ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICAS
DEPARTAMENTO DE POSTGRADOS

PROYECTO DE TITULACIÓN

PREVIO A LA OBTENCIÓN DEL TÍTULO DE:

“MAGÍSTER EN ESTADÍSTICA APLICADA”

TEMA:

Inspección de los perfiles clínico-patológicos y hematológicos de los pacientes con Neoplasias Mieloproliferativas Philadelphia negativas (PV, TE y MFP) basada en herramientas multivariantes de visualización avanzada.

AUTOR:

GEORGE VICENTE ACOSTA CHONG

Guayaquil - Ecuador

2024

RESUMEN

El objetivo de este estudio es proporcionar información sobre las Neoplasias Mieloproliferativas Philadelphia Negativas (NMP Ph-) en pacientes ecuatorianos, utilizando análisis estadísticos multivariantes y visualizaciones avanzadas. Se busca inspeccionar las características clínico-patológicas y hematológicas entre las tres categorías clásicas de NMP Ph-: Policitemia Vera (PV), Trombocitemia Esencial (TE) y Mielofibrosis Primaria (MFP).

Para el estudio se utilizó una muestra levantada por profesionales de la Sociedad de Lucha contra el Cáncer (SOLCA) Guayaquil, SOLCA Quito y Quality of Care (Guayaquil), con 111 pacientes diagnosticados con PV, TE o MFP, y 119 columnas de datos y resultados médicos. Más del 60% de las variables eran binarias, representando la presencia o ausencia de un marcador clínico. Se aplicó el análisis de Correspondencias Múltiples (ACM), la inspección del Biplot Logístico y el Análisis discriminante lineal, tras una exhaustiva preparación de los datos que incluyó limpieza, imputación y selección de características. En esta última fase, se utilizó modelos de clasificación con el fin de evaluar las variables más relevantes para explicar a los grupos, obteniendo 13 variables que fueron las utilizadas posteriormente en las técnicas multivariantes. Se editaron los aspectos estéticos de las visualizaciones multivariantes, para disminuir la complejidad de interpretación de las gráficas saturadas de elementos e información, facilitando la comunicación de resultados.

El ACM junto con las tablas de contingencia condicionadas por grupo revelaron las características clínicas más comunes de cada grupo de NMP Ph-, destacando los niveles de Hemoglobina, Plaquetas, Eritropoyetina sérica, entre otros. El Biplot Logístico y el análisis discriminante identificaron las características que diferenciaban a los grupos, obteniendo una mejor diferenciación de los pacientes con MFP. Este estudio puso de manifiesto la eficiencia de las herramientas multivariantes para analizar datos recogidos en escalas no cuantitativas. Los hallazgos contribuyen a una mejor comprensión del comportamiento clínico de los pacientes con Neoplasias Mieloproliferativas en Ecuador, aportando al conocimiento de este grupo de enfermedades poco estudiadas en la región sur del continente.

Palabras clave
Neoplasias Mieloproliferativas Philadelphia Negativas, Policitemia Vera, Trombocitemia Esencial, Mielofibrosis Primaria, Análisis Multivariante, Visualizaciones.

ABSTRACT

The objective of this study is to provide information on Philadelphia-Negative Myeloproliferative Neoplasms (NMP Ph-) in Ecuadorian patients using multivariate statistical analysis and advanced visualizations. It aims to inspect the clinical-pathological and hematological characteristics among the three classic categories of NMP Ph-: Polycythemia Vera (PV), Essential Thrombocythemia (ET), and Primary Myelofibrosis (PMF).

For the study, a sample collected by professionals from the Sociedad de Lucha contra el Cancer (SOLCA) in Guayaquil, SOLCA Quito, and Quality of Care (Guayaquil) was used, comprising 111 patients diagnosed with PV, ET, or PMF, and 119 columns of data and medical results. Over 60% of the variables were binary, representing the presence or absence of a clinical marker. Multiple Correspondence Analysis (MCA), inspection of the Logistic Biplot, and Linear Discriminant Analysis (LDA) were applied after extensive data preparation, which included cleaning, imputation, and feature selection. In this latter phase, classification models were used to evaluate the most relevant variables for explaining the groups, resulting in 13 variables that were subsequently used in multivariate techniques. The aesthetic aspects of the multivariate visualizations were edited to reduce the complexity of interpreting graphs saturated with elements and information, facilitating the communication of results.

MCA, along with contingency tables conditioned by group, revealed the most common clinical characteristics of each NMP Ph- group, highlighting Hemoglobin levels, Platelets, serum Erythropoietin, among others. The Logistic Biplot and LDA identified the characteristics that differentiated the groups, achieving better differentiation for patients with PMF. This study demonstrated the efficiency of multivariate tools for analyzing data collected on non-quantitative scales. The findings contribute to a better understanding of the clinical behavior of patients with Myeloproliferative Neoplasms in Ecuador, contributing to the understanding of this group of poorly studied diseases in the southern region of the continent.

Keywords:

Philadelphia-Negative Myeloproliferative Neoplasms, Polycythemia Vera, Essential Thrombocythemia, Primary Myelofibrosis, Multivariate Analysis, Visualizations.

DEDICATORIA

A papá.
En más de una década, su enfermedad no lo detuvo.
Con esta tesis honro su vida, su ejemplo de trabajo, su lucha y, a pesar de su condición, siempre se ha preocupado por sus hijos, familia y amigos.

A mamá.
Mujer esforzada y valiente.
Pilar fundamental en la lucha de papá y en la familia.
Con esta tesis honro su vida y anhelo algún día,
ser tan fuerte, constante y amoroso como lo ha sido ella.

AGRADECIMIENTO

"Es justo y es necesario dar gracias a Dios, siempre y en todo lugar".
Agradezco a Dios por ayudarme a seguir. Y de quien estoy seguro una vez más que, si pongo mis proyectos en sus manos desde el principio hasta el fin, saldrá mejor de lo que esperaba.

A mis padres por apoyarme siempre, desde el grado y en el postgrado.
A Arianna Ruiz Cruz, por su tiempo, su apoyo y su amor. ¡Tres cosas tan valiosas! A todos los profesores que me acompañaron en el trabajo en el CEIE-ESPOL y la Maestría, en especial a la PhD. Purificación Galindo por la oportunidad y por su trabajo en la Estadística Multivariante que nos inspira a muchos.
A mi tutor, el Dr. Fuad Huamán, por su tiempo y enseñanzas sobre las enfermedades y la salud.

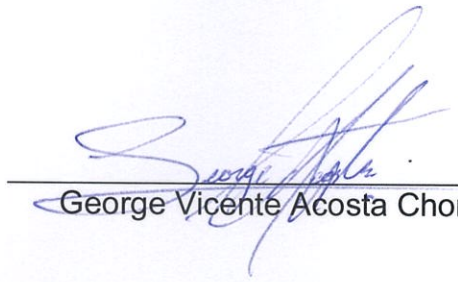
"La gratitud es la memoria del corazón". Lao Tsé.

Agradecimientos extraordinarios a los desarrolladores de ChatGPT y Tidyverse. Hoy en día podemos dedicar menos horas a la programación en estadística, y más tiempo, justamente, a explorar las capacidades del mundo de la estadística, la ciencia de datos y las matemáticas.

DECLARACIÓN EXPRESA

La responsabilidad por los hechos y doctrinas expuestas en este Proyecto de Titulación me corresponde exclusivamente y ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría. El patrimonio intelectual del mismo corresponde exclusivamente a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.



George Vicente Acosta Chong

TRIBUNAL DE GRADUACIÓN

Mariela González Narváez, Ph.D.
PRESIDENTE

Dr. Fuad Huamán Garaicoa
TUTOR

Omar Ruiz Barzola, Ph.D.
DOCENTE EVALUADOR

ABREVIATURAS O SIGLAS

NMP Ph-	Neoplasias Mieloproliferativas Philadelphia Negativas
PV	Policitemia Vera
TE	Trombocitemia Esencial
MFP	Mielofibrosis Primaria
NOS	NMP Ph- no especificada.
OMS	Organización Mundial de la Salud
ACM	Análisis de Correspondencias Múltiple
ADL	Análisis Discriminante Lineal
ELB	Biplot Logístico Externo
RL	Regresión Logística
RLM	Regresión Logística Multinomial
RF	Bosques aleatorios (Random Forest)
Hb	Hemoglobina
PLT	Plaquetas
WBC	Leucocitos
LDH	Lactato deshidrogenasa
EPO	Eritropoyetina Sérica
Retic	Reticulocitos
Leucoeritro	Presencia de Leucoeritroblastosis
(M)	Mielograma
(H)	Histopatología
(S)	Sangre
GranuloM	Granulocitos en Mielograma
GranuloH	Granulocitos en Histopatología

TABLA DE CONTENIDO

LISTADO DE FIGURAS

LISTADO DE TABLAS

CAPÍTULO 1

1. INTRODUCCIÓN

1.1. Antecedentes

En las últimas décadas se han popularizado los métodos que relacionan la matemática, la estadística y las ciencias computacionales, donde los objetivos frecuentes son la clasificación, el descubrimiento de patrones: grupos o tendencias, y predicción de eventos o casos, según el área en el que se esté realizando el análisis o investigación. El trabajo aplica métodos estadísticos avanzados, específicamente con el enfoque del análisis multivariante, aplicados en datos de salud provenientes de instituciones que estudian y tratan enfermedades relacionadas con cánceres en humanos.

Las Neoplasias Mieloproliferativas Philadelphia Negativas (NMP Ph-) son un grupo de enfermedades de la sangre que, de manera más formal, se les conoce como desórdenes clonales en la formación de las células de la sangre caracterizados por la sobreproducción celular de una o más líneas mieloides (Valladares, y otros, 2021). Las NMP Ph- son relativamente de baja frecuencia, registrando una incidencia de 1,15-4,99 por cada 100.000 hab/año según un estudio internacional (Valladares, y otros, 2021). A pesar de ser enfermedades de baja frecuencia, en los artículos científicos generalmente se incluyen a tres categorías que son las más prevalentes en los análisis médicos: la Policitemia Vera (PV), Trombocitemia Esencial (TE) y la Mielofibrosis Primaria (MFP). Todas ellas son enfermedades crónicas que afectan a las células sanguíneas y la médula ósea.

La evidencia científica sobre estudios de las NMP Ph- generalmente se centra en Europa o América del Norte, pero en el contexto Latinoamericano la información es escasa y aislada. Específicamente en Ecuador, existe una publicación del registro de tumores basada en SOLCA Quito, con información demográfica, pero a nivel país no se conocen dichas características. Tampoco existe la información clínico-patológicas de los pacientes con esta condición.

La importancia de poseer información en Ecuador sobre las NMP Ph- impacta directamente en que las condiciones epidemiológicas de nuestra población podrían tener un impacto diferente en la prevalencia, incidencia y factores de riesgo, en

comparación con otras poblaciones del mundo. El diagnóstico y tratamiento de la enfermedad debe adaptarse a las condiciones de los centros de salud del país y a las opciones de los pacientes; lo que también influye en la planificación de los servicios públicos y privados de nuestro país.

1.2. Descripción del problema

Las publicaciones científicas sobre las NMP Ph- que han demostrado el uso apropiado de técnicas estadísticas no básicas, para la generación de información médica son escasas. Adicionalmente, dado que las NMP Ph- y sus 3 entidades clásicas: PV, TE y MFP, son enfermedades consideradas poco frecuentes, analizar con mayor detalle sus similitudes y diferencias es parte de la información que se debe investigar.

Estudiar estas enfermedades en Ecuador, más allá del aporte a la comunidad científica global, tiene beneficios directos en la planificación de los centros de salud públicos y privados, la disponibilidad de atención a los pacientes ecuatorianos y el bienestar de la población general. Por tales motivos, realizar este estudio se considera importante para acreditar información de la relación entre la enfermedad y la población ecuatoriana.

Pregunta de investigación

¿Cuáles son las características clínico-patológicas y hematológicas distintivas de los pacientes con PV, TE y MFP en la muestra de pacientes ecuatorianos?

1.3. Objetivos

1.3.1. Objetivo General

Analizar las diferencias clínico-patológicas y hematológicas de pacientes de varias instituciones de atención oncológica del país, diagnosticados con NMP Ph- de tipo Policitemia Vera, Trombocitemia Esencial y Mielofibrosis Primaria, utilizando métodos multivariantes que permitan la inspección de patrones de asociación entre las variables y patrones de disimilitud entre los grupos de pacientes.

1.3.2. Objetivos específicos

1. Analizar las variables clínico-patológicas y hematológicas que caracterizan a los tres grupos de pacientes con NMP Ph-, Policitemia Vera, Trombocitemia Esencial y Mielofibrosis Primaria.

2. Analizar las variables con mayor capacidad discriminante entre los tres grupos de pacientes con NMP Ph-.
3. Elaborar visualizaciones avanzadas de los datos que faciliten la interpretación y comunicación de los hallazgos.

1.4. Hipótesis

Existe una estructura de variables correlacionadas con las variables de diagnóstico que marcan diferencias multivariantes entre los grupos de pacientes PV, TE y MFP de las instituciones oncológicas de Ecuador.

1.5. Alcance

Para definir el alcance de este trabajo se plantean las preguntas referentes a población, espacio y tiempo de la investigación.

¿Qué se va a investigar?

Las Neoplasias Mieloproliferativas Philadelphia Negativas (NMP Ph-) y sus categorías PV, TE y MFP.

¿Quiénes van a participar en la investigación?

Pacientes que poseen una condición denominada neoplasias mieloproliferativas Philadelphia negativas de 3 instituciones de salud en Ecuador.

¿Dónde se va a llevar a cabo la investigación?

La investigación se llevará a cabo en la ciudad de Guayaquil-Ecuador, utilizando datos de 3 instituciones: Solca-Guayaquil, Quality of Care (Guayaquil) y Solca-Quito.

¿Cuándo fueron tomados los datos?

El registro corresponde a pacientes mayores de 18 años, que han sido diagnosticados con esta condición en un corte de 10 años, desde el 2014 hasta el 2023.

CAPÍTULO 2

2. MARCO TEÓRICO

Este trabajo aplica la estadística sobre datos de Medicina, 2 áreas de ciencias y en cada una su propia especialidad. Por un lado, la Medicina, ubicando a las Neoplasias Mieloproliferativas en enfermedades que deben ser analizadas entre varios especialistas para el correcto diagnóstico y tratamiento de la enfermedad; y, por otro lado, la estadística con el enfoque en el análisis multivariante y los métodos que utilizan variables binarias, donde los recursos estadísticos son menores que los métodos clásicos que utilizan variables numéricas. Por tales motivos, este capítulo aborda las definiciones claves para la comprensión de la metodología usada.

2.1. Neoplasias Mieloproliferativas (NMP)

Las Neoplasias Mieloproliferativas (NMP) son un grupo de enfermedades de la sangre que se caracterizan porque en la médula ósea, que es donde se fabrican las células sanguíneas, existe sobreproducción de uno o más tipos de las células, como los glóbulos rojos, blancos o plaquetas. Aquellas NMP que son de mayor relevancia en la práctica clínica son las Neoplasias Mieloproliferativas crónicas con el cromosoma Philadelphia negativo (NMP Ph-) y la Leucemia Mieloide Crónica (LMC) que se caracteriza por presentar el cromosoma Philadelphia (Ph+) (GEMFIN, 2020). Información más técnica y precisa sobre la enfermedad se encuentra en el “Manual de recomendaciones en Neoplasias Mieloproliferativas Crónicas Filadelfia Negativas”, publicada en el 2020 por el Grupo Español de Enfermedades Mieloproliferativas Crónicas Filadelfia Negativas.

Neoplasias Mieloproliferativas Philadelphia negativas clásicas (NMP Ph-)

Entre las NMP Ph- se incluyen tres entidades principales o clásicas: Policitemia Vera (PV), Trombocitemia Esencial (TE) y Mielofibrosis Primaria (MFP). A continuación, se detalla información sobre cada una de estas enfermedades y los criterios estándares de diagnóstico que, para el contexto de este trabajo, equivalen a las variables que identifican a una enfermedad u otra. Varias de las siguientes definiciones están basadas en el “Manual de Recomendaciones en Neoplasias Mieloproliferativas Filadelfia Negativas” (GEMFIN, 2020).

2.1.1. Policitemia Vera (PV)

Esta enfermedad ocurre cuando las células madre en la médula ósea sufren mutaciones que la llevan a producir glóbulos rojos de manera descontrolada. En menor medida, también se produce un exceso de glóbulos blancos y plaquetas en comparación a condiciones normales (GEMFIN).

Los criterios para identificación de la enfermedad en los pacientes se describen en la **Tabla 1**. Para el diagnóstico se requiere la presencia de los 3 criterios mayores o los 2 mayores primeros más el criterio menor (Khoury, y otros, 2022).

Tabla 1: Criterios de la OMS 2022 para el diagnóstico de Policitemia Vera (PV)

Criterios Mayores	1	**Hemoglobina (Hb)** Hombres: > 16,5 g/dL Mujeres: > 16,0 g/dL	o **Hematocrito** Hombres: > 49% Mujeres: > 48%
	2	Biopsia de médula ósea que demuestre una hipercelularidad trilineal (panmielosis), para la edad del paciente, con proliferación prominente eritroide, granulocítica y megacariocítica, con megacariocitos plemórficos maduros.	
	3	Presencia de la mutación *JAK2* p.V617F u otra mutación activadora de *JAK2*, como las del exón 12.	
Criterios Menores	1	Eritropoyetina sérica por debajo del valor de referencia normal.	

En la Tabla 2, se detallan algunos síntomas de la enfermedad y criterios de sospecha.

Tabla 2: Criterios de Sospecha y síntomas habituales de la Policitemia Vera

Criterios de Sospecha	Síntomas habituales
Hematocrito elevado: Se recomienda investigar si el nivel de hematocrito es alto (por encima de 0.48 en mujeres o 0.49 en hombres).	La PV puede causar síntomas por su impacto en los vasos sanguíneos y la sangre. Entre los síntomas se incluyen dolores de cabeza, zumbidos en los oídos,

Anomalías en sangre: Un aumento en el número de glóbulos rojos, presencia de pequeños glóbulos rojos (microcitosis) y niveles bajos de hierro (ferropenia) pueden sugerir Policitemia Vera. Aumento de glóbulos blancos y plaquetas: Niveles elevados de glóbulos blancos y plaquetas también pueden ser signos de esta condición. Antecedentes de coágulos sanguíneos: Si alguien tiene un historial de coágulos sanguíneos, especialmente en lugares poco comunes, se debe investigar de inmediato. Trombosis en áreas inusuales: Los pacientes con coágulos sanguíneos en áreas como el hígado o el bazo pueden tener una mutación genética asociada con PV, incluso si no hay otros cambios en la sangre.	problemas visuales como visión borrosa, mareos o vértigo. También son comunes el cansancio y el picor, así como otros síntomas causados por un aumento en la actividad celular. El riesgo de coágulos puede provocar trombosis, insuficiencia cardíaca o accidentes cerebrovasculares debido a la sangre espesa y la tendencia de las plaquetas a agruparse. Además, las células madre pueden migrar desde la médula ósea al bazo, causando su crecimiento y provocando molestias abdominales o sensación de saciedad temprana al comer.

Dado a que existen otras enfermedades que presentan síntomas similares, para diagnosticar la PV normalmente hay que hacer más de una prueba (GEMFIN).

2.1.2. Trombocitemia Esencial (TE)

La Trombocitemia Esencial está caracterizada por una producción anormalmente alta de plaquetas en la médula ósea. Los criterios de diagnóstico estándar para la TE se describen en la Tabla 3 y se requiere el cumplimiento de los 4 criterios mayores o los 3 primeros mayores más 1 menor.

Tabla 3: Criterios de la OMS para el diagnóstico de Trombocitemia Esencial (TE)

Criterios Mayores	1	Trombocitosis persistente ≥ $450x10^9$/L.
	2	

		Biopsia medular con predominio de megacariocitos maduros, muy grandes y con núcleo hiperlobulado, sin incremento significativo de la granulopoyesis o de la eritropoyesis, y muy infrecuentemente incremento de las fibras de reticulina (grado 1).
	3	No evidenciar, según los criterios diagnósticos de la OMS, existencia de PV, MFP, leucemia mieloide crónica (LMC) u otra neoplasia mieloide.*
	4	Mutación *JAK2*, *CALR* o *MPL*.
Criterios Menores	1	Presencia de un marcador clonal o ausencia de evidencia de trombocitosis reactiva. O, Exclusión de trombocitosis reactiva.

** Causas reactivas de trombocitosis son: ferropenia, esplenectomía, cirugía, infección, inflamación, cáncer metástasis y síndromes linfoproliferativo.*

Para las pruebas iniciales del diagnóstico de Trombocitemia Esencial (TE), primero se deben descartar otras causas de trombocitosis y neoplasias mieloproliferativas. Luego, se realizan varias pruebas:

- Historial médico y examen físico.
- Análisis de sangre, incluyendo recuento de plaquetas y hematocrito.
- Pruebas bioquímicas.
- Serologías para VIH, hepatitis B y C.
- Estudios moleculares para buscar mutaciones en genes como *JAK2*, *CALR* o *MPL*.
- Pruebas de imagen, como radiografías de tórax y ecografías abdominales (opcional).
- Aspirado y biopsia de médula ósea.

Estas pruebas son esenciales para un diagnóstico preciso y ayudan a determinar el manejo adecuado del paciente.

2.1.3. Mielofibrosis Primaria (MFP)

En las MFP se presenta un crecimiento anormal de las células que producen plaquetas (megacariocitos) y de un tipo de glóbulos blancos que combaten infecciones (granulocitos) en la médula ósea, que se acompaña de la formación de tejido conectivo fibroso y de producción de células de la sangre (hematopoyesis) fuera de la médula ósea.

La OMS brinda dos conjuntos de criterios para el diagnóstico de Mielofibrosis primaria, tanto para su etapa inicial/prefibrótica, y para la etapa establecida/fibrótica. Ambas fases (inicial y fibrótica) se diagnostican con los 3 criterios mayores y al menos 1 menor confirmado en 2 determinaciones consecutivas. Estos criterios se describen en la Tabla 4 y Tabla 5.

*Tabla 4: Criterios de la OMS para el diagnóstico de Mielofibrosis Primaria en **fase inicial/prefibrótica***

Criterios Mayores (se requiere que se cumplan todos)	1	Biopsia medular con proliferación de megacariocitos atípicos, sin fibrosis reticulínica > grado 1, con incremento en la celularidad medular ajustada según la edad, proliferación granulocítica y con disminución de la eritropoyesis en muchos casos.
	2	No cumplir los criterios de la OMS para LMC, PV, TE, neoplasias mielodisplásicas u otras neoplasias mieloides.
	3	Mutación de *JAK2*, *CALR* o *MPL* o, en ausencia de estas mutaciones, presencia de otra mutación clonal* o ausencia de fibrosis reticulínica medular reactiva**.
Criterios Menores (se requiere al menos 1, confirmado en 2 determinaciones consecutivas)	a	Anemia no atribuible a comorbilidad.
	b	Leucocitosis ≥ 11 x 10^9/L.
	c	Esplenomegalia palpable y/o detectada por imágenes.
	d	Aumento del nivel de lactodeshidrogenasa (LDH) sérica por encima del valor superior normal de referencia para cada centro.

**Si no hay alguna de las tres mutaciones clonales mayores, la búsqueda de otras asociadas a neoplasias mieloides (p.ej., ASXL1, EZH2, TET2, IDH1/IDH2, SRSF2, SF3B1) ayuda a determinar la naturaleza clonal de la enfermedad.*

***Fibrosis reticulínica leve (grado 1) secundaria a infección, enfermedad autoinmune u otro trastorno inflamatorio crónico, tricoleucemia u otra neoplasia linfoide, cáncer metastásico o mielopatía tóxica (crónica).*

Tabla 5: Criterios de la OMS para el diagnóstico de Mielofibrosis Primaria ***en fase establecida/fibrótica***

Criterios Mayores (se requiere que se cumplan todos)	1	Biopsia medular con proliferación de megacariocitos atípicos, acompañada de fibrosis reticulínica y/o colágena grados 2 o 3.
	2	No cumplir criterios de la OMS para LMC, PV, TE, neoplasias mielodisplásicas u otras neoplasias mieloides.
	3	Mutación de *JAK2*, *CALR* o *MPL*, o presencia de otro marcador clonal* o ausencia de fibrosis reticulínica reactiva.**
Criterios Menores (se requiere al menos 1, confirmado en 2 determinaciones consecutivas)	a	Anemia no atribuible a comorbilidad.
	b	Leucocitosis ≥ 11 x 10^9/L.
	c	Esplenomegalia palpable y/o detectada por imágenes.
	d	Aumento del nivel de lactodeshidrogenasa (LDH) sérica por encima del valor superior normal de referencia para cada centro.
	e	Leucoeritroblastosis en sangre periférica.

**En ausencia de cualquiera de las tres mutaciones clonales mayores, la búsqueda de otras mutaciones asociadas a neoplasias mieloides (p.ej., ASXL1, EZH2, TET2, IDH1/IDH2, SRSF2, SF3B1) ayuda a determinar la naturaleza clonal de la enfermedad.*

***Fibrosis reticulínica leve (grado 1) secundaria a infección, enfermedad autoinmune u otro trastorno inflamatorio crónico, tricoleucemia u otra neoplasia linfoide, cáncer metastásico o mielopatía tóxica (crónica).*

A nivel de los criterios mayores, la fase inicial/prefibrótica se caracteriza por la ausencia o presencia mínima de fibrosis reticulínica en la biopsia medular, mientras que la fase establecida muestra una fibrosis reticulínica más prominente. Los criterios menores se repiten, pero en la fase establecida se aumenta el síndrome leucoeritroblástico en sangre periférica, que es la presencia simultánea de glóbulos inmaduros rojos y blancos en la sangre circulante.

Existe también la posibilidad de ser diagnosticado con una MFP posterior a una Policitemia Vera (MFP post-PV) o a una Trombocitemia Esencial (MFP post-TE), en los que se presentan otros síntomas adicionales, volviéndola una condición más agresiva para la salud del paciente.

Las pruebas iniciales para el diagnóstico de la mielofibrosis incluyen:

- Anamnesis y exploración física para detectar síntomas constitucionales y signos de otras enfermedades.
- Análisis de sangre para evaluar el balance hematológico, LDH, ácido úrico, ferritina y vitamina B12, así como pruebas básicas de coagulación.
- Serologías para detectar infecciones como VIH, VHC y VHB.
- Estudios moleculares secuenciales para detectar mutaciones, como el *JAK2* p.V617F, *CALR* (exón 9) y *MPL* (exón 10), seguido de otras mutaciones en genes mieloides.
- Aspirado y biopsia medular para estudiar la morfología y citogenética de las células.

2.2. Tipos de variables y escalas de medición

En el análisis de datos y la investigación científica, es fundamental comprender los diferentes tipos de variables y las escalas de medición. Estas clasificaciones son esenciales para el diseño de estudios y la interpretación de resultados.

Según la forma de recoger datos, las variables generalmente se clasifican en dos categorías: cualitativas y cuantitativas; y se asignan los valores a cada variable según una escala de medición que se haya definido en el estudio. A continuación, una clasificación del tipo de variable y sus escalas de medición.

Variables cualitativas: Estas variables describen categorías de clasificación y no tienen un valor numérico intrínseco. Se subdividen en:

- Nominales o categóricas: Los datos son categorías de clasificación como el sexo o el lugar de procedencia. No hay números que indiquen orden o cantidad, y las operaciones matemáticas permitidas son limitadas a conteo (Morales, 2012). Ejemplos incluyen el sexo (masculino/femenino) codificado como 1 y 2, respectivamente, donde los números son identificadores arbitrarios (Orlandoni, 2010).

- Dicotómicas: Un subtipo de las nominales con sólo dos categorías mutuamente excluyentes, como responder sí o no a una pregunta. Estas variables suelen codificarse con 1 o 0, representando presencia o ausencia de una característica (Morales, 2012).
- Ordinales: Variables que representan un orden específico entre categorías. No se conoce la distancia entre las categorías, sólo el orden. Ejemplos incluyen el nivel educativo (primario, secundario, universitario), donde se puede ordenar, pero no medir la distancia entre niveles (Orlandoni, 2010).

Variables cuantitativas: Estas variables representan cantidades numéricas y pueden ser medidas. Se subdividen en:

- De intervalo: Variables que permiten medir distancias entre valores con igualdad de intervalos, pero no tienen un cero absoluto. Un ejemplo es la temperatura en grados Celsius, donde el cero es arbitrario (Orlandoni, 2010).
- De razón: Variables que tienen un verdadero cero y permiten todas las operaciones matemáticas. Ejemplos incluyen la altura, el peso y el ingreso anual, donde el cero representa la ausencia total de la magnitud medida (Morales, 2012) (Orlandoni, 2010).

En la presente tesis, se hará un uso extensivo de variables dicotómicas para evaluar la presencia o ausencia de ciertas características clínicas en la muestra estudiada.

2.3. Estadística Multivariante

Cuando se toma una muestra de datos, es común desear analizar más de una característica simultáneamente. No obstante, los análisis de datos suelen comenzar con enfoques univariantes o bivariantes. Sin embargo, desde finales del siglo XIX, se han desarrollado diversas técnicas que permiten analizar múltiples variables de manera simultánea, lo que conduce a una comprensión más completa y detallada de los datos. Estas técnicas permiten, por ejemplo, explorar relaciones complejas entre variables o identificar similitudes y diferencias entre observaciones. Entre las técnicas clásicas se incluyen:

- Regresión Lineal Múltiple (Galton, 1886).
- Análisis de Componentes Principales [ACP] (Pearson, 1901).
- Análisis Discriminante Lineal [ADL] (Fisher, 1936).
- Análisis Factorial (Spearman, 1904).

- Análisis de Clúster (Driver & Kroeber, 1932).
- Análisis de Coordenadas Principales [PCoA] (Gower, 1966).
- Análisis de Correspondencias Múltiple [ACM] (Benzecri, 1973).
- Modelos de Ecuaciones Estructurales [SEM] (Jöreskog, 1970).

Adicionalmente, en 1971, Gabriel y otros investigadores introdujeron el concepto de "Biplot", una técnica multivariante que permite representar simultáneamente las coordenadas de las observaciones y de las variables de un conjunto de datos en un solo gráfico, proporcionando una visión integral de la estructura de la matriz de datos. Las técnicas de visualización multivariante, como el Biplot, han experimentado un desarrollo significativo desde la década de 1980, demostrando una gran capacidad para extraer y comunicar información compleja de manera eficiente. A continuación, se detalla el procedimiento de cálculo para una técnica que combina el Análisis de Coordenadas Principales (PCoA) y el modelo de Regresión Logística (LR), denominada Biplot Logístico Externo (ELB).

2.4. Biplot Logístico Externo

Esta técnica multivariante permite construir una gráfica en donde se representa simultáneamente la información de las filas (individuos) y columnas (variables) de una matriz de datos binarios (Demey, Vicente-Villardón, Galindo-Villardón, & Zambrano, 2008). Similar a otras técnicas biplots, los vectores representan a las variables, pero la diferencia en ese sentido es que la orientación de los vectores indica la dirección en la que incrementa la probabilidad de que una característica (binaria) esté presente en los individuos situados en el gráfico bidimensional. Esta gráfica tiene una gran utilidad para clasificar o determinar las variables con mayor poder discriminante cuando estas sean las más cortas entre los puntos de predicción de probabilidad 0.5 y 0.75. A continuación, se describen los pasos de cálculo que sigue este método, según Demey (2008):

1. Preparación de los datos: La técnica comienza utilizando una matriz de datos binarios **X** que pueden representar la presencia/ausencia de una variable. La matriz **X** no debe tener datos perdidos.
2. Análisis de Coordenadas Principales (PCoA). Extraer las coordenadas principales de la matriz **X** utilizando un coeficiente de similitud adecuado.

3. Regresión Logística: Calcular regresiones logísticas usando las coordenadas principales como variables independientes y cada variable dicotómica como dependiente.
4. Usar la corrección de Bonferroni para filtrar las variables con alta capacidad de discriminación.
5. Construcción del biplot. Dibujar los elementos (filas y columnas) en un gráfico biplot usando las coordenadas principales y los coeficientes de las regresiones logísticas estimadas.

Tras estos pasos, el algoritmo ELB ofrece una visualización multivariante con una visión más clara de la estructura de los datos, obteniendo interpretaciones sobre las relaciones individuo-individuo, individuo-variable y variable-variable. Esta metodología representa una mejora con respecto a los métodos de clasificación tradicionales.

CAPÍTULO 3

3. METODOLOGÍA

3.1. Diseño de la Investigación

Ahora bien, la motivación de este estudio es la aplicación de métodos estadísticos clásicos y multivariantes, para lo que se ha utilizado datos recopilados de pacientes diagnosticados con una condición específica. Se trata de una investigación de tipo exploratoria, cuantitativa, no experimental, con datos obtenidos de manera *multicéntrica* y considerados de corte *transversal*. Estos dos últimos aspectos se explican en la siguiente sección con la fuente de datos. La elección de la metodología para esta investigación se basa en que la información clínica de los pacientes fue recibida en forma tabular y también por la orientación cuantitativa de los objetivos planteados.

3.1.1. Fuente, población y muestra

Un grupo de hematólogos recopilaron los datos de varias instituciones de salud en Ecuador, lo que le dio al estudio un carácter multicéntrico. A pesar de esto, es importante mencionar dos aclaraciones que determinan a la población: 1) El diagnóstico de las NMP Ph- puede haber sido realizado dentro de dichas instituciones, o también pueden haber sido diagnosticados fuera de estas, tan solo se confirma el registro; y 2) No es posible definir que todos los individuos eran ecuatorianos, dado que esa información no fue provista. Los profesionales encargados registraron a los pacientes que atendieron con el diagnóstico de NMP Ph- en la plataforma informática https://www.nmp-ecuador.com/Login.aspx, y se cuenta con registros desde 2014 hasta 2023. Se consideró que estos datos son *transversales*, ya que en su mayoría corresponden a resultados de exámenes médicos realizados al momento del diagnóstico de cada paciente. Es preciso mencionar que los datos provistos para esta investigación incluyeron unas pocas características de seguimiento clínico, pero estas variables no incumplen los requisitos de las técnicas que se emplearon.

Dada las condiciones mencionadas, la población objeto de estudio comprende a los individuos registrados en una de estas instituciones de Ecuador, con un diagnóstico de Neoplasias Mieloproliferativas Philadelphia Negativas (NMP Ph-) en una de sus tres categorías principales: Policitemia Vera (PV), Trombocitemia Esencial (TE) y Mielofibrosis Primaria (MFP). La muestra alcanza un tamaño de 111 personas atendidas y registradas por los profesionales de estas instituciones en un período de 10 años.

3.1.2. Estrategia Metodológica

Partiendo de la estadística como una ciencia fundamental en estudios cuantitativos, se realizaron estadísticas descriptivas de unas pocas variables que previamente se tenía interés de explorar, pero por lo expuesto sobre Big Data/Alta dimensión, el proceso que se siguió en esta investigación se resume en los siguientes pasos:

1. Exploración de datos (EDA).
2. Preprocesamiento.
3. Selección de variables.
4. Imputación.
5. Métodos Multivariantes.

Ilustración 1: Secuencia metodológica

Como parte del EDA, se realizó exploraciones tanto de forma, y de algunas variables de interés. Esta etapa no incluyó completamente el análisis univariante y bivariante de todas las variables desde el inicio, debido a la alta dimensionalidad. Mas bien, a medida que se extraían resultados con los métodos multivariantes, se complementaba la investigación con análisis específicos de ciertas variables con cierta relevancia.

3.1.3. Software

Para el tratamiento y análisis de datos empleados se utilizó el software de código abierto “R” con su interfaz gráfica “R Studio”, desarrollado por la empresa “Posit” (https://posit.co/).

Con el auge de la ciencia de datos, han surgido numerosas innovaciones en los métodos de análisis de datos, los métodos estadísticos y la estadística computacional. Entre estas innovaciones destaca la colección de librerías "Tidyverse" para R, creado por Hadley

Wickham. Las funciones de estos paquetes de código optimizaron la programación en R para todos los pasos y técnicas empleadas, especialmente mediante el operador *pipe* (%>%), que simplifica la transición entre procesos consecutivos. Además, se utilizó la librería "Factoshiny", que permite realizar diversas técnicas multivariantes y visualizaciones sin necesidad de programar, a través de una segunda interfaz gráfica dinámica e intuitiva. Finalmente, se empleó programas de edición de imágenes para pulir los detalles gráficos de las visualizaciones estadísticas y facilitar la interpretación y comunicación de los resultados.

3.2. Conjunto de datos (Exploración)

La matriz de datos provista contenía 119 columnas y 111 casos de NMP Ph- se dividían en:

- 51 casos (46%) de Policitemia Vera,
- 46 casos (41%) de Trombocitemia Esencial,
- 1 de Mielofibrosis Primaria en la etapa denominada “Prefibrótica”,
- 10 de Mielofibrosis Primaria en la etapa “Fibrótica”,
- 3 Neoplasias Mieloproliferativas no clasificables (2,7%).

La matriz contiene la información clínica de cada paciente en grupos de (variables) análisis médicos que se detallan Tabla 6:

Tabla 6: Variables del conjunto de datos

Grupo	N°	Nombre de variable	Tipo / Escala de medición
Identificador	1	ID	Texto
Diagnóstico	2	DX NMP	Nominal
	3	Comentario	Texto
Epidemiológico	4	Sexo	Dicotómica
	5	FECHA NACI	Fecha
	6	18 – 30 años	Dicotómica
	7	31 - 50 años	Dicotómica
	8	51 - 65 años	Dicotómica
	9	mayor a 65 años	Dicotómica
Hábitos	10	Alcohol	Dicotómica
	11	Tabaco	Dicotómica
	12	Sedentarismo	Dicotómica
	13	Comentario	Texto
Antecedentes patológicos y comorbilidades	14	Comorbilidades	Dicotómica
	15	Diabetes mellitus	Dicotómica
	16	Hipertensión arterial	Dicotómica
	17	Hipertiroidismo	Dicotómica
	18	Hipotiroidismo	Dicotómica
	19	Comentario	Texto
	20	Esplenectomía	Dicotómica
	21	Quimioterapia previa	Dicotómica
	22	Radioterapia previa	Dicotómica
	23	Anticoagulantes	Dicotómica
Síntomas	24	Fatiga	Dicotómica
	25	Pérdida de peso	Dicotómica

	26	Sudoración nocturna	Dicotómica
	27	Prurito	Dicotómica
	28	Fiebre	Dicotómica
	29	Dolor óseo	Dicotómica
	30	Migraña	Dicotómica
Examen físico	31	Tamaño del bazo	Cuantitativa
	32	Tamaño del hígado	Cuantitativa
	33	Presencia de adenopatías	Dicotómica
	34	Tamaño de adenopatías	Ordinal
Laboratorio clínico	35	Hemoglobina	Ordinal
	36	Hematocrito	Ordinal
	37	Eritropoyetina sérica (EPO)	Dicotómica
	38	Leucocitos	Ordinal
	39	Plaquetas	Ordinal
	40	Reticulocitos	Ordinal
	41	Ferritina	Ordinal
	42	Vitamina B12	Ordinal
	43	LDH	Ordinal
Especiales - Sangre	44	Presencia de Blastos	Dicotómica
	45	Cantidad de Blastos	Cuantitativa
	46	Presencia de Leucoeritroblastosis	Dicotómica
	47	Presencia de Dacriocitos	Dicotómica
Especiales- Mielograma (Historia clínica)	48	Celularidad hematopoyética	Cuantitativa
	49	Cantidad de serie eritroide	Ordinal
	50	Cantidad de serie granulocítica	Ordinal
	51	Cantidad de serie megacariocítica	Ordinal
	52	Cantidad de Blastos	Cuantitativa
	53	Eritroides	Dicotómica
	54	Granulocíticos	Dicotómica
	55	Linfoides	Dicotómica
	56	Megacariocíticos	Dicotómica
	57	Monocitoides	Dicotómica
	58	Comentario	Texto
	59	Presencia de Displasia	Dicotómica
	60	Eritroides	Dicotómica
	61	Granulocíticos	Dicotómica
	62	Linfoides	Dicotómica
	63	Megacariocíticos	Dicotómica
	64	Monocitoides	Dicotómica
	65	Mixta	Dicotómica
	66	Comentario	Texto
Histopatología - Médula Ósea (Informe histopatológico)	67	Celularidad hematopoyética	Cuantitativa
	68	Cantidad de serie eritroide	Ordinal
	69	Cantidad de serie granulocítica	Ordinal
	70	Cantidad de serie megacariocítica	Ordinal
	71	Tamaño de megacariocitos	Ordinal
	72	Conglomerados de megacariocitos	Dicotómica
	73	Cantidad de Blastos	Cuantitativa
	74	Eritroides	Dicotómica
	75	Granulocíticos	Dicotómica
	76	Linfoides	Dicotómica
	77	Megacariocíticos	Dicotómica
	78	Monocitoides	Dicotómica
	79	Comentario	Texto

	80	Presencia de Displasia	Dicotómica
	81	Eritroides	Dicotómica
	82	Granulocíticos	Dicotómica
	83	Linfoides	Dicotómica
	84	Megacariocíticos	Dicotómica
	85	Monocitoides	Dicotómica
	86	Mixta	Dicotómica
	87	Comentario	Texto
	88	Hierro	Ordinal
	89	Fibrosis reticulínica	Ordinal
	90	Fibrosis colágena	Ordinal
	91	Osteoesclerosis	Ordinal
	92	Relación mielo:eritroide	Ordinal
	93	Uso de Histoquímica	Dicotómica
	94	Tipo de Histoquímica	Nominal
	95	Uso de Inmunohistoquímica	Dicotómica
	96	Tipo de Inmunohistoquímicas	Nominal
Complementarios Laboratorio	97	Blastos en Citometría de flujo	Cuantitativa
	98	Citogenética	Nominal
	99	Presencia de BCR-ABL1 (Biología molecular)	Nominal
	100	Presencia de JAK2 V617F (Biología molecular	Nominal
Clínica de desenlace	101	Puntaje IPSS al diagnóstico	Ordinal
	102	Puntaje DIPSS al ingreso al registro	Ordinal
	103	Aspirina	Dicotómica
	104	Esplenectomía	Dicotómica
	105	Flebotomías	Dicotómica
	106	Hidroxiurea	Dicotómica
	107	Interferón alfa	Dicotómica
	108	Ruxolitinib	Dicotómica
	109	Trasplante alogénico de progenitores hematopoyéticos	Dicotómica
	110	Otro(s)	Dicotómica
	111	Comentario	Texto
	112	Complicaciones trombóticas	Dicotómica
	113	Complicaciones hemorrágicas	Dicotómica
	114	Presencia de Transformación/progresión	Dicotómica
	115	Tipo de Transformación/progresión	Nominal
	116	Comentario	Texto
	117	Tiempo de Sobrevida	Cuantitativa
	118	Desenlace	Dicotómica
	119	Causas de muerte	Nominal

Como se observa existe heterogeneidad entre los grupos de variables y la escala de medición de todas ellas. En la Tabla 7 se resume la frecuencia de variables según el tipo o escala de medición:

Tabla 7: Distribución de variables según su tipo y escala de medida

Tipo de variable - Escala de medición	Cantidad
Cualitativa - Dicotómica	68
Cualitativa - Ordinal	23
Cualitativa - Nominal	8

Cuantitativa - Razón	9
Fecha	1
Texto	1
Identificador	1

Predominan las variables dicotómicas: dos resultados mutuamente excluyentes. Esto se da porque en su mayoría corresponden a variables que verifican si el paciente posee o no una característica clínica.

3.3. Preprocesamiento

Para garantizar la calidad y relevancia del análisis, posterior al análisis exploratorio, se llevó a cabo un exhaustivo preprocesamiento de los datos que, gracias al software R usando la notación de Tidyverse, se simplificó el orden del preprocesamiento en los siguientes pasos:

- Construcción de nuevas variables derivadas de las existentes.
- Corregir observaciones que pudieran haber sido registradas con inconsistencias.
- Consolidar diferentes formas de describir a los valores perdidos (NA).
- Remover columnas no informativas por diferentes criterios:
- Configurar formato adecuado de las variables en el software.
- Agrupar categorías de las NMP Ph- (PV, TE y MFP).
- Abreviar nombres de variables.

Algunos de estos pasos se repitieron o se ajustaron los criterios al momento de seleccionar subconjuntos de variables.

3.4. Selección de variables o características

La selección de características se realizó en varias etapas. En primer lugar, se efectuó una selección directa de las variables más pertinentes a los objetivos de estudio: “Clínico-patológicas y hematológicas”. Posteriormente, se descartaron las variables con varianza aproximada a cero o con un alto porcentaje de valores perdidos, eliminando así características que no aportaban información útil o que podrían introducir ruido en el análisis.

3.4.1. Selección directa

Se seleccionó de forma directa el conjunto de variables correspondientes a las variables de interés de los objetivos específicos. En la Tabla 8, se indica las variables y sus grupos, además de la variable con las categorías de diagnóstico de las NMP

Ph-, de tal manera que al aplicar las técnicas de exploración multivariante podamos agregar el menor ruido posible al conjunto de datos. A este subconjunto de variables se le denominó “hematopato” en el software.

Tabla 8: Conjunto de variables clínicas, hematológicas y patológicas

Diagnóstico	1	DX.NMP	Nominal
Laboratorio clínico	2	Hemoglobina	Ordinal
	3	Hematocrito	Ordinal
	4	Eritropoyetina sérica (EPO)	Dicotómica
	5	Leucocitos	Ordinal
	6	Plaquetas	Ordinal
	7	Reticulocitos	Ordinal
	8	Ferritina	Ordinal
	9	Vitamina B12	Ordinal
	10	LDH	Ordinal
Especiales - Sangre	11	Presencia de Blastos	Dicotómica
	12	Cantidad de Blastos	Cuantitativa
	13	Presencia de Leucoeritroblastosis	Dicotómica
	14	Presencia de Dacriocitos	Dicotómica
Especiales-Mielograma (Historia clínica)	15	Celularidad hematopoyética	Cuantitativa
	16	Cantidad de serie eritroide	Ordinal
	17	Cantidad de serie granulocítica	Ordinal
	18	Cantidad de serie megacariocítica	Ordinal
	19	Cantidad de Blastos	Cuantitativa
	20	Eritroides	Dicotómica
	21	Granulocíticos	Dicotómica
	22	Linfoides	Dicotómica
	23	Megacariocíticos	Dicotómica
	24	Monocitoides	Dicotómica
	25	Presencia de Displasia	Dicotómica
	26	Eritroides	Dicotómica
	27	Granulocíticos	Dicotómica
	28	Linfoides	Dicotómica
	29	Megacariocíticos	Dicotómica
	30	Monocitoides	Dicotómica
	31	Mixta	Dicotómica
Histopatología - Médula Ósea	32	Celularidad hematopoyética	Cuantitativa
	33	Cantidad de serie eritroide	Ordinal

(Informe histopatológico)	34	Cantidad de serie granulocítica	Ordinal
	35	Cantidad de serie megacariocítica	Ordinal
	36	Tamaño de megacariocitos	Ordinal
	37	Conglomerados de megacariocitos	Dicotómica
	38	Cantidad de Blastos	Cuantitativa
	39	Eritroides	Dicotómica
	40	Granulocíticos	Dicotómica
	41	Linfoides	Dicotómica
	42	Megacariocíticos	Dicotómica
	43	Monocitoides	Dicotómica
	44	Presencia de Displasia	Dicotómica
	45	Eritroides	Dicotómica
	46	Granulocíticos	Dicotómica
	47	Linfoides	Dicotómica
	48	Megacariocíticos	Dicotómica
	49	Monocitoides	Dicotómica
	50	Mixta	Dicotómica
	51	Hierro	Ordinal
	52	Fibrosis reticulínica	Ordinal
	53	Fibrosis colágena	Ordinal
	54	Osteoesclerosis	Ordinal
	55	Relación mielo:eritroide	Ordinal

Adicionalmente, se generaron otros conjuntos de datos con las variables restantes o de interés específico, para estudiarlas posterior al conjunto de variables hematológicas y patológicas.

3.4.2. Criterios de descarte

Tres criterios ayudaron a separar variables poco o nada informativas desde el punto de vista estadístico, aunque para el campo médico pudieran tener alguna información relevante. Los criterios establecidos son:

- Variables con una misma respuesta en todos los registros (varianza cero).
- Variables con muy poca variabilidad o con cerca del 100% de registros con un mismo valor (>98%).
- Variables con alto porcentaje de valores perdidos o “NA”.
- Variables de texto o comentarios.

Se probó diferentes porcentajes de valores perdidos y al examinar las variables que se retenían, luego de cumplir con los demás criterios de descarte y tolerar un máximo

de 30% de NA, se retuvieron 13 variables del subconjunto “hematopato”. Sin embargo, se permitió llegar hasta un 40% de valores perdidos, dado que permitía retener una variable adicional, “Eritropoyetina sérica (EPO)”. Esta variable es de interés, dado que forma parte de los criterios para el diagnóstico de las NMP Ph-.

3.4.3. Imputación.

Tras la depuración, las variables se sometieron a un proceso de imputación de valores perdidos, para que se pudieran emplear tres técnicas de identificación de las variables más influyentes: Regresión Logística Multinomial, Bosques aleatorios y análisis de correlación policórica. El objetivo de este paso fue seleccionar las mejores variables para clasificar o discriminar a los grupos de NMP Ph- que se consideraron en una variable de respuesta a la que se buscaba explicar. La motivación detrás de este paso fue encontrar las variables que mejor se asocien linealmente o de otro orden, para la construcción de buenos ejes latentes en un Análisis de Correspondencias Múltiples (ACM), es decir, ejes que conserven de la mejor manera la variabilidad o inercia total de la matriz de datos.

3.4.4. Selección de características usando modelos

Al final de un proceso de algunos pasos para seleccionar variables, se espera aplicar técnicas multivariantes para explorar los datos con mayor eficiencia. Los métodos multivariantes como PCA, MCA y MFA tienen en común la construcción de ejes relacionados linealmente con las variables observadas o con otras latentes. Por esta razón, es importante contar con variables que tengan relación, al menos lineal, para que puedan construir ejes que capturen la mayor cantidad de información. Por esta razón, se incluyó un paso de selección de variables que busque cumplir dicho objetivo, y los métodos usados fueron:

- Regresión logística multinomial.
- Bosques Aleatorios.
- Análisis de la matriz de correlación policórica.

Los dos primeros métodos tienen como objetivo identificar las variables más importantes para modelar o clasificar a las categorías de la enfermedad. El tercer método busca cuantificar el nivel de relación de las variables a través del coeficiente de correlación policórica, para seleccionar aquellas que están más relacionadas y que se asociarían mejor en nuevos ejes latentes.

3.5. Uso del Análisis de Correspondencias Múltiple (ACM)

El ACM se aplicó específicamente al subconjunto de variables “Hematopato”. Se realizaron dos análisis distintos: uno con un subconjunto menor de variables consideradas más importantes, identificadas a través de Regresión Logística Multinomial, Bosques aleatorios y correlación policórica; y otro análisis adicional, sobre todas las demás variables no asociadas con mielograma e histopatología, permitiendo así un análisis integral y detallado de los datos. Como parte del ACM, se revisaron detalles como la inercia de las dimensiones, biplots, y las variables y características asociadas a cada categoría de las NMP Ph-.

3.5.1. Análisis de Inercia y Visualización de Resultados

Para cada ACM realizado, se analizó la inercia de los subconjuntos de variables, permitiendo evaluar la cantidad de variabilidad explicada por cada eje latente. Los resultados se visualizaron mediante biplots, donde se representaron cada grupo de las Neoplasias Mieloproliferativas Philadelphia Negativa pintadas con colores diferentes y las variables con sus categorías. Se destacó la calidad de las variables según el hiperplano en el que se dibujaban, identificando aquellas que mejor se relacionan con cada eje y grupo de las NMP. Además, se generaron gráficos de individuos para detectar y analizar posibles observaciones atípicas.

3.5.2. Interpretación de Dimensiones

Las dimensiones obtenidas del ACM se interpretaron según las contribuciones de las variables. Se prestó atención a las variables con contribuciones altas en unos ejes y bajas en otros, ya que éstas son particularmente interesantes para entender cómo discriminan entre los diferentes grupos. Esta interpretación permitió una comprensión más profunda de las relaciones subyacentes entre las variables y los grupos de estudio.

3.5.3. Combinación de Planos

Adicionalmente, se realizó diferentes biplots combinando los ejes obtenidos del ACM y formando distintos hiperplanos para proporcionar una visión más amplia de los pacientes y las variables exploradas. Esta combinación de planos permitió obtener vistas diferentes y corroborar algunas particularidades de los individuos y sus características, demostrando la eficacia de analizar más de dos dimensiones principales.

Esta metodología aseguró una exploración rigurosa y exhaustiva, para extraer algunas conclusiones significativas y fiables a partir de los datos disponibles.

3.6. Análisis Discriminante (LDA)

Con el objetivo de obtener otros resultados de un método multivariante diseñado para discriminar observaciones o grupos, se aplicó el análisis discriminante lineal, investigando las características que más contribuyeron a la construcción de las funciones discriminantes, para explorar los resultados de discriminación de este método.

3.7. Biplot logístico Externo

Para esta técnica fue necesario contar con una matriz completamente de variables binarias. Se utilizó el aplicativo MultBiplot y el software R con la librería denominada "MultBiplotR". En R se transformó la matriz de datos en columnas de unos y ceros, donde cada columna representa una categoría específica de una variable categórica nominal u ordinal, luego se llevaron estos datos al aplicativo MultBiplot donde se realizaron los cálculos del Biplot Logístico Externo (Vicente-Villardón & Hernández-Sánchez, 2020), con el fin de interpretar los vectores de longitudes más pequeñas en el gráfico, ya que aquellos representan las variables con mayor capacidad discriminante. Posteriormente, se llevó la imagen al software de edición gráfica Adobe Illustrator para mejorar algunos detalles estéticos del biplot, que permitieron enfocar la atención en los grupos y en las variables con vectores cortos, para facilitar la interpretación del gráfico.

CAPÍTULO 4

4. RESULTADOS

4.1. Imputación

A pesar de que se toleró un porcentaje relativamente considerable de valores perdidos, no fueron muchas las agregadas, pero su interés clínico era importante. Las variables permitidas y su porcentaje de NA fueron:

- Hemoglobina: 19,8%
- Eritropoyetina sérica: 38,7%
- Presencia de Blastos: 1,8%
- Presencia de Dacriocitos: 0,9%
- Cantidad de Blastos: 5,4%.

La proporción total de observaciones perdidas en el subconjunto de variables retenidas fue del 4,76%. Cabe mencionar que se revisó si es apropiado la imputación de este tipo de variables médicas, y lo importante es el método de imputación, ya que, al analizar la única variable numérica retenida, se consideró imputar de manera univariante con la mediana de la variable sin valores perdidos. A continuación, se muestra los resultados de la variable antes y después de la imputación.

Tabla 9: Imputación de una variable numérica

Estadísticas antes y después de imputar la variable numérica "**Cantidad de Blastos**" utilizando la mediana de la variable.		
Estadísticos	**Antes**	**Después**
Min:	0,0000	0,0000
1er Cuartil (25%):	0,0000	0,0000
Mediana:	0,0000	0,0000
3er Cuartil (75%):	0,0000	0,0000
Máximo:	8,0000	8,0000
Media:	0,2762	0,2613
Cantidad de NA:	6	0

Considerando el rango de valores de esta variable, no se observa mayores cambios, solo una leve disminución en la media de la variable.

Posteriormente, tanto para las variables ordinales “Hemoglobina” y “Eritropoyetina sérica”, como para “Presencia de Blastos” y “Presencia de Dacriocitos” que son variables de presencia o ausencia del estudio médico, se utilizó el Análisis de Correspondencias Múltiple como método multivariante de imputación, que permitió encontrar la categoría más propensa a ser la correcta, considerando todas las variables estudiadas en los demás individuos que comparten características iguales o similares. Para este paso se utilizó la función “imputeMCA” del paquete “missMDA” en el software R y como ejemplo de los resultados se muestra una tabla cruzada con la categoría de la NMP Ph- para verificar con mayor detalle los resultados de la imputación.

Tabla 10: Imputación de una variable categórica

Tablas de frecuencia cruzada de "**Hemoglobina**" y las categorías de las NMP Ph-, antes y después de utilizar el ACM como método de imputación.								
	Antes				**Después**			
Categorías de Hemoglobina	MFP	PV	TE	NOS	MFP	PV	TE	NOS
3	0	0	0	1	0	0	0	1
2	6	0	1	0	6	0	1	0
1	3	0	4	0	3	0	4	0
0	1	1	31	1	1	1	31	1
4	0	34	5	1	1	50	10	1
Cantidad de NA	1	16	5	0	0	0	0	0

Todos los valores imputados se agregaron en la categoría “4” de hemoglobina, el cual representaba el nivel más alto. Esta imputación tiene un sentido médico, ya que en las NMP Ph- es común tener los niveles de hemoglobina elevados, incluso es un criterio para el diagnóstico de PV. Esta misma mecánica se realizó con otros subgrupos de variables.

Este paso de imputación de valores perdidos era necesario para usar las técnicas de selección de variables en el siguiente último paso.

4.2. Selección de características

Los resultados de este paso se muestran en la siguiente Tabla.

Tabla 11: Métodos de selección de variables

Regresión multinomial Library(caret) Top 10 variables con mayor importancia:	**Bosques aleatorios Library(caret) Top 10 variables con mayor importancia:**	**Correlación Policórica Library(polycor) 7 variables con corr > 0.7 + 1 variable numérica**
Presencia de Leucoeritroblastosis	Hemoglobina	Hemoglobina
Presencia de Dacriocitos	Plaquetas	Plaquetas
Eritropoyetina sérica	Presencia de Leucoeritroblastosis	Presencia de Blastos (S)
Hemoglobina	Presencia de Displasia (H)	Presencia de Leucoeritroblastosis
Plaquetas	Presencia de Dacriocitos	Presencia de Dacriocitos
Presencia de Displasia (H)	Eritropoyetina sérica	Granulocitos (M)
LDH	Leucocitos	Granulocitos (H)
Leucocitos	Cantidad de Blastos (M)	
Reticulocitos	LDH	
Cantidad de Blastos (M)	Reticulocitos	

(M): En mielograma
(H): En histopatología
(S): En sangre

Los dos primeros métodos dieron exactamente las mismas 10 variables importantes, aunque ordenadas de manera diferente. La tercera técnica identificó siete variables que fueron las más asociadas según el nivel de correlación policórica, para valores mayores a 0.7. En estas siete variables, coincidieron cinco que los dos métodos anteriores identificaron. Finalmente, se eligió las 10 variables seleccionadas en los primeros métodos para usarlas en los métodos multivariantes.

4.3. Estadísticas descriptivas

A continuación, un resumen de frecuencias de las categorías de las variables seleccionadas.

Distribución por Grupo NMP Ph-

Grupo	Frecuencia
PV	51
TE	46
MFP	11
NOS	3

Hemoglobina (Hb)

Hb	Frecuencia
Muy Baja	1
Baja	7
Mod. Baja	7
Normal	34
Elevada	62

EPO

EPO	Frecuencia
Inadecuada	76
Adecuada	35

Leucocitos (WBC)

WBC	Frecuencia
Muy Bajos	2
Bajos	2
Normal	77
Elevados	31

Plaquetas (PLT)

PLT	Frecuencia
Extrema Baja	1
Muy Bajas	1
Baja Leve	5
Normal	48
Elevadas	56

Reticulocitos (Retic)

Reticulocitos	Frecuencia
Normal	65
Medio Alto	45
Elevados	1

LDH

LDH	Frecuencia
Normal	33
Medio Alto	29
Muy Alto	49

Leucoeritroblastosis (Leucoeritro)

Leucoeritro	Frecuencia
No	100
Sí	11

Dacriocitos

Dacriocitos	Frecuencia
No	102
Sí	9

Displasia

Displasia	Frecuencia
No	78
Sí	33

Presencia de Blastos

Blastos	Frecuencia
0	105
1	6

Granulocitos Mielograma

GranulocitosM	Frecuencia
0	106
1	5

Granulocitos Histopatología

GranulocitosH	Frecuencia
0	105
1	6

4.4. Análisis de Correspondencias Múltiple (ACM)

Debido a que se tenía la mayor parte de variables de tipo categóricas y un tamaño muestral considerable, donde cada registro es independiente de los demás, se decidió aplicar el Análisis de Correspondencias Múltiple dadas las condiciones de los datos, utilizando las variables seleccionadas en bosques aleatorios y regresión multinomial.

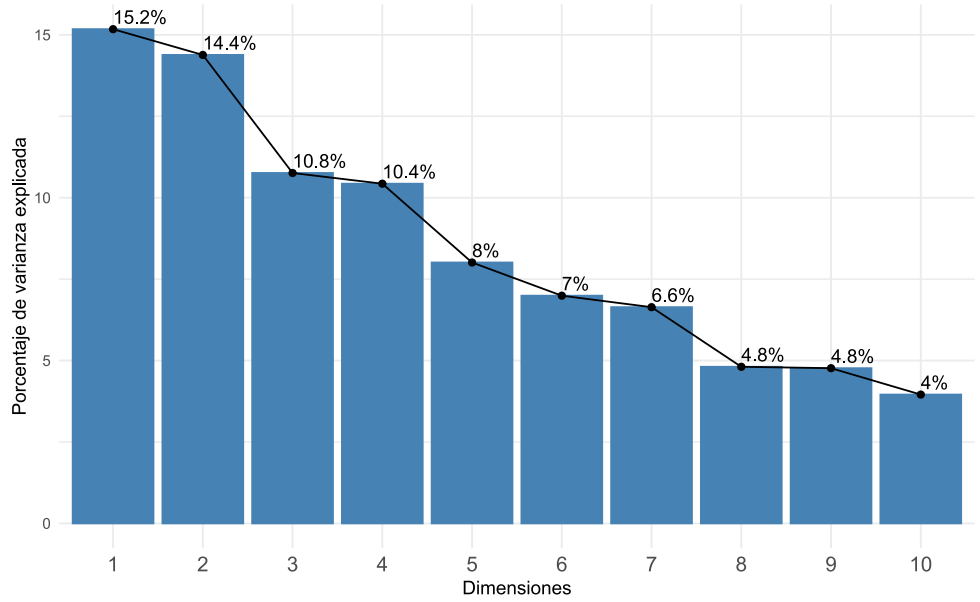
4.4.1. Determinación de la cantidad de dimensiones a analizar

Tabla 12: Varianza explicada por las dimensiones extraídas en el ACM

N° de la dimensión	Valor propio	Porcentaje de varianza	Porcentaje acumulado de varianza
Dim. 1	0.32	15.18	15.18
Dim. 2	0.30	14.38	29.56
Dim. 3	0.23	10.76	40.32
Dim. 4	0.22	10.43	50.75
Dim. 5	0.17	8.01	58.76
Dim. 6	0.15	6.99	65.75
Dim. 7	0.14	6.64	72.39
Dim. 8	0.10	4.81	77.20
Dim. 9	0.10	4.77	81.96
Dim. 10	0.08	3.96	85.92
Dim. 11	0.07	3.27	89.18
Dim. 12	0.05	2.56	91.75
Dim. 13	0.04	2.08	93.83
Dim. 14	0.03	1.63	95.45
Dim. 15	0.03	1.46	96.92
Dim. 16	0.03	1.23	98.15
Dim. 17	0.02	0.96	99.10
Dim. 18	0.02	0.90	100.00
Dim. 19	0.00	0.00	100.00

El porcentaje de varianza acumulado en las dos primeras dimensiones es del 30% de la inercia total, como se observa en la Tabla 12 y se visualiza en el gráfico de sedimentación.

Ilustración 2: Gráfico de sedimentación (descomposición de la varianza explicada por los ejes)



4.4.2. Interpretación de ejes

La contribución de cada variable para la construcción de los ejes dimensionales permite interpretarlos por tener características similares o con un sentido práctico. A continuación, se describe a cada dimensión por las variables y categorías que le contribuyen a su construcción, según la Ilustración 3.

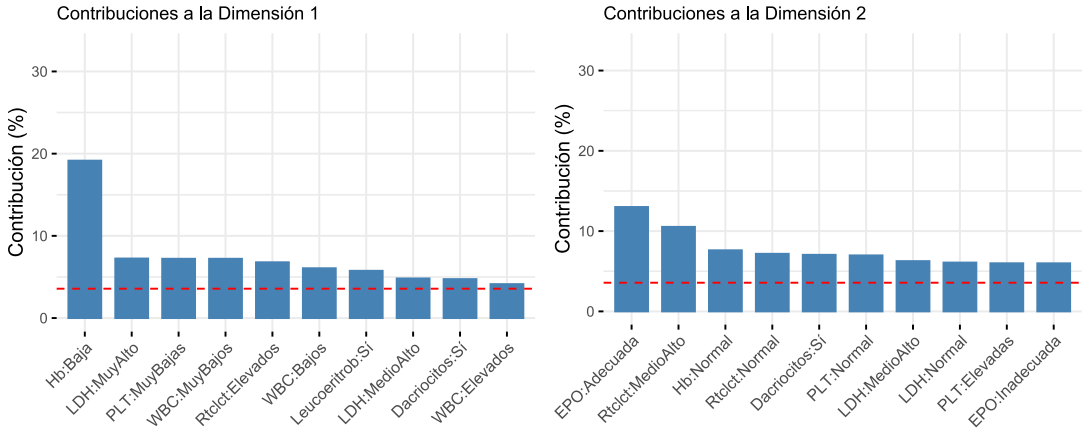
Primera dimensión (15,18% de la varianza)

Según el gráfico de contribuciones de las variables al eje 1, la primera dimensión está conformado fuertemente con la información de la Hemoglobina con código “2” que significa entre 6 - 8,9 mg/dL y se interpreta como un valor “medio-bajo”. Además, esta categoría tiene contribuciones bajas en las 3 dimensiones posteriores. Otras características médicas con valores bajos son los leucocitos y plaquetas. Por otra parte, este eje tiene valores altos en LDH y reticulocitos.

Segunda dimensión (14,38% de la varianza)

Diferente al eje 1, donde predominaba una categoría, la dimensión 2 comparte su inercia entre algunas variables que, además, tienen contribuciones bajas en los otros 3 ejes, y que en detalle son marcadores más estables, como la Eritropoyetina sérica mayor o igual a 125, equivalente a “Adecuado”. También están presentes los valores medios y bajos de Reticulocitos; Hemoglobina y plaquetas normales. Por último, este eje se ubican los resultados positivos para “Presencia de Dacriocitos”.

Ilustración 3: Contribución de las variables-categorías a la conformación de los nuevos ejes



4.4.3. Visualizaciones del Análisis de Correspondencias Múltiple

Previo a inspeccionar las variables y categorías que se relacionan con cada grupo de NMP Ph-, se presenta la Ilustración 4 y 5, de todos los individuos y de las variables-categorías, respectivamente, proyectados en las 2 primeras dimensiones.

Ilustración 4: Proyección de los pacientes con NMP Ph- en las 2 primeras dimensiones

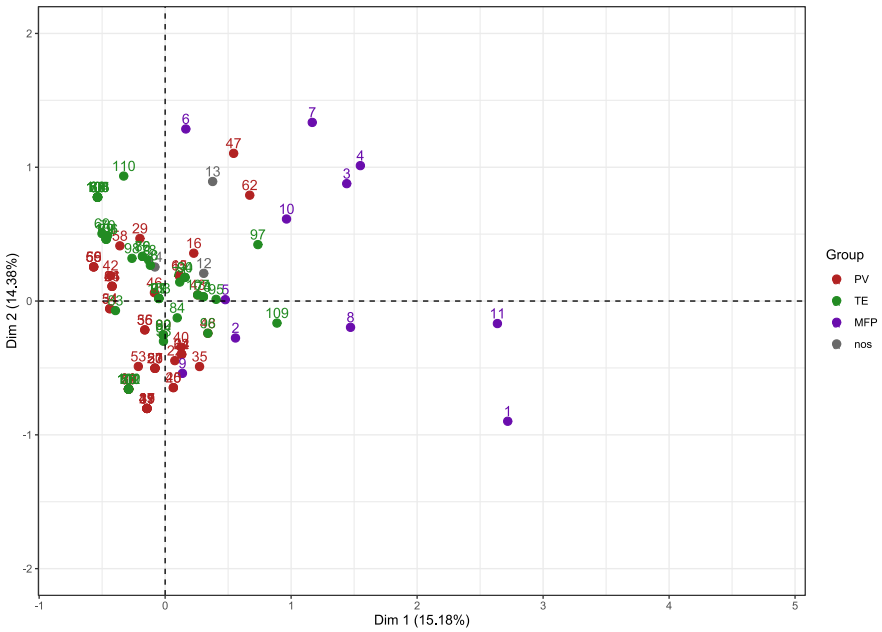
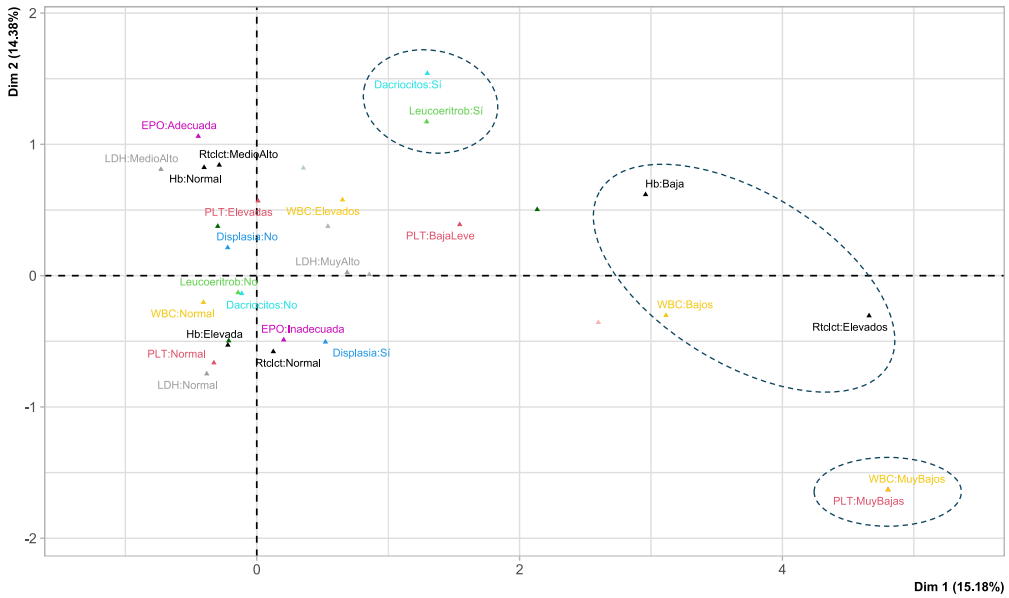


Ilustración 5: Categorías de las variables con una calidad de representación mínima de 0.1 (cos2)



Los puntos morados corresponden a los pacientes con MFP y se observa que tienden a separarse de los demás individuos en el sentido positivo del eje 1, indicando que existen características que los diferencia según las variables que construyen a dicho eje. Además, los pacientes con TE, PV y los no-clasificados (grises) no demuestran un patrón claro de separación, la interpretación indica que existen características similares en estos grupos, pero también es importante el análisis univariante de las variables-categorías cercanas a los puntos para confirmar similitudes y diferencias en los grupos que se estudia.

En la Ilustración 5, se proyectaron todas las categorías que tenían una calidad de representación (cos^2) de al menos 0.1, de tal manera, no están dibujadas todas las categorías, pero sí las más *precisas* en ese plano. Además, por color se pueden observar las categorías correspondientes a una misma variable; así, se proyectan los niveles de Leucocitos, en amarillo, en diferentes ubicaciones del primer plano. En esta ilustración se observan algunas relaciones en las categorías, por ejemplo, la presencia tanto de Leucoeritroblastosis y de Dacriocitos, o también los niveles "Muy bajos" de plaquetas y leucocitos. En otras categorías no se muestran relaciones específicas.

A continuación, se describen las características más representativas y asociadas a los pacientes según cada una de sus NMP Ph-.

Características de los pacientes con Policitemia Vera (PV)

Se analizaron conjuntamente las etiquetas o categorías del gráfico 6 que están más cerca de los puntos correspondientes a los pacientes con PV, y las frecuencias de las variables categóricas, condicionadas por el tamaño del grupo PV. De tal manera, se presentan las categorías más representativas de los pacientes con PV. Al final se destacan las categorías que tuvieron mayor frecuencia en el grupo:

Ilustración 6: Análisis de Correspondencias Múltiple - Policitemia Vera

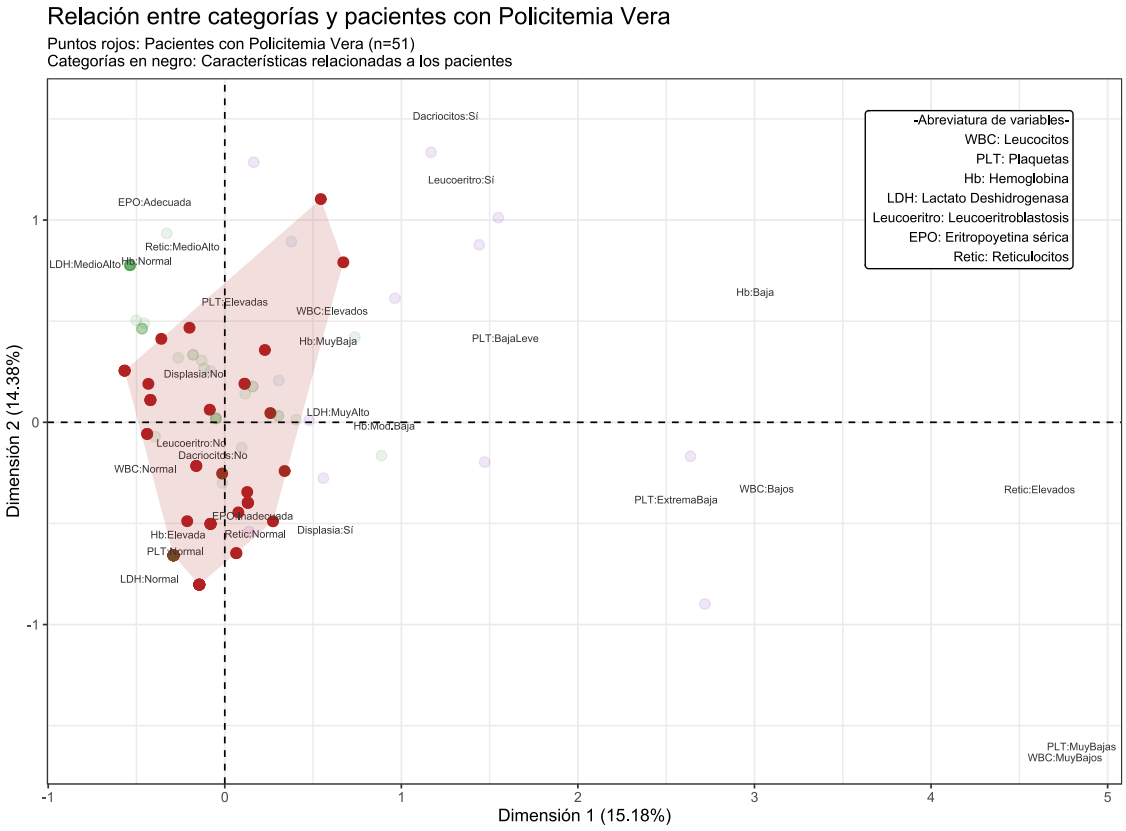


Tabla 13: Categorías cercanas a los pacientes con PV de la gráfica 6

Variable–Categoría (medidas)	Cant. pacientes PV (%) n=51
Leucocitos: Elevados (> 10.000 cel./mm^3)	11 (21,6%)
Plaquetas: Elevadas (> 400.000 cel./microL)	13 (25,5%)
Hemoglobina: Muy baja (< 6 mg/dL)	0 (0%)
Ausencia de Displasia	32 (62,7%)
LDH: Muy alto (> 200 UI/L)	20 (39,2%)
Ausencia de Leucoeritroblastosis	49 (96,1%)
Ausencia de Dacriocitos	48 (94,1%)
Leucocitos: Normal (5.000-10.000 cel./mm^3)	40 (78,4%)
EPO: Inadecuada (< 125 mU/mL)	41 (80,4%)
Hemoglobina: Elevada (> 16 mg/dL)	50 (98,0%)
Plaquetas: Normal (150.000-400.000 cel./microL)	36 (70,6%)

Reticulocitos: Normal (0 - 0,4 %)	33 (64,7%)
LDH: Normal (< 100 UI/L)	19 (37,3%)

Según la frecuencia, los pacientes con PV se resumen en tener Hemoglobina elevada, eritropoyetina sérica menor a 125 o inadecuada, pero tienen valores normales en Leucocitos, Plaquetas y Reticulocitos, y finalmente no poseen Leucoeritroblastosis, Dacriocitos ni Displasia. Este resumen se basa las características más frecuentes del grupo, pero también debe considerarse que algunos pacientes pueden tener otras características, aunque su frecuencia en el grupo sea menor. Un ejemplo de esto puede visualizarse con los niveles de la variable “Plaquetas”, ya que la categoría “Normal” se encuentra cerca de un grupo de individuos y la categoría “Elevadas” se encuentra cerca de otros pocos. Luego de analizar las tablas de frecuencias condicionadas al grupo, se inspeccionó que el nivel de plaquetas en el 70,6% de pacientes PV son normales, mientras que el 25% los tiene elevados. Allí la importancia de acompañar la visualización multivariante con el análisis univariante de las variables de interés.

Características de los pacientes con Trombocitemia Vera (TE)

En un principio se observó que existen pacientes con los mismos valores en las variables seleccionadas, por eso sus coordenadas eran las mismas en el espacio de dimensiones reducidas. Las categorías fueron revisadas con detalle para analizar solo aquellas que sus coordenadas tenían cercanía a la de los pacientes con TE. De la misma manera que antes, se interpretan las categorías más frecuentes en el grupo TE:

Ilustración 7: Análisis de Correspondencias Múltiple - Trombocitemia Esencial

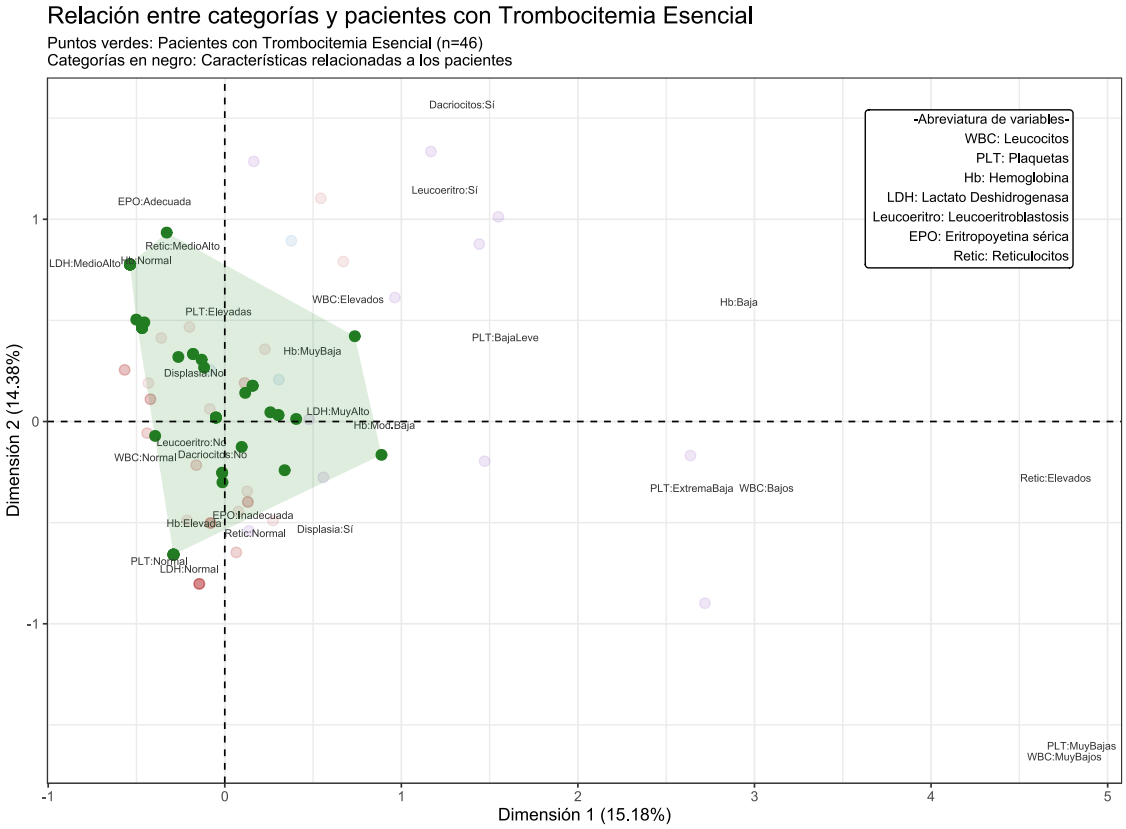


Tabla 14: Categorías cercanas a los pacientes con TE de la gráfica 7

Variable–Categoría (medidas)	Cant. pacientes TE (%) n=46
Reticulocitos: Medio Alto (0,5 – 2,9 %)	21 (45,7%)
LDH: Medio Alto (100-200 UI/L)	16 (34,8%)
Hemoglobina: Normal (12-16 mg/dL)	31 (67,4%)
Plaquetas: Elevadas (> 400.000 cel./microL)	39 (84,8%)
Leucocitos: Elevados (> 10.000 cel./mm^3)	11 (23,9%)
Hemoglobina: Muy baja (< 6 mg/dL)	0 (0%)
Ausencia de Displasia	40 (87%)
LDH: Muy alto (> 200 UI/L)	18 (39,1%)
Hemoglobina: Moderadamente Baja (9-11,9 mg/dL)	4 (8,7%%)
Ausencia de Leucoeritroblastosis	46 (100%)
Ausencia de Dacriocitos	45 (97,8%)
Leucocitos: Normal (5.000-10.000 cel./mm^3)	34 (73,9%)
Hemoglobina: Elevada (> 16 mg/dL)	10 (21,7%)
EPO: Inadecuada (< 125 mU/mL)	27 (58,7%)
Plaquetas: Normal (150.000-400.000 cel./microL)	7 (15,2%)
Reticulocitos: Normal (0 - 0,4 %)	25 (54,3%)
LDH: Normal (< 100 UI/L)	12 (26,1%)

En resumen, los pacientes con TE se caracterizan porque en su mayoría no hay presencia de Leucoeritroblastosis, Dacriocitos ni Displasia; un gran grupo de individuos (84,8%) poseen plaquetas elevadas, pero leucocitos y hemoglobina en niveles normales. Este resumen también se basa las características más frecuentes del grupo, y de igual manera que antes, existen pacientes que tienen otras características, aunque su frecuencia en el grupo sea menor. Por ejemplo, al observar la cercanía de algunas categorías de hemoglobina, se analizó que el 67,4% la tenía en un nivel normal, el 21,7% tenía la hemoglobina elevada, y el 10,8% la tenía de baja a moderada baja.

Características de los pacientes con Mielofibrosis Primaria (MFP)

Todos los individuos con MFP se ubicaron en el lado positivo del eje 1 (Ilustración 8), así, además de analizar las categorías proyectadas de manera cercana a los individuos, se exploró las categorías en este eje con coordenadas positivas que más contribuyeron a la dimensión. De igual manera que antes, se interpretan las características que fueron más frecuentes en los pacientes:

Ilustración 8 Análisis de Correspondencias Múltiple - Mielofibrosis Primaria

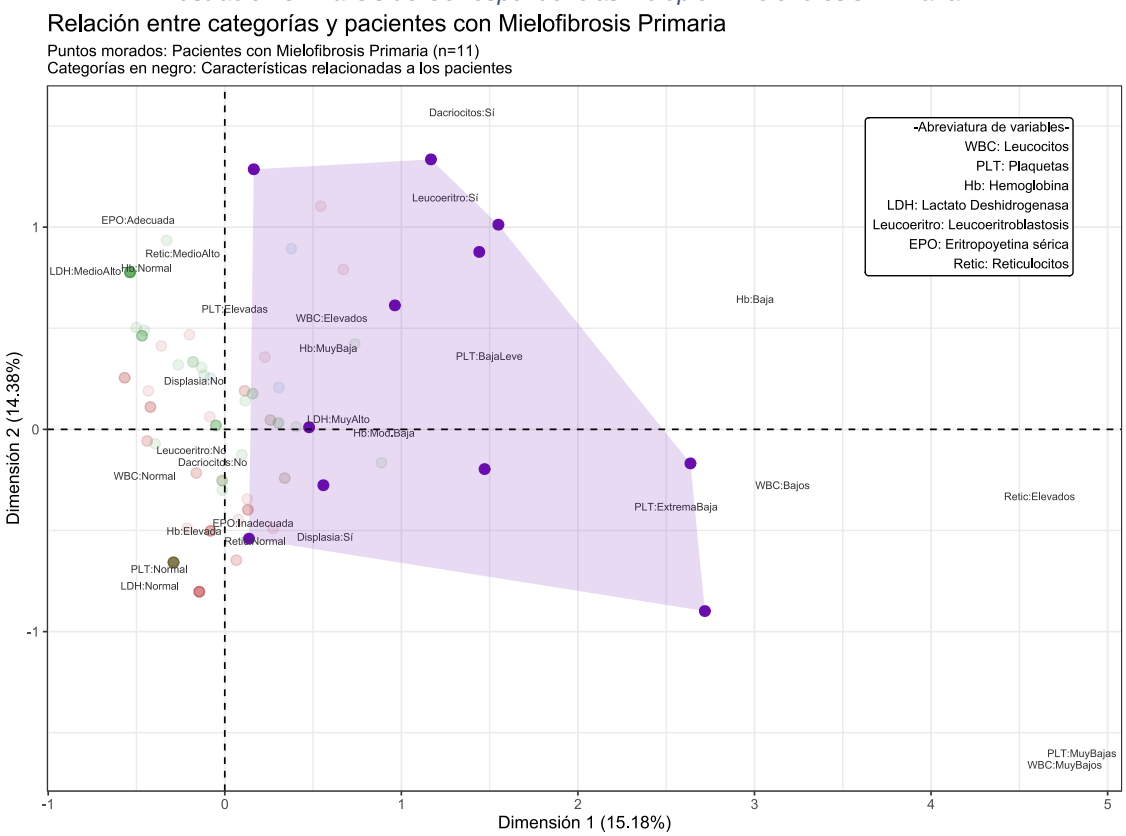


Tabla 15: Categorías cercanas a los pacientes con MFP de la gráfica 8

Variable–Categoría (medidas)	Cant. pacientes MFP (%)

	n=11
Presencia de Dacriocitos	5 (45,5%)
Presencia de Leucoeritroblastosis	8 (72,7%)
Leucocitos: Elevados (> 10.000 cel./mm^3)	6 (54,5%)
Hemoglobina: Baja (6,0-8,9 mg/dL)	6 (54,5%)
Plaquetas: Elevadas (>400.000 cel./microL)	3 (27,3%)
Plaquetas: Baja leve (100.000-149.999 cel./microL)	2 (18,2%)
Hemoglobina: Muy baja (< 6 mg/dL)	0 (0%)
LDH: Muy alto (> 200 UI/L)	9 (81,8%)
Hemoglobina: Moderadamente baja (9-11,9 mg/dL)	3 (27,3%)
Plaquetas: Extrema baja (< 10.000 cel./microL)	1 (9,1%)
Leucocitos: Bajos (3.000-4.999 cel./mm^3)	1 (9,1%)
Reticulocitos: Elevados (≥ 3 %)	1 (9,1%)
EPO: Inadecuada (< 125 mU/mL)	8 (72,7%)
Presencia de Displasia	7 (63,6%)
Reticulocitos: Normal (0 - 0,4 %)	6 (54,5%)
Plaquetas: Muy bajas (10.000-49.999 cel./microL)	1 (9,1%)
Leucocitos: Muy bajos (< 3.000 cel./mm^3)	1 (9,1%)

Pero además de lo frecuente, se encontró particularidades en aquellas categorías más alejadas en el sentido positivo del eje 1, por ejemplo, el individuo 1, el más alejado del grupo, es el único de grupo, y de todos los demás, que posee la combinación de características “Muy bajas” en plaquetas y leucocitos, a pesar de que “la mayoría” (>60%) presenta niveles de normales a elevados en estas variables, siendo una observación atípica en toda la muestra. Luego de explorar a ese paciente, se identificó que de los 11 pacientes con Mielofibrosis Primaria, aquella persona era la única que se encontraba en la fase “inicial” de la MFP, mientras que los otros 10 ya estaban en fase “establecida”. En dicha fase, los niveles de algunas variables varían. En los pacientes con MFP hubo 2 pacientes, de todos los registros, los únicos con un 8% de cantidad de blastos, que además es el valor máximo de la variable, y también atípico.

De esta manera se encontró otras observaciones atípicas dentro del mismo grupo, que a pesar de ser un grupo de menor tamaño en comparación a PV y TE, tiene características más variadas dentro de las variables de laboratorio clínico, hematológicas y patológicas.

En resumen, las características comunes de los pacientes analizados con MFP son la presencia de Leucoeritroblastosis y Displasias; niveles medios a bajos de

hemoglobina; altos niveles de lactodeshidrogenasa en estos pacientes, y, por último, existieron observaciones con características o niveles de variables “únicas” que los diferenciaba un poco del resto de pacientes.

Características de los pacientes no clasificados en las NMP Ph- principales (NOS)

En este grupo solo había 3 observaciones por estudiar. Exhaustivamente se encontró que en aquellos 3 pacientes todos cumplían con:

- Leucocitos: (>10.000) Elevados – 100%
- EPO: mayor o igual a 125=Adecuada – 100%
- Ausencia de Dacriocitos – 100%

Más allá de generalizar al grupo con estas variables, se los menciona como casos de estudio particulares. Su etiqueta de “no clasificados” es porque según los criterios de la OMS tienen similitud con más de una de las NMP Ph- clásicas, pero no terminan de identificarse completamente como alguna específica.

4.5. Análisis de las variables clínico-patológicas y hematológicas que caracterizan a los grupos de pacientes con NMP Ph-

Como se expuso en las Tablas 1, 2, 3 y 4, la OMS emite criterios estándares que rigen el diagnóstico de las categorías Policitemia Vera, Trombocitemia Esencial y Mielofibrosis Primarias en su fase prefibrótica o establecida, respectivamente de las Neoplasias Mieloproliferativas Philadelphia Negativas. Las variables más frecuentes que intervienen en dichos criterios son:

- Hemoglobina
- Hematocrito
- Hipercelularidad
- Prueba clínica de la mutación *JAK2* p.V617F
- Eritropoyetina sérica
- Trombocitosis
- Presencia de megacariocitos
- Fibrosis reticulínica y colágena

Posterior a esta caracterización clínica, se analizó en el conjunto de pacientes cuáles fueron las características (variables y categorías) más frecuentes en cada uno de los grupos de las NMP Ph-, considerando los resultados del ACM y las características con frecuencias mayores al 60% (Tabla 13, 14 y 15) en cada grupo. Como resumen de los hallazgos, se presenta la Tabla 16 con las categorías más frecuentes de cada variable para cada grupo de pacientes con las NMP Ph-.

Tabla 16 Características más frecuentes en cada grupo de NMP Ph-

Variables\Grupos	PV	TE	MFP	NOS
Hemoglobina	Elevada	Normal	Baja o Moderada Baja	
Leucocitos	Normal	Normal		Elevados
Plaquetas	Normal	Elevadas		
Reticulocitos	Normal			
LDH			Alto	
Leucoeritroblastosis	No	No	Sí	
Dacriocitos	No	No		No
Displasias	No	No	Sí	
EPO	Inadecuada		Inadecuada	Adecuada
Cantidad de Blastos			Valores mayores a lo normal.	

Al considerar solo las características de la Tabla 16 se perciben los perfiles de pacientes con Policitemia Vera y Trombocitemia Esencial muy similares, con indicadores normales y ausencia de anomalías hematológicas, excepto por sus diferencias en hemoglobina y plaquetas que, por definición médica de las enfermedades, coincide con ser lo más característico y lo que a la vez los diferencia. Mientras que la Mielofibrosis Primaria se mostró en los pacientes con anomalías como Leucoeritroblastosis y Displasias, una cantidad de blastos un poco más elevada de lo normal y la hemoglobina tiende a ser más baja. En resumen, son características que muestran a la enfermedad con más afectaciones en los pacientes. Este grupo es el más diverso y aleja sus características de la Policitemia Vera y la Trombocitemia Esencial.

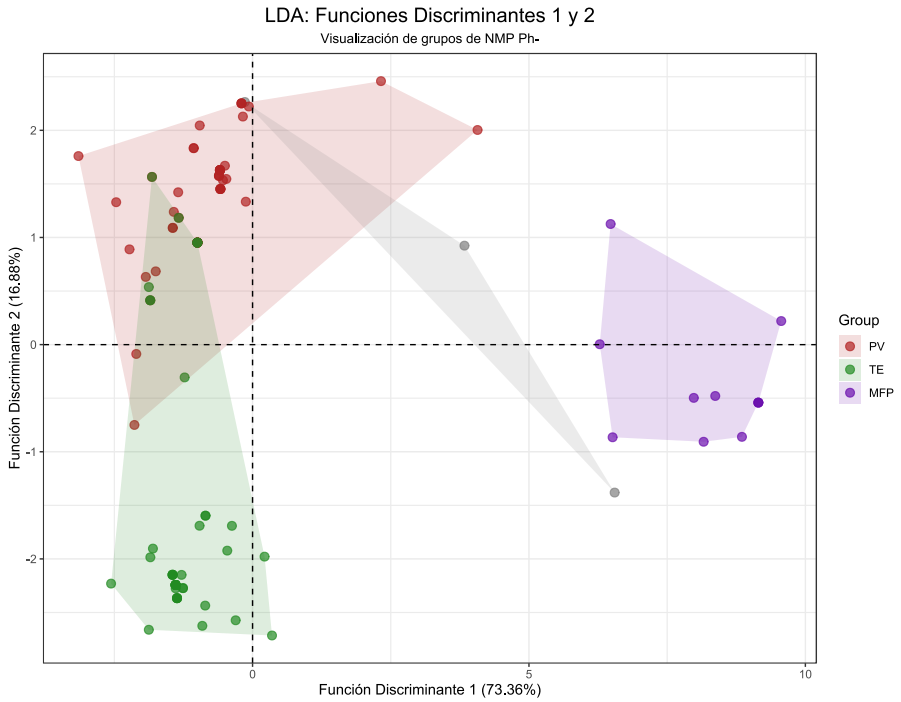
4.6. Análisis Discriminante

El análisis discriminante logró resumir el 90,2% de información en sus dos primeras dimensiones o funciones discriminantes, logrando una representación visual más eficiente de los grupos separados, en comparación a los métodos multivariantes utilizados anteriormente, donde los nuevos ejes se construyeron con otros criterios, como por ejemplo maximizar la variabilidad en sus componentes, pero no buscaban maximizar la distancia entre grupos.

En la Ilustración 9 se presenta el gráfico bidimensional del análisis discriminante, con los tres grupos de pacientes por el tipo de NMP Ph-. Se ha sombreado con

colores las áreas correspondientes a cada grupo para mejorar la visualización y facilitar la interpretación. Es notable cómo los pacientes con Mielofibrosis Primaria se separan de los otros dos grupos, en las coordenadas del eje horizontal. Mientras que el eje vertical logra distinguir mejor a los pacientes con Policitemia Vera y Trombocitemia Esencial, aunque algunos pacientes muestran coordenadas que se superponen con el otro grupo, indicando que, según la información de las variables utilizadas, existen pacientes TE con características similares a los de PV, y viceversa.

Ilustración 9: Gráfica del Análisis Discriminante Lineal



Además, dos individuos con NMP no clasificadas se proyectan cerca del grupo de Mielofibrosis Primaria y el otro de ellos se ubica con características muy similares a los pacientes con PV.

Para esta investigación, son cruciales las variables y categorías que más aportan a discriminar a los grupos, por tal motivo se examinaron los coeficientes de cada variable en las funciones discriminantes y se presentan en la Tabla 17. Estos valores se utilizaron para construir las gráficas de barras 10 y 11, para visualizar de manera más clara las variables y categorías que destacan en cada eje.

Tabla 17: Coeficiente de cada variable-categoría en cada función discriminante

Variable: Categoría	FD1	FD2
PLT: Elevadas	-7,91	-1,88
Hb: Normal	-7,26	-2,16
Retic: Elevados	8,59	0,67
Hb: Elevada	-7,74	0,62
Hb: Baja	-5,17	-2,98
Hb: Mod.Baja	-5,78	-1,87
Leucoeritro: Sí	7,07	-0,5
PLT: Baja Leve	-6,99	-0,37
PLT: Normal	-6,64	-0,84
Blastos: Sí	-2,97	-0,37
Dacriocitos: Sí	-0,77	1,62
WBC: Bajos	-1,49	-0,24
GranulocitosH: Sí	-1,17	0,01
Displasia: Sí	0,41	0,68
EPO: Adecuada	0,82	0,34
Retic: Medio Alto	-0,48	0,38
LDH: Muy Alto	0,42	0,5
GranulocitosM: Sí	-0,38	0,33
Cant. Blastos	0,65	0,05
LDH: Medio Alto	0,06	-0,09
WBC: Elevados	0,12	0,06
PLT: Muy Bajas	0,12	0,06
WBC: Normal	0,01	-0,04

Ilustración 10: Diagrama de barras de los coeficientes de la Función Discriminante 1

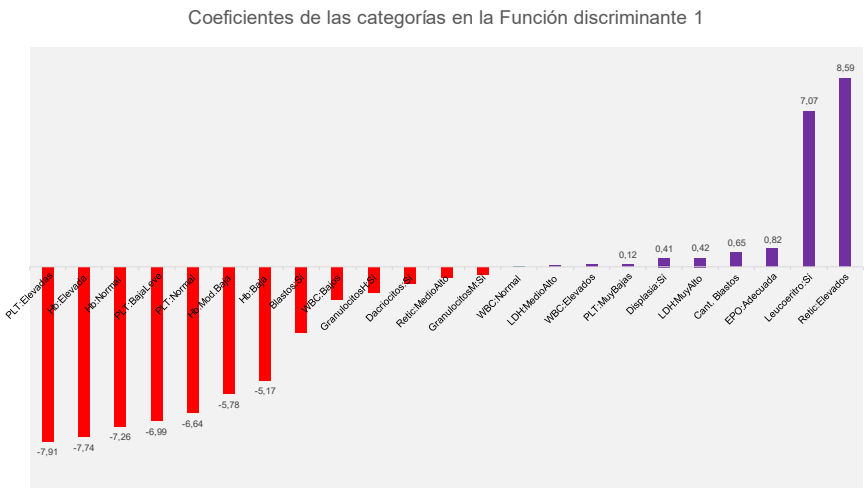
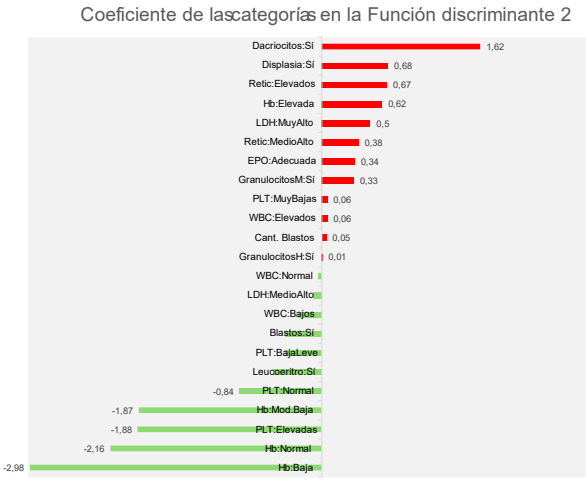


Ilustración 11: Diagrama de barras de los coeficientes de la Función Discriminante 2



Conforme a las magnitudes más grandes de las categorías, se observa que las características que más ayudaron a discriminar los grupos en el eje 1 fueron: Reticulocitos elevados, presencia de Leucoeritroblastosis, el nivel de plaquetas (elevadas, bajas y normales), también el nivel de hemoglobina (elevada, normal o baja). Mientras que en el eje 2 destacaron: la presencia de dacriocitos y displasias, nuevamente los niveles elevados de reticulocitos y de hemoglobina, aunque esta última variable, contrastó con la dirección de los pacientes que tenían hemoglobina baja y normal. Por último, también tomó peso el nivel elevado de plaquetas.

4.7. Biplot logístico Externo

La bondad de ajuste global del biplot logístico como porcentaje de clasificaciones correctas fue del 91,6%. La gráfica se presenta en la Ilustración 12. En este biplot, se enfatizó las variables que obtuvieron un R^2 mayor a 0.6 con el objetivo de tener una mejor imagen de la matriz de datos. Al resto de variables se les opacó su color, pero se las mantuvo en el gráfico para entender su ubicación y dirección. Acompañando a la visualización, se presenta la Tabla 18 con los doce vectores más cortos del gráfico describiendo su longitud y su R^2.

Ilustración 12: Biplot logístico externo de las variables-categorías clínicas

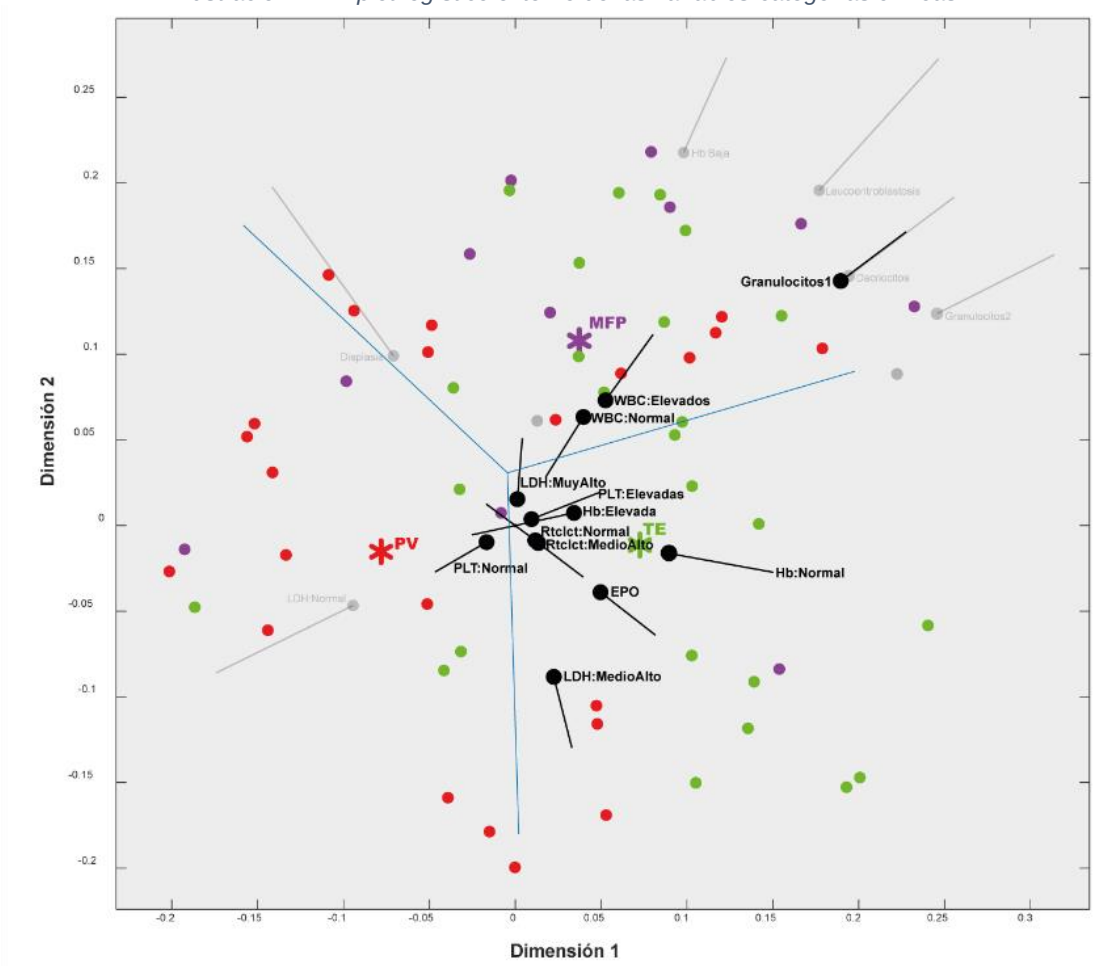


Tabla 18: Longitud de los vectores con R^2 > 0.6 en el Biplot logístico

Categoría	Longitud del vector	R2
Plaquetas: Normales	0,0347	0,854
LDH: Medio Alto	0,0427	0,834
Reticulocitos: Medio Alto	0,0329	0,829
Eritropoyetina sérica	0,0407	0,827
Leucocitos: Normales	0,0414	0,815
Reticulocitos: Normal	0,0356	0,805
LDH: Muy Alto	0,0358	0,798
Plaquetas: Elevadas	0,0426	0,781
Leucocitos: Elevados	0,0473	0,763
Granulocitos Mielograma	0,0478	0,7
Hemoglobina: Elevada	0,0609	0,639
Hemoglobina: Normal	0,0630	0,631

Del gráfico 12 y la tabla 18 se obtienen los siguientes hallazgos:

- Reticulocitos: específicamente quienes tenían “Medio alto” en el resultado de esta variable, fueron más pacientes del grupo TE.

- Plaquetas: cuando el nivel de plaquetas fue “normal” hubo más pacientes PV, y cuando estaban “elevadas” se proyectan más pacientes con TE.
- LDH: la mayoría de los pacientes con MFP tenían niveles muy altos de LDH.
- Eritropoyetina Sérica: Siendo Inadecuada frecuentemente en PV y MFP.
- Leucocitos: la mayoría de los pacientes con PV y TE tenían niveles “normales” de este marcador clínico.
- Hemoglobina: Elevada para PV y normales para TE.

Otras observaciones del Biplot Logístico son:

- Considerando los centroides de los grupos de PV y TE, se forma un gradiente horizontal, de tal manera que las variables que están en ese sentido discriminan mejor a estos grupos. Ejemplo: Plaquetas y Hemoglobina.
- Y los vectores más verticales que no están lejos del origen de coordenadas, podrían describir mejor a pacientes con MFP. Ejemplo: LDH: Muy alto.
- EPO y Granulocitos son las únicas variables binarias con un R^2 mayor a 0.6, todas las demás variables resaltadas cumplen con tres particularidades:
    - 1) corresponden a variables ordinales, logrando un buen resultado al menos para dos niveles de cada uno, considerando que originalmente el gráfico se diseñó para variables puramente binarias, como presencia o ausencia de una característica;
    - 2) son niveles “normales” o “altos” de sus marcadores clínicos, lo cual indica que las observaciones se pueden clasificar mejor si se encuentran con niveles normales o altos, sugiriendo que los pacientes con niveles bajos en dichas variables clínicas no se explorarían con mucha precisión en los gráficos o análisis multivariantes de este tipo;
    - 3) se dibujan en direcciones opuestas en el gráfico, ratificando que hay una mejor separación de los pacientes con niveles altos o normales de las características clínicas exploradas.

Según este análisis, debido a las magnitudes de R^2 y la orientación las variables que mejor ayudan a discriminar a los grupos fueron Hemoglobina, Plaquetas, LDH y Eritropoyetina sérica. Por otra parte, los niveles de leucocitos no separan a un grupo en particular, pero es frecuente que los pacientes PV y TE, tienen generalmente valores “normales” de este marcador.

4.8. Análisis de las características que más discriminan entre los grupos de pacientes con NMP Ph-

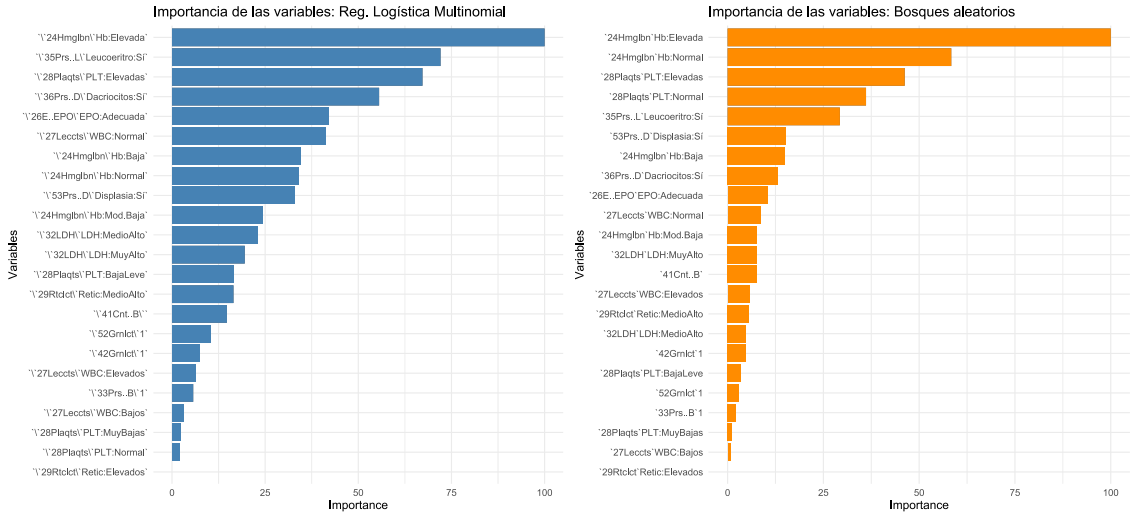
La idea de discriminar es encontrar lo que es característico del grupo y que no característico en los demás. Es decir, lo que los hace diferentes. Está bastante bien definidas las características exclusivas de los grupos de PV y de TE, pero respecto a discriminar al grupo MFP se puede decir que: 1) las características identificadas con la técnica de BL: el LDH muy alto y la EPO inadecuada, tienen alta frecuencia en el grupo, pero también estas categorías están presentes en PV y TE (Tabla 13 y 14), es decir, no son características exclusivas de MFP; 2) las características identificadas con el ADL: presencia de Leucoeritroblastosis en el grupo, sí es una característica común en el grupo MFP (72,7%) y que no resalta en los otros 2 grupos. No obstante, "reticulocitos elevados", aunque solo tenga 1 observación en toda la muestra de los 3 grupos (característica atípica), dicha observación pertenece al grupo de MFP, y fue la que mayor peso obtuvo en la primera función discriminante del ADL, por lo que podríamos notar que la técnica, aunque precisa, se sesga con una característica atípica; 3) Por último, el ACM identifica la categoría "Baja" o "Moderadamente Baja" de la hemoglobina, que en conjunto representan al 81,8% de los pacientes del grupo, pero la categoría "Baja" es más exclusiva para MFP, siendo buena característica para discriminar.

Las dos primeras técnicas usadas para selección de características (Tabla 19), antes que todo son métodos de clasificación, es decir, construyen un modelo que encuentre las mejores variables con la información para clasificar a cada grupo y explicar sus características. Por lo que sus resultados también pueden ser tomados en cuenta para discriminar a los tipos de NMP Ph-.

Tabla 19: Matrices de confusión de los modelos de clasificación que se usaron para selección de variables

		Regresión Logística Multinomial Balanced Accuracy: 89,19%				**Bosques Aleatorios** Balanced Accuracy: 87,1%			
		PV	**TE**	**MFP**	**NOS**	**PV**	**TE**	**MFP**	**NOS**
Categorías Reales	**PV**	49	2	0	0	50	1	0	0
	TE	10	36	0	0	10	36	0	0
	MFP	0	0	11	0	1	0	10	0
	NOS	0	0	0	3	1	1	0	1

Ilustración 13: Importancia de las variables en los modelos a) RLM y b) RF



El modelo de Regresión Logística Multinomial (RLM), obtuvo una mayor métrica de precisión (89,19%) de clasificación en comparación al de Bosques aleatorios (87,1%). Además, RLM fue mejor para predecir específicamente a la categoría de MFP. Por su parte, el modelo de Bosques aleatorios (RF) fue levemente mejor para predecir la categoría de PV, pero tuvo más errores en MFP y en las 3 observaciones NOS. Y otro punto a favor, es que las características que el modelo RLM consideró más importantes son mejores para discriminar los 3 grupos: Hemoglobina elevada para PV, plaquetas elevadas para TE y Presencia de Leucoeritroblastosis y de Dacriocitos para MFP, gráfico 13 a). Mientras que RF entre sus más importantes consideró dos variables características de PV, también dos características de TE y solo una de MFP, dando un poco menos de peso para explicar a MFP, cuando desfavorablemente, es el grupo más diverso entre los tres.

De tal manera, se puede concluir como las mejores variables-categorías para discriminar a los tres grupos de NMP Ph- a: Hemoglobina elevada, Plaquetas elevadas, Presencia de Leucoeritroblastosis, Hemoglobina baja y Presencia de Dacriocitos. Estos hallazgos proporcionan conocimientos para comprender mejor estas enfermedades en el contexto ecuatoriano.

CAPÍTULO 5

5. CONCLUSIONES Y RECOMENDACIONES

- Los análisis multivariantes aplicados lograron revelar las características más comunes, similitudes y diferencias entre los perfiles clínicos de los pacientes de los tres grupos de Neoplasias Mieloproliferativas Philadelphia Negativas (NMP Ph-) clásicas. Al considerar solo las características de la Tabla 16 se perciben los perfiles de pacientes con Policitemia Vera y Trombocitemia Esencial muy similares, excepto por las condiciones que los caracterizan, niveles elevados de hemoglobina (PV) y de plaquetas (TE). Mientras que la Mielofibrosis Primaria, se mostró más agresiva en los pacientes con algunas anomalías hematológicas. Este último grupo, siendo el más pequeño, fue el más diverso de los tres alejando su comportamiento del de PV y TE.
- La variable de hemoglobina es la que mejor discrimina a los 3 grupos, teniendo el nivel de “elevada” para pacientes con Policitemia Vera, “normal” para Trombocitemia Esencial y una tendencia de niveles bajos para pacientes con Mielofibrosis Primaria. La siguiente variable con poder para discriminar o separar a los grupos es el nivel de Plaquetas. Para cualificar a los pacientes con MFP, específicamente, se considera la presencia de Leucoeritroblastosis y de Dacriocitos, y la hemoglobina en niveles bajos.
- Con la técnica del análisis discriminante lineal fue más sencillo explorar las diferencias de los grupos de NMP Ph- en comparación con el método del Biplot Logístico (BL), sin embargo, el BL es más completo porque incluye elementos adicionales que permiten obtener simultáneamente más interpretaciones con el mismo análisis. De todas maneras, ambos métodos coincidieron en que los niveles de hemoglobina y de plaquetas forman parte de las que discriminan mejor.
- Cada técnica utilizada en esta investigación aportó con información para determinar las características que ayudan a discriminar al grupo de los MFP. Por otra parte, se observó que la técnica del Biplot Logístico puede ser utilizada desde el principio con fines exploratorios de la matriz de datos ya que simultáneamente se pueden obtener interpretaciones de correlaciones, similitudes, para caracterizar y para discriminar los grupos también.

- Dado el enfoque multivariante, el hecho de poseer tantos elementos en los gráficos puede dificultar la interpretación de individuos, grupos, o categorías cercanas, por tal motivo, parte del trabajo de mejorar las visualizaciones fue pintar los polígonos que conforman los individuos en los planos bidimensionales, sin embargo no tienen interpretación estadística alguna, tampoco puede considerarse que un nuevo individuo diagnosticado con uno de estos tipos de NMP Ph- necesariamente se proyectará dentro del mismo polígono, ya que la zona no representa una inferencia a la población. Dicho tratamiento a los gráficos solo tuvo fines ilustrativos para resaltar de mejor manera la zona en la que los individuos se proyectan en la dimensión reducida.

Recomendaciones

- Considerar analizar otras dimensiones adicionales. El hecho de que la primera y segunda componente no alcancen un mayor porcentaje de explicación de la varianza total indica que las categorías no tienen una estructura de asociación tan fuerte para que la gráfica del ACM la pueda representar en su primer plano.
- Analizar datos atípicos multivariantes. El hecho de poder identificar una observación como atípica de manera multivariante es una ganancia cuando la cantidad de variables a analizar es elevada. Sin embargo, por su presencia, los cálculos del ACM podrían haber sido distorsionados, complicando la estimación de las nuevas dimensiones y, por consiguiente, sus gráficas.
- Incrementar los esfuerzos en estudiar las enfermedades de baja prevalencia como las NMP Ph-, por medio de proyectos que planteen la toma eficiente de datos con las escalas apropiadas y sin valores omitidos, que permitan realizar análisis más completos y precisos.

Bibliografía

Benzecri, J. (1973). L'analyse des données. *Population*, 1190. Obtenido de https://www.persee.fr/doc/pop_0032-4663_1975_num_30_6_15911

Demey, J. R., Vicente-Villardón, J. L., Galindo-Villardón, M. P., & Zambrano, A. Y. (2008). Identifying molecular markers associated with classification of genotypes by External Logistic Biplots. *Bioinformatics*, 2832–2838. doi:doi:10.1093/bioinformatics/btn552

Driver, H. E., & Kroeber, A. L. (1932). Quantitative Expression of Cultural Relationships. *University of California Publications in American Archaeology and Ethnology*, 211-256. Obtenido de https://web.archive.org/web/20201206053117/https://dpg.lib.berkeley.edu/webdb/anthpubs/search?all=&volume=31&journal=1&item=5

Escobar Montes, K., Vicente Villardón, J. L., Alarcón Cano, D. F., & Siteneski, A. (2022). Clinical related factors to neuroendocrine tumors in Ecuadorian patients: a logistic biplot approach. *Invest Clin*, 63(1) 19-31. doi:https://doi.org/10.54817/IC.v63n1a02

Fisher, R. (1936). THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. *Annals of Eugenics*, 179-188. doi:https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Galton, S. F. (1886). Regression Towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 246-263. doi:https://doi.org/10.2307/2841583

GEMFIN. (2020). *Manual 2020 de recomendaciones en Neoplasias Mieloproliferativas Crónicas Filadelfia Negativas.* España: Grupo Español de Enfermedades Mieloproliferativas Crónicas Filadelfia Negativas.

GEMFIN. (s.f.). *GEMFIN: Documentos.* Obtenido de GEMFIN: https://www.gemfin.org/documentos-no-profesionales/

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 325-338. doi:https://doi.org/10.1093/biomet/53.3-4.325

Jöreskog, K. G. (1970). A General Method for Estimating a Linear Structural Equation System. *ETS Research Bulletin Series*. doi:https://doi.org/10.1002/j.2333-8504.1970.tb00783.x

Khoury, J. D., Solary, E., Abla, O., Akkari, Y., Alaggio, R., Apperley, J. F., & Bejar, R. (2022). The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Myeloid and Histiocytic/Dendritic Neoplasms. *Leukemia*, 36:1703–1719. doi:https://doi.org/10.1038/s41375-022-01613-1

Morales, P. (2012). Tipos de variables y sus implicaciones en el diseño de una investigación. En P. Morales, *Estadística aplicada a las Ciencias Sociales.* Madrid.

Orlandoni, G. (2010). Escalas de medición en Estadística. *TELOS. Revista de Estudios Interdisciplinarios en Ciencias Sociales*, 243-247.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The Philosophical Magazine*, 559-572. doi:https://doi.org/10.1080/14786440109462720

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2), 201-293. doi:https://doi.org/10.2307/1412107

Valladares, X., Benavente, R., Rojas, C., Peña, C., Valenzuela, R., Monardes, V., . . . Abarca, M. (2021). Características clínicas y epidemiológicas de las neoplasias mieloproliferativas Philadelphia negativas en el sistema público de salud de Chile. *Rev Med Chile*, 149: 1532-1538. doi:http://dx.doi.org/10.4067/S0034-98872021001101532

Vicente-Villardón, J. L., & Hernández-Sánchez, J. (2020). External Logistic Biplots for Mixed Types of Data. En *Advanced Studies in Classification and Data Science* (págs. 169-183). Springer.