicando la ecuación 4.2.

$$Z'_{ij} = \frac{Z_{ij} - \text{mín}_{k=0..n}\{Z_{kj}\}}{\text{máx}_{k=0..n}\{Z_{kj}\} - \text{mín}_{k=0..n}\{Z_{kj}\}} \tag{4.2}$$

La matriz de datos normalizada, Z', se obtiene calculando Z'_{ij} para todos los i, j.

4.3.5. Reducción de la Dimensionalidad

Cuando el número de atributos es demasiado alto no es posible aprender de los datos directamente, la calidad de la solución es pobre o el tiempo necesario para obtener la solución no es práctica. Con frecuencia, la situación mencionada se produce por un fenómeno conocido como la "maldición de la dimensionalidad". Esto se produce cuando la cantidad de atributos o dimensiones es alta y por tanto es complejo diferenciar si dos observaciones son cercanas o distantes en el espacio. Una solución a este problema es reducir el número de dimensiones. Esta reducción se puede lograr de muchas maneras, por ejemplo, descartando aquellos atributos que son irrelevantes. Otros métodos combinan la información de los atributos formando factores que condensan la información en menos atributos como análisis de componentes principales (PCA), que ha sido detallado y explicado en la sección 2.4.1.

En este caso, ejecutamos la conocida técnica llamada PCA como método para reducir la cantidad de atributos para algunos de los algoritmos aplicados en esta investigación (MLR y SVR). Mediante la aplicación de PCA, la información original contenidas en 2150 dimensiones, se concentraron y redujeron a tan solo 3 características. Estas corresponden a transformaciones lineales de los atributos originales y son ortogonales, explicando el 95 % de la varianza original.

La figura 4.4 muestra un gráfico del codo para un conjunto de datos de trigo de Santa Rosa con régimen hídrico de riego completo medido en la etapa fenológica de llenado de grano para el año 2011. El eje x representa cada uno de los componentes principales obtenidos después de aplicar el algoritmo PCA y el eje y muestra la proporción de la varianza total explicada por cada componente. Como se aprecia se produce una inflexión notoria en el tercer componente principal y afortunadamente para este proyecto, en cada conjunto de datos estudiado se mantuvo la misma característica.

4.4. Modelado Predictivo

Para cada conjunto de datos se utilizaron 4 modelos de predicción: Multiple Linear Regression (MLR), Partial Least Squares Regression (PLS), Ridge Regression y Support Vector Regression (SVR). El funcionamiento de cada uno de los algoritmos fueron explicados en las secciones 2.3.2, 2.3.3, 2.3.4 y 2.3.5 respectivamente. En esta

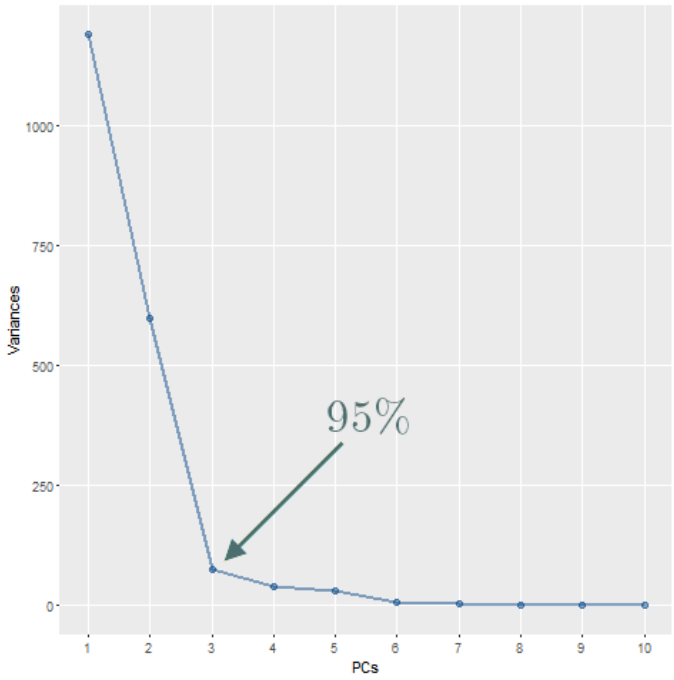


Figura 4.4: Gráfico del codo de los componentes principales de un conjunto de datos

sección, explicaremos como cada uno de estos modelos fueron ajustados y utilizados para predecir en esta investigación.

Multiple Linear Regression Como parte de la preparación de los datos, se utilizo PCA para reducir la cantidad de variables a utilizar en la predicción. De los componentes resultantes del proceso, estos fueron ordenados desde el que más aporta información para explicar los datos hasta el que menos aporta. De esta forma, en base a una primera inspección visual mediante el gráfico del codo, se seleccionaron los 3 primeros componentes principales para generar el modelo.

Partial Least Squares Regression A diferencia de MLR, PLS no requiere una previa reducción de la dimensionalidad de los datos, pues este tiene un mecanismo interno basado en PCA que realiza esta simplificación. PLS por otro lado, construye un conjunto de combinaciones lineales de predictores en el espacio $\mathcal{X}$ que explican y. Esta técnica también pondera los componentes resultantes por la relevancia para y [14] y utilizamos los 3 primeros componentes para la generación del modelo.

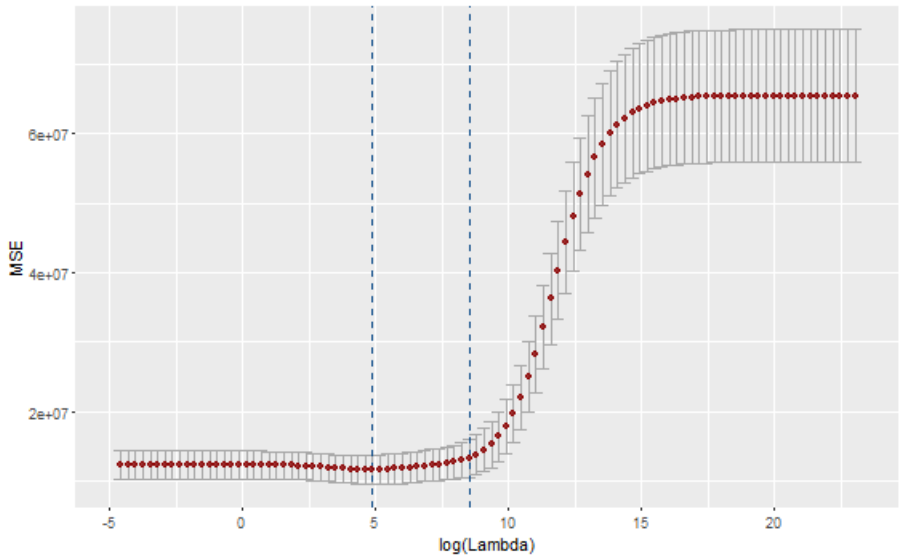


Figura 4.5: Gráfico de optimización del parámetro λ

Ridge Regression Ridge Regression a diferencia de los dos modelos anteriores, requiere de la optimización de un parámetro λ. La utilización de este parámetro se utiliza para penalizar el tamaño de los coeficientes del modelo [14]. Para la optimización del parámetro λ, se utilizo la metodología descrita por los autores de [17], en la que mediante una grilla de 100 valores posibles para λ en un rango de $[10^{-2}, 10^{10}]$, se realizaron 10 veces validación cruzada. Después de todas estas ejecución, el mejor λ es identificado y utilizado como parámetro para el modelo.

En la figura 4.5, se muestran todos los posibles valores de lambda junto con su error cuadrático medio (MSE). Como podemos apreciar en la figura 4.5 el mejor valor de lambda esta asociado a su vez con el MSE más pequeño.

Support Vector Regression Support Vector Regression transforma los datos en una nueva dimensión mediante el uso de un "kernel trick", por lo que es posible encontrar un hiperplano que separa los datos en grupos [14]. En este sentido, probamos para cada conjunto de datos 4 "kernels" diferentes (lineal, polinomial, radial y sigmoidal) y seleccionamos en cada caso el mejor basado la minimización del error y la maximización del coeficiente de determinación R^2. El modelo también considera la optimización de 2 parámetros llamados *Costo* y *Epsilon*, los cuales fueron utilizados con su valores por defecto en 1 y 0.1 respectivamente.

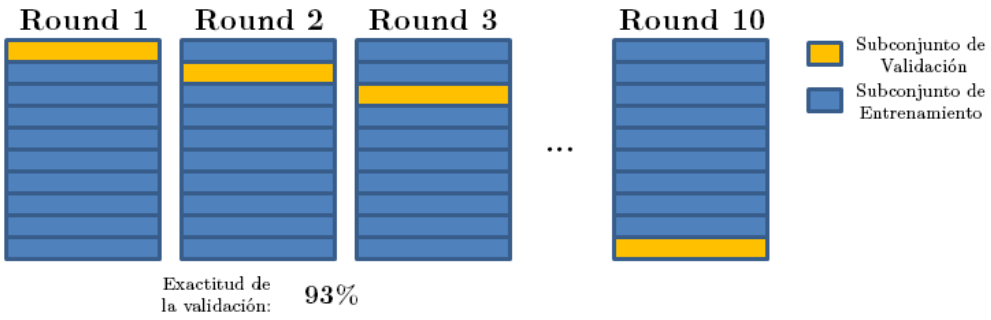


Figura 4.6: Ejemplificación del funcionamiento de 10 veces validación cruzada

4.5. Validación del Modelo

De forma indistinta al modelo utilizado en una investigación, medir el rendimiento de este es importante porque nos da una medida cuantitativa de la calidad del modelo seleccionado. Después de la etapa de entrenamiento y ajuste de un modelo podemos calcular fácilmente todos los estadísticos que queramos a partir de los valores reales (y) y predichos ($\hat{y}$) por el modelo. En este contexto, estos estadísticos se conocen como "sobre-estimados" y no son capaces de demostrar la efectividad de la predicción de un modelo. Es de importancia para nuestro estudio medir la efectividad de un modelo respecto de su capacidad para predecir datos de pruebas independientes, que corresponden a instancias que son desconocidas para el modelo y que no fueron utilizadas durante el proceso de entrenamiento.

Cuando el número de instancias para ajustar el modelo es pequeño, es difícil estimar el error de prueba del modelo, por lo que un número de técnicas se pueden utilizar para este fin [17]. En particular en este proyecto, se utilizo 10 veces validación cruzada.

Se divide el conjunto de datos en 10 subconjuntos iguales. En cada validación, un subconjunto será usado como prueba y los restantes 9 como entrenamiento. Con esta metodología estamos completamente seguros de que una observación participará como prueba en una validación y formará parte del subconjunto de entrenamiento en nueve validaciones como se ejemplifica en la figura 4.6.

En cada proceso de validación dejaremos evidentemente 1 subconjunto como prueba, mientras que con los 9 restantes optimizaremos y ajustaremos el modelo seleccionado. Hecho esto, podemos calcular y/o estimar los $\hat{y}$ para el subconjunto de

prueba. Una vez que hemos repetido el proceso anterior 10 veces, cada una de las observaciones x_i habrá sido estimada por un modelo ajustado y por tanto conocemos su $\hat{y}$ respectivo. Con dichos datos, ahora es posible calcular cualquier estadístico que sea de importancia para el estudio, en particular RSE, R^2 e IA.

5. Resultados

Las tablas 5.1 y 5.2 muestran un análisis descriptivo para el desempeño de todos los rasgos evaluados con los cuatro modelos: MLR, PLS, Regresión Ridge y SVR, para el 2011 y 2012, respectivamente. Debido al gran número de genotipos y los tres ambientes con disponibilidad de agua contrastante, fue posible obtener una gran amplitud en el rendimiento de los rasgos evaluados, tanto dentro de cada ambiente particular como para el entorno combinado.

Los rasgos que presentaron mejores resultados promedios de R^2 fueron: Rendimiento (t/h), IAF y $\Delta 13C$ para los 4 modelos analizados, con valores de 0,73, 0,63 y 0,52 respectivamente para el año 2011. En general con SVR se obtienen los mejores resultados promedios, sin embargo con Regresión Ridge también se obtienen resultados promedios cercanos a los obtenidos con SVR con diferencias de 0,04. En comparación con MLR y PLS, la diferencia de los resultados obtenidos versus SVR es más notoria y se producen diferencias de 0,14 y 0,11 puntos para el rendimiento (t/h).

El mismo experimento realizado con datos para el 2012 mantiene la tendencia de los resultados de los rasgos, sin embargo están por debajo de lo esperado. En la tabla 5.2 se presentan el análisis descriptivo de los resultados obtenidos para el periodo 2012. Los valores promedios de R^2 corresponden a 0,4, 0,39 y 0,43 para rendimiento (t/h), IAF y $\Delta 13C$, respectivamente. Además de obtener resultados promedios que están muy por debajo de lo obtenido en comparación al año 2011, la desviación de los resultados mismos aumenta considerablemente, obteniendo diferencias de 0,82, 0,28 y 0,84 entre los valores mínimos y máximos para los rasgos mencionados anteriormente.

Cuadro 5.1: Media y desviación estándar del estadístico R^2 para los cuatro modelos con datos del 2011

	MLR				**PLS**				**Ridge**				**SVR**			
Rasgo	$\bar{x}$	σ	Min	Max	$\bar{x}$	σ	Min	Max	$\bar{x}$	σ	Min	Max	$\bar{x}$	σ	Min	Max
SPAD 1	0.09	0.03	0.05	0.13	0.13	0.05	0.07	0.18	0.21	0.08	0.11	0.30	0.21	0.09	0.07	0.33
SPAD 2	0.19	0.15	0.03	0.42	0.25	0.14	0.06	0.42	0.38	0.18	0.12	0.53	0.40	0.21	0.12	0.62
SPAD 3	0.21	0.22	0.06	0.36	0.32	0.22	0.17	0.48	0.45	0.16	0.33	0.56	0.41	0.14	0.31	0.51
Rendimiento (t/h)	0.59	0.18	0.38	0.87	0.62	0.17	0.44	0.88	0.69	0.19	0.45	0.93	**0.73**	0.17	0.52	0.93
IAF	0.54	0.17	0.27	0.73	0.55	0.17	0.28	0.73	0.59	0.19	0.27	0.75	**0.63**	0.17	0.36	0.76
CHOa mg/g	0.11	0.08	0.05	0.25	0.11	0.07	0.04	0.25	0.11	0.08	0.02	0.26	0.10	0.07	0.03	0.20
CHOm mg/g	0.05	0.05	0.02	0.15	0.05	0.05	0.02	0.16	0.07	0.08	0.01	0.22	0.06	0.09	0.00	0.24
CHOa mg/tallo	0.10	0.05	0.05	0.17	0.11	0.04	0.06	0.18	0.13	0.04	0.09	0.19	0.11	0.03	0.08	0.15
CHOm mg/tallo	0.05	0.03	0.01	0.08	0.06	0.03	0.01	0.09	0.08	0.04	0.01	0.13	0.06	0.05	0.01	0.14
$\Delta 13C$	0.43	0.28	0.10	0.80	0.47	0.29	0.11	0.80	0.48	0.34	0.12	0.89	**0.52**	0.34	0.14	0.90
Espigas/m^2	0.31	0.09	0.15	0.44	0.33	0.09	0.17	0.44	0.43	0.09	0.30	0.53	0.39	0.10	0.24	0.52
Granos/espiga	0.09	0.10	0.01	0.29	0.12	0.08	0.06	0.30	0.17	0.09	0.10	0.36	0.16	0.10	0.08	0.36
Peso 100G	0.24	0.15	0.04	0.43	0.25	0.15	0.04	0.43	0.37	0.20	0.01	0.55	0.39	0.20	0.03	0.57

Cuadro 5.2: Media y desviación estándar del estadístico R^2 para los cuatro modelos con datos del 2012

	MLR				**PLS**				**Ridge**				**SVR**			
	$\bar{x}$	σ	Min	Max	$\bar{x}$	σ	Min	Max	$\bar{x}$	σ	Min	Max	$\bar{x}$	σ	Min	Max
SPAD 1	0.16	0.21	0.00	0.51	0.23	0.20	0.06	0.55	0.27	0.21	0.07	0.61	0.27	0.23	0.07	0.66
SPAD 2	0.13	0.16	0.00	0.42	0.18	0.15	0.04	0.43	0.20	0.15	0.07	0.44	0.20	0.17	0.05	0.51
SPAD 3	0.03	0.03	0.01	0.07	0.13	0.03	0.11	0.17	0.20	0.06	0.12	0.25	0.17	0.05	0.11	0.23
Rendimiento (t/h)	0.26	0.31	0.02	0.83	0.30	0.31	0.06	0.85	0.37	0.36	0.06	0.92	0.40	0.34	0.11	0.93
IAF	0.24	0.08	0.14	0.36	0.30	0.09	0.16	0.41	0.33	0.10	0.17	0.45	0.39	0.09	0.23	0.51
CHOa mg/g	0.14	0.14	0.02	0.37	0.17	0.15	0.03	0.41	0.23	0.19	0.07	0.49	0.22	0.20	0.06	0.49
CHOm mg/g	0.01	0.01	0.00	0.03	0.02	0.01	0.01	0.04	0.04	0.02	0.01	0.07	0.03	0.02	0.01	0.06
CHOa mg/tallo	0.19	0.14	0.03	0.40	0.22	0.19	0.03	0.46	0.29	0.22	0.07	0.56	0.27	0.22	0.06	0.56
CHOm mg/tallo	0.02	0.03	0.00	0.08	0.03	0.03	0.00	0.09	0.05	0.04	0.02	0.13	0.04	0.03	0.02	0.08
$\Delta 13C$	0.27	0.35	0.01	0.81	0.31	0.36	0.02	0.85	0.40	0.41	0.10	0.93	0.43	0.40	0.10	0.94
Espigas/m^2																
Granos/espiga	0.08	0.08	0.01	0.21	0.10	0.07	0.01	0.21	0.14	0.07	0.06	0.24	0.15	0.08	0.05	0.26
Peso 100G	0.15	0.10	0.02	0.29	0.17	0.08	0.06	0.29	0.23	0.09	0.08	0.33	0.25	0.11	0.07	0.38

De los 4 modelos de regresión ajustados en esta investigación el que presento las mejores predicciones para distintos rasgos se lograron utilizando SVR y regresión Ridge, especialmente en el ambiente combinado (tabla 5.1). Por el contrario, las predicciones obtenidas con los modelos MLR y PLS no obtuvieron resultados tan satisfactorios en casi todas las características analizadas en esta investigación.

En base a todos los resultados obtenidos que se detallan en las tablas A.1 y A.2 del anexo A se generó un subconjunto de las tablas para las 3 características que presentaron los mejores resultados promedios de R^2 y cuyos valores son superiores a 0,5. Rendimiento, IAF y $\Delta 13C$ mostraron los mejores resultados de predicciones (R^2, IA y RMSE) tanto en ambientes individuales como en el ambiente combinado. Las tablas 5.3 y 5.4 presentan todos los resultados obtenidos para cada una de los conjuntos de datos analizados para estos 3 rasgos tanto para el año 2011 como el 2012, respectivamente.

Todos los modelos de regresión usando los datos espectrales de la fase de llenado de grano (GF, del inglés *Grain Filling*) en ambiente combinado mostraron $R^2 > 0{,}87$ para rendimiento durante el 2011 y $R^2 > 0{,}83$ para la misma característica en el año 2012 (tablas 5.3 y 5.4). Sin embargo, SVR y regresión Ridge presentaron el mejor R^2 (0,93) cuando se utilizan los datos espectrales de la fase de llenado de grano en un ambiente combinado, tanto para el 2011 como el 2012. Para el ambiente individual, la predicción de rendimiento en MWS mostró $R^2 > 0{,}79$ y $R^2 > 0{,}58$ con los datos de llenado de grano y antesis respectivamente, durante el 2011. FI, por otro lado, presentó el R^2 más bajo en ambos estadios fenológicos (tablas 5.3 y 5.4).

Las predicciones de IAF para cualquier método de regresión utilizado presentó un $R^2 > 0{,}57$ en ambos estadios fenológicos, con MWS y ambiente combinado para el 2011. Nuevamente SVR mostró los mejores resultados con un $R^2 > 0{,}70$ (tabla 5.3). El mismo experimento con datos espectrales para el año 2012 no obtuvo los mismos resultados para IAF. Durante el 2012, solo se obtuvo un $R^2 > 0{,}5$ con los datos de reflectancia espectral en fase de antesis para el ambiente combinado (tabla 5.4).

Cuadro 5.3: Detalle de los resultados obtenidos para los 3 rasgos con mejores rendimiento durante el 2011

			MLR			**PLS**			**Ridge**			**SVR**		
*	**	**Ambiente**	**RMSE**	**R2**	**IA**	**RMSE**	**R2**	**IA**	**RMSE**	**R2**	**IA**	**RMSE**	**R2**	**IA**
Rendimiento (t/h)	AN	SWS												
		MWS	1.37	**0.58**	0.85	1.23	**0.66**	0.89	1.11	**0.74**	0.93	0.95	**0.80**	0.94
		FI	1.29	0.38	0.74	1.23	0.44	0.77	1.19	**0.54**	0.85	1.12	**0.57**	0.86
		Combinado	1.52	**0.51**	0.81	1.47	**0.54**	0.83	1.01	**0.79**	0.94	0.92	**0.82**	0.95
	GF	SWS	0.43	**0.54**	0.84	0.43	**0.54**	0.83	0.49	0.48	0.83	0.46	**0.52**	0.85
		MWS	0.96	**0.79**	0.94	0.93	**0.81**	0.94	0.79	**0.86**	0.96	0.71	**0.89**	0.97
		FI	1.23	0.44	0.77	1.21	0.46	0.79	1.35	0.45	0.82	1.10	**0.57**	0.87
		Combinado	1.03	**0.87**	0.97	1.02	**0.88**	0.97	0.77	**0.93**	0.98	0.74	**0.93**	0.98
IAF	AN	SWS												
		MWS	0.87	**0.66**	0.89	0.84	**0.68**	0.90	0.77	**0.75**	0.93	0.75	**0.76**	0.93
		FI	0.86	0.38	0.73	0.85	0.40	0.75	0.89	0.44	0.81	0.81	0.49	0.83
		Combinado	0.89	**0.57**	0.85	0.88	**0.58**	0.85	0.75	**0.71**	0.92	0.71	**0.73**	0.92
	GF	SWS												
		MWS	0.78	**0.73**	0.92	0.78	**0.73**	0.92	0.83	**0.71**	0.92	0.76	**0.75**	0.93
		FI	0.91	0.27	0.65	0.90	0.28	0.66	1.02	0.27	0.72	0.89	0.36	0.76
		Combinado	0.84	**0.60**	0.86	0.84	**0.60**	0.86	0.81	**0.65**	0.89	0.74	**0.70**	0.91
$\Delta 13C$	AN	SWS												
		MWS	0.89	**0.57**	0.85	0.76	**0.68**	0.90	0.83	**0.66**	0.90	0.71	**0.74**	0.92
		FI	0.66	0.17	0.53	0.66	0.18	0.54	0.81	0.12	0.61	0.75	0.14	0.62
		Combinado	0.85	0.46	0.78	0.81	**0.52**	0.82	0.74	**0.62**	0.88	0.69	**0.65**	0.89
	GF	SWS	0.58	0.10	0.43	0.57	0.11	0.47	0.69	0.12	0.61	0.62	0.15	0.63
		MWS	0.70	**0.74**	0.92	0.69	**0.75**	0.92	0.63	**0.80**	0.94	0.52	**0.85**	0.96
		FI	0.66	0.21	0.58	0.66	0.21	0.58	0.90	0.14	0.61	0.70	0.21	0.67
		Combinado	0.80	**0.80**	0.94	0.79	**0.80**	0.94	0.60	**0.89**	0.97	0.56	**0.90**	0.97

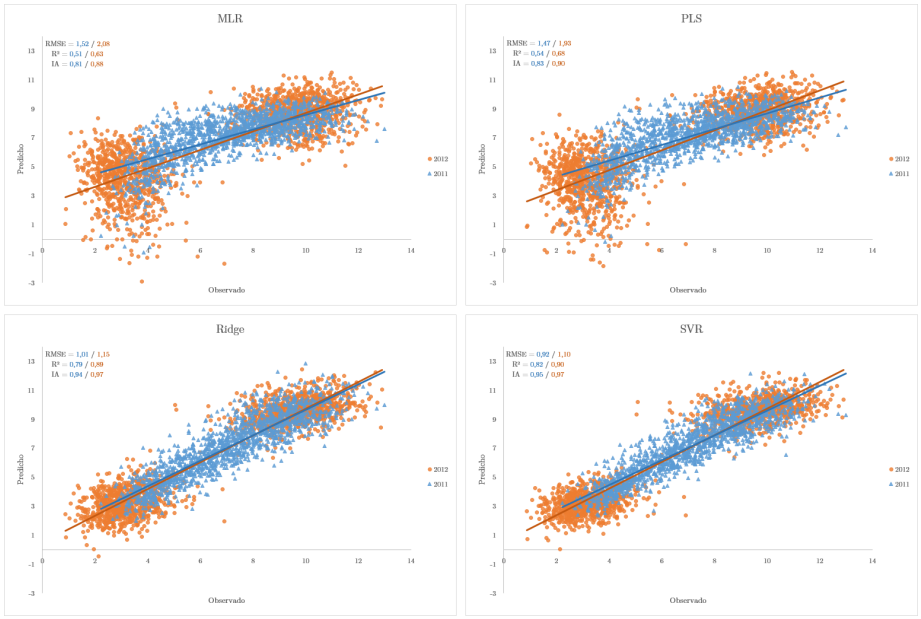
Cuadro 5.4: Detalle de los resultados obtenidos para los 3 rasgos con mejores rendimiento durante el 2012

			MLR			PLS			Ridge			SVR		
*	**	**Ambiente**	**RMSE**	**R2**	**IA**	**RMSE**	**R2**	**IA**	**RMSE**	**R2**	**IA**	**RMSE**	**R2**	**IA**
Rendimiento (t/h)	AN	SWS	0.86	0.07	0.35	0.84	0.11	0.46	0.95	0.14	0.63	0.85	0.19	0.65
		MWS	1.13	0.03	0.27	1.12	0.06	0.36	1.33	0.06	0.55	1.18	0.11	0.59
		FI	1.15	0.02	0.23	1.11	0.10	0.41	1.32	0.16	0.64	1.08	0.23	0.68
		Combinado	2.08	**0.63**	0.88	1.93	**0.68**	0.90	1.15	**0.89**	0.97	1.10	**0.90**	0.97
	GF	SWS	0.71	0.33	0.69	0.69	0.36	0.72	0.68	0.48	0.83	0.59	0.56	0.86
		MWS	1.08	0.11	0.43	1.07	0.12	0.46	1.28	0.13	0.62	1.15	0.17	0.65
		FI	1.11	0.09	0.40	1.10	0.11	0.45	1.23	0.14	0.63	1.19	0.15	0.63
		Combinado	1.44	**0.83**	0.95	1.32	**0.85**	0.96	0.98	**0.92**	0.98	0.89	**0.93**	0.98
IAF	AN	SWS												
		MWS	0.78	0.20	0.56	0.76	0.24	0.62	0.89	0.27	0.72	0.71	0.37	0.77
		FI	0.78	0.36	0.72	0.75	0.41	0.76	0.85	0.37	0.78	0.75	0.45	0.81
		Combinado	0.81	0.29	0.66	0.77	0.37	0.73	0.76	0.45	0.81	0.68	**0.51**	0.83
	GF	SWS												
		MWS	0.81	0.14	0.49	0.80	0.16	0.52	0.98	0.17	0.65	0.84	0.23	0.69
		FI	0.84	0.26	0.63	0.80	0.33	0.70	0.85	0.36	0.77	0.82	0.37	0.77
		Combinado	0.86	0.21	0.57	0.82	0.27	0.65	0.83	0.36	0.77	0.76	0.41	0.79
$\Delta 13C$	AN	SWS	0.52	0.05	0.31	0.50	0.10	0.42	0.60	0.10	0.59	0.53	0.15	0.61
		MWS												
		FI	0.48	0.01	0.17	0.48	0.02	0.24	0.58	0.11	0.60	0.49	0.12	0.59
		Combinado	1.21	**0.63**	0.87	1.13	**0.68**	0.90	0.60	**0.91**	0.98	0.57	**0.92**	0.98
	GF	SWS	0.49	0.12	0.47	0.48	0.16	0.52	0.56	0.23	0.69	0.44	0.34	0.75
		MWS												
		FI	0.48	0.02	0.20	0.48	0.03	0.26	0.56	0.10	0.59	0.52	0.10	0.58
		Combinado	0.86	**0.81**	0.95	0.77	**0.85**	0.96	0.52	**0.93**	0.98	0.49	**0.94**	0.98

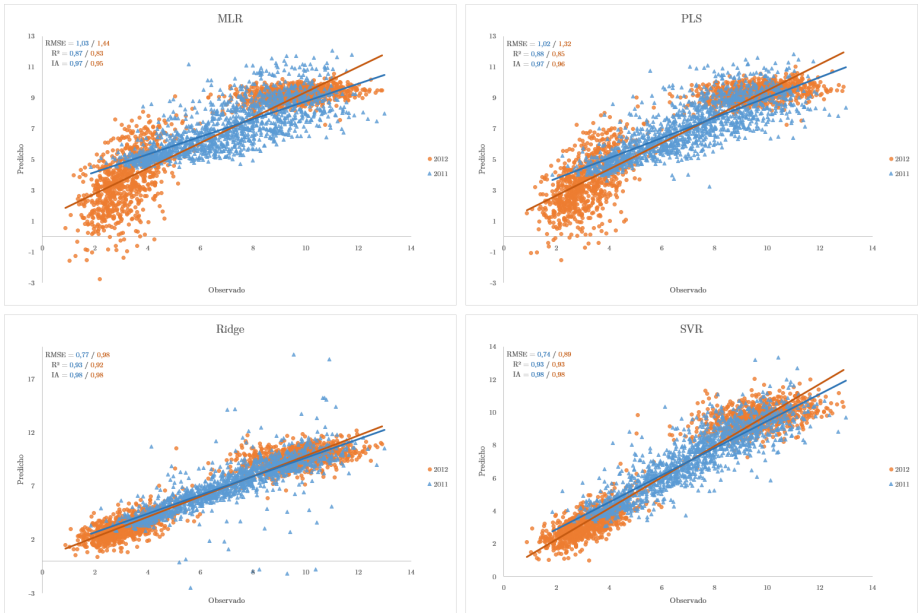
Los resultados de $\Delta 13C$ para el 2011 con SVR en ambiente combinado, tuvieron valores de R^2 para AN y GF de 0,65 y 0,90, respectivamente (muy similar a lo obtenido con regresión Ridge). Del mismo modo, el $\Delta 13C$ predicho por cualquiera de los métodos de regresión, excepto MLR, mostró $R^2 > 0{,}52$, siendo SVR el mejor resultado con $R^2 = 0{,}90$ con los datos espectrales de llenado de grano. El ambiente con estrés hídrico medio presentó para todos los casos de prueba resultados de $R^2 > 0{,}57$, siendo el más alto con $R^2 = 0{,}85$. En el ambiente de riego completo se obtuvieron los valores de predicción más bajos ($R^2 < 0{,}21$) para esta característica (tabla 5.3). Por otro lado, a pesar de que los resultados en los ambientes individuales son pésimos con los datos recolectados durante el 2012, en un ambiente combinado los modelos resultan mejor ajustados para $\Delta 13C$ con un $R^2 > 0{,}63$ (tabla 5.4).

Las figuras 5.1, 5.2 y 5.3 muestran los resultados del ambiente combinado usando datos espectrales medidos en antesis y llenado del grano, durante el 2011 y 2012. En cada figura se presentan de forma separada los resultados para los cuatro modelos evaluados (MLR, PLS, Ridge y SVR) diferenciando de acuerdo al color y la forma los resultados para cada año. Del mismo modo, en las figuras se muestran los estadísticos obtenidos en cada caso y la línea de tendencia de los resultados.

A partir de lo observado en las figuras 5.1, 5.2 y 5.3, se identifica que tanto MLR y PLS muestran un rendimiento similar, del mismo modo que Ridge y SVR presentan un comportamiento análogo. En general, la línea de tendencia que se muestra claramente marcada en cada una de las figuras 5.1, 5.2 y 5.3, reafirman los resultados obtenidos de R^2 en los ambientes combinados.

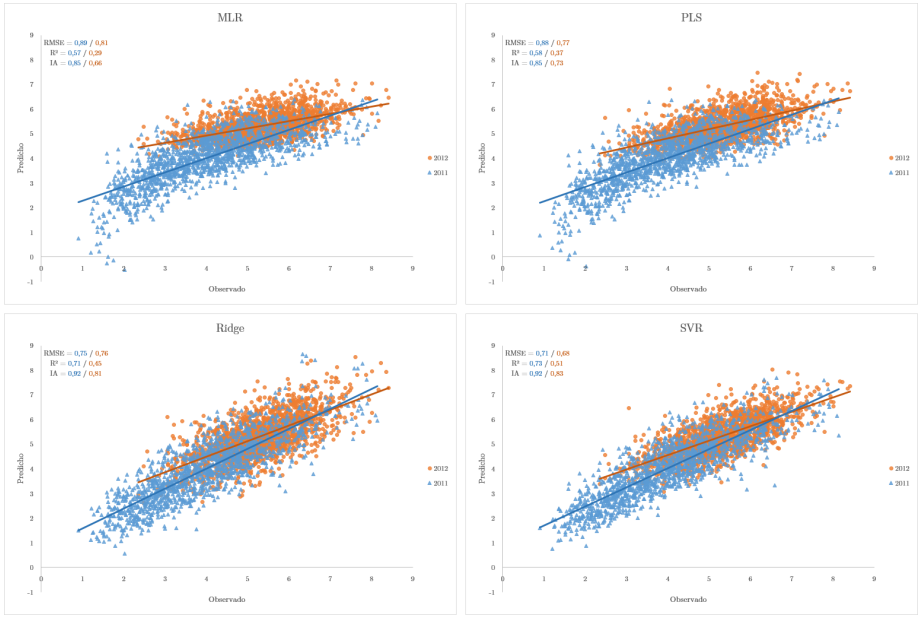


(a) Modelos realizados con mediciones de reflectancia tomadas en la etapa de antesis

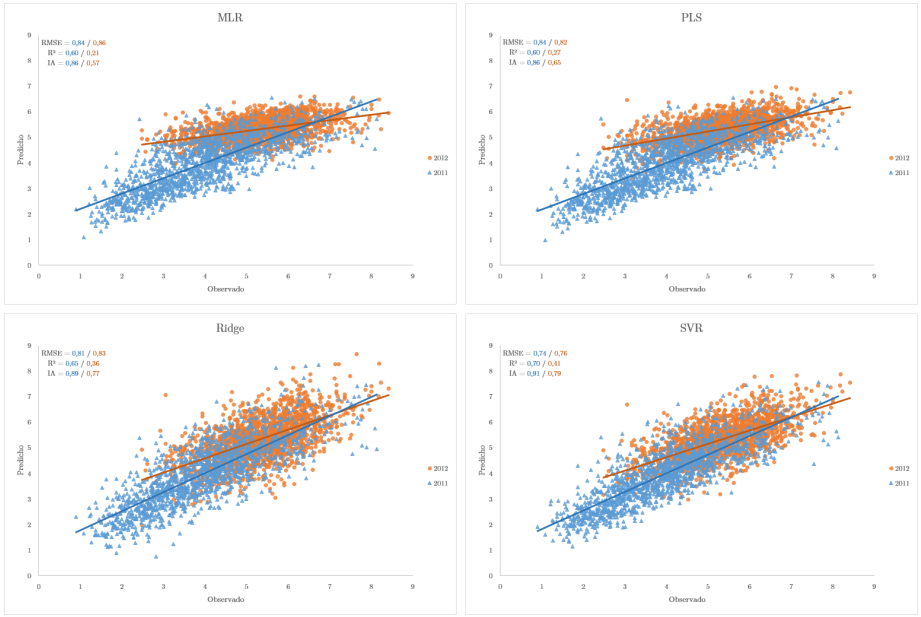


(b) Modelos realizados con mediciones de reflectancia tomadas en la etapa de llenado del grano

Figura 5.1: Datos predichos vs observados después de la validación cruzada para rendimiento

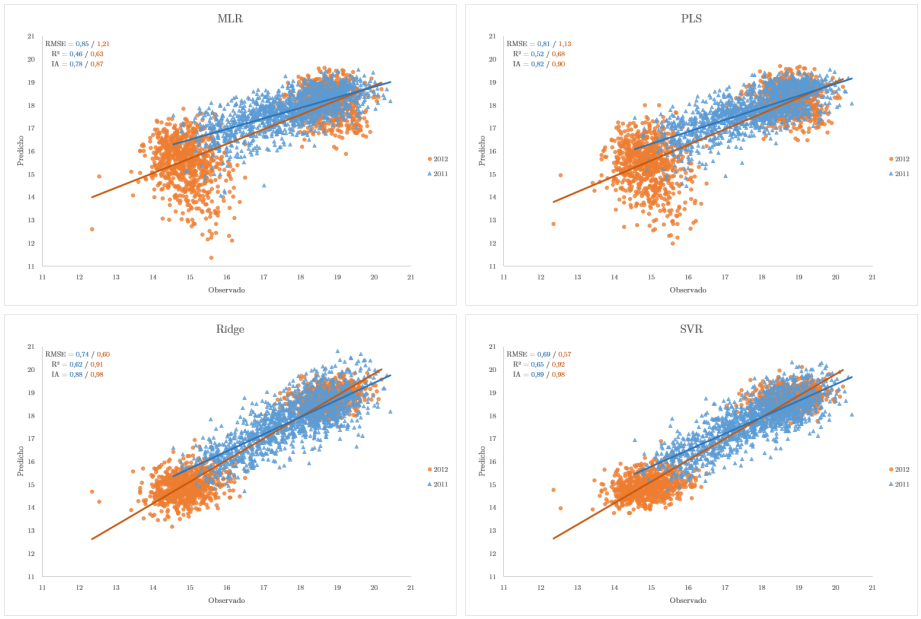


(a) Modelos realizados con mediciones de reflectancia tomadas en la etapa de antesis

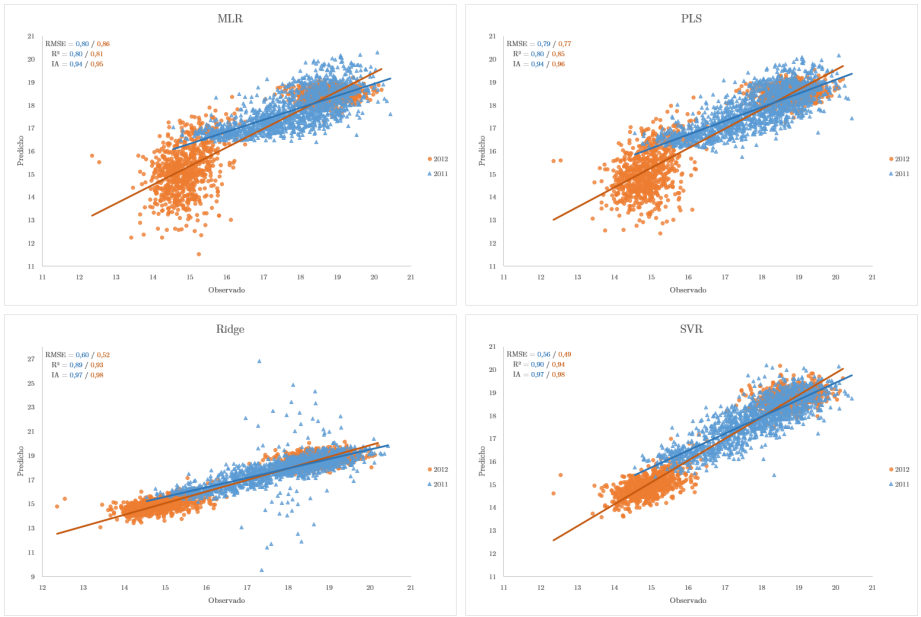


(b) Modelos realizados con mediciones de reflectancia tomadas en la etapa de llenado del grano

Figura 5.2: Datos predichos vs observados después de la validación cruzada para IAF



(a) Modelos realizados con mediciones de reflectancia tomadas en la etapa de antesis



(b) Modelos realizados con mediciones de reflectancia tomadas en la etapa de llenado del grano

Figura 5.3: Datos predichos vs observados después de la validación cruzada para $\Delta 13C$

6. Conclusiones

En este estudio se analizaron una serie de rasgos del trigo utilizando modelos de predicción. De todos los componentes sometidos al análisis que destacaron por sus resultados fueron el rendimiento, IAF y $\Delta 13C$, evaluados a partir de 384 genotipos del trigo de primavera CIMMYT, Uruguay y Chile. Se compararon cuatro modelos, que fueron capaces de discriminar tanto la variabilidad genotípica y ambiental en tres entornos con diferentes cantidades de disponibilidad de agua.

Este diseño experimental permitió tener una alta dispersión en los datos (Anexo A), tanto entre los entornos, especialmente entre FI y SWS, y entre los genotipos dentro de un mismo entorno. La baja diferencia entre FI y SWS mostrada por SPAD y CHOa y CHOm, hace evidente que el estrés hídrico, especialmente en SWS, no se expresó fuertemente sobretodo en la etapa de antesis. La concentración de CHO de alto tallo muestra que la movilización hacia el grano no ha comenzado todavía (no presente todavía en una etapa de antesis), también se ha demostrado que en condiciones de estrés hay una mayor concentración de CHO en el tallo [16], [28], como ocurrió en este estudio en el que SWS presentó mayores valores de CHO. Esto confirma que en las condiciones en las que se desarrollaron las pruebas se mostró un estrés tardío, típico de los climas mediterráneos. Del mismo modo, vemos que las mayores diferencias se producen en CHOm mg/g (medido en la madurez) y rendimiento, lo que implica que la expresión de estrés hídrico es más fuerte en la madurez.

$\Delta 13C$ y los componentes de rendimiento muestran diferencias porcentuales entre FI y SWS mayores de 21 %. Debido a los amplios rangos de datos entre entornos individuales, el entorno combinado, formado por los tres ambientes individuales juntos, en primer lugar, presenta un número mucho mayor de observaciones y segundo

alcanza una mayor variabilidad en los datos (Anexo A), lo que podría explicar el mayor R^2 obtenido en el ambiente combinado en la mayoría de los casos.

Si vemos los resultados de las predicciones de entornos individuales, se observó que se obtuvieron niveles similares de ajuste en MWS y ambiente combinado, especialmente para IAF. En este caso específico, los modelos espectrales construidos con datos tanto de antesis como de llenado de grano tienen mejor ajuste en el ambiente MWS que en combinados, FI o SWS con cualquiera de las cuatro metodologías de modelado ($0{,}66 < R^2 < 0{,}76$) datos espectrales medidos en antesis y $0{,}73 < R^2 < 0{,}75$ datos espectrales medidos en llenado de grano) (5.3).

El desempeño mostrado por las cuatro técnicas de modelado evaluadas mostró un comportamiento diferente dependiendo del rasgo evaluado. Todas las metodologías mostraron los niveles más altos de ajuste para rendimiento, IAF $\Delta 13C$ sobre MLR y PLS, respectivamente, con R^2 hasta 0,93 para rendimiento usando SVR. Regresión Ridge y SRV tuvieron un rendimiento muy similar, presentando menos de 8 % diferencias en R^2 en su rendimiento (tabla 5.3).

Existen diferencias muy notorias en los resultados obtenidos en los ambientes individuales entre el 2011 y 2012. Dichas diferencias se deben a dificultades durante el proceso, en el que no se pudo controlar de forma óptima la cantidad de agua presente en los ambientes individuales durante el 2012. Lo anterior, que supone diferencias en los experimentos y resultados que afectaron los ajustes de los modelos, impidieron realizar predicciones de los datos del 2012 utilizando los modelos generados a partir de los datos de reflectancia espectral medidos durante el 2011.

PLS, Ridge regresión y MLR han sido ampliamente utilizados como métodos de modelado de datos espectrales para predecir rasgos fenotípicos, porque son fáciles de usar, rápidos y permiten la interpretación de datos espectrales y las relaciones entre la variable dependiente. Sin embargo, el potencial presentado por SVR, que queda claramente expresado en los resultados obtenidos, demuestra que puede ser muy útil debido a su capacidad para encontrar soluciones globales no lineales y su capacidad para trabajar con vectores de entrada de alta dimensionalidad. El mayor logro de este algoritmo radica en la supremacía de los resultados por sobre PLS, dado que este último es uno de los modelos más utilizados en el área de fenomica.

La principal desventaja de SVR para la calibración de modelos predictivos es la poca o ninguna capacidad de interpretar biológicamente los datos, ya que la función del núcleo que SVR utiliza modifica la matriz de datos original modificando su

tamaño y dimensionalidad, perdiendo sentido biológico y actuando como una "caja negra". Este punto posterior es en realidad una de las capacidades más fuertes de métodos tales como PLS o MLR, que proporcionan pesos diferenciados para cada componente del modelo sin perder el significado original de los datos, permitiendo el reconocimiento de áreas del espectro que podrían proporcionar información relevante para explicar el rasgo medido.

Bibliografía

[1] JL Araus, T Amaro, J Casadesus, A Asbati, and MM Nachit. Relationships between ash content, carbon isotope discrimination and yield in durum wheat. *Functional Plant Biology*, 25(7):835–842, 1998.

[2] P. Yvonne Barnes, Edward Early, and Albert Parr. Nist measurement services: Spectral reflectance. *NASA*, (19980204706), 1998.

[3] Marcus Borengasser, William S Hungate, and Russell Watkins. *Hyperspectral remote sensing: principles and applications*. Crc Press, 2007.

[4] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.

[5] Llorenç Cabrera-Bosquet, José Crossa, Jarislav von Zitzewitz, María Dolors Serret, and José Luis Araus. High-throughput phenotyping and genomic selection: The frontiers of crop breeding converge. *Journal of integrative plant biology*, 54(5):312–320, 2012.

[6] Daniel F. Calderini and Gustavo A. Slafer. Changes in yield and yield stability in wheat during the 20th century. *Field Crops Research*, 57(3):335–347, 1998.

[7] Anthony G Condon, RA Richards, GJ Rebetzke, and GD Farquhar. Breeding for high water-use efficiency. *Journal of experimental botany*, 55(407):2447–2460, 2004.

[8] Corinna Cortes and Vladimir Vapnik. Support vector machine. *Machine learning*, 20(3):273–297, 1995.

[9] Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161, 1997.

[10] Alejandra Engler and Alejandro del Pozo. Assessing long-and short-term trends in cereal yields: the case of chile between 1929 and 2009. *Ciencia e Investigación Agraria*, 40:55–67, 2013.

[11] JP Ferrio, D. Villegas, J. Zarco, N. Aparicio, JL Araus, and C. Royo. Assessment of durum wheat yield using visible and near-infrared reflectance spectra of canopies. *Field Crops Research*, 94(2):126–148, 2005.

[12] Andy Field, Jeremy Miles, and Zoe Field. *Discovering statistics using R*. Sage publications, 2012.

[13] Food and Agriculture Organization of the United Nations [FAO]. The state of food insecurity in the world. meeting the 2015 international hunger targets: Taking stock of uneven progress, 2015.

[14] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[15] David M Gates, Harry J Keegan, John C Schleter, and Victor R Weidner. Spectral properties of plants. *Applied optics*, 4(1):11–20, 1965.

[16] Javier Hernandez, Gustavo A. Lobos, Iván Matus, Alejandro del Pozo, Paola Silva, and Mauricio Galleguillos. Using ridge regression models to estimate grain yield from field spectral data in bread wheat (triticum aestivum l.) grown under three water regimes. *Remote Sensing*, 7(2):2109–2126, 2015.

[17] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Springer, 2013.

[18] Edward B Knipling. Physical and physiological basis for the reflectance of visible and near-infrared radiation from vegetation. *Remote sensing of environment*, 1(3):155–159, 1970.

[19] Fei Li, Bodo Mistele, Yuncai Hu, Xinping Chen, and Urs Schmidhalter. Reflectance estimation of canopy nitrogen content in winter wheat using optimised hyperspectral

spectral indices and partial least squares regression. *European Journal of Agronomy*, 52:198–209, 2014.

[20] Zhan-Yu Liu, Hong-Feng Wu, and Jing-Feng Huang. Application of neural networks to discriminate fungal infection levels in rice panicles using hyperspectral reflectance and principal components analysis. *Computers and Electronics in Agriculture*, 72(2):99–106, 2010.

[21] David B. Lobell, Marshall B. Burke, Claudia Tebaldi, Michael D. Mastrandrea, Walter P. Falcon, and Rosamond L. Naylor. Prioritizing climate change adaptation needs for food security in 2030. *Science*, 319(5863):607–610, 2008.

[22] Gustavo A. Lobos, Iván Matus, Alejandra Rodriguez, Sebastián Romero-Bravo, José Luis Araus, and Alejandro del Pozo. Wheat genotypic variability in grain yield and carbon isotope discrimination under mediterranean conditions assessed by spectral reflectance. *Journal of integrative plant biology*, 56(5):470–479, 2014.

[23] Oded Maimon and Lior Rokach. *Data mining and knowledge discovery handbook*, volume 2. Springer, 2005.

[24] Juan L. Minetti, Walter M. Vargas, Arnobio G. Poblete, L. R. Acuña, and Guillermo A. Casagrande. Non-linear trends and low frequency oscillations in annual precipitation over argentina and chile, 1931-1999. *Atmósfera*, 16(2):119–135, 2003.

[25] Dimitrios Moshou, Cedric Bravo, Jonathan West, Stijn Wahlen, Alastair McCartney, and Herman Ramon. Automatic detection of 'yellow rust'in wheat using reflectance measurements and neural networks. *Computers and electronics in agriculture*, 44(3):173–188, 2004.

[26] AJD Pask, J Pietragalla, DM Mullan, PN Chavez-Dulanto, and MP Reynolds. Fitomejoramiento fisiologico ii: una guia de campo para la caracterizacion fenotipica de trigo. 2013.

[27] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[28] MP Reynolds, AJD Pask, DM Mullan, and PN Chavez-Dulanto. *Fitomejoramiento fisiológico I: enfoques interdisciplinarios para mejorar la adaptación del cultivo*. CIMMYT, 2013.

[29] Vali Rasooli Sharabian, Noboru Noguchi, and Kazunobu Ishi. Significant wavelengths for prediction of winter wheat growth status and grain yield using multivariate analysis. *Engineering in Agriculture, Environment and Food*, 7(1):14–21, 2014.

[30] Lincoln Taiz, Eduardo Zeiger, Ian Max Møller, and Angus Murphy. *Plant physiology and development*. Sinauer Associates, Incorporated, 2015.

[31] David Tilman, Christian Balzer, Jason Hill, and Belinda L Befort. Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences*, 108(50):20260–20264, 2011.

[32] Govindan Velu and Ravi Prakash Singh. Phenotyping in wheat breeding. In *Phenotyping for Plant Breeding*, pages 41–71. Springer, 2013.

[33] Cort J Willmott. On the validation of models. *Physical geography*, 2(2):184–194, 1981.

ANEXOS

A. Resultados 2011 y 2012

Cuadro A.1: Resultados estadísticos de la predicción de los rasgos evaluados en los 4 ambientes, para los 2 estados fenológicos medidos durante el 2011 para los 4 modelos de regresión evaluados

			MLR			PLS			Ridge			SVR		
*	**	**Ambiente**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**
SPAD 1	AN	SWS												
		MWS	2.95	0.13	0.46	2.91	0.15	0.50	3.12	0.23	0.69	2.87	0.25	0.70
		FI	3.05	0.11	0.44	2.93	0.18	0.55	2.91	0.29	0.73	2.96	0.28	0.72
		Combinado	3.08	0.12	0.44	3.03	0.15	0.49	2.91	0.30	0.73	2.75	0.33	0.74
	GF	SWS	4.51	0.09	0.41	4.35	0.15	0.52	4.82	0.11	0.57	5.22	0.07	0.55
		MWS	3.07	0.06	0.34	2.92	0.15	0.50	3.02	0.18	0.64	2.88	0.20	0.62
		FI	3.17	0.05	0.31	3.14	0.07	0.36	3.38	0.12	0.59	3.28	0.10	0.56
		Combinado	3.59	0.05	0.30	3.55	0.07	0.36	3.44	0.21	0.66	3.21	0.24	0.64

Continúa en la siguiente página...

Cuadro A.1 – *Continuación de la página anterior*

			MLR			PLS			Ridge			SVR		
*	**	**Ambiente**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**
SPAD 2	AN	SWS												
		MWS	10.23	0.22	0.60	9.69	0.30	0.67	8.68	0.48	0.82	8.51	0.49	0.83
		FI	3.59	0.06	0.36	3.36	0.18	0.56	3.37	0.26	0.71	3.43	0.27	0.72
		Combinado	8.56	0.20	0.55	8.03	0.29	0.66	6.95	**0.50**	0.83	6.58	**0.53**	0.83
	GF	SWS												
		MWS	8.84	0.42	0.77	8.81	0.42	0.77	8.25	**0.53**	0.85	7.18	**0.62**	0.88
		FI	3.62	0.03	0.25	3.57	0.06	0.34	3.85	0.12	0.61	3.71	0.12	0.59
		Combinado												
SPAD 3	AN	SWS												
		MWS												
		FI	9.52	0.06	0.33	8.95	0.17	0.52	8.43	0.33	0.75	8.67	0.31	0.74
		Combinado												
	GF	SWS												
		MWS												
		FI	7.67	0.36	0.72	6.92	0.48	0.80	6.42	**0.56**	0.86	6.96	**0.51**	0.84
		Combinado												
Rendimiento (t/h)	AN	SWS												
		MWS	1.37	0.58	0.85	1.23	**0.66**	0.89	1.11	**0.74**	0.93	0.95	**0.80**	0.94
		FI	1.29	0.38	0.74	1.23	0.44	0.77	1.19	**0.54**	0.85	1.12	**0.57**	0.86
		Combinado	1.52	**0.51**	0.81	1.47	**0.54**	0.83	1.01	**0.79**	0.94	0.92	**0.82**	0.95
	GF	SWS	0.43	**0.54**	0.84	0.43	**0.54**	0.83	0.49	0.48	0.83	0.46	**0.52**	0.85
		MWS	0.96	**0.79**	0.94	0.93	**0.81**	0.94	0.79	**0.86**	0.96	0.71	**0.89**	0.97
		FI	1.23	0.44	0.77	1.21	0.46	0.79	1.35	0.45	0.82	1.10	**0.57**	0.87
		Combinado	1.03	**0.87**	0.97	1.02	**0.88**	0.97	0.77	**0.93**	0.98	0.74	**0.93**	0.98

Continúa en la siguiente página...

Cuadro A.1 – *Continuación de la página anterior*

			MLR			PLS			Ridge			SVR		
*	**	**Ambiente**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**
IAF	AN	SWS												
		MWS	0.87	**0.66**	0.89	0.84	**0.68**	0.90	0.77	**0.75**	0.93	0.75	**0.76**	0.93
		FI	0.86	0.38	0.73	0.85	0.40	0.75	0.89	0.44	0.81	0.81	0.49	0.83
		Combinado	0.89	**0.57**	0.85	0.88	**0.58**	0.85	0.75	**0.71**	0.92	0.71	**0.73**	0.92
	GF	SWS												
		MWS	0.78	**0.73**	0.92	0.78	**0.73**	0.92	0.83	**0.71**	0.92	0.76	**0.75**	0.93
		FI	0.91	0.27	0.65	0.90	0.28	0.66	1.02	0.27	0.72	0.89	0.36	0.76
		Combinado	0.84	**0.60**	0.86	0.84	**0.60**	0.86	0.81	**0.65**	0.89	0.74	**0.70**	0.91
CHOa mg/g	AN	SWS												
		MWS	52.41	0.05	0.28	52.47	0.04	0.28	54.10	0.02	0.30	55.77	0.03	0.40
		FI	48.63	0.07	0.33	48.22	0.09	0.39	47.97	0.11	0.47	53.84	0.05	0.46
		Combinado	52.59	0.07	0.32	52.53	0.07	0.33	52.93	0.08	0.43	53.74	0.08	0.45
	GF	SWS	47.49	0.25	0.62	47.54	0.25	0.63	47.45	0.26	0.64	52.11	0.19	0.64
		MWS	53.09	0.05	0.29	53.14	0.05	0.29	53.44	0.05	0.34	53.28	0.07	0.41
		FI	46.30	0.14	0.47	46.30	0.14	0.48	46.59	0.13	0.48	49.96	0.08	0.49
		Combinado	51.76	0.15	0.48	51.70	0.15	0.48	52.13	0.16	0.56	50.73	0.20	0.56
CHOm mg/g	AN	SWS												
		MWS	23.95	0.02	0.21	23.95	0.02	0.22	24.02	0.02	0.22	27.23	0.00	0.31
		FI	7.20	0.02	0.22	7.19	0.02	0.23	7.30	0.01	0.27	9.11	0.00	0.40
		Combinado	19.32	0.05	0.28	19.17	0.06	0.32	19.21	0.11	0.50	19.18	0.10	0.45
	GF	SWS	23.07	0.04	0.29	23.01	0.04	0.32	23.23	0.05	0.43	25.72	0.04	0.48
		MWS	23.86	0.03	0.23	23.78	0.03	0.25	23.33	0.08	0.41	24.09	0.05	0.39
		FI	7.22	0.03	0.25	7.21	0.03	0.27	7.35	0.02	0.34	8.37	0.01	0.43
		Combinado	20.24	0.15	0.48	20.12	0.16	0.49	19.44	0.22	0.61	19.49	0.24	0.56

Continúa en la siguiente página...

Cuadro A.1 – *Continuación de la página anterior*

			MLR			PLS			Ridge			SVR		
*	**	**Ambiente**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**
CHOa mg/tallo	AN	SWS												
		MWS	154.13	0.07	0.29	154.12	0.07	0.29	151.34	0.10	0.41	155.87	0.08	0.41
		FI	97.92	0.14	0.48	97.54	0.14	0.49	97.04	0.16	0.55	109.70	0.08	0.53
		Combinado	132.18	0.10	0.38	132.07	0.11	0.39	131.76	0.12	0.46	132.71	0.12	0.47
	GF	SWS	84.85	0.12	0.46	85.11	0.11	0.46	85.42	0.12	0.52	95.74	0.09	0.56
		MWS	154.58	0.06	0.27	154.48	0.06	0.28	152.80	0.09	0.39	153.69	0.09	0.37
		FI	93.36	0.17	0.53	92.93	0.18	0.54	92.31	0.19	0.57	98.04	0.15	0.59
		Combinado	128.29	0.05	0.25	126.57	0.07	0.31	122.79	0.13	0.49	122.44	0.14	0.46
CHOm mg/tallo	AN	SWS												
		MWS	34.08	0.05	0.29	34.07	0.05	0.30	34.57	0.04	0.31	38.35	0.01	0.33
		FI	10.10	0.08	0.37	10.09	0.08	0.38	10.17	0.07	0.42	12.11	0.02	0.46
		Combinado	26.66	0.06	0.31	26.52	0.07	0.33	26.70	0.10	0.48	26.68	0.09	0.41
	GF	SWS	19.86	0.01	0.19	19.83	0.01	0.21	20.03	0.01	0.29	22.22	0.01	0.41
		MWS	32.52	0.05	0.30	32.37	0.06	0.32	32.01	0.09	0.43	32.88	0.08	0.43
		FI	10.06	0.08	0.39	10.01	0.09	0.41	10.21	0.10	0.53	10.96	0.08	0.54
		Combinado	25.56	0.01	0.14	25.36	0.03	0.21	24.44	0.13	0.53	24.28	0.14	0.43
$\Delta 13C$	AN	SWS												
		MWS	0.89	**0.57**	0.85	0.76	**0.68**	0.90	0.83	**0.66**	0.90	0.71	**0.74**	0.92
		FI	0.66	0.17	0.53	0.66	0.18	0.54	0.81	0.12	0.61	0.75	0.14	0.62
		Combinado	0.85	0.46	0.78	0.81	**0.52**	0.82	0.74	**0.62**	0.88	0.69	**0.65**	0.89
	GF	SWS	0.58	0.10	0.43	0.57	0.11	0.47	0.69	0.12	0.61	0.62	0.15	0.63
		MWS	0.70	**0.74**	0.92	0.69	**0.75**	0.92	0.63	**0.80**	0.94	0.52	**0.85**	0.96
		FI	0.66	0.21	0.58	0.66	0.21	0.58	0.90	0.14	0.61	0.70	0.21	0.67
		Combinado	0.80	**0.80**	0.94	0.79	**0.80**	0.94	0.60	**0.89**	0.97	0.56	**0.90**	0.97

Continúa en la siguiente página...

Cuadro A.1 – *Continuación de la página anterior*

			MLR			PLS			Ridge			SVR		
*	**	**Ambiente**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**
Espigas/m^2	AN	SWS												
		MWS	108.12	0.38	0.73	107.24	0.39	0.74	96.82	**0.50**	0.82	103.55	0.46	0.81
		FI	106.38	0.27	0.64	105.97	0.28	0.65	94.41	0.43	0.78	104.38	0.36	0.77
		Combinado	106.15	0.34	0.70	105.52	0.34	0.70	92.84	**0.50**	0.82	95.51	0.48	0.81
	GF	SWS	84.91	0.33	0.70	85.08	0.33	0.70	85.53	0.32	0.71	97.53	0.24	0.70
		MWS	116.20	0.29	0.66	112.65	0.33	0.70	107.68	0.39	0.75	110.89	0.37	0.75
		FI	114.66	0.15	0.52	113.53	0.17	0.54	104.39	0.30	0.69	107.25	0.30	0.72
		Combinado	108.50	0.44	0.77	107.95	0.44	0.77	99.39	**0.53**	0.83	100.43	**0.52**	0.83
Granos/espiga	AN	SWS												
		MWS	6.23	0.05	0.31	6.10	0.09	0.40	6.22	0.16	0.63	6.26	0.15	0.61
		FI	6.24	0.02	0.22	6.14	0.06	0.33	6.23	0.12	0.58	6.70	0.10	0.57
		Combinado	6.50	0.03	0.22	6.37	0.06	0.33	6.14	0.19	0.64	6.22	0.18	0.63
	GF	SWS	6.05	0.11	0.45	6.06	0.11	0.45	6.32	0.10	0.55	6.64	0.10	0.57
		MWS	6.12	0.09	0.41	6.02	0.12	0.47	6.32	0.13	0.58	6.13	0.14	0.58
		FI	6.04	0.01	0.15	5.78	0.09	0.39	5.93	0.12	0.57	6.37	0.08	0.54
		Combinado	6.19	0.29	0.66	6.13	0.30	0.68	5.93	0.36	0.74	5.90	0.36	0.73
Peso 100G	AN	SWS												
		MWS	5.97	0.34	0.71	5.87	0.37	0.72	5.74	0.47	0.82	5.34	**0.51**	0.84
		FI	5.84	0.15	0.50	5.82	0.16	0.51	6.28	0.23	0.69	5.87	0.26	0.71
		Combinado	6.40	0.24	0.60	6.37	0.24	0.61	5.37	**0.50**	0.83	5.20	**0.51**	0.84
	GF	SWS	5.46	0.04	0.29	5.47	0.04	0.30	5.90	0.01	0.38	6.30	0.03	0.49
		MWS	5.81	0.38	0.73	5.65	0.41	0.76	5.08	**0.54**	0.85	4.87	**0.57**	0.85
		FI	5.99	0.11	0.44	5.92	0.13	0.48	5.48	0.32	0.74	5.41	0.31	0.73
		Combinado	6.00	0.43	0.77	5.99	0.43	0.77	5.37	**0.55**	0.85	5.31	**0.55**	0.84

Cuadro A.2: Resultados estadísticos de la predicción de los rasgos evaluados en los 4 ambientes, para los 2 estados fenológicos medidos durante el 2012 para los 4 modelos de regresión evaluados

			MLR			**PLS**			**Ridge**			**SVR**		
*	**	**Ambiente**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**
SPAD 1	AN	SWS	3.51	0.00	0.07	3.34	0.09	0.40	3.69	0.09	0.57	3.72	0.10	0.58
		MWS	3.25	0.11	0.43	3.01	0.24	0.61	3.39	0.17	0.64	3.34	0.18	0.65
		FI	3.12	0.07	0.36	2.95	0.17	0.53	3.00	0.21	0.66	3.04	0.20	0.66
		Combinado	3.76	0.48	0.80	3.59	**0.53**	0.82	3.44	**0.58**	0.86	3.36	**0.59**	0.87
	GF	SWS	3.49	0.01	0.16	3.28	0.12	0.46	3.19	0.26	0.70	3.16	0.25	0.69
		MWS	3.33	0.07	0.35	3.26	0.11	0.44	3.55	0.13	0.61	3.76	0.11	0.60
		FI	3.20	0.02	0.23	3.15	0.06	0.33	3.41	0.07	0.53	3.59	0.07	0.54
		Combinado	3.63	**0.51**	0.82	3.50	**0.55**	0.84	3.29	**0.61**	0.88	3.03	**0.66**	0.90

Continúa en la siguiente página...

Cuadro A.2 – *Continuación de la página anterior*

			MLR			PLS			Ridge			SVR		
*	**	**Ambiente**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**
SPAD 2	AN	SWS	10.26	0.00	0.12	10.07	0.04	0.27	10.37	0.07	0.48	10.78	0.05	0.46
		MWS	3.35	0.06	0.34	3.15	0.17	0.54	3.52	0.13	0.60	3.57	0.12	0.59
		FI	3.47	0.07	0.36	3.33	0.14	0.50	3.51	0.14	0.60	3.78	0.11	0.58
		Combinado	8.43	0.34	0.69	8.17	0.38	0.72	8.01	0.42	0.78	7.96	0.44	0.74
	GF	SWS	10.13	0.04	0.25	10.08	0.05	0.29	10.67	0.11	0.58	10.12	0.11	0.51
		MWS	3.39	0.04	0.28	3.29	0.10	0.42	3.59	0.14	0.62	3.60	0.14	0.62
		FI	3.51	0.04	0.28	3.41	0.10	0.42	3.64	0.13	0.60	3.87	0.12	0.60
		Combinado	7.95	0.42	0.75	7.84	0.43	0.76	8.01	0.44	0.80	7.52	**0.51**	0.77
SPAD 3	AN	SWS												
		MWS	5.70	0.07	0.36	5.40	0.17	0.52	5.25	0.25	0.68	5.46	0.23	0.68
		FI	6.55	0.01	0.17	6.15	0.13	0.47	6.12	0.19	0.63	6.24	0.18	0.62
		Combinado												
	GF	SWS												
		MWS	5.84	0.02	0.21	5.53	0.12	0.43	5.55	0.23	0.69	5.99	0.18	0.66
		FI	6.36	0.04	0.28	6.13	0.11	0.44	6.61	0.12	0.58	6.74	0.11	0.57
		Combinado												
Rendimiento (t/h)	AN	SWS	0.86	0.07	0.35	0.84	0.11	0.46	0.95	0.14	0.63	0.85	0.19	0.65
		MWS	1.13	0.03	0.27	1.12	0.06	0.36	1.33	0.06	0.55	1.18	0.11	0.59
		FI	1.15	0.02	0.23	1.11	0.10	0.41	1.32	0.16	0.64	1.08	0.23	0.68
		Combinado	2.08	**0.63**	0.88	1.93	**0.68**	0.90	1.15	**0.89**	0.97	1.10	**0.90**	0.97
	GF	SWS	0.71	0.33	0.69	0.69	0.36	0.72	0.68	0.48	0.83	0.59	**0.56**	0.86
		MWS	1.08	0.11	0.43	1.07	0.12	0.46	1.28	0.13	0.62	1.15	0.17	0.65
		FI	1.11	0.09	0.40	1.10	0.11	0.45	1.23	0.14	0.63	1.19	0.15	0.63
		Combinado	1.44	**0.83**	0.95	1.32	**0.85**	0.96	0.98	**0.92**	0.98	0.89	**0.93**	0.98

Continúa en la siguiente página...

Cuadro A.2 – *Continuación de la página anterior*

			MLR			PLS			Ridge			SVR		
*	**	**Ambiente**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**
IAF	AN	SWS												
		MWS	0.78	0.20	0.56	0.76	0.24	0.62	0.89	0.27	0.72	0.71	0.37	0.77
		FI	0.78	0.36	0.72	0.75	0.41	0.76	0.85	0.37	0.78	0.75	0.45	0.81
		Combinado	0.81	0.29	0.66	0.77	0.37	0.73	0.76	0.45	0.81	0.68	**0.51**	0.83
	GF	SWS												
		MWS	0.81	0.14	0.49	0.80	0.16	0.52	0.98	0.17	0.65	0.84	0.23	0.69
		FI	0.84	0.26	0.63	0.80	0.33	0.70	0.85	0.36	0.77	0.82	0.37	0.77
		Combinado	0.86	0.21	0.57	0.82	0.27	0.65	0.83	0.36	0.77	0.76	0.41	0.79
CHOa mg/g	AN	SWS	51.59	0.02	0.18	50.74	0.05	0.29	50.33	0.09	0.46	51.75	0.10	0.52
		MWS												
		FI	46.45	0.10	0.41	45.61	0.13	0.47	44.52	0.19	0.58	48.83	0.11	0.56
		Combinado	58.04	0.23	0.58	55.32	0.30	0.66	47.61	0.49	0.81	48.22	0.48	0.80
	GF	SWS	51.29	0.02	0.20	51.14	0.03	0.24	50.87	0.07	0.45	52.69	0.06	0.45
		MWS												
		FI	46.77	0.08	0.36	46.73	0.08	0.38	47.28	0.09	0.47	49.82	0.10	0.55
		Combinado	52.51	0.37	0.72	51.06	0.41	0.75	49.26	0.46	0.81	47.86	0.49	0.80
CHOm mg/g	AN	SWS	22.69	0.03	0.23	22.61	0.04	0.26	22.77	0.06	0.44	24.56	0.03	0.43
		MWS												
		FI	32.78	0.00	0.11	32.77	0.01	0.12	32.62	0.01	0.16	33.92	0.01	0.28
		Combinado	28.65	0.01	0.14	28.59	0.02	0.17	28.80	0.03	0.32	28.98	0.03	0.31
	GF	SWS	23.24	0.03	0.22	23.16	0.03	0.25	22.98	0.07	0.42	23.75	0.06	0.45
		MWS												
		FI	32.87	0.00	0.05	32.49	0.02	0.15	32.75	0.02	0.21	33.94	0.03	0.33
		Combinado	28.77	0.00	0.05	28.63	0.01	0.11	29.13	0.04	0.38	28.52	0.05	0.32

Continúa en la siguiente página...

Cuadro A.2 – *Continuación de la página anterior*

			MLR			PLS			Ridge			SVR		
*	**	**Ambiente**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**
CHOa mg/tallo	AN	SWS	138.53	0.03	0.25	138.47	0.04	0.26	136.86	0.07	0.39	148.67	0.06	0.49
		MWS												
		FI	76.73	0.21	0.57	75.95	0.23	0.59	72.52	0.30	0.69	78.77	0.22	0.66
		Combinado	140.56	0.29	0.64	125.49	0.43	0.77	111.42	**0.55**	0.84	114.69	**0.53**	0.83
	GF	SWS	138.87	0.03	0.24	138.73	0.03	0.25	135.01	0.10	0.47	144.07	0.07	0.49
		MWS												
		FI	79.36	0.15	0.48	79.37	0.15	0.49	79.00	0.18	0.58	84.12	0.16	0.61
		Combinado	129.88	0.40	0.74	122.93	0.46	0.79	111.16	**0.56**	0.84	112.10	**0.56**	0.84
CHOm mg/tallo	AN	SWS	33.56	0.03	0.22	33.35	0.04	0.27	33.47	0.06	0.41	35.40	0.05	0.45
		MWS												
		FI	44.06	0.00	0.10	44.04	0.00	0.12	43.96	0.02	0.20	45.16	0.02	0.29
		Combinado	40.00	0.02	0.16	39.96	0.02	0.19	40.06	0.03	0.30	40.68	0.03	0.30
	GF	SWS	32.94	0.08	0.38	32.80	0.09	0.40	32.43	0.13	0.51	34.21	0.08	0.47
		MWS												
		FI	43.88	0.00	0.07	43.61	0.01	0.14	43.83	0.02	0.23	45.34	0.02	0.31
		Combinado	39.99	0.01	0.10	39.82	0.01	0.15	39.60	0.06	0.38	39.35	0.06	0.34
$\Delta 13C$	AN	SWS	0.52	0.05	0.31	0.50	0.10	0.42	0.60	0.10	0.59	0.53	0.15	0.61
		MWS												
		FI	0.48	0.01	0.17	0.48	0.02	0.24	0.58	0.11	0.60	0.49	0.12	0.59
		Combinado	1.21	**0.63**	0.87	1.13	**0.68**	0.90	0.60	**0.91**	0.98	0.57	**0.92**	0.98
	GF	SWS	0.49	0.12	0.47	0.48	0.16	0.52	0.56	0.23	0.69	0.44	0.34	0.75
		MWS												
		FI	0.48	0.02	0.20	0.48	0.03	0.26	0.56	0.10	0.59	0.52	0.10	0.58
		Combinado	0.86	**0.81**	0.95	0.77	**0.85**	0.96	0.52	**0.93**	0.98	0.49	**0.94**	0.98

Continúa en la siguiente página...

Cuadro A.2 – *Continuación de la página anterior*

			MLR			PLS			Ridge			SVR		
*	**	**Ambiente**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**	**RMSE**	R^2	**IA**
Espigas/m^2	AN	SWS												
		MWS												
		FI												
		Combinado												
	GF	SWS												
		MWS												
		FI												
		Combinado												
Granos/espiga	AN	SWS	7.02	0.01	0.15	7.00	0.01	0.20	7.63	0.06	0.53	7.71	0.05	0.51
		MWS												
		FI	6.36	0.03	0.26	6.26	0.07	0.35	6.28	0.13	0.59	6.21	0.16	0.61
		Combinado	6.87	0.15	0.49	6.82	0.16	0.51	7.01	0.20	0.66	6.72	0.22	0.64
	GF	SWS	6.86	0.05	0.33	6.82	0.07	0.35	7.10	0.13	0.60	6.88	0.14	0.60
		MWS												
		FI	6.37	0.03	0.25	6.24	0.07	0.36	6.61	0.07	0.52	7.26	0.05	0.51
		Combinado	6.64	0.21	0.58	6.62	0.21	0.59	6.76	0.24	0.68	6.45	0.26	0.66
Peso 100G	AN	SWS	6.57	0.06	0.34	6.50	0.08	0.39	6.81	0.08	0.50	7.16	0.07	0.52
		MWS												
		FI	5.58	0.29	0.66	5.57	0.29	0.66	5.91	0.30	0.73	5.74	0.31	0.73
		Combinado	6.34	0.21	0.58	6.33	0.21	0.58	6.38	0.26	0.69	6.17	0.28	0.69
	GF	SWS	6.73	0.02	0.24	6.59	0.06	0.35	6.58	0.21	0.67	5.98	0.27	0.69
		MWS												
		FI	6.03	0.16	0.51	6.01	0.17	0.52	6.51	0.20	0.66	6.50	0.19	0.66
		Combinado	6.48	0.17	0.53	6.44	0.18	0.55	6.08	0.33	0.74	5.65	0.38	0.74