UNIVERSIDAD DE TALCA

FACULTAD DE INGENIERÍA

ESCUELA DE INGENIERÍA CIVIL DE MINAS

APLICACIÓN DE TÉCNICAS DE APRENDIZAJE DE MÁQUINAS A LA MODELACIÓN DE VARIABLES GEOLÓGICAS DE INDICADORES

MEMORIA PARA OPTAR AL TÍTULO
DE INGENIERO CIVIL DE MINAS

JUAN IGNACIO MELLA VERGARA

PROFESOR GUÍA

Mg. FRANCISCO RIVAS SALDAÑA

MIEMBROS DE LA COMISIÓN

Mg. CARLOS MORAGA CRUZ

Sr. BRANCO ESPINOZA VARGAS

CURICÓ-CHILE

2022

CONSTANCIA

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

Curicó, 2023

RESUMEN

Dentro de la minería una de las etapas del proceso minero que mayor incertidumbre produce es dentro de la estimación de recursos, y para poder realizar estimaciones se hace uso de la geoestadística. Nuevas tecnologías están naciendo, con mayor capacidad de análisis de datos y procesamiento de ellos, y vienen a contribuir en la búsqueda de bajar la incertidumbre geológica, que se produce en la modelación de variables geológicas.

El objetivo de la presente investigación de memoria plantea utilizar métodos alternativos a los ya utilizados en la minería, como por ejemplo, el Kriging de indicadores para la estimación de variables nominales.

Por lo anterior, es que se plantean dos métodos del aprendizaje de máquinas, las redes neuronales artificiales y la regresión logística, y se utiliza una base de datos con distintos tipos de litologías, con el objetivo de realizar estimaciones a variables categóricas, y compararlas con el Kriging de indicadores.

De los métodos propuestos, las redes neuronales artificiales resultan ser las que mejor estiman variables nominales de acuerdo con métricas de desempeño, en segundo lugar se encuentra la regresión logística y muy de cerca el Kriging de indicadores.

Se concluye que a pesar del buen rendimiento de los métodos del aprendizaje de máquinas según los estadísticos de exactitud, precisión, sensibilidad y especificidad, no se puede asegurar que son significativamente mejores que el representante geoestadístico, es decir, el Kriging de indicadores, y por lo tanto queda en manos del experto que herramientas prefiere utilizar para modelar las variables geológicas.

ABSTRACT

Within mining, one of the stages of the mining process that produces the greatest uncertainty is within the estimation of resources, and in order to make estimates, geostatistics are used. New technologies are being born, with greater capacity for data analysis and data processing, and they come to contribute in the search to lower geological uncertainty.

The objective of the present memory research proposes to use alternative methods to those already used in mining, such as, for example, the Kriging of indicators for the estimation of nominal variables.

Therefore, it is that two methods of machine learning are proposed, artificial neural networks and logistic regression, and a database with different types of lithologies is used, in order to make estimates to categorical variables, and compare them with the Kriging of indicators.

Of the proposed methods, artificial neural networks turn out to be the ones that best estimate nominal variables, followed by logistic regression and very closely the Kriging of indicators.

It is concluded that despite the good performance of the machine learning methods according to the statistics of accuracy, precision, sensitivity and specificity, it cannot be assured that they are significantly better than the geostatistical representative, that is, the Kriging of indicators, and therefore, it is up to the expert which tools he prefers to use to model geological variables.

AGRADECIMIENTOS

Agradezco a mi profesor guía por la paciencia y los consejos para lograr desarrollar esta memoria.

A mis principales 3 amigos durante la carrera, Joaquín, Nicolás y Branco y a tantos otros buenos compañeros.

A mis hermanos por apoyarme y darme consejos de vida.

Y finalmente a mis padres que sin ellos nada sería posible, a mi Padre le agradezco darme educación e inculcarme siempre que debía estudiar para tener un mejor futuro, y creer en mis capacidades, y a mi madre le agradezco por enseñarme a ser perseverante, y por siempre apoyarme en los malos momentos y comprenderme como hijo.

Dedicado a

mi Mamá, Papá y Hermanos.

TABLA DE CONTENIDOS

CAPÍTULO 1: INTRODUCCIÓN 1
1.1 Antecedentes y motivación 1
1.2 Descripción del problema 2
1.3 Solución propuesta 2
1.4 Objetivo general 3
1.5 Objetivos específicos 3
1.6 Alcances 3
CAPÍTULO 2: MARCO TEÓRICO 4
2.1 Estudio exploratorio de datos 4
2.1.1 Medidas de posición 5
2.1.2 Medidas de dispersión 5
2.1.3 Histograma y diagrama de caja 6
2.2 Definición de Geoestadística y conceptos básicos de variables regionalizadas 7
2.2.1 Variable regionalizada 8
2.2.2 Función aleatoria 9
2.2.3 Momentos de una función aleatoria 9
2.2.4 Hipótesis de estacionaridad 11
2.3 Herramientas geoestadísticas 12
2.3.1 Variograma experimental 12
2.3.1.1 Propiedades del variograma experimental 13
2.3.2 Variograma modelado 13
2.3.2.1 Propiedades variograma teórico 14
2.3.2.2 Comportamiento para distancias muy grandes 14
2.3.3 Modelos elementales de variograma 15
2.3.4 Anisotropías 19
2.4 Kriging 20
2.4.1 Construcción del Kriging 20
2.4.2 Kriging con media conocida (Kriging Simple) 21
2.4.2.1 Hipótesis 21
2.4.2.2 Determinación del estimador 22
2.4.2.3 Varianza de Kriging 24

2.4.3 Kriging con media desconocida (Kriging Ordinario) *25*
2.4.3.1 Hipótesis 25
2.4.3.2 Determinación del estimador 25
2.4.3.3 Varianza de Kriging 27
2.4.4 Kriging multivariable (Co-Kriging) *28*
2.4.5 Kriging no lineal *29*
2.4.6 Función indicadora *29*
2.4.7 Kriging de Indicadores *30*
2.4.8 Validación cruzada *31*
2.5 Definición de Aprendizaje de Máquinas 32
2.5.1 Tipos de tareas del Aprendizaje de Máquinas *32*
2.5.2 Aprendizaje supervisado, no supervisado y semisupervisado *33*
2.6 Redes Neuronales 33
2.6.1 Arquitectura de redes neuronales artificiales *36*
2.6.2 Algoritmo Backpropagation *38*
2.7 Regresión Logística 39
2.8 Overfitting y Underfitting de los modelos 44
2.9 Medidas de desempeño de modelos de estimación 45
CAPÍTULO 3: METODOLOGÍA 48
CAPÍTULO 4: DESARROLLO 49
4.1 Descripción de la base de datos 49
4.2 Análisis exploratorio de datos 49
4.3 Compositación de Sondajes 53
4.4 Nueva base de datos Post Compositación de sondajes 56
4.5 Elección de los métodos de estimación 59
4.6 Despliegue de los sondajes en SGeMS 59
4.7 Variogramas 63
4.8 Métodos de Machine Learning 68
4.8.1. Redes neuronales artificiales *68*
4.8.2. Regresión logística *70*
CAPÍTULO 5: RESULTADOS Y DISCUSIÓN 71
5.1 Estimación Kriging de Indicadores 71
5.2. Estimación Redes Neuronales Artificiales 77
5.3. Estimación Regresión Logística 85

5.4. Comparación de métodos de estimación 93
Conclusión 104
Referencias bibliográficas 107
Apéndice A: Análisis exploratorio de datos 109
Apéndice B: Variogramas experimentales y modelados 112
Apéndice C: Resultados de estimaciones 121
Apéndice D: Matriz de confusión y resultados estadísticos litologías 132

ÍNDICE DE ILUSTRACIONES

Ilustración 1: Histograma de las concentraciones de cobalto (Emery,2013) 6

Ilustración 2: Diagrama de caja para las concentraciones de cobalto (Emery,2013) 6

Ilustración 3: Esquema sintético de los conceptos e hipótesis que sustentan el modelo geoestadístico (Emery,2013). 11

Ilustración 4: Ejemplo de Variograma con meseta y alcance (Emery,2013) 15

Ilustración 5: Ejemplo de variograma con efecto pepita (Emery,2013) 16

Ilustración 6: Ejemplo de variograma esférico (Emery,2013.) 16

Ilustración 7: Ejemplo de modelo exponencial (Emery,2013) 17

Ilustración 8: Ejemplo de modelo gaussiano (Emery,2013) 18

Ilustración 9: Ejemplo de modelo anidado (Emery,2013) 18

Ilustración 10: Ejemplo de anisotropía geométrica (izquierda) y anisotropía zonal (derecha) (Emery,2013) 19

Ilustración 11: Estructura de una neurona artificial y su analogía con una neurona biológica (Caparrini,2018). 34

Ilustración 12: Estructura de una red neuronal artificial (Matich,2001) 35

Ilustración 13: Ejemplo de una red neuronal con conexión feedforward (Flórez & Fernández,2008) 37

Ilustración 14: Ejemplos de una red neuronal con conexión feedforward (Flórez & Fernández,2008) 37

Ilustración 15: Ejemplo de regresión logística (Hosmer & Lemeshow,2000) 42

Ilustración 16: Ejemplo gráfico de overfitting y underfitting (Rodríguez-Sahagún 2018). 45

Ilustración 17: Histograma Ley de Cobre (Elaboración propia) 51

Ilustración 18: Boxplot Ley de Cobre (Elaboración propia) 51

Ilustración 19: Sondajes leyes de Cobre (Vulcan) 54

Ilustración 20: Sondajes con leyenda de las leyes de Cobre (Vulcan) 54

Ilustración 21: Compositación de sondajes con leyenda de las leyes de Cobre (Vulcan) 55

Ilustración 22: cantidad porcentual de datos para las litologías (elaboración propia) 57

Ilustración 23: Sondajes desplegados junto con bloque de estimación para litología AND (elaborado en SGeMS). 61

Ilustración 24: Sondajes desplegados junto con bloque de estimación para litología S2 (elaborado en SGeMS). 61

Ilustración 25: Sondajes desplegados junto con bloque de estimación para litología HBX (elaborado en SGeMS). 62

Ilustración 26: Sondajes desplegados junto con bloque de estimación para litología Mixto (elaborado en SGeMS). 62

Ilustración 27: Variograma experimental 0° para litología AND (elaborado en SGeMS). 64

Ilustración 28: Variograma experimental 45° para litología AND (elaborado en SGeMS). 65

Ilustración 29: Variograma experimental 90° para litología AND (elaborado en SGeMS). 65

Ilustración 30: Variograma experimental 135° para litología AND (elaborado en SGeMS). 66

Ilustración 31: Variograma experimental omnidireccional para litología AND (elaborado en SGeMS). 66

Ilustración 32: Variograma modelado omnidireccional para litología AND (elaborado en SGeMS). 67

Ilustración 33: Flujo modelo de redes neuronales artificiales (elaborado en Orange Canvas). 69

Ilustración 34: Flujo modelo de regresión logística (elaborado en Orange Canvas). 70

Ilustración 35: Vista de frente Kriging de indicadores para litología AND (elaborado en SGeMS). 71

Ilustración 36: Vista en planta Kriging de indicadores para litología AND (elaborado en SGeMS). 72

Ilustración 37: Varianza Kriging de indicadores para litología AND (elaborado en SGeMS). 73

Ilustración 38: Vista de frente Kriging de indicadores para litología S2 (elaborado en SGeMS). 73

Ilustración 39: Vista en planta Kriging de indicadores para litología S2 (elaborado en SGeMS). 74

Ilustración 40: Varianza Kriging de indicadores para litología S2 (elaborado en SGeMS). 74

Ilustración 41: Plano XY redes neuronales datos de entrenamiento para litología AND (elaborado en Orange Canvas). 77

Ilustración 42: Plano XZ redes neuronales datos de entrenamiento para litología AND (elaborado en Orange Canvas). 78

Ilustración 43: Plano XY redes neuronales datos de prueba para litología AND (elaborado en Orange Canvas). 79

Ilustración 44: Plano XZ redes neuronales datos de prueba para litología AND (elaborado en Orange Canvas). 80

Ilustración 45: Plano XY redes neuronales datos de entrenamiento para litología S2 (elaborado en Orange Canvas 81

Ilustración 46: Plano XZ redes neuronales datos de entrenamiento para litología S2 (elaborado en Orange Canvas). 82

Ilustración 47: Plano XY redes neuronales datos de prueba para litología S2 (elaborado en Orange Canvas). 83

Ilustración 48: Plano XZ redes neuronales datos de prueba para litología S2 (elaborado en Orange Canvas). 84

Ilustración 49: Plano XY regresión logística datos de entrenamiento para litología AND (elaborado en Orange Canvas). 85

Ilustración 50: Plano XZ regresión logística datos de entrenamiento para litología AND (elaborado en Orange Canvas). **86**

Ilustración 51: Plano XY regresión logística datos de prueba para litología AND (elaborado en Orange Canvas). **87**

Ilustración 52: Plano XZ regresión logística datos de prueba para litología AND (elaborado en Orange Canvas) **88**

Ilustración 53: Plano XY regresión logística datos de entrenamiento para litología S2 (elaborado en Orange Canvas). **89**

Ilustración 54: Plano XZ regresión logística datos de entrenamiento para litología S2 (elaborado en Orange Canvas). **89**

Ilustración 55: Plano XY regresión logística datos de prueba para litología S2 (elaborado en Orange Canvas). **91**

Ilustración 56: Plano XZ regresión logística datos de prueba para litología S2 (elaborado en Orange Canvas). **91**

Ilustración 57: Gráfico de barras de la litología de tipo AND, para los 3 tipos de métodos (elaboración propia). **94**

Ilustración 58: Gráfico de barras de la litología de tipo S2, para los 3 tipos de métodos (elaboración propia). **95**

Ilustración 59: Gráfico de barras de la litología de tipo HBX, para los 3 tipos de métodos (elaboración propia). **96**

Ilustración 60: Gráfico de barras de la litología de tipo Mixto, para los 3 tipos de métodos (elaboración propia). **97**

Ilustración 61: Gráfico de barras de Métricas de Desempeño para la litología de tipo AND, para los 3 tipos de métodos (elaboración propia). **99**

Ilustración 62: Gráfico de barras de Métricas de Desempeño para la litología de tipo S2, para los 3 tipos de métodos (elaboración propia) **99**

Ilustración 63: Gráfico de barras de Métricas de Desempeño para la litología de tipo HBX, para los 3 tipos de métodos (elaboración propia). **100**

Ilustración 64: Gráfico de barras de Métricas de Desempeño para la litología de tipo Mixto, para los 3 tipos de métodos (elaboración propia). **101**

Ilustración 65: Gráfico de barras de Métricas de Desempeño para el promedio de todos los tipos de litología, para los 3 tipos de métodos (elaboración propia). **102**

Ilustración A.1: Histograma Ley de Plata (Elaboración propia) **109**

Ilustración A.2: Boxplot Ley de Plata (Elaboración propia) **110**

Ilustración A.3: Histograma Ley de Oro (Elaboración propia) **111**

Ilustración A.4: Boxplot Ley de Oro (Elaboración propia) **111**

Ilustración B.1: Variograma experimental 0° para litología S2 (elaborado en SGeMS). **112**

Ilustración B.2: Variograma experimental 45° para litología S2 (elaborado en SGeMS). **112**

Ilustración B.3: Variograma experimental 90° para litología S2 (elaborado en SGeMS)..... 113
Ilustración B.4: Variograma experimental 135° para litología S2 (elaborado en SGeMS)..... 113
Ilustración B.5: Variograma experimental omnidireccional para litología S2 (elaborado en SGeMS). 114
Ilustración B.6: Variograma modelado para litología S2 (elaborado en SGeMS). 114
Ilustración B.7: Variograma experimental 0° para litología HBX (elaborado en SGeMS). 115
Ilustración B.8: Variograma experimental 45° para litología HBX (elaborado en SGeMS). 115
Ilustración B.9: Variograma experimental 90° para litología HBX (elaborado en SGeMS). 116
Ilustración B.10: Variograma experimental 135° para litología HBX (elaborado en SGeMS). .. 116
Ilustración B.11: Variograma experimental omnidireccional para litología HBX (elaborado en SGeMS). 117
Ilustración B.12: Variograma modelado para litología HBX (elaborado en SGeMS)..... 117
Ilustración B.13: Variograma experimental 0° para litología Mixto (elaborado en SGeMS)..... 118
Ilustración B.14: Variograma experimental 45° para litología Mixto (elaborado en SGeMS).... 118
Ilustración B.15: Variograma experimental 90° para litología Mixto (elaborado en SGeMS).... 119
Ilustración B.16: Variograma experimental 135° para litología Mixto (elaborado en SGeMS). 119
Ilustración B.17: Variograma experimental omnidireccional para litología Mixto (elaborado en SGeMS). 120
Ilustración B.18: Variograma modelado para litología Mixto (elaborado en SGeMS). 120
Ilustración C.1: Vista de frente Kriging de indicadores para litología HBX (elaborado en SGeMS). 121
Ilustración C.2: Vista en planta Kriging de indicadores para litología HBX (elaborado en SGeMS). 121
Ilustración C.3: Varianza Kriging de indicadores para litología HBX (elaborado en SGeMS). 122
Ilustración C.4: Vista de frente Kriging de indicadores para litología Mixto (elaborado en SGeMS). 122
Ilustración C.5: Vista en planta Kriging de indicadores para litología Mixto (elaborado en SGeMS). 123
Ilustración C.6: Varianza Kriging de indicadores para litología Mixto (elaborado en SGeMS). 123
Ilustración C.7: Plano XY redes neuronales datos de entrenamiento para litología HBX (elaborado en Orange Canvas)..... 124
Ilustración C.8: Plano XZ redes neuronales datos de entrenamiento para litología HBX (elaborado en Orange Canvas)..... 124
Ilustración C.9: Plano XY redes neuronales datos de prueba para litología HBX (elaborado en Orange Canvas). 125
Ilustración C.10: Plano XZ redes neuronales datos de prueba para litología HBX (elaborado en Orange Canvas). 125

Ilustración C.11: Plano XY redes neuronales datos de entrenamiento para litología Mixto (elaborado en Orange Canvas)..... 126

Ilustración C.12: Plano XZ redes neuronales datos de entrenamiento para litología Mixto (elaborado en Orange Canvas)..... 126

Ilustración C.13: Plano XY redes neuronales datos de prueba para litología Mixto (elaborado en Orange Canvas). 127

Ilustración C.14: Plano XZ redes neuronales datos de prueba para litología Mixto (elaborado en Orange Canvas). 127

Ilustración C.15: Plano XY regresión logística datos de entrenamiento para litología HBX (elaborado en Orange Canvas)..... 128

Ilustración C.16: Plano XZ regresión logística datos de entrenamiento para litología HBX (elaborado en Orange Canvas)..... 128

Ilustración C.17: Plano XY regresión logística datos de prueba para litología HBX (elaborado en Orange Canvas). 129

Ilustración C.18: Plano XZ regresión logística datos de prueba para litología HBX (elaborado en Orange Canvas). 129

Ilustración C.19: Plano XY regresión logística datos de entrenamiento para litología Mixto (elaborado en Orange Canvas)..... 130

Ilustración C.20: Plano XZ regresión logística datos de entrenamiento para litología Mixto (elaborado en Orange Canvas)..... 130

Ilustración C.21: Plano XY regresión logística datos de prueba para litología Mixto (elaborado en Orange Canvas). 131

Ilustración C.22: Plano XZ regresión logística datos de prueba para litología Mixto (elaborado en Orange Canvas). 131

ÍNDICE DE TABLAS

Tabla 1: Ejemplo de Matriz de confusión para caso de 2x2 (elaboración Propia)..... 45
Tabla 2: Estadísticas básicas ley de Cobre (Elaboración propia)..... 50
Tabla 3: Estadísticas básicas ley de Cobre sin datos atípicos (Elaboración propia)..... 52
Tabla 4: Estadísticas básicas ley de Cobre compositos (Elaboración propia)..... 56
Tabla 5: cantidad de datos para litologías agrupadas (elaboración propia). 56
Tabla 6: cantidad de datos para litologías agrupadas (elaboración propia). 57
Tabla 7: Ejemplo de método One Hot Encoding a la base de datos (Elaboración propia) 58
Tabla 8: Mínimos, máximos y rango según coordenada (elaboración propia). 59
Tabla 9: Dimensiones y cantidad de bloques según coordenada (elaboración propia). 60
Tabla 10: Parámetros del paso de Variogramas experimentales (elaboración propia)..... 63
Tabla 11: Parámetros de las direcciones de los Variogramas experimentales (elaboración propia). 64
Tabla 12: Parámetros redes neuronales artificiales (elaboración propia). 68
Tabla 13: Matriz de confusión Kriging de indicadores para litología AND (elaboración propia).75
Tabla 14: Resultados estadísticos Kriging de indicadores para litología AND (elaboración propia). 75
Tabla 15: Matriz de confusión Kriging de indicadores para litología S2 (elaboración propia). ... 76
Tabla 16: Resultados estadísticos Kriging de indicadores para litología S2 (elaboración propia). 76
Tabla 17: Matriz de confusión datos de entrenamiento para litología AND (elaboración propia). 78
Tabla 18: Resultados estadísticos datos de entrenamiento para litología AND (elaboración propia). 79
Tabla 19: Matriz de confusión datos de prueba para litología AND (elaboración propia)..... 80
Tabla 20: Resultados estadísticos datos de prueba para litología AND (elaboración propia)..... 81
Tabla 21: Matriz de confusión datos de entrenamiento para litología S2 (elaboración propia).... 82
Tabla 22: Resultados estadísticos datos de entrenamiento para litología S2 (elaboración propia). 83
Tabla 23: Matriz de confusión datos de prueba para litología S2 (elaboración propia). 84
Tabla 24: Resultados estadísticos datos de prueba para litología S2 (elaboración propia). 85
Tabla 25: Matriz de confusión datos de entrenamiento para litología AND (elaboración propia). 86
Tabla 26: Resultados estadísticos datos de entrenamiento para litología AND (elaboración

propia). 86
Tabla 27: Matriz de confusión datos de prueba para litología AND (elaboración propia). 88
Tabla 28: Resultados estadísticos datos de prueba para litología AND (elaboración propia). 88
Tabla 29: Matriz de confusión datos de entrenamiento para litología S2 (elaboración propia).... 90
Tabla 30: Resultados estadísticos datos de entrenamiento para litología S2 (elaboración propia). 90
Tabla 31: Matriz de confusión datos de prueba para litología S2 (elaboración propia). 92
Tabla 32: Resultados estadísticos datos de prueba para litología S2 (elaboración propia). 92
Tabla 33: Matriz de confusión comparación métodos de estimación para litología AND (elaboración propia). 93
Tabla 34: Matriz de confusión comparación métodos de estimación para litología S2 (elaboración propia). 95
Tabla 35: Matriz de confusión comparación métodos de estimación para litología HBX (elaboración propia). 96
Tabla 36: Matriz de confusión comparación métodos de estimación para litología Mixto (elaboración propia). 97
Tabla 37: Resultados estadísticos comparación métodos de estimación para todas las litologías (elaboración propia). 98
Tabla 38: Resultados estadísticos comparación métodos de estimación para todas las litologías (elaboración propia). 102
Tabla A.1: Estadísticas Básicas Ley de Plata (Elaboración propia) 109
Tabla A.2: Estadísticas Básicas Ley de Oro (Elaboración propia) 110
Tabla D.1: Matriz de confusión Kriging de indicadores para litología HBX (elaboración propia). 132
Tabla D.2: Resultados estadísticos Kriging de indicadores para litología HBX (elaboración propia). 132
Tabla D.3: Matriz de confusión Kriging de indicadores para litología Mixto (elaboración propia). 132
Tabla D.4: Resultados estadísticos Kriging de indicadores para litología Mixto (elaboración propia). 133
Tabla D.5: Matriz de confusión datos de entrenamiento para litología HBX (elaboración propia). 133
Tabla D.6: Resultados estadísticos datos de entrenamiento para litología HBX (elaboración propia). 133
Tabla D.7: Matriz de confusión datos de prueba para litología HBX (elaboración propia). 133
Tabla D.8: Resultados estadísticos datos de prueba para litología HBX (elaboración propia). .. 134
Tabla D.9: Matriz de confusión datos de entrenamiento para litología Mixto (elaboración

propia). 134

Tabla D.10: Resultados estadísticos datos de entrenamiento para litología Mixto (elaboración propia). 134

Tabla D.11: Matriz de confusión datos de prueba para litología Mixto (elaboración propia). ... 134

Tabla D.12: Resultados estadísticos datos de prueba para litología Mixto (elaboración propia). 134

Tabla D.13: Matriz de confusión datos de entrenamiento para litología HBX (elaboración propia). 135

Tabla D.14: Resultados estadísticos datos de entrenamiento para litología HBX (elaboración propia). 135

Tabla D.15: Matriz de confusión datos de prueba para litología HBX (elaboración propia). 135

Tabla D.16: Resultados estadísticos datos de prueba para litología HBX (elaboración propia). 135

Tabla D.17: Matriz de confusión datos de entrenamiento para litología Mixto (elaboración propia). 136

Tabla D.18: Resultados estadísticos datos de entrenamiento para litología Mixto (elaboración propia). 136

Tabla D.19: Matriz de confusión datos de prueba para litología Mixto (elaboración propia). ... 136

Tabla D.20: Resultados estadísticos datos de prueba para litología Mixto (elaboración propia). 136

ÍNDICE DE ECUACIONES

Ecuación 1: valor esperado de una función aleatoria (Emery,2013) 9
Ecuación 2: varianza de una función aleatoria (Emery,2013) 9
Ecuación 3: covarianza de una función aleatoria (Emery,2013) 10
Ecuación 4: Correlograma de una función aleatoria (Emery,2013) 10
Ecuación 5: Semi-variograma de una función aleatoria (Emery,2013) 10
Ecuación 6: fórmula variograma experimental (Emery,2013) 12
Ecuación 7: propiedad 1 variograma experimental (Emery,2013) 13
Ecuación 8: propiedad 2 variograma experimental (Emery,2013) 13
Ecuación 9: propiedad 1 variograma teórico (Emery,2013) 14
Ecuación 10: propiedad 2 variograma teórico (Emery,2013) 14
Ecuación 11: propiedad 3 variograma teórico (Emery,2013) 14
Ecuación 12: fórmula modelo efecto pepita (Emery,2013) 15
Ecuación 13: fórmula modelo esférico (Emery,2013) 16
Ecuación 14: fórmula modelo exponencial (Emery,2013). 17
Ecuación 15: fórmula modelo gaussiano (Emery,2013) 17
Ecuación 16: ejemplo de modelos anidados (Emery,2013) 18
Ecuación 17: restricción de linealidad para la construcción del Kriging (Emery,2013) 20
Ecuación 18: restricción de Insesgo para la construcción del Kriging (Emery,2013) 21
Ecuación 19: restricción de optimalidad para la construcción del Kriging (Emery,2013) 21
Ecuación 20: hipótesis para Kriging con media conocida (Emery,2013) 21
Ecuación 21: linealidad del Kriging simple (Emery,2013) 22
Ecuación 22: Insesgo del Kriging simple (Emery,2013) 22
Ecuación 23: propiedad de valor esperado del Kriging simple (Emery,2013) 22
Ecuación 24: optimalidad del Kriging simple (Emery,2013) 22
Ecuación 25: Inicio de reglas de cálculo para el Kriging simple (Emery,2013) 22
Ecuación 26: Desarrollo de reglas de cálculo para Kriging simple (Emery,2013) 23
Ecuación 27: continuación 1 desarrollo de reglas de cálculo Kriging simple (Emery,2013) 23
Ecuación 28: continuación 2 desarrollo de reglas de cálculo Kriging simple (Emery,2013) 23
Ecuación 29: sistema de ecuaciones de reglas de cálculo Kriging simple (Emery,2013) 23
Ecuación 30: fórmula final para el estimador Kriging simple (Emery,2013). 24

Ecuación 31: varianza para Kriging simple (Emery,2013) 24
Ecuación 32: comparación varianza de Kriging versus varianza a priori para Kriging simple (Emery,2013)..... 24
Ecuación 33: hipótesis para Kriging con media desconocida (Emery,2013) 25
Ecuación 34: linealidad del Kriging ordinario (Emery,2013) 25
Ecuación 35: Insesgo del Kriging ordinario (Emery,2013) 25
Ecuación 36: valor esperado nulo del Kriging ordinario (Emery,2013) 25
Ecuación 37: optimalidad del Kriging ordinario (Emery,2013) 26
Ecuación 38: Inicio de reglas de cálculo para el Kriging ordinario (Emery,2013) 26
Ecuación 39: desarrollo de reglas de cálculo para el Kriging ordinario (Emery,2013)..... 26
Ecuación 40: sistema de ecuaciones de reglas de cálculo Kriging ordinario (Emery,2013) 27
Ecuación 41: desarrollo de sistema de ecuaciones de reglas de cálculo Kriging ordinario (Emery,2013)..... 27
Ecuación 42: sistema de ecuaciones final para Kriging ordinario (Emery,2013)..... 27
Ecuación 43: varianza para Kriging ordinario (Emery,2013) 27
Ecuación 44: variograma cruzado entre dos variables (Emery,2013)..... 28
Ecuación 45: inferencia para variograma cruzado entre dos variables (Emery,2013)..... 28
Ecuación 46: función indicadora para valores continuos (Journel,1983) 29
Ecuación 47: función indicadora para valores categóricos (Journel,1983). 30
Ecuación 48: propiedades para la función indicadora (Journel,1983). 30
Ecuación 49: fórmula para Kriging de indicadores (Journel,1983) 30
Ecuación 50: Ejemplo de variable dicotómica regresión logística (Hosmer & Lemeshow, 2000). 39
Ecuación 51: Probabilidad para regresión logística (Hosmer & Lemeshow, 2000)..... 39
Ecuación 52: Máxima verosimilitud para regresión logística (Hosmer & Lemeshow, 2000) 40
Ecuación 53: otra forma de expresar máxima verosimilitud regresión logística (Hosmer & Lemeshow, 2000) 40
Ecuación 54: Logaritmo de máxima verosimilitud regresión logística (Hosmer & Lemeshow, 2000)..... 40
Ecuación 55: ecuaciones de verosimilitud regresión logística (Hosmer & Lemeshow, 2000) 41
Ecuación 56: punto de corte para las probabilidades de regresión logística (Hosmer & Lemeshow, 2000) 41
Ecuación 57: Métrica de desempeño exactitud para matriz de confusión (elaboración propia). 46
Ecuación 58: Métrica de desempeño precisión para matriz de confusión (elaboración propia). 47
Ecuación 59: Métrica de desempeño sensibilidad para matriz de confusión (elaboración propia). 47

Ecuación 60: Métrica de desempeño especificidad para matriz de confusión (elaboración propia) 47

CAPÍTULO 1: INTRODUCCIÓN

1.1 Antecedentes y motivación

La minería es una actividad económica que se caracteriza por ser interdisciplinaria, en la cual se debe interactuar con la geología y ser un puente hacia la metalurgia, es por esto que es relevante que se lleve a cabo de buena forma las principales etapas del proceso minero-metalúrgico ,que consiste en la estimación de recursos, planificación minera, operación minera, procesamiento de minerales y terminando en el beneficio económico, con el objetivo de maximizar el valor actual neto del proyecto minero. Debido a lo anterior, una de las etapas que mayor incertidumbre representa para la minería es la estimación de recursos, la cual consiste en cuantificar los recursos disponibles del yacimiento minero, en donde es preciso utilizar herramientas de estimación, es así como aparece la geoestadística para dar respuesta a esta problemática. La Geoestadística es una rama de la estadística que mediante la utilización de diferentes técnicas, se centra en la determinación de variables numéricas que se distribuyen a través del espacio, es decir, a localizaciones específicas mediante coordenadas. Por lo anterior, es que esta memoria nace con la motivación de aportar al estudio de la incertidumbre que se provoca en las estimaciones de los recursos minerales al modelar variables geológicas. Normalmente los estudios de variables geológicas se hacen a variables continuas, tal como las leyes minerales, y se deja de lado el estudio a variables de tipo nominal o categóricas, como por ejemplo la litología, tipo de roca, textura, mineralización, alteración, dichas variables también tienen una importancia fundamental en procesos posteriores de la planificación, desde etapas geotécnicas hasta etapas metalúrgicas. Por lo tanto, la memoria tiene como finalidad ser un aporte al conocimiento de estimaciones para variables nominales, ya que al lograr tener un mayor conocimiento previo a la estimación de variables continuas, se podría lograr un mayor control geológico del proyecto minero y ser una ayuda a la toma de decisiones.

1.2 Descripción del problema

En la minería, los valores son provenientes de los sondajes, en donde se recolectan en una base de datos, que luego son estudiados a través de las técnicas de la geoestadística. Usualmente dichos datos son para variables continuas, pero que sucede cuando se está en presencia de variables nominales, es así como es necesario acudir a una de las técnicas de la geoestadística que es la estimación por Kriging de Indicadores, por lo que en esta memoria se pretende buscar una alternativa a dicho método, la cual consiste en el uso del aprendizaje de máquinas o también llamado Machine Learning. El aprendizaje de máquinas o machine learning es una disciplina que pertenece a la inteligencia artificial que se centra en la capacidad de otorgarle algoritmos a una máquina con la finalidad de identificar patrones y realizar predicciones de nuevos valores, sin ser necesario tener que nuevamente ser expresamente programada, obteniendo así un aprendizaje automático.

1.3 Solución propuesta

Ahora bien, la solución que se propone en esta memoria es aplicar técnicas del aprendizaje de máquinas a la estimación de variables nominales, específicamente Redes Neuronales y Regresión Logística utilizando el software de uso libre Orange Canvas.

Para el caso geoestadístico las estimaciones se harán mediante el Kriging de Indicadores, en la cual se utilizará el software de modelamiento geoestadístico SGeMS y se complementará con gvSIG.

Luego se realizará una comparación estadística entre las estimaciones realizadas para cada método propuesto, mediante el uso de matriz de confusión y sus métricas de desempeño.

Finalmente se discute si el uso de técnicas de Machine Learning sirven como una opción a las estimaciones de variables nominales por parte del Kriging de indicadores.

1.4 Objetivo general

Analizar la capacidad de las técnicas de aprendizaje de máquinas para la modelación de variables geológicas de indicadores.

1.5 Objetivos específicos

- Analizar la base de datos mediante el uso de la estadística clásica, con la finalidad de entender la información entregada.
- Planificar que tipo de técnicas se usarán del aprendizaje de máquinas y de la geoestadística.
- Aplicar las respectivas técnicas escogidas a la base de datos de cada método propuesto.
- Evaluar los resultados obtenidos para cada método propuesto.
- Comparar estadísticamente los resultados obtenidos para cada método propuesto.

1.6 Alcances

- No se proponen nuevos códigos de programación, debido al uso de programas especializados en Geoestadística y Machine Learning.
- No se realizará una planificación para un proyecto minero con los datos estimados para los métodos propuestos.
- No se considera la categorización de recursos y reservas
- No se aplicará un análisis económico a los métodos de estimación.
- No se llevará a cabo un estudio de caso para la correlación entre variables litológicas.
- El estudio de los datos se hará utilizando valores reales de la minería, y no para valores simulados mediante programación.
- Para el estudio de métodos de machine learning se consideran solo métodos de estimación con aprendizaje supervisado, para el caso de no supervisado, no serán aplicados en esta memoria.

CAPÍTULO 2: MARCO TEÓRICO

En el presente capítulo se presentan los antecedentes teóricos que son utilizados para el desarrollo de la investigación, centrándose en explicar conceptos de la estadística clásica, terminología básica y métodos de la Geoestadística, y finalmente que se entiende por Machine Learning y los métodos de clasificación que se utilizarán.

2.1 Estudio exploratorio de datos

Para comenzar un proyecto en el que se trabaja con datos, se hace necesario analizar el comportamiento estadístico de dichos datos, para lo cual se utilizan herramientas de la estadística clásica como lo son medidas de posición, medidas de dispersión, histogramas o diagramas de caja.

Para Emery (2013) la geoestadística tiene como objetivo estudiar una o más variables regionalizadas conocidas a través de los datos provenientes de los sondajes. Primeramente a la aplicación de los métodos geoestadísticos es conveniente realizar un análisis exploratorio de los datos que se tienen, con la finalidad de cumplir 3 objetivos:

- Analizar a través de herramientas simples de la estadística la cantidad, la calidad y la ubicación de los datos que se tienen.
- Definir las zonas del campo de estudio las cuales pueden ser en subzonas, debido a que puede ser relevante observar cambios radicales en el comportamiento espacial de los datos que se puede ver afectado por anomalías o fenómenos geológicos.
- Anticipar complejidades o problemas que se pueden presentar en la fase de estimación local, como por ejemplo, la presencia de valores atípicos.

Para Emery (2013) en el estudio de datos es conveniente calcular algunas estadísticas básicas de la distribución de los valores, entre las más elementales se encuentran las siguientes:

2.1.1 Medidas de posición

- ***Media*:** Es el promedio aritmético de los valores.
- ***Cuantiles o percentiles*:** Valores que dividen la población de estudio en partes de igual cantidad de datos.
- ***Mínimo y máximo*:** establecen el rango en el cual los valores se distribuyen.

2.1.2 Medidas de dispersión

- ***Varianza*:** es el promedio aritmético de la desviación cuadrática que existe entre cada valor y la media.
- ***Desviación estándar*:** es la raíz cuadrada de la varianza, la cual se expresa en la misma unidad dimensional que la variable de estudio.
- ***Coeficiente de variación:*** es la razón entre la desviación estándar y la media, destacar que es adimensional y para variables positivas.
- ***Rango Intercuartil*:** es el ancho del intervalo entre el primer y el tercer cuartil que contiene la mitad de los datos.

2.1.3 Histograma y diagrama de caja

El histograma es una representación gráfica de las frecuencias con la que se reitera o repite un valor dentro de un intervalo de valores. Es una herramienta útil de aplicar para detectar valores atípicos o también para visualizar tempranamente la homogeneidad de los valores (Emery, 2013).

En la ilustración 1 se da un ejemplo gráfico de un histograma:

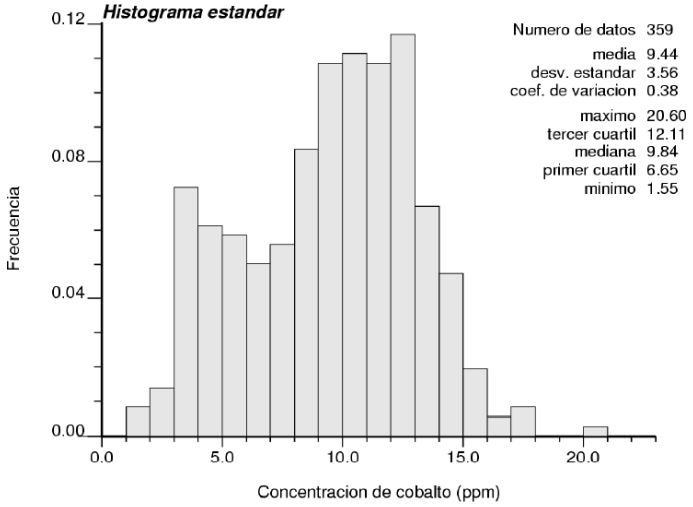

Ilustración 1: Histograma de las concentraciones de cobalto (Emery,2013)

El diagrama de caja es un gráfico que se utiliza para representar una variable numérica que permite visualizar mediante los cuartiles, como se distribuyen los datos, ya sea su grado de asimetría, los valores extremos, la posición de la mediana o la dispersión (Emery, 2013).

En la ilustración 2 se da un ejemplo de diagrama de caja:

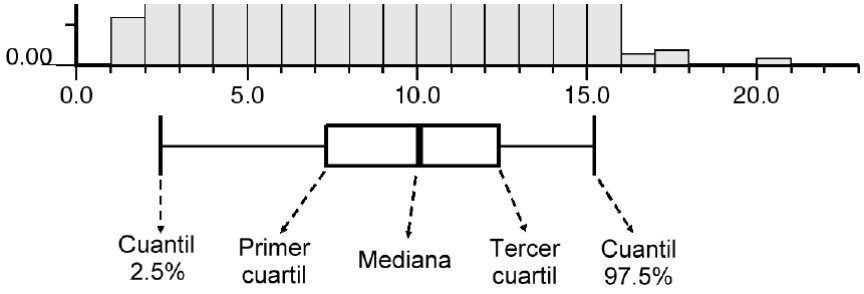

Ilustración 2: Diagrama de caja para las concentraciones de cobalto (Emery,2013)

2.2 Definición de Geoestadística y conceptos básicos de variables regionalizadas

El prefijo “geo” alude a las ciencias de la tierra, en donde mayormente se ha aplicado esta disciplina. El término “estadística” se refiere a la aplicación de herramientas estadísticas y de probabilidades. La geoestadística toma en cuenta las dependencias que existen entre las observaciones disponibles, considerando que ellas están ubicadas en el espacio (Emery,2013).

Para Matheron (1971) la Geoestadística es la aplicación de la teoría de las variables regionalizadas a la estimación de los depósitos mineros. De manera general, diremos que un fenómeno es regionalizado cuando se desplaza en el espacio, manifestando una cierta estructura. Las ciencias de la tierra, entre otras, nos proporcionan numerosos ejemplos.

Los campos de aplicación actuales son de múltiples disciplinas (Emery,2013), entre las que se encuentran:

- Evaluación de recursos naturales
- Ciencias del suelo
- Ciencias ambientales
- Topografía
- Oceanografía
- Geofísica
- Agricultura
- Análisis de imágenes

2.2.1 Variable regionalizada

Una variable regionalizada es una función que representa la variación en el espacio de una cierta magnitud asociada a un fenómeno natural (Alfaro,2007).

En Geoestadística se utiliza la notación condensada, un punto del espacio se representa por la letra **x**. Por ejemplo la ley en el punto x se representa por z(**x**) (Alfaro,2007). Por consiguiente, z(**x**) puede significar:

z(**x**): si el problema es unidimensional (1-D)

z($\mathbf{x}_1$, $\mathbf{x}_2$): si el problema es bidimensional (2-D)

z($\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$): si el problema es tridimensional (3-D)

El objeto sobre el cual se trabaja será una descripción matemática del fenómeno regionalizado, que miden ciertas propiedades o atributos relacionados (Emery,2013).

Según Emery (2013) se pueden encontrar los siguientes:

- La ley de un mineral
- La potencia de una veta
- Tipo de roca
- Densidad de la roca
- Recuperación metalúrgica
- Porosidad y permeabilidad de la roca en un reservorio de petróleo o acuífero
- La concentración de un elemento contaminante en la atmosfera o el suelo
- El número de árboles y su diámetro promedio en áreas de observación de un bosque

Debido a que un fenómeno regionalizado nunca tiene una extensión infinita, la variable regionalizada se estudia sólo dentro de un dominio limitado **D** llamado **campo** de la variable. Este **campo** puede representar una zona natural, fuera de la cual la variable no está definida o también puede tratarse de un dominio particular en donde la variable interesa (Emery,2013).

2.2.2 Función aleatoria

Los modelos geoestadísticos toman en cuenta el valor z(**x**) de la variable regionalizada en un sitio **x** del **campo D** como una realización de una variable aleatoria Z(**x**). Cuando **x** recorre **D**, se obtiene un conjunto de variables aleatorias $Z = \{Z(\mathbf{x}), \mathbf{x} \in D\}$ la cual constituye una función aleatoria. Por lo tanto, la variable regionalizada $Z = \{Z(\mathbf{x}), \mathbf{x} \in D\}$ es una realización de la función aleatoria Z (Emery,2013).

2.2.3 Momentos de una función aleatoria

En muchos casos se puede reducir la complejidad de caracterizar la función aleatoria, al tomar en cuenta solo algunos parámetros descriptivos o momentos de las distribuciones, resumiendo la información más importante (Emery,2013).

Según Emery (2013) los que se presentan a continuación:

Valor esperado:

En cada sitio para **x**, m(**x**) representa la media alrededor de la cual se distribuyen los valores tomados por las realizaciones de la función aleatoria.

$$m(\mathbf{x}) = E\,[Z\,(\mathbf{x})]$$

Ecuación 1: valor esperado de una función aleatoria (Emery,2013).

Varianza:

Es una cantidad positiva, en la que su raíz cuadrada se llama desviación estándar. La varianza y la desviación estándar constituyen medidas de la dispersión de Z(**x**) en torno a su valor medio m(**x**) y cuantifican su carácter aleatorio.

$$\sigma^2\,(\mathbf{x}) = \mathrm{var}\,[Z\,(\mathbf{x})]$$
$$\sigma^2\,(\mathbf{x}) = E\,\{[Z\,(\mathbf{x}) - m(\mathbf{x})]^2\}$$
$$\sigma^2\,(\mathbf{x}) = E\,[Z\,(\mathbf{x})^2] - m(\mathbf{x})^2$$

Ecuación 2: varianza de una función aleatoria (Emery,2013).

Covarianza:

La covarianza está centrada entre dos variables aleatorias Z($\mathbf{x}_1$) y Z($\mathbf{x}_2$), en la cual la covarianza da una visión elemental entre Z($\mathbf{x}_1$) y Z($\mathbf{x}_2$).

$$C(\mathbf{x}_1, \mathbf{x}_2) = \text{cov}[Z(\mathbf{x}_1), Z(\mathbf{x}_2)]$$
$$C(\mathbf{x}_1, \mathbf{x}_2) = E\{[Z(\mathbf{x}_1) - m(\mathbf{x}_1)][Z(\mathbf{x}_2) - m(\mathbf{x}_2)]\}$$
$$C(\mathbf{x}_1, \mathbf{x}_2) = E[Z(\mathbf{x}_1) Z(\mathbf{x}_2)] - m(\mathbf{x}_1) m(\mathbf{x}_2)$$

Ecuación 3: covarianza de una función aleatoria (Emery,2013).

Correlograma:

Es el coeficiente de correlación lineal entre dos variables aleatorias Z($\mathbf{x}_1$) y Z($\mathbf{x}_2$). Además, es adimensional y toma valores entre el intervalo [-1,1].

$$\rho(\mathbf{x}_1, \mathbf{x}_2) = \text{corr}[Z(\mathbf{x}_1), Z(\mathbf{x}_2)] = \frac{\text{cov}[Z(\mathbf{x}_1), Z(\mathbf{x}_2)]}{\sqrt{\text{var}[Z(\mathbf{x}_1)]\ \text{var}[Z(\mathbf{x}_2)]}}$$

Ecuación 4: Correlograma de una función aleatoria (Emery,2013).

Semi-Variograma:

Es la desviación cuadrática media entre dos variables que da un indicio de que tan diferentes son dos variables aleatorias.

$$\gamma(x_1, x_2) = \frac{1}{2}\text{var}[Z(x_1) - Z(x_2)]$$

Ecuación 5: Semi-variograma de una función aleatoria (Emery,2013).

De ahora en adelante, se omitirá el prefijo semi y se considerará llamarlo solamente Variograma.

2.2.4 Hipótesis de estacionaridad

Consiste en postular que la distribución espacial de la función aleatoria no cambia por traslación, en otras palabras, que las propiedades de un conjunto de datos no dependen de su posición absoluta en el espacio, sino más bien que solamente dependen de sus posiciones relativas (Emery,2013).

Para Emery (2013) conlleva las siguientes simplificaciones:

- La distribución univariable no depende del sitio considerado
- La esperanza y la varianza son constantes en el espacio
- La distribución bivariable solo depende de la separación entre los sitios que se consideran, así también para la covarianza, Correlograma y el variograma.

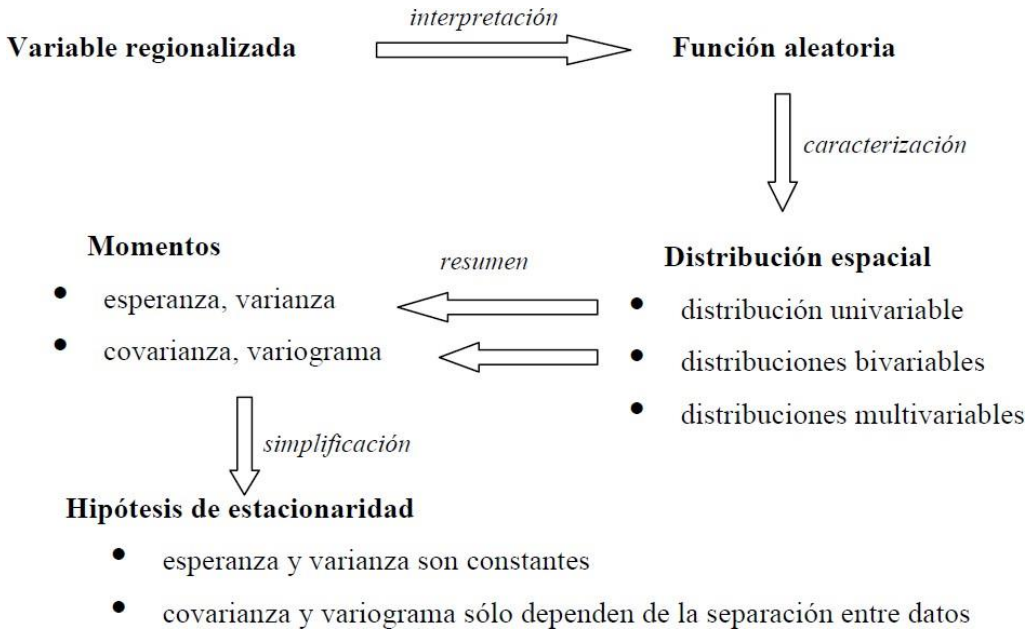

Ilustración 3: Esquema sintético de los conceptos e hipótesis que sustentan el modelo geoestadístico (Emery,2013).

2.3 Herramientas geoestadísticas

Una característica importante dentro del modelamiento geoestadístico es establecer medidas cuantitativas de la variabilidad espacial para poder ser aplicadas luego en una estimación (Rossi & Deutsch,2013).

Para lograr realizar la estimación de una variable regionalizada que es no muestreada, Emery (2013) considera las siguientes dos etapas:

- Análisis estructural: en esta etapa se describe la correlación que existe entre la variable de interés con los puntos en el espacio con la ayuda del variograma.
- Estimación de la variable: se debe tomar en cuenta la variable de interés en los sitios sin muestras con la ayuda del método de Kriging.

2.3.1 Variograma experimental

Representa el valor esperado de la diferencia al cuadrado de dos variables separadas por una distancia **h**. El variograma no es dependiente de las coordenadas de las muestras y es una medida de la variabilidad, por lo tanto crece a medida que las muestras son más diferentes entre ellas (Rossi & Deutsch,2013).

Es posible estimar el variograma con la información disponible con el estimador llamado variograma experimental (Emery,2013) para un vector de separación **h** dado, el cual queda definido por:

$$\hat{\gamma}(\mathrm{h}) = \frac{1}{2\,|N(\mathrm{h})|} \sum_{i=1}^{N(h)} [\mathrm{Z}(\mathbf{x}_\alpha) - \mathrm{Z}(\mathbf{x}_\beta)]^2$$

Ecuación 6: fórmula variograma experimental (Emery,2013).

en donde $N(h) = \{(\alpha,\beta) \text{ tal que } \mathbf{x}_\alpha - \mathbf{x}_\beta = \mathbf{h}\}$ y $|N(\mathbf{h})|$ es el número de pares contenidos en el conjunto N(h).

2.3.1.1 Propiedades del variograma experimental

El variograma experimental presenta las siguientes propiedades según Emery (2013):

- Es un estimador **insesgado** del variograma teórico:

$$E[\hat{\gamma}(\mathbf{h})] = \gamma(\mathbf{h})$$

Ecuación 7: propiedad 1 variograma experimental (Emery,2013).

- Su varianza relativa es un indicador de la **robustez** de $\hat{\gamma}(\mathbf{h})$

$$\text{var}[\hat{\gamma}(\mathbf{h})]/[\gamma(\mathbf{h})]^2$$

Ecuación 8: propiedad 2 variograma experimental (Emery,2013).

Mientras más elevada se encuentre su varianza, es más proclive a que el variograma experimental se acerque a su valor esperado, es decir, el variograma teórico (Emery,2013).

2.3.2 Variograma modelado

No es posible aplicar de forma directa el variograma experimental, debido a que está definido para ciertas distancias y direcciones, por lo que no es completo. Además, está limitado al número de datos y a los parámetros de tolerancia realizados en los cálculos.

Para resolver esta problemática, se hace necesario ajustar un modelo teórico de variograma basado en el variograma experimental la idea es ajustar un modelo teórico de variograma en torno al variograma experimental (Emery,2013).

2.3.2.1 Propiedades variograma teórico

Según Emery (2013) la función de un variograma debe presentar las siguientes propiedades:

- Paridad:

$$\gamma(\mathbf{h}) = \gamma(-\mathbf{h})$$

Ecuación 9: propiedad 1 variograma teórico (Emery,2013).

- Nulidad en el origen:

$$\gamma(\mathbf{0}) = 0$$

Ecuación 10: propiedad 2 variograma teórico (Emery,2013).

- Positividad:

$$\gamma(\mathbf{h}) \geq 0$$

Ecuación 11: propiedad 3 variograma teórico (Emery,2013).

- Comportamiento al infinito igual a cero.

- Función de tipo negativo condicional

2.3.2.2 Comportamiento para distancias muy grandes

Usualmente el variograma crece desde el origen y permanece estable a partir de una distancia **a** en torno a una **meseta**. Ambas variables aleatorias Z(x) y Z(**x**+**h**) estarán correlacionadas si la longitud del vector de separación **h** es inferior a la distancia **a**, que se denomina el **alcance**. El alcance es la zona de influencia, más allá de |**h**| = **a**, el variograma es constante e igual a su **meseta** y las variables *Z(***x***)* y *Z(***x**+**h***)* son independientes (Emery,2013).

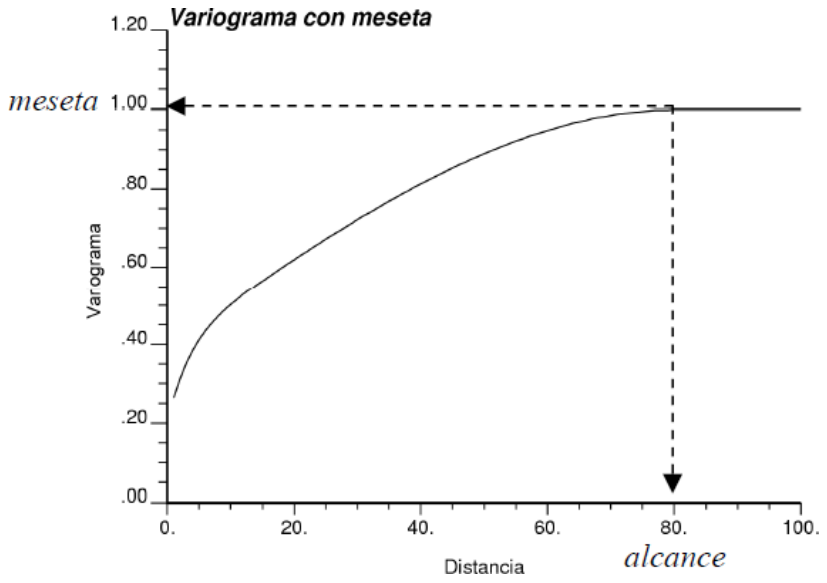

Ilustración 4: Ejemplo de Variograma con meseta y alcance (Emery,2013).

2.3.3 Modelos elementales de variograma

A continuación veremos los tipos de modelos de Variogramas más elementales que se presentan (Emery,2013).

Efecto pepita:

El variograma pepítico de meseta *C* se define como:

$$\gamma(\mathbf{h}) = \begin{cases} 0 \text{ si } \mathbf{h} = \mathbf{0} \\ C \text{ en casocontrario} \end{cases}$$

Ecuación 12: fórmula modelo efecto pepita (Emery,2013).

Este modelo representa que no hay correlación espacial entre los datos (Emery,2013).

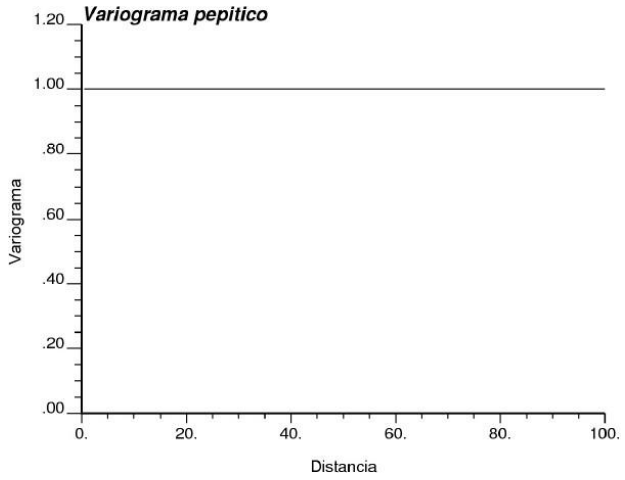

Ilustración 5: Ejemplo de variograma con efecto pepita (Emery,2013).

Modelo esférico:

El variograma esférico de alcance *a* y meseta *C* se define como:

$$\gamma(\mathbf{h})=\begin{cases} C\left\{\frac{3}{2}\frac{|\mathbf{h}|}{a}-\frac{1}{2}\left(\frac{|\mathbf{h}|}{a}\right)^3\right\} & \text{si } |\mathbf{h}|\le a \\ C \text{ en caso contrario} \end{cases}$$

Ecuación 13: fórmula modelo esférico (Emery,2013).

Se caracteriza por tener un crecimiento rápido y de forma lineal en el origen, Además, representa fenómenos continuos (Emery,2013).

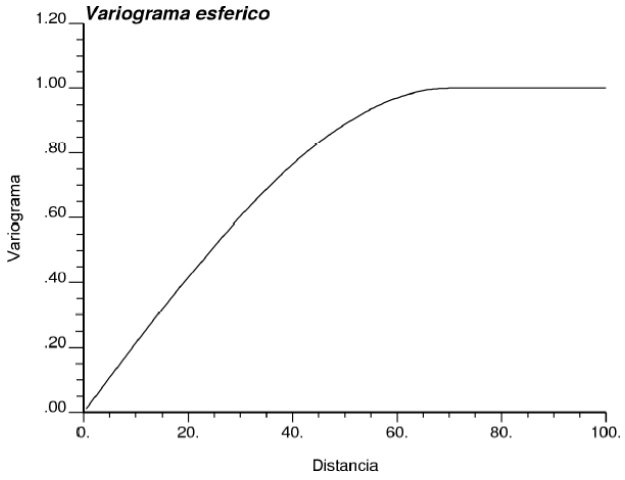

Ilustración 6: Ejemplo de variograma esférico (Emery,2013.)

Modelo exponencial

El variograma exponencial de parámetro *a* y meseta *C* se define como:

$$\gamma(\mathbf{h}) = C\left\{1 - \exp\left(-\frac{|\mathbf{h}|}{a}\right)\right\}$$

Ecuación 14: fórmula modelo exponencial (Emery,2013).

El modelo exponencial alcanza la meseta C solo en forma asintótica, en cambio el modelo esférico llega a la meseta exacta para |**h**| = **a** (Emery,2013).

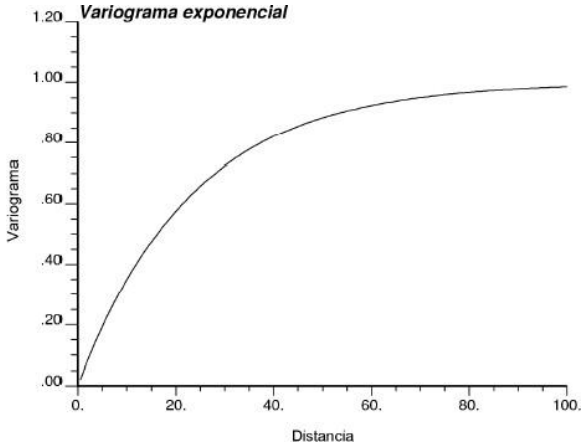

Ilustración 7: Ejemplo de modelo exponencial (Emery,2013).

Modelo gaussiano:

El variograma Gaussiano de parámetro *a* y meseta *C* se define como:

$$\gamma(\mathbf{h}) = C\left\{1 - \exp\left(-\frac{|\mathbf{h}|^2}{a^2}\right)\right\}$$

Ecuación 15: fórmula modelo gaussiano (Emery,2013).

Se caracteriza por ser de forma parabólica cercana al origen. Además, su meseta se alcanza asintóticamente y el alcance practico se considera en $a\sqrt{3}$ (Emery,2013).

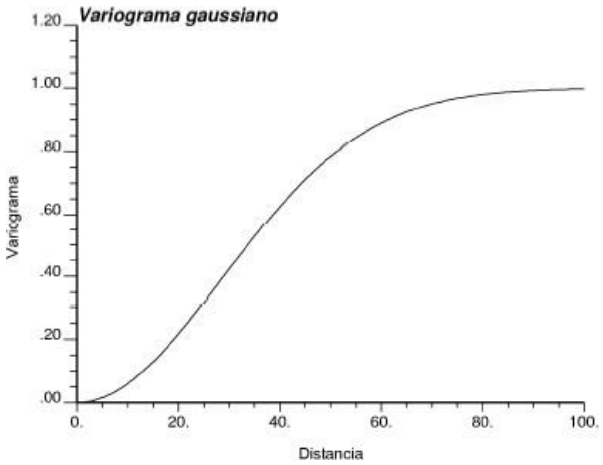

Ilustración 8: Ejemplo de modelo gaussiano (Emery,2013).

Modelos anidados:

El variograma puede modelarse como la suma de varios modelos elementales denominados modelos anidados. Cada escala de observación integra todas las estructuras de los niveles inferiores *(*Emery,2013):

$$\gamma(\mathbf{h}) = \gamma_1(\mathbf{h}) + \gamma_2(\mathbf{h}) + ... + \gamma_S(\mathbf{h})$$

Ecuación 16: ejemplo de modelos anidados (Emery,2013).

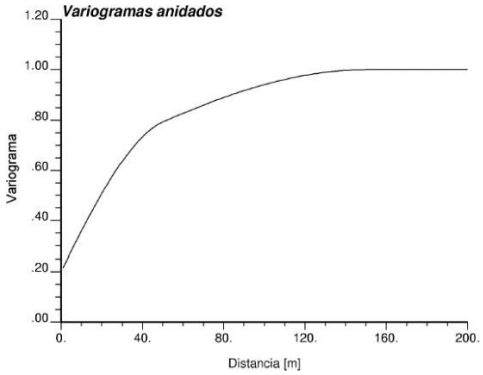

Ilustración 9: Ejemplo de modelo anidado (Emery,2013)

2.3.4 Anisotropías

Las anisotropías se provocan cuando el variograma es diferente según las direcciones del espacio. En ausencia de anisotropía, $\gamma(\mathbf{h})$ sólo dependería de $|\mathbf{h}|$, no de su orientación. En la realidad las anisotropías se pueden identificar al comparar los variogramas experimentales calculados a lo largo de varias direcciones del espacio.

Se distingue varios tipos de anisotropía, en especial, la anisotropía geométrica y la anisotropía zonal (Emery,2013):

- **Anisotropía geométrica:** Es aquella en que el variograma calculado en distintas direcciones presenta la misma meseta, pero con rangos diferentes (Emery,2013).
- **Anisotropía zonal:** Es aquella en que el variograma calculado en distintas direcciones presenta el mismo rango, pero con mesetas diferentes (Emery,2013).

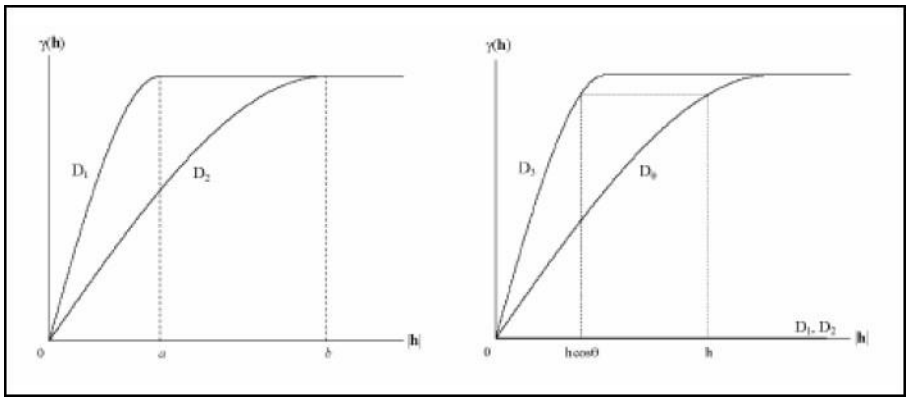

Ilustración 10: Ejemplo de anisotropía geométrica (izquierda) y anisotropía zonal (derecha) (Emery,2013)

2.4 Kriging

Dentro de la geoestadística existen múltiples métodos con los que se pueden realizar estimaciones o predicciones, como por ejemplo, interpolación por el vecino más cercano, inverso a la distancia, método de Sibson, método Baricéntrico, entre otros, pero el estimador por excelencia es el llamado Kriging, el cual debe su nombre a Danie Krige, uno de los precursores de la geoestadística (Emery,2013).

Todos los métodos anteriormente nombrados consideran la información de tipo geométrica para realizar las estimaciones, en cambio el Kriging considera la continuidad espacial de la variable en estudio, resumida a través del modelo de variograma (Emery,2013).

Según Emery (2013) el Kriging busca mejorar la interpolación de los datos al considerar:

- Sus distancias al sitio a estimar.
- Las redundancias entre los datos por posibles agrupamientos
- La continuidad espacial de la variable regionalizada, es decir, el variograma.

Por lo tanto el Kriging permite cuantificar la precisión de la estimación mediante una varianza que mide la dispersión del error potencial cometido en la estimación (Emery,2013).

2.4.1 Construcción del Kriging

Para que se pueda desarrollar el Kriging es necesario que cumpla con 3 propiedades, las cuales se presentan a continuación (Emery,2013):

- **Restricción de linealidad**

El estimador tiene que ser una combinación lineal ponderada de los datos (Emery,2013), que se expresa en la siguiente ecuación:

$$Z^*(\mathbf{x}_0) = a + \sum_{\alpha=1}^{n} \lambda_\alpha Z(\mathbf{x}_\alpha)$$

Ecuación 17: restricción de linealidad para la construcción del Kriging (Emery,2013).

donde $\mathbf{x}_0$ es el sitio donde se busca tener una estimación, $\{\mathbf{x}_\alpha, \alpha = 1... n\}$ son los sitios con datos, mientras que los ponderadores $\{\lambda_\alpha, \alpha = 1... n\}$ y el coeficiente *a* son las incógnitas del problema de Kriging.

- **Restricción de insesgo**

Consiste en expresar que el error de estimación tiene esperanza nula (Emery,2013), lo que se expresa en la siguiente ecuación:

$$E[Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0)] = 0$$

Ecuación 18: restricción de Insesgo para la construcción del Kriging (Emery,2013).

- **Restricción de optimalidad**

Al superar las etapas anteriores, el estimador está sometido a una o varias restricciones pero no está totalmente especificado. La última etapa consiste en buscar los ponderadores que minimizan la varianza del error de estimación (Emery,2013), la cual se expresa en la siguiente ecuación:

$$\mathrm{var}[Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0)]$$

Ecuación 19: restricción de optimalidad para la construcción del Kriging (Emery,2013).

2.4.2 Kriging con media conocida (Kriging Simple)

2.4.2.1 Hipótesis

Se supone que la variable regionalizada *z* es la realización de una función aleatoria *Z*

estacionaria tal que (Emery,2013):

$$\begin{cases} \forall \mathbf{x} \in V, E[Z(\mathbf{x})] = m \quad \textbf{conocida} \\ \forall \mathbf{x}, \mathbf{x}+\mathbf{h} \in V, \mathrm{cov}[Z(\mathbf{x}+\mathbf{h}), Z(\mathbf{x})] = C(\mathbf{h}) \end{cases}$$

Ecuación 20: hipótesis para Kriging con media conocida (Emery,2013).

en donde V representa la vecindad de Kriging.

2.4.2.2 Determinación del estimador

A continuación se verá una a una las etapas del Kriging Simple (Emery,2013):

- **Linealidad:** se asegura esta restricción al tomar como estimador en $\mathbf{x}_0$

$$Z^*(\mathbf{x}_0)=a+\sum_{\alpha=1}^{n}\lambda_\alpha Z(\mathbf{x}_\alpha)$$

Ecuación 21: linealidad del Kriging simple (Emery,2013).

- **Insesgo:** el valor esperado del error de estimación es:

$$E[Z^*(\mathbf{x}_0)-Z(\mathbf{x}_0)]=a+\sum_{\alpha=1}^{n}\lambda_\alpha \underbrace{E[Z(\mathbf{x}_\alpha)]}_{=m}-\underbrace{E[Z(\mathbf{x}_0)]}_{=m}=a+(\sum_{\alpha=1}^{n}\lambda_\alpha-1)m$$

Ecuación 22: Insesgo del Kriging simple (Emery,2013).

Este valor esperado es nulo si se cumple que:

$$a=(1-\sum_{\alpha=1}^{n}\lambda_\alpha)m$$

Ecuación 23: propiedad de valor esperado del Kriging simple (Emery,2013).

- **Optimalidad:** debemos calcular la varianza del error de estimación:

$$\mathrm{var}[Z^*(\mathbf{x}_0)-Z(\mathbf{x}_0)]=\mathrm{var}\{\sum_{\alpha=1}^{n}\lambda_\alpha Z(\mathbf{x}_\alpha)-Z(\mathbf{x}_0)\}$$

Ecuación 24: optimalidad del Kriging simple (Emery,2013).

El término constante *a* no influye en la varianza, por lo cual se omite de la expresión. Ahora, se tiene la siguiente regla de cálculo (Emery,2013):

$$\mathrm{var}\{\lambda_1 Z_1\}=\lambda_1^2\,\mathrm{cov}\{Z_1,Z_1\}$$

$$\mathrm{var}\{\lambda_1 Z_1+\lambda_2 Z_2\}=\lambda_1^2\,\mathrm{cov}\{Z_1,Z_1\}+\lambda_2^2\,\mathrm{cov}\{Z_2,Z_2\}+2\lambda_1\lambda_2\,\mathrm{cov}\{Z_1,Z_2\}$$

:

Ecuación 25: Inicio de reglas de cálculo para el Kriging simple (Emery,2013).

Generalizando:

$$\mathrm{var}\{\sum_{k=1}^{K}\lambda_k Z_k\}=\sum_{k=1}^{K}\sum_{k'=1}^{K}\lambda_k\lambda_{k'}\,\mathrm{cov}\{Z_k,Z_{k'}\}$$

Ecuación 26: Desarrollo de reglas de cálculo para Kriging simple (Emery,2013).

Aplicando esta fórmula, se obtiene:

$$\mathrm{var}[Z^*(\mathbf{x}_0)-Z(\mathbf{x}_0)]$$

$$=\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n}\lambda_\alpha\lambda_\beta\,\mathrm{cov}\{Z(\mathbf{x}_\alpha),Z(\mathbf{x}_\beta)\}+\mathrm{cov}\{Z(\mathbf{x}_0),Z(\mathbf{x}_0)\}-2\sum_{\alpha=1}^{n}\lambda_\alpha\,\mathrm{cov}\{Z(\mathbf{x}_\alpha),Z(\mathbf{x}_0)\}$$

$$=\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n}\lambda_\alpha\lambda_\beta C(\mathbf{x}_\alpha-\mathbf{x}_\beta)+C(\mathbf{0})-2\sum_{\alpha=1}^{n}\lambda_\alpha C(\mathbf{x}_\alpha-\mathbf{x}_0)$$

Ecuación 27: continuación 1 desarrollo de reglas de cálculo Kriging simple (Emery,2013).

en donde $C(.)$ es la covarianza de la función aleatoria Z. El mínimo de esta expresión se obtiene anulando sus derivadas parciales con respecto a las incógnitas $\{\lambda_\alpha,\ \alpha = 1\ldots n\}$. Se obtiene finalmente el sistema de ecuaciones (Emery,2013):

$$\sum_{\beta=1}^{n}\lambda_\beta C(\mathbf{x}_\alpha-\mathbf{x}_\beta)=C(\mathbf{x}_\alpha-\mathbf{x}_0)\ \forall\ \alpha=1\ldots n.$$

Ecuación 28: continuación 2 desarrollo de reglas de cálculo Kriging simple (Emery,2013).

Es un sistema lineal, en el cual el número de ecuaciones y de incógnitas es igual a la cantidad de datos utilizados. En escritura matricial, este sistema es (Emery,2013):

$$\begin{pmatrix} C(\mathbf{x}_1-\mathbf{x}_1) & \cdots & C(\mathbf{x}_1-\mathbf{x}_n)\\ \vdots & & \vdots\\ C(\mathbf{x}_n-\mathbf{x}_1) & \cdots & C(\mathbf{x}_n-\mathbf{x}_n)\end{pmatrix}\begin{pmatrix}\lambda_1\\ \vdots\\ \lambda_n\end{pmatrix}=\begin{pmatrix} C(\mathbf{x}_1-\mathbf{x}_0)\\ \vdots\\ C(\mathbf{x}_n-\mathbf{x}_0)\end{pmatrix}$$

Ecuación 29: sistema de ecuaciones de reglas de cálculo Kriging simple (Emery,2013).

lo que permite determinar los ponderadores de kriging $\{\lambda_\alpha,\ \alpha = 1\ldots n\}$.

Es interesante notar que, debido a la condición de insesgo, el estimador se pone bajo la forma (Emery,2013):

$$Z^*(\mathbf{x}_0)=\sum_{\alpha=1}^{n}\lambda_\alpha Z(\mathbf{x}_\alpha)+(1-\sum_{\alpha=1}^{n}\lambda_\alpha)m\,,$$

Ecuación 30: fórmula final para el estimador Kriging simple (Emery,2013).

de modo que el valor de la media aparece como si fuera un dato adicional, al cual se asigna una ponderación igual al complemento de la ponderación acumulada de los otros datos (Emery,2013).

2.4.2.3 Varianza de Kriging

La varianza mínima del error de estimación en el sitio $\mathbf{x}_0$, llamada varianza de kriging, se simplifica de la siguiente forma (Emery,2013):

$$\sigma^2_{KS}(\mathbf{x}_0)=\sigma^2-\sum_{\alpha=1}^{n}\lambda_\alpha C(\mathbf{x}_\alpha-\mathbf{x}_0)$$

Ecuación 31: varianza para Kriging simple (Emery,2013).

en donde $\sigma^2 = C(\mathbf{0})$ es la varianza a priori de la función aleatoria *Z*. Se puede mostrar que la varianza de kriging simple siempre es menor o igual a la varianza a priori *(Emery,2013)*:

$$\sigma^2_{KS}(\mathbf{x}_0)\leq\sigma^2$$

Ecuación 32: comparación varianza de Kriging versus varianza a priori para Kriging simple (Emery,2013).

2.4.3 Kriging con media desconocida (Kriging Ordinario)

2.4.3.1 Hipótesis

Se supone ahora que la variable regionalizada es la realización de una función aleatoria

Z estacionaria tal que:

$$\begin{cases} \forall \mathbf{x} \in V,\ E[Z(\mathbf{x})] = m \quad \textbf{desconocida} \\ \forall \mathbf{x}, \mathbf{x}+\mathbf{h} \in V,\ \mathrm{cov}[Z(\mathbf{x}+\mathbf{h}), Z(\mathbf{x})] = C(\mathbf{h}) \end{cases}$$

Ecuación 33: hipótesis para Kriging con media desconocida (Emery,2013).

en donde V representa la vecindad de kriging.

2.4.3.2 Determinación del estimador

A continuación se verán una a una las etapas del Kriging Ordinario (Emery,2013):

- **Linealidad:** se asegura esta restricción al tomar como estimador en $\mathbf{x}_0$

$$Z^*(\mathbf{x}_0) = a + \sum_{\alpha=1}^{n} \lambda_\alpha Z(\mathbf{x}_\alpha)$$

Ecuación 34: linealidad del Kriging ordinario (Emery,2013).

- **Insesgo:** el valor esperado del error de estimación es

$$E[Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0)] = a + \sum_{\alpha=1}^{n} \lambda_\alpha \underbrace{E[Z(\mathbf{x}_\alpha)]}_{=m} - \underbrace{E[Z(\mathbf{x}_0)]}_{=m} = a + (\sum_{\alpha=1}^{n} \lambda_\alpha - 1)m$$

Ecuación 35: Insesgo del Kriging ordinario (Emery,2013).

Como se desconoce el valor de la media *m*, este valor esperado es nulo si:

$$a = 0 \ \text{y} \ \sum_{\alpha=1}^{n} \lambda_\alpha = 1$$

Ecuación 36: valor esperado nulo del Kriging ordinario (Emery,2013).

La igualdad sobre la suma de los ponderadores asegura que, en el caso en que todos los datos son iguales a una misma constante, el valor estimado restituirá esta constante (Emery,2013).

- **Optimalidad:** como en el caso del kriging simple, la varianza del error de estimación es:

$$\mathrm{var}[Z^*(\mathbf{x}_0)-Z(\mathbf{x}_0)]=\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n}\lambda_\alpha\lambda_\beta C(\mathbf{x}_\alpha-\mathbf{x}_\beta)+C(\mathbf{0})-2\sum_{\alpha=1}^{n}\lambda_\alpha C(\mathbf{x}_\alpha-\mathbf{x}_0)$$

Ecuación 37: optimalidad del Kriging ordinario (Emery,2013).

Se necesita minimizar esta expresión bajo la condición de insesgo, que impone que la suma de las incógnitas es igual a 1. Esto se logra introduciendo una incógnita adicional llamada multiplicador de Lagrange, que denotaremos como μ, se escribe (Emery,2013):

$$\mathrm{var}[Z^*(\mathbf{x}_0)-Z(\mathbf{x}_0)]$$
$$=C(\mathbf{0})+\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n}\lambda_\alpha\lambda_\beta C(\mathbf{x}_\alpha-\mathbf{x}_\beta)-2\sum_{\alpha=1}^{n}\lambda_\alpha C(\mathbf{x}_\alpha-\mathbf{x}_0)+2\mu\underbrace{\left(\sum_{\alpha=1}^{n}\lambda_\alpha-1\right)}_{=0}$$

Ecuación 38: Inicio de reglas de cálculo para el Kriging ordinario (Emery,2013).

y se minimiza la función de las n+1 variables $\lambda_1,\ldots\ \lambda_n$, μ. Calculando las n+1 derivadas parciales de esta función y luego anulándolas, se obtiene el sistema (Emery,2013):

$$\begin{cases}\dfrac{\partial}{\partial\lambda_\alpha}=0:\sum_{\beta=1}^{n}\lambda_\beta C(\mathbf{x}_\alpha-\mathbf{x}_\beta)+\mu=C(\mathbf{x}_\alpha-\mathbf{x}_0) & \forall\alpha=1\ldots n\\ \dfrac{\partial}{\partial\mu}=0:\sum_{\alpha=1}^{n}\lambda_\alpha=1 & \text{(condición de insesgo)}\end{cases}$$

Ecuación 39: desarrollo de reglas de cálculo para el Kriging ordinario (Emery,2013).

Este sistema contiene una incógnita y una ecuación más que el sistema de kriging simple. Se puede escribir en notación matricial (Emery,2013):

$$\begin{pmatrix} C(\mathbf{x}_1-\mathbf{x}_1) & \cdots & C(\mathbf{x}_1-\mathbf{x}_n) & 1 \\ \vdots & & \vdots & \vdots \\ C(\mathbf{x}_n-\mathbf{x}_1) & \cdots & C(\mathbf{x}_n-\mathbf{x}_n) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{pmatrix} = \begin{pmatrix} C(\mathbf{x}_1-\mathbf{x}_0) \\ \vdots \\ C(\mathbf{x}_n-\mathbf{x}_0) \\ 1 \end{pmatrix}$$

Ecuación 40: sistema de ecuaciones de reglas de cálculo Kriging ordinario (Emery,2013).

Este kriging se denomina kriging ordinario. Siendo el variograma una herramienta equivalente a la covarianza, a partir de la relación $\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$, se puede elegir utilizarlo en lugar de la función de covarianza. Las ecuaciones de kriging pasan a ser (Emery,2013):

$$\begin{cases} \sum_{\beta=1}^{n} \lambda_\beta \gamma(\mathbf{x}_\alpha-\mathbf{x}_\beta) - \mu = \gamma(\mathbf{x}_\alpha-\mathbf{x}_0) \quad \forall \alpha = 1 \ldots n \\ \sum_{\alpha=1}^{n} \lambda_\alpha = 1 \end{cases}$$

Ecuación 41: desarrollo de sistema de ecuaciones de reglas de cálculo Kriging ordinario (Emery,2013).

Esto es:

$$\begin{pmatrix} \gamma(\mathbf{x}_1-\mathbf{x}_1) & \cdots & \gamma(\mathbf{x}_1-\mathbf{x}_n) & 1 \\ \vdots & & \vdots & \vdots \\ \gamma(\mathbf{x}_n-\mathbf{x}_1) & \cdots & \gamma(\mathbf{x}_n-\mathbf{x}_n) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ -\mu \end{pmatrix} = \begin{pmatrix} \gamma(\mathbf{x}_1-\mathbf{x}_0) \\ \vdots \\ \gamma(\mathbf{x}_n-\mathbf{x}_0) \\ 1 \end{pmatrix}$$

Ecuación 42: sistema de ecuaciones final para Kriging ordinario (Emery,2013).

2.4.3.3 Varianza de Kriging

La varianza de kriging ordinario (varianza del error cometido en el sitio $\mathbf{x}_0$) se expresa de la siguiente forma (Emery,2013):

$$\sigma^2_{KO}(\mathbf{x}_0) = \sigma^2 - \sum_{\alpha=1}^{n} \lambda_\alpha C(\mathbf{x}_\alpha-\mathbf{x}_0) - \mu$$

$$= \sum_{\alpha=1}^{n} \lambda_\alpha \gamma(\mathbf{x}_\alpha-\mathbf{x}_0) - \mu$$

Ecuación 43: varianza para Kriging ordinario (Emery,2013).

en donde $\sigma^2 = C(\mathbf{0})$ es la varianza *a priori* de la función aleatoria *Z*, o sea, la meseta de su variograma. Ahora, la segunda igualdad muestra que la varianza de kriging no depende de este valor σ^2, por lo cual el kriging ordinario sigue aplicable incluso cuando el variograma no presenta meseta (Emery,2013).

2.4.4 Kriging multivariable (Co-Kriging)

Se trata de la versión multivariable del kriging, en la cual se busca estimar el valor de una variable tomando en cuenta los datos de esta variable y de otras variables correlacionadas. La puesta en marcha del Co-Kriging requiere tener los modelos variográficos de cada variable en estudio, así como también los variogramas cruzados entre las distintas variables, para poder medir la correlación existente entre las variables (Emery,2013).

El variograma cruzado entre dos variables (Z_1 y Z_2) se define con la siguiente ecuación (Emery,2013):

$$\gamma_{12}(\mathbf{h}) = \frac{1}{2}\mathrm{cov}\{Z_1(\mathbf{x}+\mathbf{h}) - Z_1(\mathbf{x}), Z_2(\mathbf{x}+\mathbf{h}) - Z_2(\mathbf{x})\}$$

Ecuación 44: variograma cruzado entre dos variables (Emery,2013).

y se puede inferir a partir de los datos disponibles al plantear la siguiente ecuación (Emery,2013):

$$\hat{\gamma}_{12}(\mathbf{h}) = \frac{1}{2|N(\mathbf{h})|}\sum_{N(\mathbf{h})}[z_1(\mathbf{x}_\alpha) - z_1(\mathbf{x}_\beta)][z_2(\mathbf{x}_\alpha) - z_2(\mathbf{x}_\beta)]$$

Ecuación 45: inferencia para variograma cruzado entre dos variables (Emery,2013).

en donde $N(\mathbf{h}) = \{(\alpha,\beta)$ tal que $\mathbf{x}_\alpha - \mathbf{x}_\beta = \mathbf{h}$, siendo ambas variables z_1 y z_2 medidas en $\mathbf{x}_\alpha$ y $\mathbf{x}_\beta\}$. Para poder determinar el variograma cruzado experimental, se necesita tener los datos de las diferentes variables en los mismos sitios (Emery,2013).

2.4.5 Kriging no lineal

Los métodos de Kriging no lineal consisten en aplicar kriging a una transformada (no lineal) de la variable *Z*, luego en volver a esta variable. Esta etapa de transformación de vuelta no es trivial, pues requiere introducir correcciones para que el estimador final no tenga sesgo (Emery,2013).

Otros métodos de kriging no lineal buscan caracterizar el valor desconocido $Z(\mathbf{x}_0)$ no por un valor estimado, sino que por una distribución de probabilidad, entre los que destacan los siguientes (Emery,2013):

- **El *kriging de indicadores*:** basado en una codificación de la variable *Z* en un conjunto de variables binarias o indicadores, según si *Z* sobrepasa o no determinados umbrales.
- **El *kriging disyuntivo:*** también llamado Co-Kriging de indicadores.
- **El *kriging multi-Gaussiano*:** basado en una transformación de la variable original *Z* en una variable de distribución Gaussiana.

2.4.6 Función indicadora

Una función indicadora trata sobre una transformación binaria de la variable aleatoria original Z(x) que, para un determinado valor $z_k \in \mathbb{R}$, y se expresa en la siguiente ecuación (Journel,1983):

$$I_Z(x; z_k) = I(Z(x) \leq z_k) \begin{cases} 1, & si \quad Z(x) \leq z_k \\ 0, & si \quad Z(x) > z_k \end{cases}$$

Ecuación 46: función indicadora para valores continuos (Journel,1983).

La función indicadora trata en discretizar el rango de valores de la variable continua Z, R_Z,

en una serie de K valores de corte z_k, k = 1, …, K (Journel,1983).

En esta investigación de memoria, según lo ya explicado anteriormente, se estudiará la estimación de variables nominales, las cuales permiten codificar un conjunto de dominios que subdividen el espacio, como por ejemplo, dominios por tipos de litologías, roca, textura, o representar variables con un número limitado de categorías.

Sea Z una variable categórica o categorizable, en K categorías, s_k, $k = 1, \ldots, K$. Estas categorías deben ser exhaustivas y mutuamente excluyentes. Para una variable categórica o nominal su función indicadora es la siguiente (Journel,1983):

$$I(x; s_k) = I(Z(x) = s_k) \begin{cases} 1, & si\ Z(x) = s_k \\ 0, & si\ Z(x) \neq s_k \end{cases}.$$

Ecuación 47: función indicadora para valores categóricos (Journel,1983).

El indicador para este caso se expone como la probabilidad de que una categoría predomine en una ubicación particular. Las propiedades de exclusividad y exhaustividad implican las siguientes relaciones (Journel,1983):

$$I(x; s_k) \cdot I(x; s_{k'}) = 0, \forall k \neq k',\ \text{y}$$

$$\sum_{k=1}^{K} i(x; s_k) = 1.$$

Ecuación 48: propiedades para la función indicadora (Journel,1983).

2.4.7 Kriging de Indicadores

El Kriging de indicadores es un método no lineal y no paramétrico que consiste en que los valores son convertidos a 0 y 1 según su relación con una categoría establecida. Es posible construir una función de distribución condicional acumulada mediante la unión de K estimadores tipo Kriging de indicadores. Esta función representa un modelo probabilístico sobre la incertidumbre de los valores Z(x) no muestreados (Journel,1983).

Para estimar $E\{I(x; s_k)|(n)\}$ se aplica el siguiente ponderador lineal (Journel,1983):

$$I^*(x; s_k) = E^*\{I(x; s_k)|(n)\} = \sum_{\alpha=1}^{n} \lambda_\alpha(x; s_k) I(x; s_k)$$

Ecuación 49: fórmula para Kriging de indicadores (Journel,1983).

Cuando se tienen diversas categorías k, el sistema es llamado usualmente Kriging de indicadores múltiple. Los ponderadores y la función de distribución acumulada condicional dependen tanto de la ubicación como, así también del número de categorías s_k, con k = 1, …, K. Luego, hay un variograma indicador $\gamma_I(h; s_k)$ y un sistema Kriging por categoría (Journel,1983).

Las etapas para aplicar el Kriging de indicadores son las siguientes (Journel,1983):

- Elegir las categorías s_k.
- Para k = 1, …, K
  - ❖ Codificar los datos en indicadores $I(x, s_k)$.
  - ❖ Realizar análisis variográfico,
  - ❖ Realizar Kriging del indicador.
  - ❖ Procesar estimaciones para obtener distribución condicional válida.

2.4.8 Validación cruzada

Se puede verificar la adecuación entre los datos y los parámetros adoptados, utilizando la llamada técnica de la validación cruzada. Consiste en estimar sucesivamente, mediante Kriging, cada dato, considerando sólo los datos restantes. Se puede calcular entonces el error de estimación en cada sitio con dato y realizar un análisis estadístico de los errores cometidos en todos los sitios con datos (Emery,2013).

Según Emery (2013) la validación cruzada se presenta habitualmente mediante pruebas gráficas, las cuales pueden ser:

- La nube de correlación entre los valores de los datos y los valores estimados.
- El histograma de los errores estandarizados.
- La nube de correlación entre los errores estandarizados y los valores.
- El mapa de ubicación de los datos, donde se localiza los datos mal estimados, es decir aquellos cuyos errores estandarizados salen del intervalo.

Estos criterios permiten comprobar el desempeño del Kriging y comparar la calidad de diferentes ajustes posibles para el variograma. Una técnica similar a la validación cruzada es el jack-knife, el cual no considera una reposición de los datos si no que se divide los datos en dos grupos y se estima los datos de un grupo a partir de los datos del otro grupo (Emery,2013).

2.5 Definición de Aprendizaje de Máquinas

Según Murphy (2012) el aprendizaje de máquinas o machine learning es un conjunto de métodos que pueden detectar automáticamente patrones en los datos y luego usar los patrones descubiertos para predecir datos futuros o para realizar otros tipos de toma de decisiones en condiciones de incertidumbre.

Para Rodríguez-Sahagún (2018) el aprendizaje de máquinas consiste en que un algoritmo es capaz de aprender automáticamente por la ayuda del análisis de los datos. Se dice que un ordenador o programa informático aprende de unos determinados datos o experiencias E, con respecto a una determinada tarea T, y un medidos de eficiencia EF de la realización de dicha tarea, si su rendimiento al realizar la tarea T, medida por EF, mejora con la experiencia E.

2.5.1 Tipos de tareas del Aprendizaje de Máquinas

Según los explicando anteriormente, existen múltiples tipos de tareas que se pueden realizar con la aplicación del aprendizaje de máquinas, dentro de las cuales se encuentran las siguientes (Rodríguez-Sahagún 2018):

- Clasificación
- Regresión
- Traducción
- Detección de anomalías
- Estimación de la función de densidad de probabilidad

2.5.2 Aprendizaje supervisado, no supervisado y semisupervisado

Se pueden presentar distintos tipos de aprendizajes dependiendo del tipo de dato que se está en estudio, los cuales pueden ser (Rodríguez-Sahagún 2018):

- **Aprendizaje supervisado:** los algoritmos que utilizan este tipo de aprendizaje utilizan datos que contienen características, pero que llevan asociadas una etiqueta. El termino proviene de que al algoritmo se le enseña la etiqueta correcta para cada paquete de datos determinado, supervisándolo (Rodríguez-Sahagún 2018).

- **Aprendizaje no supervisado:** este tipo de algoritmos analiza una serie de datos que pueden contener numerosas características, y del propio algoritmo aprender las propiedades de la serie de datos. Un ejemplo de este tipo de aprendizaje podría ser el denominado clustering o agrupamiento, que consiste en dividir un conjunto de datos en conjuntos de observaciones similares (Rodríguez-Sahagún 2018).
- **Aprendizaje semisupervisado:** este tipo de aprendizaje ha experimentado un boom en los últimos años debido a la aplicación de un tipo muy específico de arquitectura denominada GAN (Generative Adversarial Network) la cual combina una pequeña porción de datos con etiquetas asociadas con una gran parte de datos sin etiquetas, cuya estructura tiene que aprender el algoritmo (Rodríguez-Sahagún 2018).

2.6 Redes Neuronales

Las redes neuronales son redes interconectadas de forma masiva en paralelo de elementos simples y con cierta organización jerárquica, las cuales interactúan entre sí emulando el sistema nervioso del ser humano. Las redes, al interconectar las neuronas, generan 3 tipos de capas que son relevantes: la capa de entrada, capa oculta y capa de salida. Cada neurona se denomina nodo y existen funciones básicas que determinan el comportamiento (Rodríguez-Sahagún 2018).

La principal similitud entre una neurona biológica y una neurona artificial es que ambas adquieren conocimiento a través del aprendizaje. En la siguiente ilustración muestra un modelo de neurona artificial *i* junto con la analogía existente con una neurona biológica (Caparrini,2018):

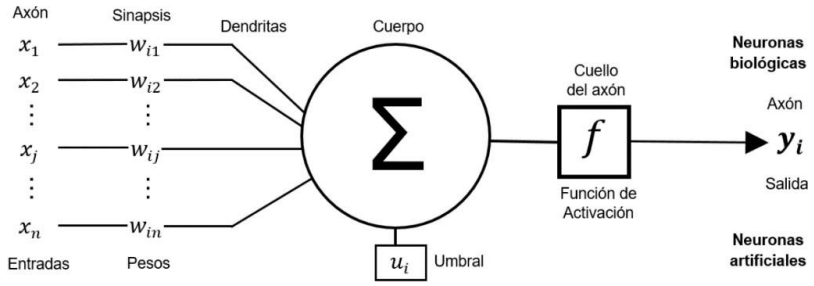

Ilustración 11: Estructura de una neurona artificial y su analogía con una neurona biológica (Caparrini,2018).

La neurona artificial anteriormente vista tiene las siguientes particularidades (Caparrini,2018):

- Un conjunto de entradas $x1, . . , xn.$
- Pesos sinápticos $w_1, \ldots, wn$ correspondientes a cada entrada.
- Una función de agregación, $\sum$.
- Una función de activación, f.
- Una salida y.

La neurona artificial puede ser vista como un diagrama donde la neurona se representa por un nodo y las líneas junto con los pesos son las conexiones entre las entradas y la salida de la neurona. La neurona recibe información en forma de vector, cada entrada x_n, donde n representa el número de entradas, es multiplicada por un correspondiente peso w_n, estos pesos representan la fuerza de las interconexiones de la red. Las entradas ponderadas por sus respectivos pesos son incorporadas a la neurona mediante la función de agregación $\sum$, que comúnmente corresponde a la sumatoria de estas (Matich,2001).

Al igual que las neuronas biológicas, las neuronas artificiales tienen diferentes estados de activación, los cuales son calculados con la función de activación, transformando la entrada global, menos el umbral u_i, en un valor de activación, cuyo rango normalmente va de (0 a 1) o de (- 1 a 1), pudiendo estar totalmente inactiva (0 o -1) o totalmente activa (1) .Entre las funciones de activación más utilizadas se pueden nombrar de tipo Sigmoide,tangente hiperbólica ReLu,entre otras.(Matich,2001).

Las neuronas se distribuyen dentro de la red formando capas, con un número determinado de dichas neuronas en cada una de ellas. En la siguiente ilustración se puede ver 3 tipos de capas de acuerdo de su posición en la red (Matich,2001):

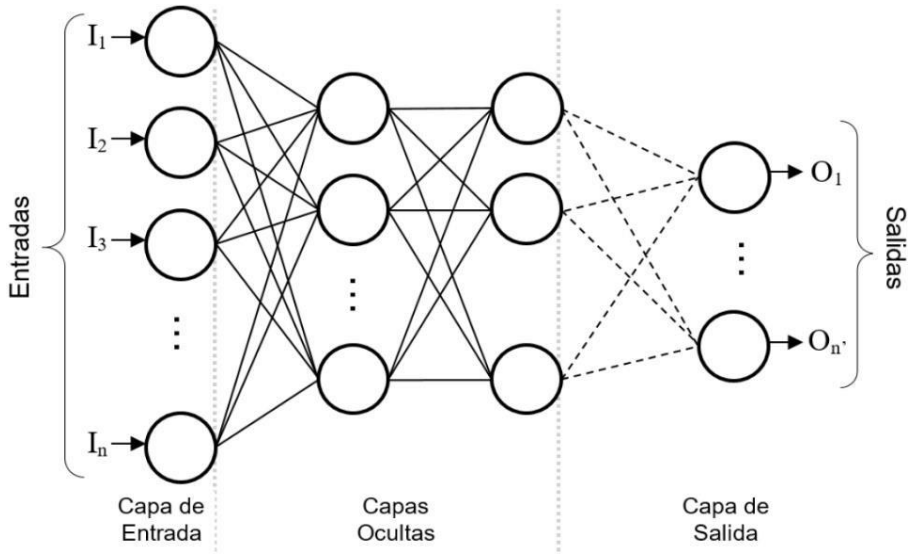

Ilustración 12: Estructura de una red neuronal artificial (Matich,2001).

- Capa de entrada: es aquella que recibe información proveniente de las fuentes externas de la red.
- Capas ocultas: son internas a la red y no tienen contacto directo con el entorno exterior. El número de capas ocultas puede estar entre cero y un número elevado. Las neuronas de las capas ocultas pueden estar interconectadas de distintas maneras, lo que determina las distintas arquitecturas de redes neuronales.
- Capa de salida: transfieren información de la red hacia el exterior.

La finalidad de que las redes neuronales procesen datos de entrada es que se obtenga una salida deseada. Semejante a una red neuronal biológica, las redes neuronales artificiales son capaces de detectar y aprender patrones complejos y características dentro de los datos, aprendiendo de la experiencia y aplicando tal conocimiento a la resolución de problemas nuevos, es decir, una red neuronal debe aprender a calcular la salida correcta para cada vector de entrada de un conjunto de registros. El proceso de aprendizaje se conoce como entrenamiento y consiste en alimentar la red con un conjunto de datos de entrenamiento y adaptar los pesos de acuerdo con una regla de aprendizaje, ya con las redes neuronales artificiales entrenadas son capaces de hacer predicciones, clasificaciones y segmentaciones (Matich,2001).

2.6.1 Arquitectura de redes neuronales artificiales

La arquitectura de una red neuronal consiste cómo se organizan las neuronas para formar capas. Los parámetros fundamentales de la red son: el número de capas, el número de neuronas por capa y el tipo de conexiones entre neuronas (Matich,2001).

Existen dos tipos de organización para las capas de una red neuronal artificial (Matich,2001):

- Redes Monocapa: Se establecen conexiones entre neuronas que pertenecen a una única capa que constituye la red. Se utilizan en tareas relacionadas con lo autoasociación (regenerar información de entrada que se presenta de forma incompleta).
- Redes Multicapa: Se disponen de conjuntos de neuronas agrupadas en varias capas. Usualmente, todas las neuronas de una capa reciben señales de entrada de la capa antecesora y envían señales de salida a la capa inmediatamente posterior, estas conexiones se denominan conexiones hacia adelante o feedforward. Sin embargo, existe la posibilidad de conectar la salida de neuronas de capas posteriores a la entrada de capas anteriores, estas conexiones se denominan conexiones hacia atrás o feedback.

Y para el caso de las conexiones se dan de dos tipos (Flórez & Fernández,2008):

- Redes Feedforward: es cuando las conexiones entre neuronas fluyen en un único sentido hacia delante, desde las neuronas de entrada a la capa o capas de procesamiento, hasta llegar a la capa de salida.

En la siguiente ilustración se encuentra un ejemplo:

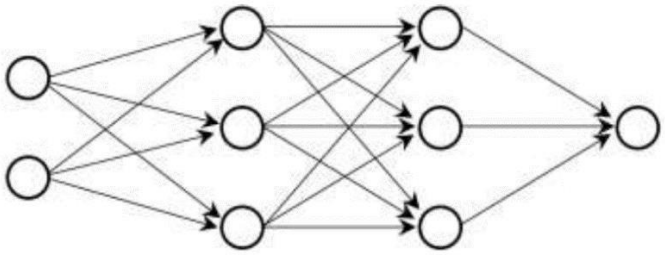

Ilustración 13: Ejemplo de una red neuronal con conexión feedforward (Flórez & Fernández,2008).

- Redes Feedback: es cuando existen conexiones hacia atrás, en donde la conexión puede ser entre una misma neurona (a), entre neuronas de una misma capa (b) y entre neuronas de una capa a una capa anterior (c).

En la siguiente Ilustración se da un ejemplo de lo anteriormente dicho:

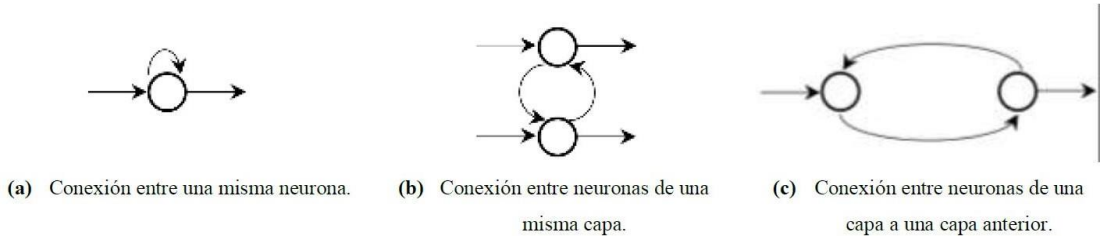

Ilustración 14: Ejemplos de una red neuronal con conexión feedforward (Flórez & Fernández,2008).

2.6.2 Algoritmo Backpropagation

Es uno de los modelos de redes neuronales más populares y utilizados y se debe a su capacidad de solucionar problemas de alta complejidad. Este modelo fue diseñado para entrenar una red multicapa con conectividad total. El algoritmo se basa en propagar el error de la capa de salida hacia atrás, ya que este es el único que puede calcular de forma exacta, y así estimar el error en las salidas de las neuronas en las capas ocultas, este proceso se lleva a cabo hasta que todas las neuronas de las capas ocultas tengan el error que aportaron relativo al error total, con la finalidad de modificar sus pesos sinápticos para que en una próxima ocasión la salida seas más cercana a la esperada. El algoritmo se detiene una vez que su condición de parada sea verificada, como que el error de la salida sea reducido a un valor permitido, o el número de iteraciones propuesto haya sido superado (Flórez & Fernández,2008).

De acuerdo con la presente investigación de memoria se considerará trabajar con redes neuronales artificiales de tipo multicapa con conexiones hacia atrás, las cuales son entrenadas por medio de mecanismos de aprendizaje supervisado, específicamente el algoritmo anteriormente visto, backpropagation o también llamado de retropropagación hacia atrás.

2.7 Regresión Logística

La regresión logística consiste en el uso de herramientas estadísticas para describir la relación entre una variable objetivo y un conjunto de variables explicativas. La regresión logística se utiliza cuando la variable que se desea modelar es dicotómica, es decir, del tipo Sí/No, Bueno/Malo, Presente/Ausente, etc. y busca modelar la influencia de la aparición de las variables explicativas en la ocurrencia del fenómeno dicotómico En la práctica, para aplicar este modelo se crea una variable binaria ficticia cuya estructura es (Hosmer & Lemeshow, 2000):

$$y_i = \begin{cases} 1 & Cuando\ el\ fenómeno\ ocurre \\ 0 & Cuando\ el\ fenómeno\ no\ ocurre \end{cases}$$

Ecuación 50: Ejemplo de variable dicotómica regresión logística (Hosmer & Lemeshow, 2000).

En donde i representa cada observación que se posee.

Por lo cual, la regresión logística es un modelo estadístico de clasificación binaria que entrega la probabilidad de pertenencia a uno de los dos grupos definidos, utilizando para ello un conjunto de regresores xi $\in n$ con i = {1 … N} y N el número de observaciones (Hosmer & Lemeshow, 2000).

La probabilidad de pertenencia se obtiene mediante la siguiente ecuación (Hosmer & Lemeshow, 2000):

$$p(x_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1' x_i)}}$$

Ecuación 51: Probabilidad para regresión logística (Hosmer & Lemeshow, 2000).

En donde $\beta_1 \in \mathrm{R}^n$.

Para la estimación de los parámetros, en otras palabras, la calibración del modelo se utiliza el método de máxima verosimilitud, en el cual se busca maximizar la probabilidad estimada de obtener los resultados categorizados según yi. La función de verosimilitud es la siguiente (Hosmer & Lemeshow, 2000):

$$\mathcal{L}(\beta) = \prod_i f(x_i, \beta)$$

Ecuación 52: Máxima verosimilitud para regresión logística (Hosmer & Lemeshow, 2000).

Donde f(xi, β) corresponde a la función de densidad de probabilidad de xi que en este caso correspondería al modelo de regresión logística. En otras palabras, la función de verosimilitud puede ser expresada como (Hosmer & Lemeshow, 2000):

$$\mathcal{L}(\beta) = \prod_i p(x_i)^{y_i}[1 - p(x_i)]^{1-y_i}$$

Ecuación 53: otra forma de expresar máxima verosimilitud regresión logística (Hosmer & Lemeshow, 2000).

Típicamente se trabaja con el logaritmo de la función de verosimilitud, pues es más simple de abordar matemáticamente, la expresión es la siguiente (Hosmer & Lemeshow, 2000):

$$L(\beta) = \ln[\mathcal{L}(\beta)] = \sum_i \{y_i \ln[p(x_i)] + (1 - y_i)\ln[1 - p(x_i)]\}$$

Ecuación 54: Logaritmo de máxima verosimilitud regresión logística (Hosmer & Lemeshow, 2000).

Los estimadores de máxima verosimilitud se calculan al aplicar condiciones de primer orden a la función de verosimilitud. Las ecuaciones obtenidas, conocidas como ecuaciones de verosimilitud son (Hosmer & Lemeshow, 2000):

$$\sum [y_i - p(x_i)] = 0$$

$$\sum x_i [y_i - p(x_i)] = 0$$

Ecuación 55: ecuaciones de verosimilitud regresión logística (Hosmer & Lemeshow, 2000).

Al resolver el problema de optimización se obtienen como resultados un conjunto de estimadores asintóticamente eficientes, insesgados y distribuidos normalmente $\hat{\beta_0}$ corresponde al punto de corte en el eje de las ordenadas y los demás estimadores $\hat{\beta_i}$, $i \neq 0$ corresponden a los coeficientes asociados a las variables explicativas. Cabe destacar que los parámetros de la regresión logística se distribuyen asintóticamente siguiendo una distribución normal de parámetros (βi, $\sigma\beta$i) (Hosmer & Lemeshow, 2000).

En la práctica, el uso que se le da a este modelo consiste en seleccionar un punto de corte para el valor de la probabilidad, tal que para valores mayores a ese punto de corte se determine que el valor esperado para la variable en estudio sea 1 y en caso contrario se asigna valor 0. De esta manera se logra la clasificación. En términos matemáticos (Hosmer & Lemeshow, 2000):

$$p(x_i) \geq punto\ de\ corte \Rightarrow\ y_i = 1$$
$$p(x_i) < punto\ de\ corte \Rightarrow\ y_i = 0$$

Ecuación 56: punto de corte para las probabilidades de regresión logística (Hosmer & Lemeshow, 2000).

En la ilustración 15 se da un ejemplo de regresión logística:

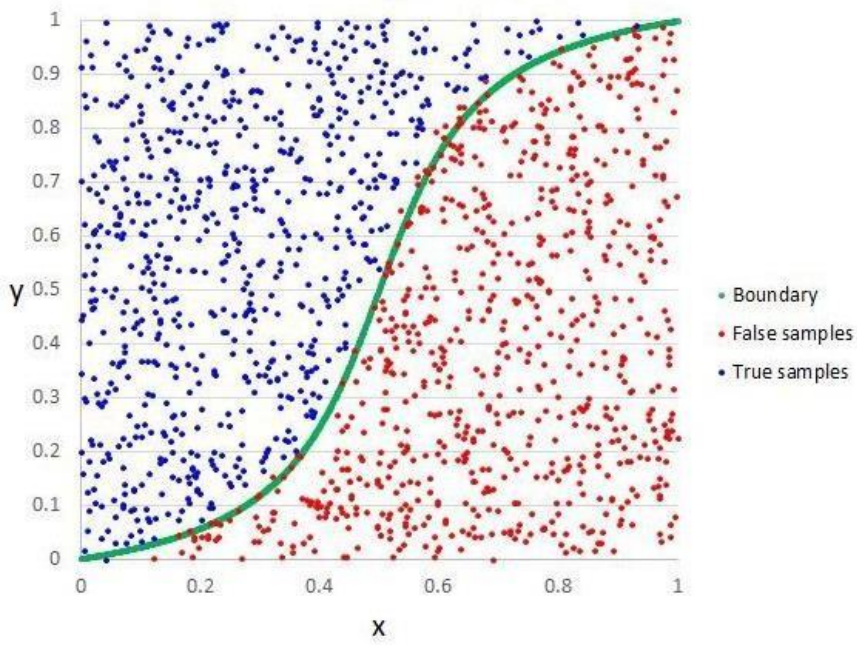

Ilustración 15: Ejemplo de regresión logística (Hosmer & Lemeshow,2000).

Existen dos tipos de regulación que se pueden utilizar en una regresión logística, y se verán a continuación:

- **Regularización Lasso (L1):** En la regularización Lasso, también llamada L1, la complejidad C se mide como la media del valor absoluto de los coeficientes del modelo. Además, Lasso será de ayuda cuando se sospeche que varios de los atributos de entrada sean irrelevantes. Al usar Lasso, se está fomentando que la solución sea poco densa. Es decir, se favorece que algunos de los coeficientes acaben valiendo 0. Esto puede ser útil para descubrir cuáles de los atributos de entrada son relevantes y, en general, para obtener un modelo que generalice mejor. Lasso es de gran ayuda al hacer la selección de atributos de entrada. Lasso funciona mejor cuando los atributos no están muy correlacionados entre ellos (Hosmer & Lemeshow, 2000).

- **Regularización Ridge (L2):** En la regularización Ridge, también llamada L2, la complejidad C se mide como la media del cuadrado de los coeficientes del modelo. Al igual que ocurría en Lasso, la regularización Ridge se puede aplicar a varias técnicas de aprendizaje automático. Ridge será de ayuda cuando se sospeche que varios de los atributos de entrada estén correlacionados entre ellos. Ridge hace que los coeficientes acaben siendo más pequeños. Esta disminución de los coeficientes minimiza el efecto de la correlación entre los atributos de entrada y hace que el modelo generalice mejor. Ridge funciona mejor cuando la mayoría de los atributos son relevantes (Hosmer & Lemeshow, 2000).

2.8 Overfitting y Underfitting de los modelos

El objetivo central del machine learning es que el algoritmo rinda bien frente a nuevos datos de entrada, no solo con los cuales el algoritmo ha sido entrenado. La habilidad para obtener un buen rendimiento con datos no observados durante la etapa de entrenamiento se denomina generalización (Rodríguez-Sahagún 2018).

Usualmente, cuando un algoritmo propio del machine learning aprende y se entrena, este tiene acceso a una parte de los datos llamado set de entrenamiento, obteniendo un error de entrenamiento, siendo el objetivo primordial reducir dicho error. A su vez también se busca reducir el error al generalizar, también denominado error de prueba. Esto se consigue normalmente dividiendo el set inicial de datos en datos de entrenamiento y datos de prueba, de manera que el algoritmo solo se entrena con el primer paquete de datos, y luego se pone a prueba con el segundo, en donde típicamente se distribuye el 80% de los datos para entrenamiento y el 20% restante para datos de prueba (Rodríguez-Sahagún 2018).

Los factores determinantes de lo bien que funciona un algoritmo son el ser capaz de hacer que el error durante el entrenamiento sea pequeño, y a su vez ser capaz de hacer que la diferencia entre el error en el entrenamiento y en la prueba sean pequeñas. Dichos factores están ligados a los dos desafíos u objetivos centrales del machine learning: overfitting y underfitting. El parámetro principal del modelo que afecta a estos dos desafíos es su capacidad o complejidad, de manera que si el modelo es demasiado simple se tendera a alcanzar un escenario de underfitting, donde tanto el error de entrenamiento como el de generalización o test se mantienen altos, y si el modelo resulta ser demasiado complejo para el problema, se tendrá la situación de overfitting, donde el error en los datos de entrenamiento seguirá disminuyendo, pero el del test ira aumentando, lo que quiere decir que el modelo se está adaptando extremadamente bien a los datos de entrenamiento, pero que su capacidad para generalizar nuevas observaciones no será la ideal (Rodríguez-Sahagún 2018).

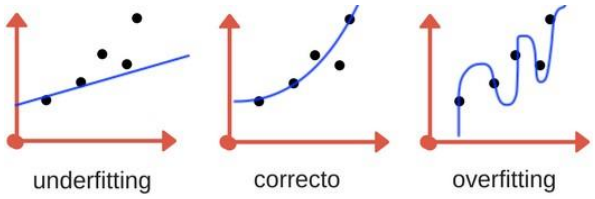

Ilustración 16: Ejemplo gráfico de overfitting y underfitting (Rodríguez-Sahagún 2018).

2.9 Medidas de desempeño de modelos de estimación

De acuerdo con lo que se busca determinar en la presente memoria, es que nace un problema con respecto a cómo medir estadísticamente los distintos métodos de estimación para el mismo conjunto de datos. Por lo tanto se hace uso como medida de desempeño, la matriz de confusión y que se verá de que trata a continuación:

Una matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo de clasificación. Contiene información acerca de la clasificación real y la predicha, es de tamaño *mxm*, donde *m* es el número de diferentes categorías. Además, las medidas de desempeño permiten evaluar de manera cuantitativa si uno de los modelos ajustados es mejor que otro. Estas medidas son calculadas en función de la matriz de confusión asociada al modelo (Kohavi & Provost,1998).

En la Figura siguiente se muestra una matriz de confusión para el caso de $m = 2$:

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Tabla 1: Ejemplo de Matriz de confusión para caso de 2x2 (elaboración Propia).

Las casillas en la matriz de confusión tienen los siguientes significados (Kohavi & Provost,1998):

- Verdaderos positivos (VP): El valor real es positivo y la prueba predijo que también era positivo.
- Verdaderos negativos (VN): El valor real es negativo y la prueba predijo que también el resultado era negativo.
- Falsos positivos (FP): El valor real es positivo, y la prueba predijo que el resultado es negativo. Esto es lo que en estadística se conoce como error tipo II.
- Falsos negativos (FN): El valor real es negativo, y la prueba predijo que el resultado es positivo. Esto es lo que en estadística se conoce como error tipo I.

A partir de las 4 opciones anteriormente vistas, surgen las métricas de la matriz de confusión y que se verán a continuación:

La Exactitud:

Se refiere a lo cerca que está el resultado de una medición del valor verdadero. En términos estadísticos, la exactitud está relacionada con el sesgo de una estimación. Se representa como la proporción de resultados verdaderos (tanto verdaderos positivos (VP) como verdaderos negativos (VN)) dividido entre el número total de casos examinados (verdaderos positivos, falsos positivos, verdaderos negativos, falsos negativos) (Kohavi & Provost,1998).

Se calcula como:

VP+VN) /(VP+FP+FN+VN)

Ecuación *57*: Métrica de desempeño exactitud para matriz de confusión (elaboración propia).

La Precisión:

Es la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión. Se representa por la proporción de verdaderos positivos dividido entre todos los resultados positivos (tanto verdaderos positivos, como falsos positivos) (Kohavi & Provost,1998).

Se calcula como:

$$VP/(VP+FP)$$

Ecuación *58*: Métrica de desempeño precisión para matriz de confusión (elaboración propia).

La Sensibilidad:

También se conoce como Tasa de Verdaderos Positivos, es la proporción de casos positivos que fueron correctamente identificadas por el algoritmo (Kohavi & Provost,1998).

Se calcula como:

$$VP/(VP+FN)$$

Ecuación *59*: Métrica de desempeño sensibilidad para matriz de confusión (elaboración propia).

La Especificidad:

También conocida como la Tasa de Verdaderos Negativos, se trata de los casos negativos que el algoritmo ha clasificado correctamente. Expresa cuan bien puede el modelo detectar esa clase (Kohavi & Provost,1998).

Se calcula como:

$$VN/(VN+FP)$$

Ecuación 60: Métrica de desempeño especificidad para matriz de confusión (elaboración propia).

CAPÍTULO 3: METODOLOGÍA

La metodología propuesta para tratar el presente problema consiste en el análisis de un caso de estudio a una base de datos, con datos para una variable continua, que en este caso es la ley mineral y otra para una variable nominal, que sería el tipo de roca. La base de datos se analizará mediante el uso de paquetes del software Orange Canvas y el software SGeMS.

El procedimiento que se efectúa con la base de datos se contempla en los siguientes pasos a seguir:

- Realizar una descripción en detalle de cada variable de la base de datos.
- Realizar un análisis descriptivo del comportamiento estadístico de la base de datos.
- Estimar las variables para el caso de variable nominal haciendo uso de algoritmos de los métodos tradicionales de la geoestadística.
- Estimar las variables para caso de variable nominal mediante las técnicas de aprendizaje de máquinas.
- Validar los modelos para el conjunto de datos de los métodos aplicados mediante geoestadística.
- Validar los modelos para el conjunto de datos de las técnicas de aprendizaje de máquinas.
- Realizar un análisis descriptivo para las predicciones obtenidas tanto para los métodos tradicionales de la geoestadística y así como también para las técnicas de aprendizaje de máquinas.
- Comparar estadísticamente para ambos métodos cuales obtienen mejores resultados.
- Se finaliza con un análisis crítico de los resultados para los métodos propuestos y se discute sobre posibles mejoras y trabajos futuros.

CAPÍTULO 4: DESARROLLO

4.1 Descripción de la base de datos

Para mantener la confidencialidad de la mina en estudio es que no se dará a conocer el nombre de la empresa de donde provienen los datos, pero si se puede dar a conocer que es una *junior* que tiene proyectos en la zona centro sur de Chile y los datos corresponden a un proyecto de exploración de un depósito epitermal de oro - plata- cobre ubicado en la Cordillera de la Costa.

Ahora bien, con respecto a la base de datos, consiste en 4640 muestras de sondajes que provienen de un yacimiento de oro- plata- cobre.

Además, se cuenta con información de la litología. A continuación, se nombran los 7 tipos de litologías que se utilizarán:

- HBX: Brecha Hidrotermal.
- S2: Zona de Falla 2.
- MQV: Veta de Cuarzo Masiva.
- AND: Andesita.
- STW: Stockwork.
- S1: Zona de Falla 1.
- TBX: Brecha Tobácea.

4.2 Análisis exploratorio de datos

La base de datos con la que se cuenta fue entregada sin un mayor análisis, por lo tanto se hace necesario estudiar su comportamiento estadístico, con el fin de averiguar si existen datos sin registros o datos extremos que pueden afectar las estadísticas posteriores.

De los 4640 datos que se tienen al proceder a realizar las estadísticas básicas, nos encontramos que no es posible de realizar, debido a que existen 145 datos con un valor asignado de -999, esto significa que no se posee información de ese dato y se le marco con ese número, dichos valores no se pueden dejar ya que afectan las estadísticas de las leyes, puesto que no pueden dar leyes con valores negativos, por lo tanto se determina eliminar dichos valores.

Luego de la eliminación de datos sin un registro correcto, se cuenta ahora con 4495 datos, se procede a realizar las estadísticas básicas de dichos datos, con el fin de encontrar posibles valores atípicos o extremos.

La siguiente tabla y gráficos representan una aproximación inicial a las estadísticas básicas para la ley de Cobre. Para el caso de ley de Plata y Oro, se encuentran en el apéndice A.

Ley de Cobre (%)	
Media	0.24
Error Típico	0.01
Mediana	0.02
Moda	0.01
Desviación estándar	0.60
Varianza de la muestra	0.36
Curtosis	256.68
Coeficiente de asimetría	11.12
Rango	19
Mínimo	0.00
Máximo	19
Suma	1057.2
Datos	4495

Tabla 2: Estadísticas básicas ley de Cobre (Elaboración propia)

Se puede apreciar en la tabla 2 que la ley media de cobre es de 0.24 en un comienzo con esta cantidad de datos, también se puede apreciar el mínimo y máximo de los datos, en donde llama la atención el valor máximo de 19% de ley de cobre. Por lo anterior, es que se procede a realizar un histograma y un Boxplot para ver la distribución de los datos.

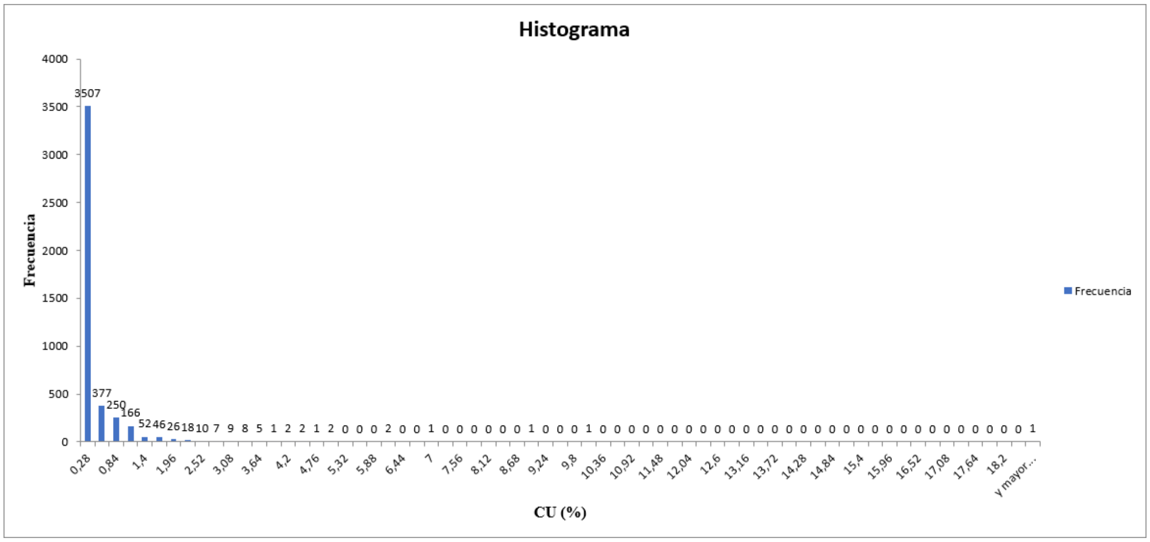

Ilustración 17: Histograma Ley de Cobre (Elaboración propia)

Se puede ver en el histograma del cobre, la frecuencia con la que se distribuyen las leyes, en donde se nota un particular agrupamiento de los datos en leyes por debajo del 1%, en donde la mayor cantidad de los datos se encuentra en el intervalo entre 0 y 0.28, también se puede notar que existen leyes medianamente altas y algunas con valores un tanto atípicos con una frecuencia de 1 o 2 valores con dichas leyes. Por lo tanto, se procede a realizar un Boxplot para tener un análisis con mayores detalles.

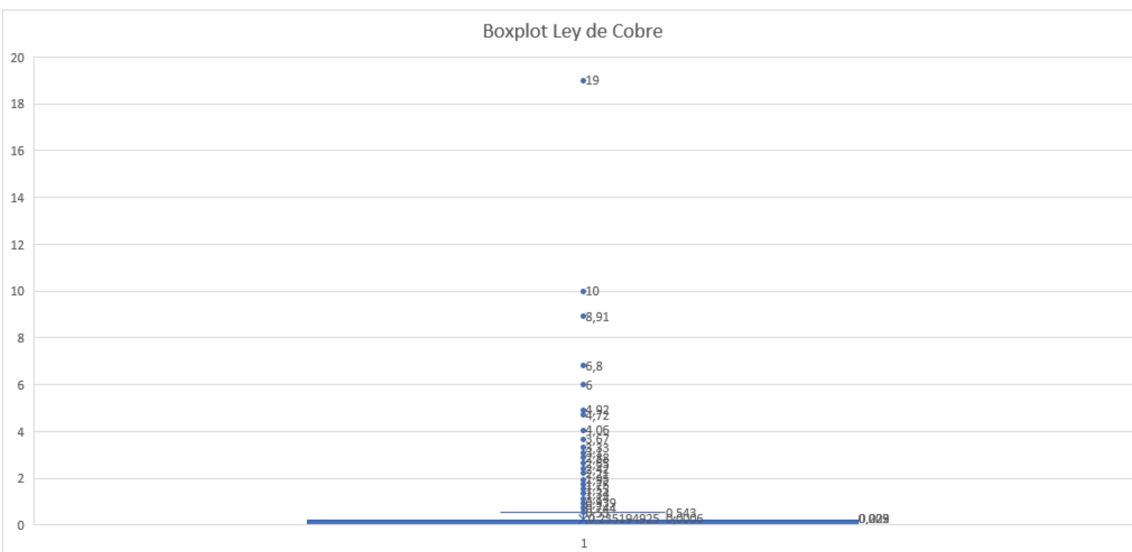

Ilustración 18: Boxplot Ley de Cobre (Elaboración propia)

En la ilustración 18 se aprecia un gran nivel de agrupamiento de los datos en leyes más bajas, por lo que se hace complejo graficar de la forma más correcta, pero si se logra identificar los posibles valores atípicos, tal como el 19% y entre otros.

Ahora bien, con el análisis preliminar realizado es que se procede a eliminar valores que sean atípicos de las muestras. Se elige el cobre como representante de los 3 tipos de leyes por tener valores menos aberrantes y además como se explicará en un futuro en la presente memoria si se consideran los 3 tipos de leyes para trabajar en las estimaciones se tendrían demasiadas variables y por lo tanto entorpeciendo las estimaciones. Se considera como criterio, que todas aquellas leyes mayores a 4, serán eliminadas ya que representan una muestra muy pequeña del total y lo único que hacen es afectar las estadísticas básicas de la ley de cobre.

Luego de la eliminación de los datos atípicos, la base de datos queda con 4480 valores, es decir, se eliminaron 15 valores extremos.

En la tabla 3 se aprecia las estadísticas básicas de la ley de cobre luego de la eliminación de datos atípicos:

Ley de Cobre (%)	
Media	0.22
Error Típico	0.01
Mediana	0.02
Moda	0.01
Desviación estándar	0.43
Varianza de la muestra	0.19
Curtosis	15.60
Coeficiente de asimetría	3.48
Rango	3.67
Mínimo	0.00
Máximo	3.67
Suma	967.44
Datos	4480

Tabla 3: Estadísticas básicas ley de Cobre sin datos atípicos (Elaboración propia)

Antes de la eliminación de los datos atípicos se tenía una ley media de cobre de 0.24 % y un valor máximo de 19%, posterior a la eliminación de datos atípicos la base de datos queda con una ley media de 0.22% y un valor máximo de 3.67%, es evidente la reducción de la ley media y sobre todo la reducción en el valor máximo, lo que hacía que afectara las estadísticas básicas de la ley de cobre.

4.3 Compositación de Sondajes

Ahora bien, se hace necesario aplicar el procedimiento de compositación para la base de datos, con el fin de obtener leyes más representativas de las muestras. Además, la base de datos cuenta con 4 archivos, que son:

- Assay: contiene las leyes de las muestras.
- Collar: contiene las coordenadas de las muestras en las 3 direcciones norte, este y elevación.
- Lito: contiene la litología de las muestras.
- Survey: contiene el ángulo de azimuth y el de dip de las muestras

Hasta ahora se ha trabajo con el Assay, es decir, las leyes de las muestras, pero en esta memoria el objetivo es trabajar con variables categóricas, en nuestro caso la litología. Por lo tanto es que se deben agrupar todos los archivos en uno solo que contenga tanto la ley de cobre con sus respectivas coordenadas, así como también contenga el tipo de litología en cada coordenada de la muestra de cobre y sus respectivos ángulos de azimuth y dip.

Por lo anterior es que se procede a trabajar con el software minero Maptek Vulcan, ya que es capaz de agrupar todos los archivos en uno solo, para luego poder hacer la Compositación que es lo que se busca lograr.

En la ilustración 19 se aprecia la distribución de los sondajes en el software Vulcan, y en la ilustración 20 se puede ver con mayor detalle la leyenda con el rango de valores para las leyes de cobre. Cabe destacar que gran parte de las leyes se encuentra en valores bajo los 0.3 %, en este caso representadas por el color rojo.

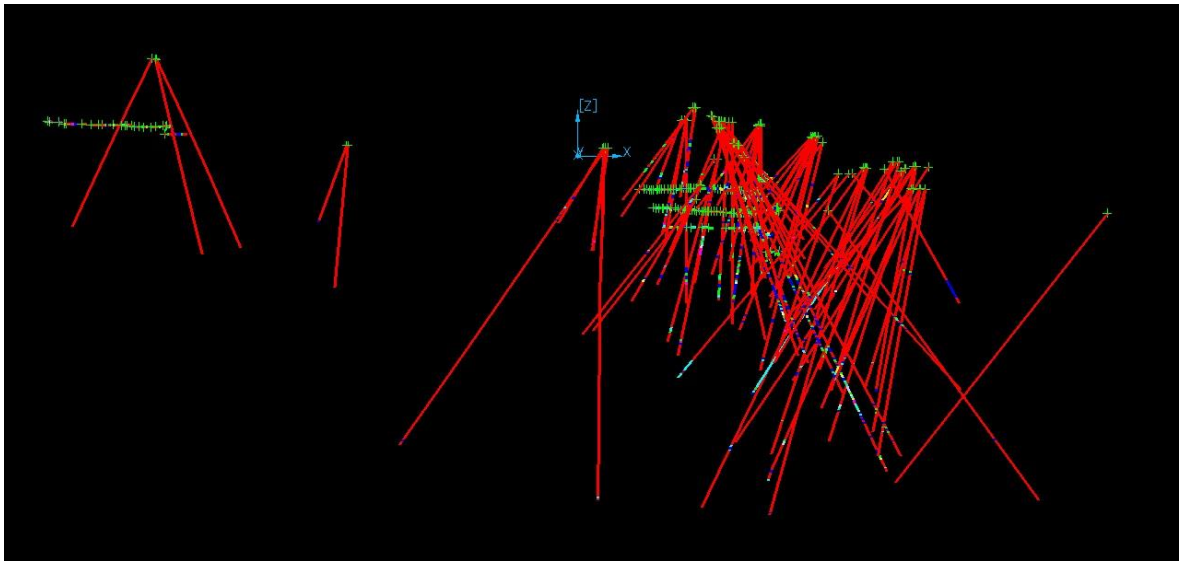

Ilustración 19: Sondajes leyes de Cobre (Vulcan)

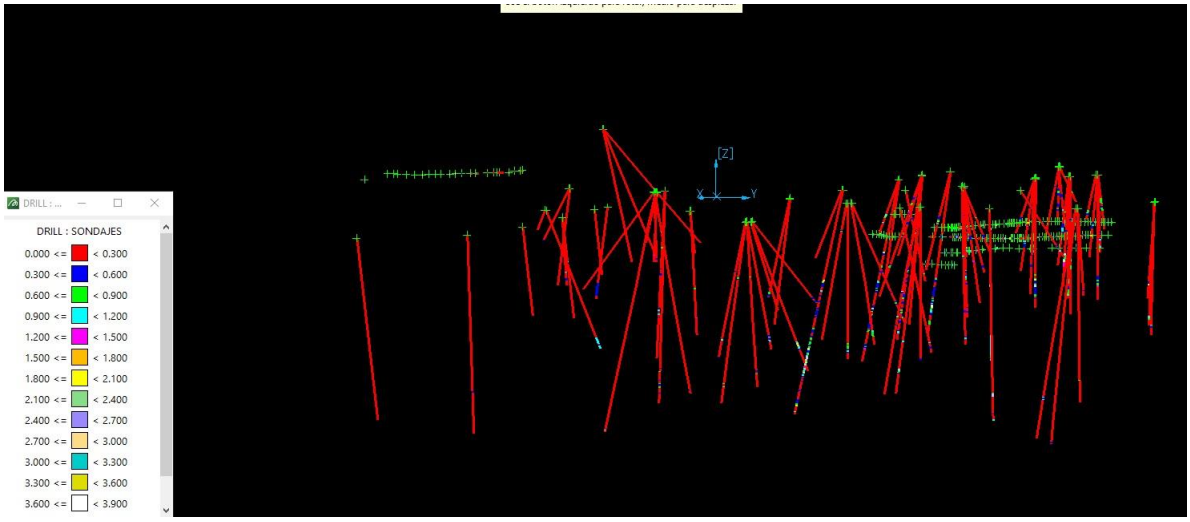

Ilustración 20: Sondajes con leyenda de las leyes de Cobre (Vulcan)

En la ilustracion 21 esta el resultado de la compositacion de los sondajes,en donde se buscaba tener una mejor representacion de los datos.

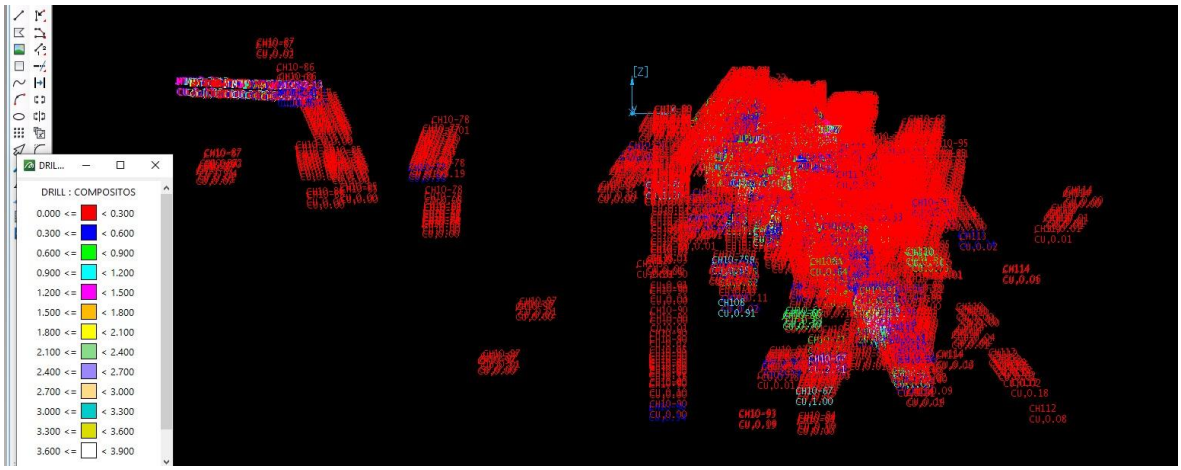

Ilustración 21: Compositación de sondajes con leyenda de las leyes de Cobre (Vulcan)

4.4 Nueva base de datos Post Compositación de sondajes

La nueva base de datos posterior al procedimiento de agrupamiento de los 4 archivos y compositación elaborado en Maptek Vulcan, es de 2832 datos, ahora ya con sus respectivas coordenadas, leyes de cobre y litología para cada muestra. En la tabla 4 se puede ver las estadísticas básicas luego de realizado todo lo anterior.

Ley de Cobre (%)	
Media	0.20
Error típico	0.01
Mediana	0.03
Moda	0.01
Desviación estándar	0.38
Varianza de la muestra	0.15
Curtosis	13.67
Coeficiente de asimetría	3.21
Rango	3.67
Mínimo	0
Máximo	3.67
Suma	575.48
Datos	2832

Tabla 4: Estadísticas básicas ley de Cobre compositos (Elaboración propia)

Ya con los 2832 datos con los que se cuenta, se continua buscando posibles valores faltantes o anómalos, entre los que se encuentra que existen 671 valores que no tienen asociado un valor de tipo de litología y por lo tanto se eliminan, ya que la presente memoria tiene como objetivo trabajar con valores categóricos, y el tener coordenadas y leyes sin litologías asociadas es un problema, ya que serían múltiples valores faltantes a la hora de hacer las estimaciones por Kriging de indicadores, así como también para la parte de machine learning.

La siguiente tabla da cuenta de cómo queda la base de datos para las litologías:

Litología	Datos	Ley media Cu (%)
AND	1729	0.12
S2	145	0.05
HBX	86	0.25
STW	81	0.1
TBX	67	0.05
MQV	41	0.21
S1	12	0.04
Total	2161	0.12

Tabla 5: cantidad de datos para litologías agrupadas (elaboración propia).

La ilustración 22 muestra la cantidad de datos medida en porcentaje para cada tipo de litologías, en donde destaca la gran cantidad de datos asociados a la litología AND.

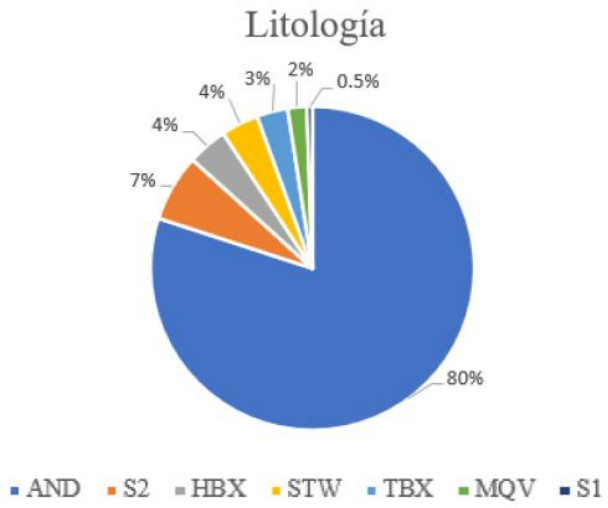

Ilustración 22: cantidad porcentual de datos para las litologías (elaboración propia).

Ahora bien, se toma la determinación de agrupar 4 tipos de litologías, las cuales son, STW, TBX, MQV, S1.Debido a lo anterior, es que esta decisión se toma ya que en los capítulos siguientes se tendrán las estimaciones realizadas para los distintos tipos de litologías, y el hecho de tener tantos tipos de litologías para la estimación por parte de Kriging de indicadores es un problema que pasa por costo de tiempo, debido a que para estimar el Kriging de indicadores se debe realizar para cada tipo de litología, lo cual implica tener más variogramas, y por lo tanto induce a tener más posibles errores, según las características del variograma experimental y modelado.

La tabla 6 da a conocer la nueva cantidad de datos con 4 litologías agrupadas:

Litología	
Tipo	Cantidad
AND	1729
S2	145
HBX	86
Mixto	201
Total	2161

Tabla 6: cantidad de datos para litologías agrupadas (elaboración propia).

Se puede notar en la tabla 6, que al agrupar las 4 litologías STW, TBX, MQV y S1, la sumatoria de sus datos, pasa a ser 201 valores con litologías, de un total de 2161 valores para toda la base de datos que contiene las distintas litologías asociadas.

Antes de trabajar con la base de datos en el software SGeMS, se debe realizar una codificación de los datos de litología, puesto que fueron entregados como un valor en texto, y además tanto para SGeMS como la parte de machine learning, los algoritmos no leen texto para sus cálculos, sino que leen valores numéricos ya sean continuos o discretos, por lo tanto se debe realizar de forma externa antes de entrar al software.

Se utilizará el método One Hot Encoding, para toda la base de datos de excel, que luego será usada tanto para la parte geoestadística como la de machine learning.

En la tabla 7 se da un ejemplo de los primeros valores de las columnas de la base de datos con la aplicación del método One Hot Encoding:

X	Y	Z	AND	S2	HBX	Mixto
240722.39	6081953.4	141.795	0	0	1	0
240721.4	6081952.6	140.286	0	0	1	0
240720.4	6081951.7	138.776	0	1	0	0
240719.41	6081950.8	137.267	0	1	0	0
240718.42	6081950	135.758	0	0	0	1
240716.84	6081948.6	133.343	0	0	0	1
240715.85	6081947.7	131.833	1	0	0	0
240714.86	6081946.9	130.324	1	0	0	0
240713.87	6081946	128.814	1	0	0	0
240712.88	6081945.2	127.305	1	0	0	0

Tabla 7: Ejemplo de método One Hot Encoding a la base de datos (Elaboración propia)

A modo de una breve explicación, el método One Hot Encoding, lo que hace es transformar el valor de una variable nominal a una variable numérica, es decir, si se toma la variable litológica AND, el método lo que hará es en donde exista la presencia de la variable AND, le asignará un valor número 1 y si no tiene presencia le asignará un 0,y este proceso se lleva a cabo para todas las litologías asociadas a la base de datos en excel, así luego será posible ingresar dichos valores a los programas SGeMS y Orange Canvas.

Una vez aplicado el método a toda la base de datos, se procede a separarla de forma aleatoria en conjunto de entrenamiento y prueba, donde se usó como criterio un 80% para entrenamiento y el 20% restante para el conjunto de prueba. Para el conjunto de entrenamiento se tiene 1729 datos y para el conjunto de prueba el restante del total que seria 432 datos.

Para que el método de geoestadística que se usará en SGeMS este en iguales condiciones con respecto a los métodos de machine learning, es que se decide utilizar el 100% de los datos de entrenamiento y prueba de los métodos de machine learning, y ellos compararlos con los utilizados en SGeMS, con la finalidad de tener la misma cantidad de datos para ambos métodos de estimación.

4.5 Elección de los métodos de estimación

Ya con la base de datos limpiada y sabiendo su distribución de datos a trabajar, es que se hace necesario elegir con que métodos se realizaran los cálculos para lograr obtener estimaciones. El estimador representante para la parte de geoestadística será el Kriging de indicadores, debido a como se explicó en el marco teórico, es un buen estimador para variables nominales. Para la parte de aprendizaje de máquinas o machine learning el estimador representante serán dos, por una parte serán las redes neuronales artificiales y por otra parte se utilizará la regresión logística: Los métodos anteriores se eligieron por ser buenos estimadores para variables dicotómicas, es decir, para valores 1 o 0, como es el caso de la base de datos que se tiene.

4.6 Despliegue de los sondajes en SGeMS

Ahora bien, ya comenzado con el trabajo a realizar, se debe contar con los mínimos, máximos y rangos para las coordenadas, con el objetivo de ingresarlas al programa SGeMS.

En la tabla 8 se puede apreciar los mínimos, máximos y rangos para las 3 coordenadas de los ejes.

	Coordenada Este (m)	Coordenada Norte (m)	Cota (m)
Mínimo	240395.337	6081396.09	-72.68
Máximo	240845.37	6082042.88	159.13
Rango	450.0	646.8	231.8

Tabla 8: Mínimos, máximos y rango según coordenada (elaboración propia).

De acuerdo a lo que se observa en la tabla 8, en el eje para las coordenadas Este o coordenada en el eje X, el valor mínimo es de aproximadamente 240395 y el valor máximo 240845 ambos medidos en metros, lo que si se calcula su rango, que no es más que la diferencia entre ellos, resulta un valor de 450 metros para la coordenada Este, para el caso de la coordenada Norte se tiene un mínimo de 6081396 y un máximo de 6082042 aproximadamente, y su rango es de 646 metros, finalmente para la Cota se tiene un mínimo de -72 y un máximo de 159 aproximadamente, y con un rango de 231 metros. Por lo anterior, es que se decide agregar 50 metros a cada coordenada, para poder abarcar más espacio y no queden posibles valores fuera de la grilla.

En el programa SGeMS, se debe además ingresar las dimensiones de los bloques y la cantidad de bloques por cada coordenada de los ejes, por lo cual se debe asignar el tamaño de bloque que se quiere trabajar según coordenada, se elige un valor de 2.5 metros para las 3 dimensiones del espacio, es decir, para Este, Norte y la Cota. La cantidad de bloques se obtiene calculando la división del rango entre la dimensión del bloque, en la tabla 9 se puede ver la cantidad de bloques según coordenada, en donde se tiene 200,280 y 120 metros según coordenada Este, Norte y cota respectivamente.

Dimensión bloque coordenada Este (m)	2.5
Dimensión bloque coordenada Norte (m)	2.5
Dimensión bloque Cota (m)	2.5
Cantidad de bloques coordenada Este	200
Cantidad de bloques coordenada Norte	280
Cantidad de bloques Cota	120

Tabla 9: Dimensiones y cantidad de bloques según coordenada (elaboración propia).

En las ilustraciones 23 y 24 se puede apreciar el despliegue de los sondajes según los datos ingresados en las tablas anteriores. Cabe destacar que dichos sondajes son con datos de litologías, a diferencia de lo que se vieron en un principio en el programa Maptek Vulcan, los cuales eran sondajes desplegados para leyes de cobre.

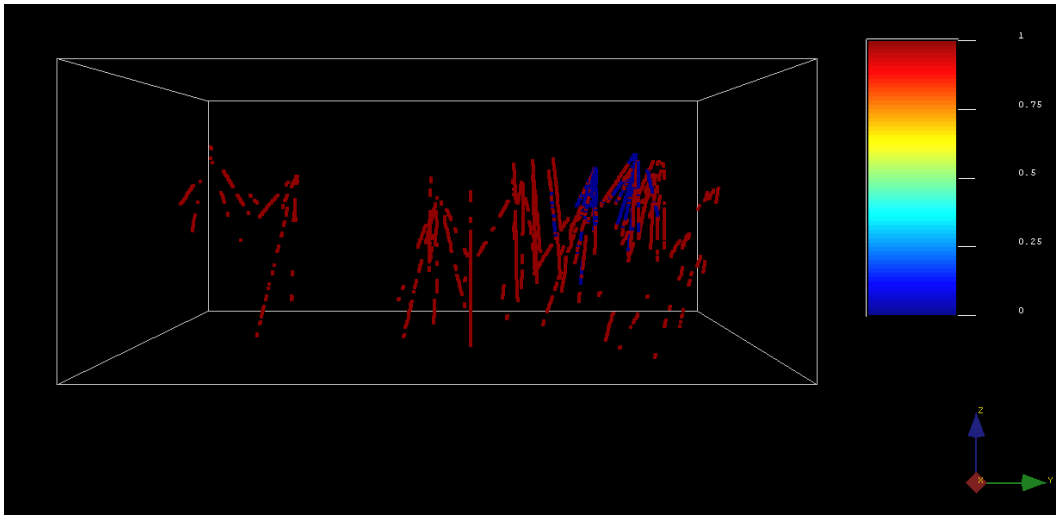

Ilustración 23: Sondajes desplegados junto con bloque de estimación para litología AND (elaborado en SGeMS).

Como se puede ver en la ilustración 23, los sondajes son desplegados para la litología AND, en donde se puede notar que se asigna un color azul para el valor 0 y un color rojo para el valor 1, esto significa que el valor 0 se da cuando no existe presencia de litología AND, en cambio, el valor 1 se da cuando existe presencia de la litología AND, y por lo tanto se da un color rojo. Cabe destacar como se vio en la base de datos, existe una alta cantidad de datos para la litología AND, por lo mismo en la ilustración 23 se nota una gran cantidad de sondajes con la presencia de dicha litología.

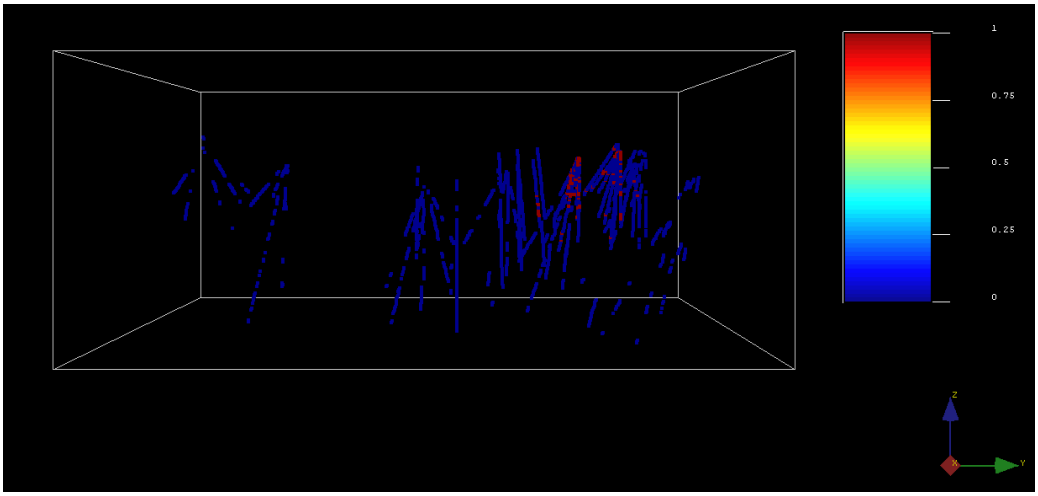

Ilustración 24: Sondajes desplegados junto con bloque de estimación para litología S2 (elaborado en SGeMS).

En la ilustración 24 se puede ver en color rojo un pequeño grupo de sondajes de color rojo, esto quiere decir, que se está en presencia de la litología S2 y todo los demás en color azul es para los otros tipos de litologías.

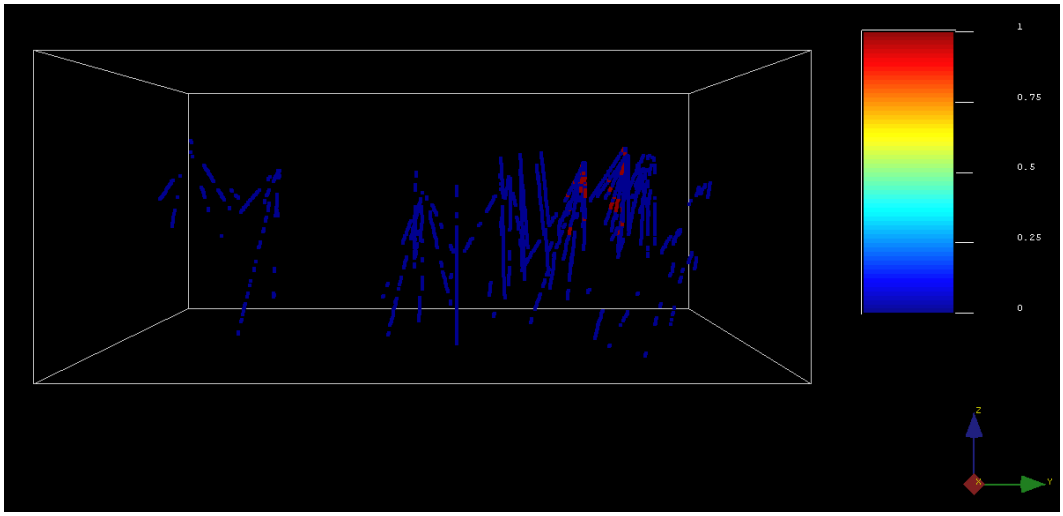

Ilustración 25: Sondajes desplegados junto con bloque de estimación para litología HBX (elaborado en SGeMS).

Para la ilustración 25 sucede algo parecido con respecto a la ilustración 24, ya que se nota en color rojo un grupo de sondajes en rojo que representan la presencia de la litología HBX, y el resto de los sondajes en color azul, representa la no presencia de la litología HBX.

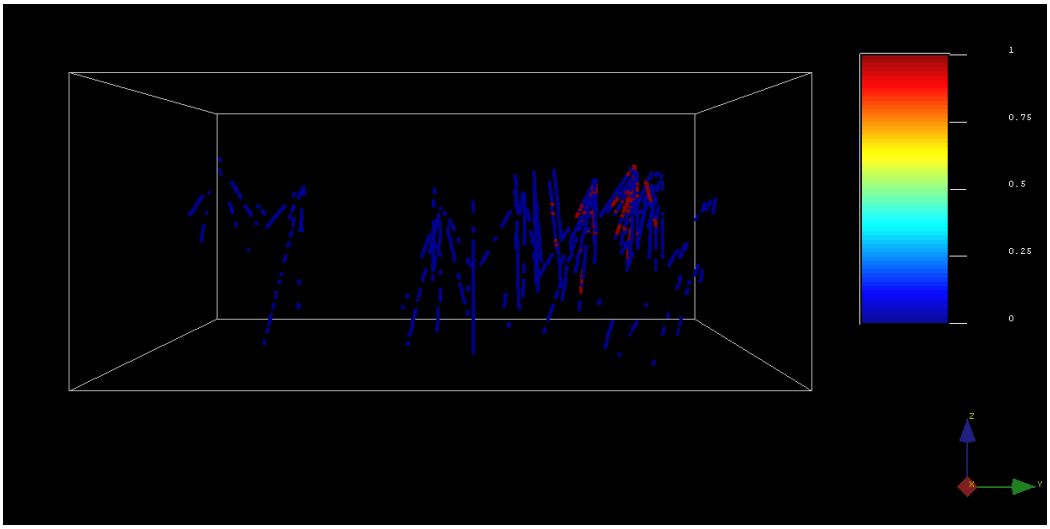

Ilustración 26: Sondajes desplegados junto con bloque de estimación para litología Mixto (elaborado en SGeMS).

Finalmente para la ilustración 26, se tiene el despliegue de sondajes para la litología Mixto, es decir, la litología conformada por un grupo de litologías, las cuales son STW, TBX, MQV y S1. Se puede ver en color rojo la presencia de la litología Mixto y en color azul el resto de las litologías que no tienen presencia, es decir, las litologías AND, S2 y HBX.

4.7 Variogramas

Para lograr analizar el comportamiento espacial de las variables en estudio se necesitan los variogramas experimentales y su correspondiente variograma modelado, por lo cual, se definieron ciertos parámetros para poder crear los variogramas experimentales, tales como, el número de pasos, separación del paso, tolerancia de paso, números de direcciones, azimuth, dip, tolerancia y ancho de banda.

Debido a la naturaleza de los datos irregularmente espaciados se pueden dar distintos tipos de fenómenos:

- Suceder que no existan valores de la variable a la distancia h.
- Suceder que no existan valores de la variable en una respectiva dirección.

En las siguientes tablas 10 y 11 se puede ver los parámetros con sus respectivos valores, en donde destacar la separación del paso de 100 metros y la tolerancia del paso que se obtiene como la mitad de la separación del paso, con un valor de 50 metros, se eligieron dichos valores para poder abarcar la mayor cantidad de datos de la base de datos de las litologías. Además, se realizó un análisis para 4 direcciones, las cuales son,0,45,90 y 135 grados. Cabe destacar que el programa SGeMS, entrega también el variograma experimental omnidireccional, el cual representa la suma de todas las direcciones estudiadas del variograma experimental.

Cantidad de pasos	Separación de paso (m)	Tolerancia de paso (m)
20	100	50

Tabla 10: Parámetros del paso de Variogramas experimentales (elaboración propia).

Azimuth (°)	Dip (°)	Tolerancia (°)	Ancho de banda
0	0	22.5	500000
45	0	22.5	500000
90	0	22.5	500000
135	0	22.5	500000

Tabla 11: Parámetros de las direcciones de los Variogramas experimentales (elaboración propia).

A continuación se presentan los Variogramas experimentales para la litología de tipo AND, en el caso de los otros tipos de litología, se encuentran en el apéndice B.

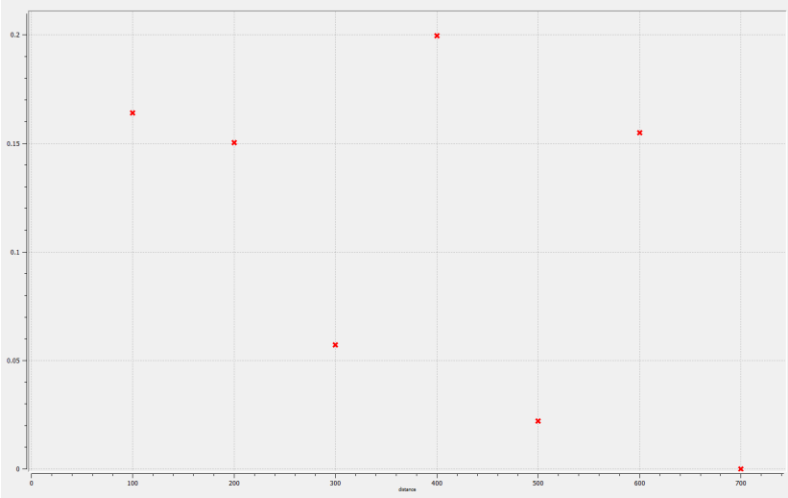

Ilustración 27: Variograma experimental 0° para litología AND (elaborado en SGeMS).

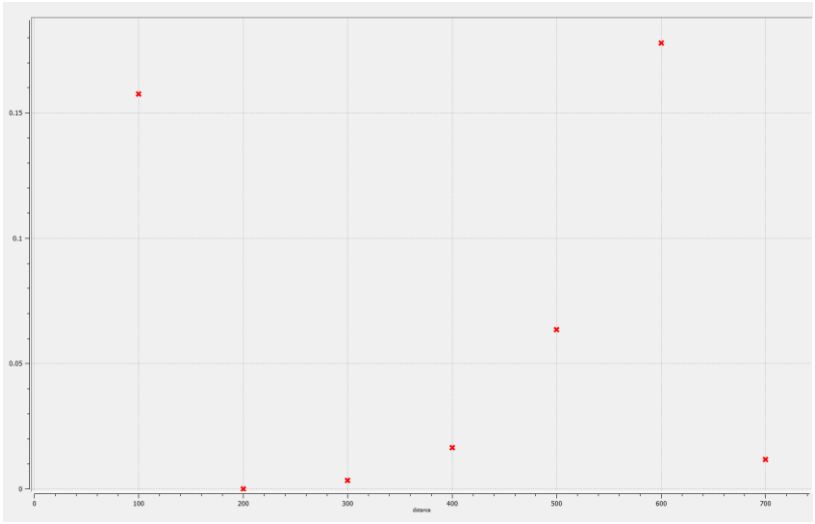

Ilustración 28: Variograma experimental 45° para litología AND (elaborado en SGeMS).

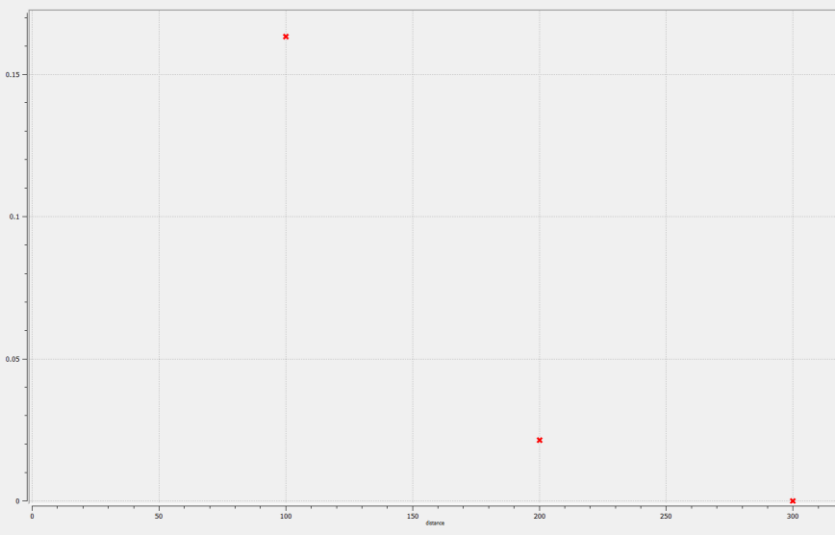

Ilustración 29: Variograma experimental 90° para litología AND (elaborado en SGeMS).

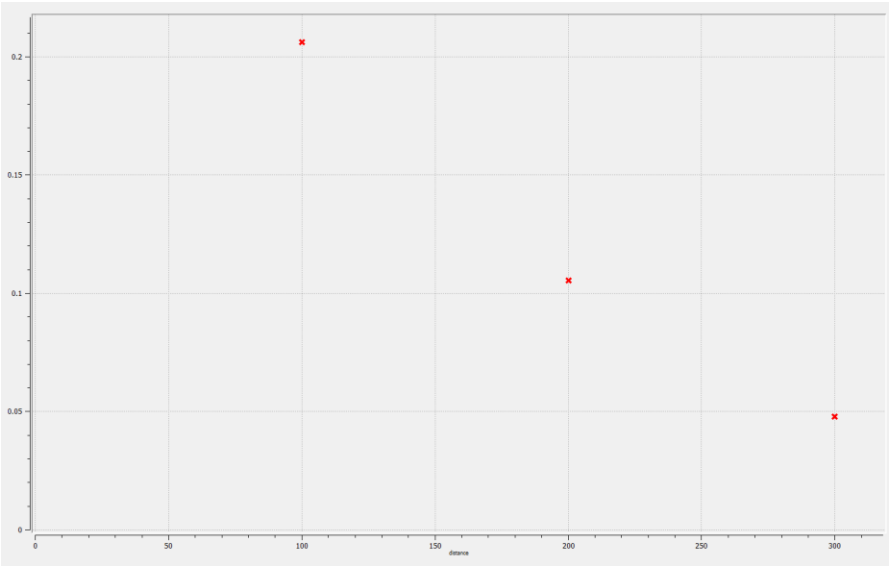

Ilustración 30: Variograma experimental 135° para litología AND (elaborado en SGeMS).

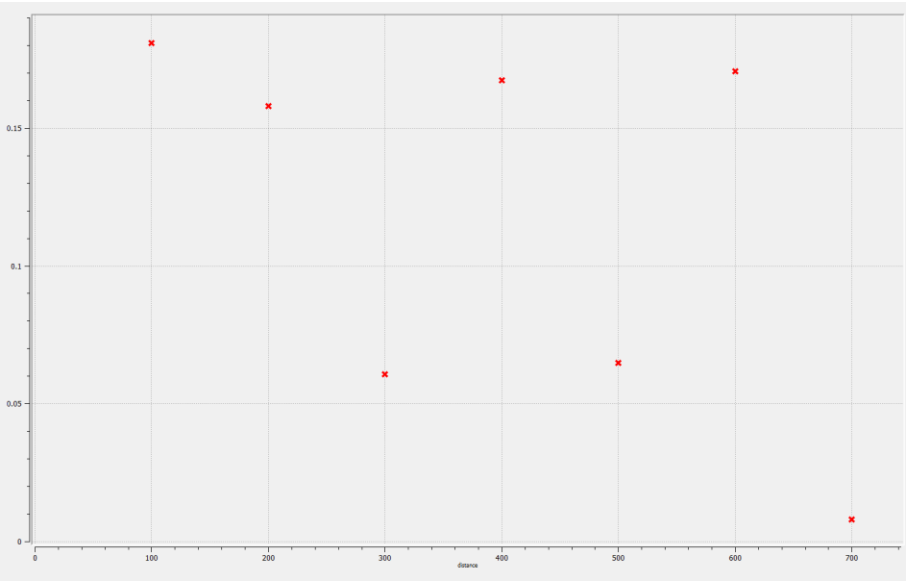

Ilustración 31: Variograma experimental omnidireccional para litología AND (elaborado en SGeMS).

Según se observa en los distintos variogramas experimentales anteriores, se puede observar para los ángulos de 0,45,90 y 135 grados, que existe una amplia movilidad del valor gamma producto de la naturaleza de los datos, ya que lo que se tiene son datos con ceros y unos ,por lo tanto es complejo elegir uno de estos variogramas experimentales para convertirlo en un variograma modelado, es así, que se toma la decisión de utilizar el variograma experimental omnidireccional, ya que muestra una correlación más pareja entre los datos y además representa la suma de los demás variogramas experimentales según cada ángulo.

En la Ilustración 32 se puede apreciar como queda el variograma modelado:

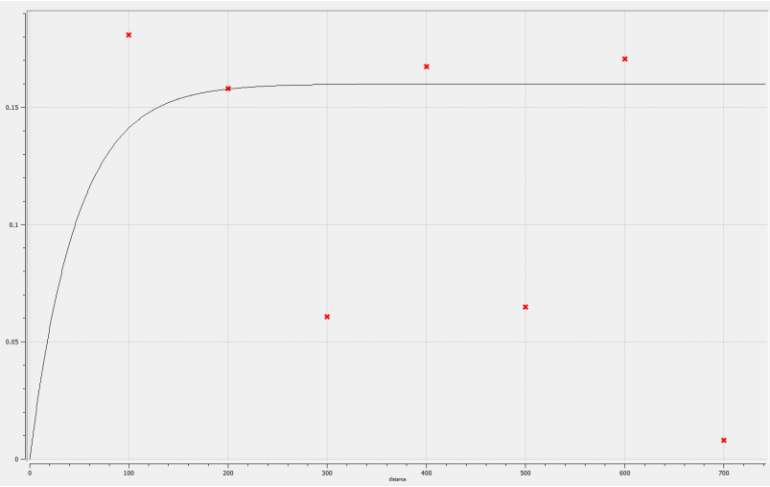

Ilustración 32: Variograma modelado omnidireccional para litología AND (elaborado en SGeMS).

Para el variograma modelado omnidireccional de la litología AND, se ajustaron sus rangos, en donde el rango menor se ajustó con un valor de 40, el rango medio con 100 y el rango mayor con un valor de 140.Además, se ajustó con un sill de 0.16.

4.8 Métodos de Machine Learning

En el presente apartado se verán los dos métodos elegidos para realizar las estimaciones de las litologías para los métodos de aprendizaje de máquinas o machine learning. Primeramente se tiene el método de las redes neuronales artificiales y luego en segunda parte se tiene el método de regresión logística. Cabe destacar que ambos métodos se llevaron a cabo en el programa de uso libre Orange Canvas en su tercera versión.

4.8.1. Redes neuronales artificiales

Para lograr desarrollar el modelo el programa Orange Canvas, necesita parámetros propios de las redes neuronales, los cuales deben ser ingresados para poder llevar a cabo las estimaciones, estos parámetros son número de neuronas por capa oculta, cantidad de capas ocultas, máximo de iteraciones, función de activación, optimizador, tasa de aprendizaje, los cuales se pueden apreciar con mayor detalle en la tabla 12.

Neuronas por capa oculta	25
Capas ocultas	5
Máximo de iteraciones	1000
Función de activación	ReLu
Optimizador	Adam
Tasa de aprendizaje	0.1

Tabla 12: Parámetros redes neuronales artificiales (elaboración propia).

Se escoge la función ReLU, ya que es más sencilla y no posee regiones de saturación, como otras funciones tales como las funciones tangentes hiperbólica y sigmoidal, que provocan mayor demora durante el proceso de entrenamiento de los datos de la red neuronal artificial.

Cabe destacar, que se eligieron tanto valores altos como bajos en los parámetros, para ir entrenando la red neuronal artificial, con la finalidad de analizar la influencia en los resultados de las estimaciones.

En la ilustración 33 se puede apreciar el flujo de trabajo realizado en el programa Orange Canvas para las redes neuronales artificiales:

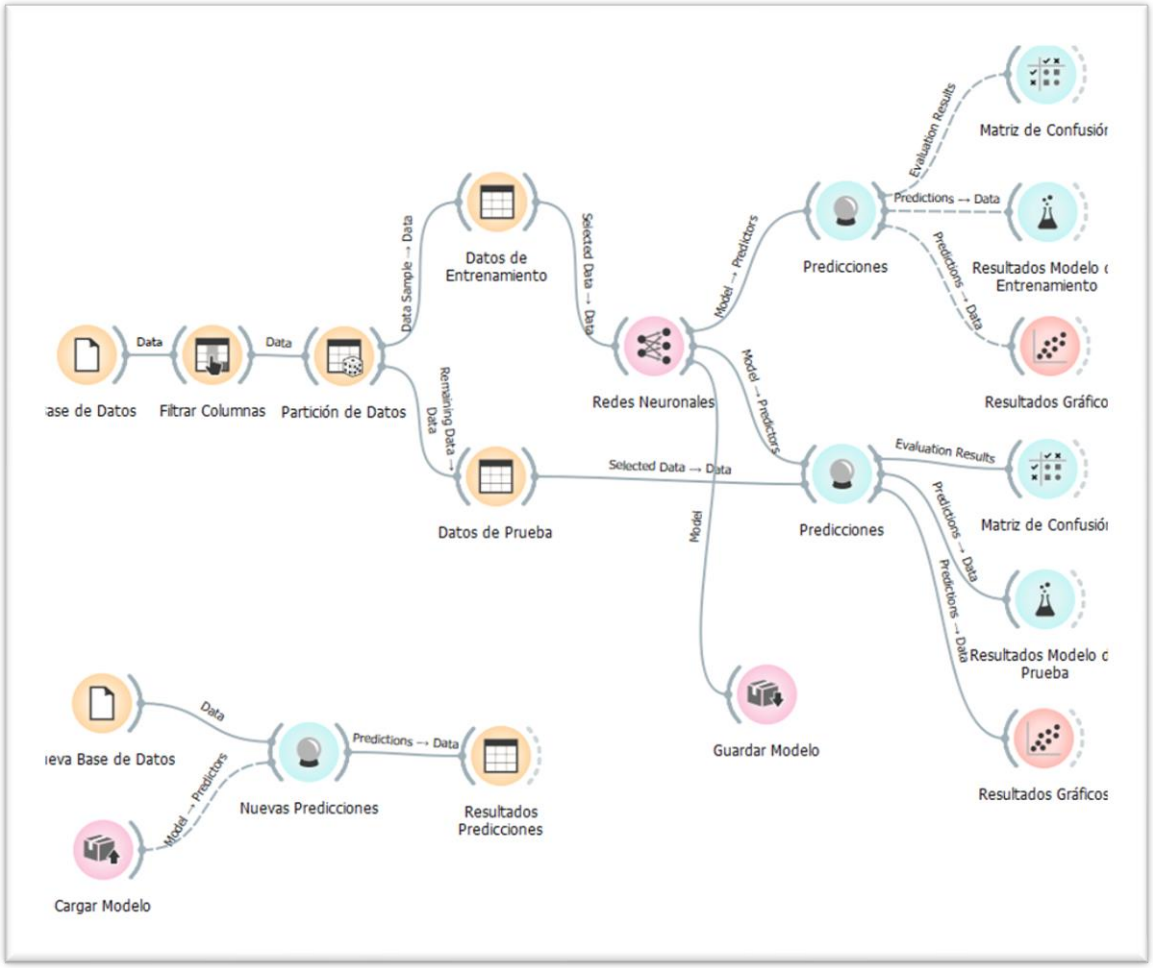

Ilustración 33: Flujo modelo de redes neuronales artificiales (elaborado en Orange Canvas).

4.8.2. Regresión logística

La regresión logística resulta ser un método mucho más sencillo y amigable con respecto a las redes neuronales, ya que necesita menos cantidades de parámetros y realiza los cálculos en menos cantidad de tiempo.

En el programa Orange Canvas se hace necesario 2 parámetros a ingresar para lograr realizar las estimaciones de la regresión logística. Se necesita el tipo de regularización y la complejidad C. En nuestro caso se escogió luego de varias iteraciones del modelo, una regularización de tipo Ridge y un valor de complejidad C igual a 0.180.

En la ilustración 34 se puede apreciar el flujo de trabajo realizado en el programa Orange Canvas para la regresión logística:

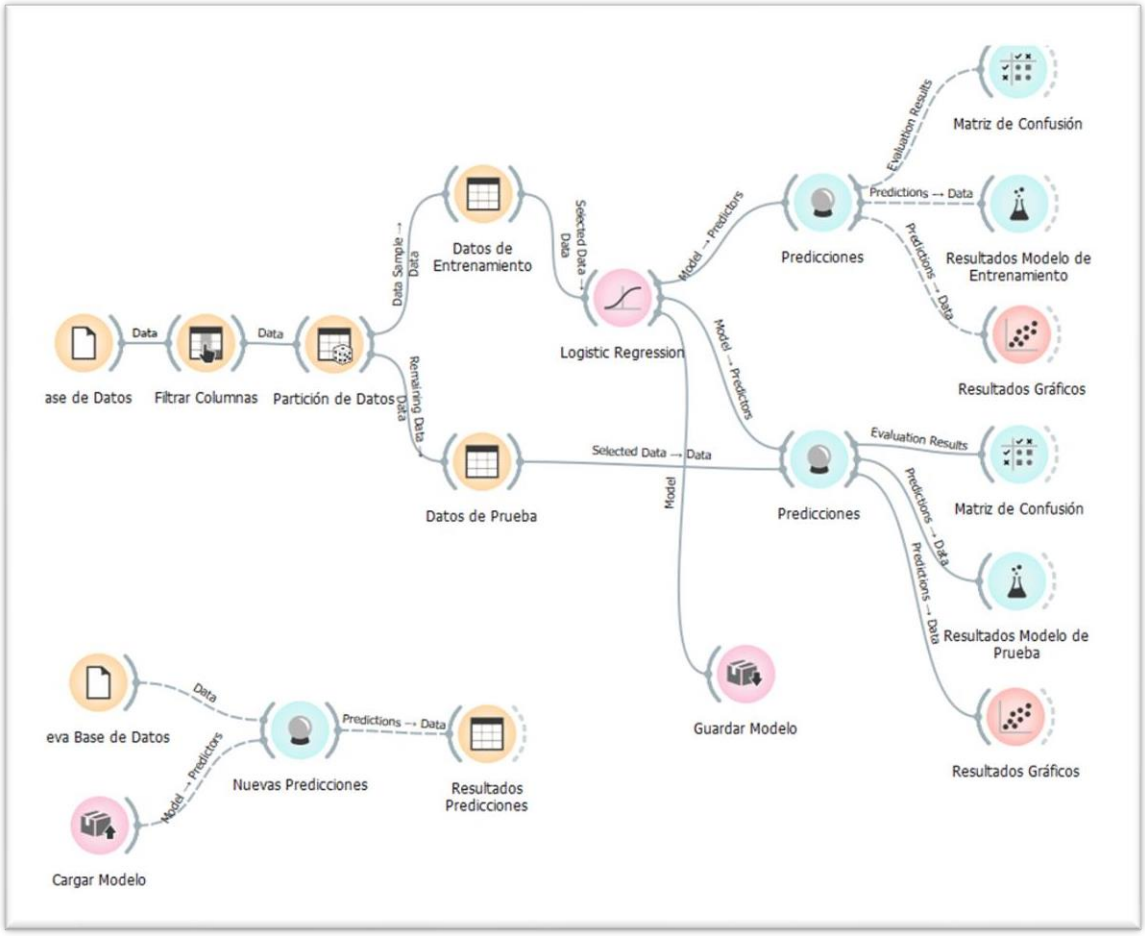

Ilustración 34: Flujo modelo de regresión logística (elaborado en Orange Canvas).

CAPÍTULO 5: RESULTADOS Y DISCUSIÓN

En el presente capitulo se muestran los resultados obtenidos para los 3 tipos de métodos de estimación, tanto como para el representante de la geoestadística, como también los dos representantes del aprendizaje de máquinas o machine learning. Además, de una posterior discusión y comparación de los 3 métodos.

5.1 Estimación Kriging de Indicadores

Para litología AND:

En la ilustración 35 se muestra el resultado de la estimación mediante el método geoestadístico Kriging de indicadores para la variable litológica AND.

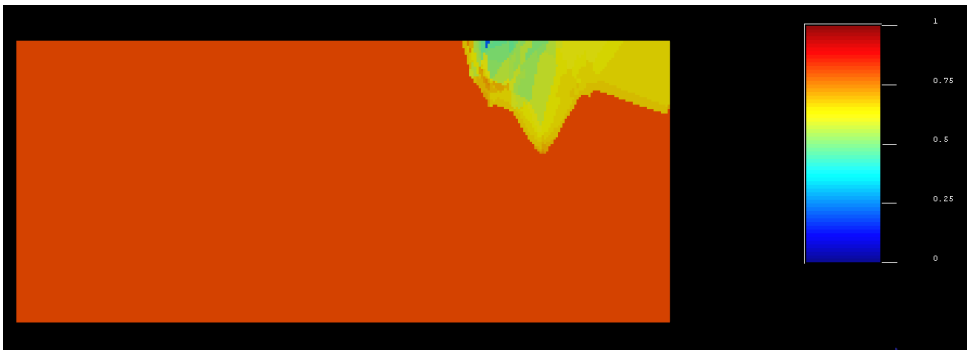

Ilustración 35: Vista de frente Kriging de indicadores para litología AND (elaborado en SGeMS).

Como se puede ver en la ilustración 35 según la leyenda de colores asignada, el valor 1 representa el color rojo y el valor 0 representa al color azul, de acuerdo con esto, podemos notar la alta presencia de un color bastante cercano al rojo, lo cual está correcto, ya que es la base de datos para la litología AND, y como se ha visto anteriormente es la que cuenta con la mayor cantidad de datos, y es esperable que su Kriging muestre una gran repartición de las estimaciones como litología AND.

También se puede comentar que se nota en el costado derecho de la ilustración 35, una variación del espectro de colores según la leyenda asociada, dichos colores representan la no presencia probable de litología AND, y por lo tanto la presencia de otros tipos de litologías, las cuales podrían ser de tipo S2, HBX o el grupo de litologías Mixto.

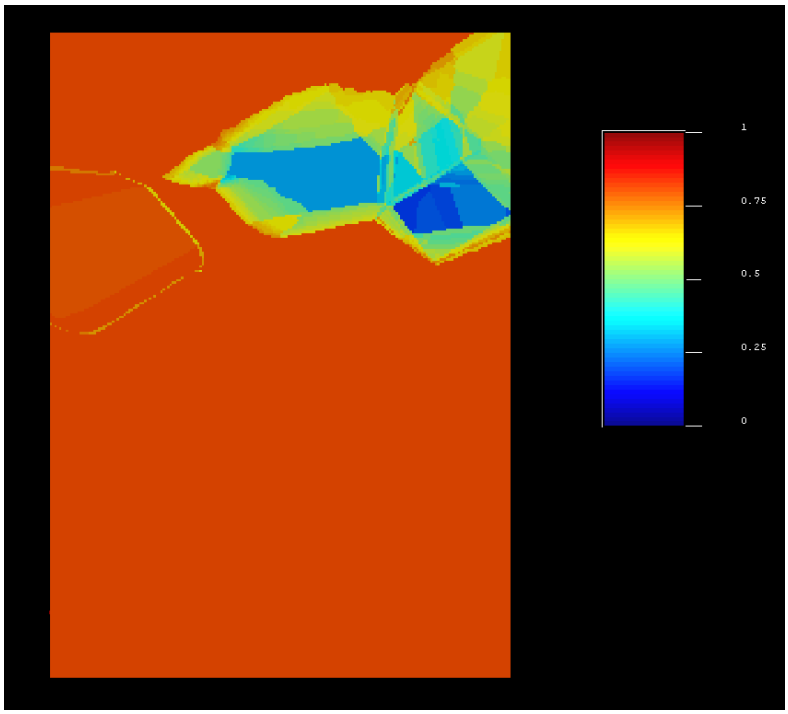

Ilustración 36: Vista en planta Kriging de indicadores para litología AND (elaborado en SGeMS).

En la ilustración 36 se tiene una vista en planta de la estimación mediante Kriging de indicadores para la litología AND. De igual manera de cómo se vio en la ilustración 35, se tiene una alta presencia de litología AND, y en su costado derecho se nota la presencia de otros tipos de litologías.

La ilustración 37 muestra la varianza del Kriging de indicadores para la litología AND.

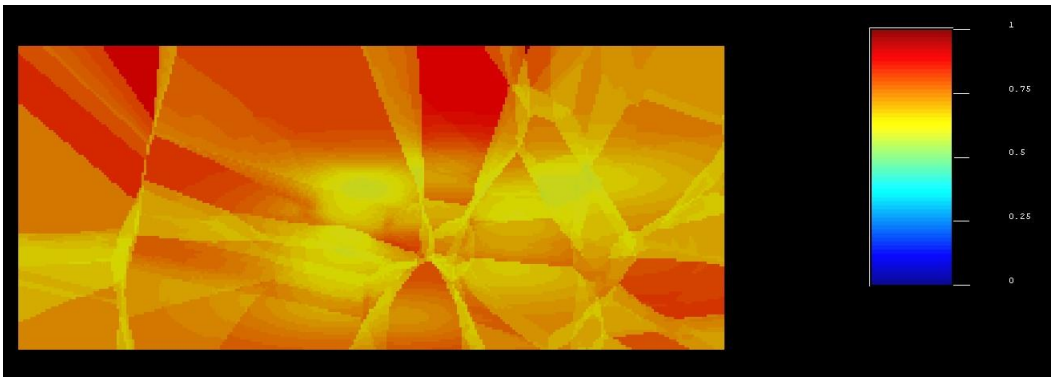

Ilustración 37: Varianza Kriging de indicadores para litología AND (elaborado en SGeMS).

Para litología S2:

En la ilustración 38 se puede apreciar el resultado de la estimación por Kriging de indicadores para la litología S2, cabe recordar que dicha litología es la segunda con más presencia de datos después de la litología AND. Como se puede ver el bloque resulta en su mayoría de color azul, es decir. se le asigno un valor 0 según la leyenda de colores, esto significa que todo aquello de color azul es la no presencia probable de litología S2, y por lo tanto el resto que se tiene en el costado derecho aproximadamente, es la presencia probable de la litología S2.

Ilustración 38: Vista de frente Kriging de indicadores para litología S2 (elaborado en SGeMS).

En la ilustración 39 se tiene una vista en planta para presencia probable de litología S2. Se puede observar un fuerte color rojo rodeados de otros colores más suaves, esto significa que el color rojo es la estimación probable de litología S2 y los otros tonos que le acompañan son el resto de los tipos de litologías, ya sea AND, HBX o Mixto.

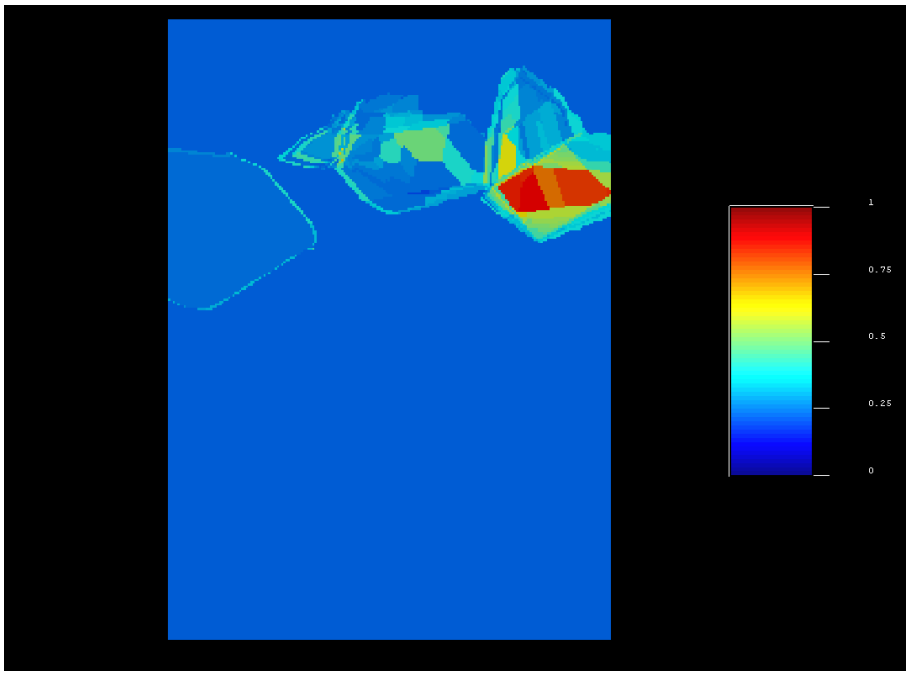

Ilustración 39: Vista en planta Kriging de indicadores para litología S2 (elaborado en SGeMS).

La ilustración 40 muestra la varianza del Kriging de indicadores para la litología S2.

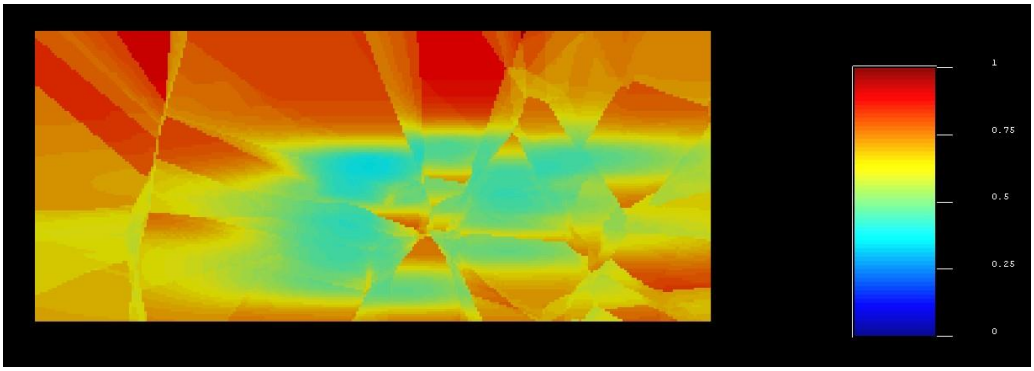

Ilustración 40: Varianza Kriging de indicadores para litología S2 (elaborado en SGeMS).

Para el caso del Kriging de indicadores para las litologías HBX y Mixto, se encuentran en el apéndice C.

Se usa un código de Python para georreferenciar los valores estimados por SGeMS, y luego mediante el uso del software gvSIG, se hacen las validaciones cruzadas y además se extrae el valor real versus el estimado para los valores que se tenía información.

Debido a que se deben hacer comparaciones entre los 3 métodos propuestos, es que se buscó de una medida estadística que sirviera para los 3 métodos, es así que se decidió utilizar la matriz de confusión y las medidas de desempeño exactitud, precisión, sensibilidad y especificidad.

En la tabla 13 se muestra el resultado de la matriz de confusión y los resultados estadísticos del Kriging de indicadores parar la litología AND:

		Predicción		
		0	1	Total
Real	0	323	160	483
	1	175	1503	1678
	Total	498	1663	2161

Tabla 13: Matriz de confusión Kriging de indicadores para litología AND (elaboración propia).

Recordando lo visto en el marco teórico sobre matriz de confusión, la casilla verde representa los aciertos positivos y la casilla naranja representa los aciertos negativos y el resto de casillas los falsos aciertos tanto negativos como positivos. Si se analiza la tabla 13, se puede notar una alta cantidad de aciertos para la litología AND, con un valor de 1503 aciertos positivos, en otras palabras, quiere decir que cuando el valor era real la predicción también coincidió con el valor real, y con esto generándose una estimación verdadera positiva.

En la tabla 14 se tiene los resultados estadísticos utilizando las medidas de desempeño de la matriz de confusión:

Kriging de Indicadores				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
AND	0.84	0.90	0.89	0.66

Tabla 14: Resultados estadísticos Kriging de indicadores para litología AND (elaboración propia).

A continuación se muestra el resultado de la matriz de confusión y los resultados estadísticos del Kriging de indicadores parar la litología S2 (para la litología HBX y Mixto se encuentran en el apéndice D).

		Predicción		
		0	1	Total
Real	0	1907	76	1983
	1	125	53	178
	Total	2032	129	2161

Tabla 15: Matriz de confusión Kriging de indicadores para litología S2 (elaboración propia).

Analizando la tabla 15, es evidente la disminución de aciertos positivos que posee, y la gran cantidad de aciertos negativos con un valor de 1907, esto se explica ya que en la base de datos se contaba con 145 datos para la litología S2,y por lo tanto al hacer la estimación por Kriging de indicadores, detecta pocos valores como aciertos positivos y detecta una gran cantidad de aciertos negativos, ya que como se tiene conocimiento la base de datos cuenta en gran medida con valores para la litología AND.

En la tabla 16 se tiene los resultados estadísticos utilizando las medidas de desempeño de la matriz de confusión:

Kriging de Indicadores				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
S2	0.90	0.41	0.29	0.96

Tabla 16: Resultados estadísticos Kriging de indicadores para litología S2 (elaboración propia).

5.2. Estimación Redes Neuronales Artificiales

Para litología AND:

En la ilustración 41 se tiene el resultado desarrollado en el programa Orange Canvas para la estimación mediante el método de redes neuronales artificiales de la litología AND para datos de entrenamiento:

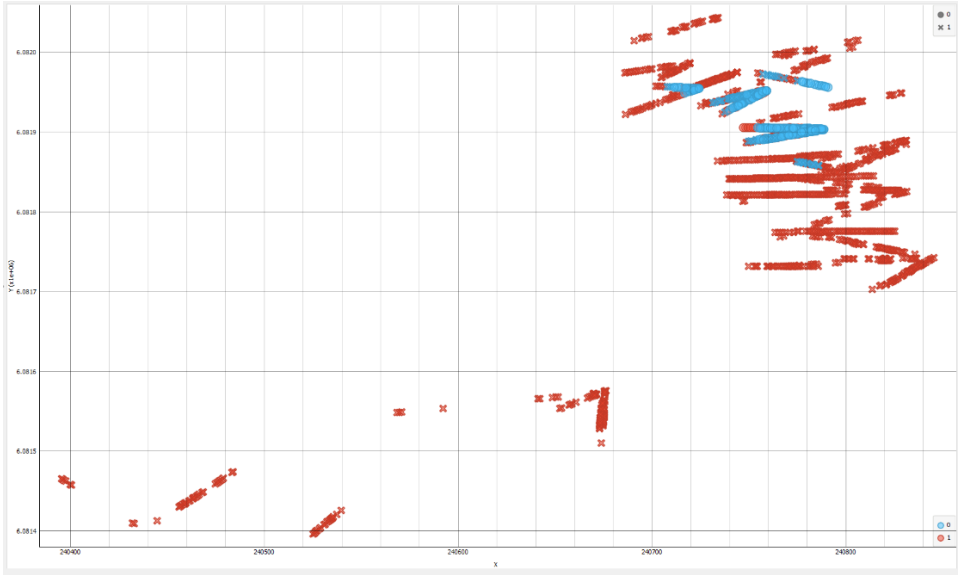

Ilustración 41: Plano XY redes neuronales datos de entrenamiento para litología AND (elaborado en Orange Canvas).

De acuerdo con la ilustración 41 se puede ver en color rojo la estimación probable de presencia de litología AND, y en color azul la no presencia probable de litología AND, es decir, la presencia probable de otras litologías. Cabe destacar que los resultados visualmente son distintos a los obtenidos primeramente en el programa SGeMS, esto se debe a que son desarrollados en Orange Canvas.

En la ilustración 42 se puede ver una vista alternativa del suceso de estimación para datos de entrenamiento de la litología AND, es evidente la gran cantidad de datos de los sondajes con la presencia de litología AND.

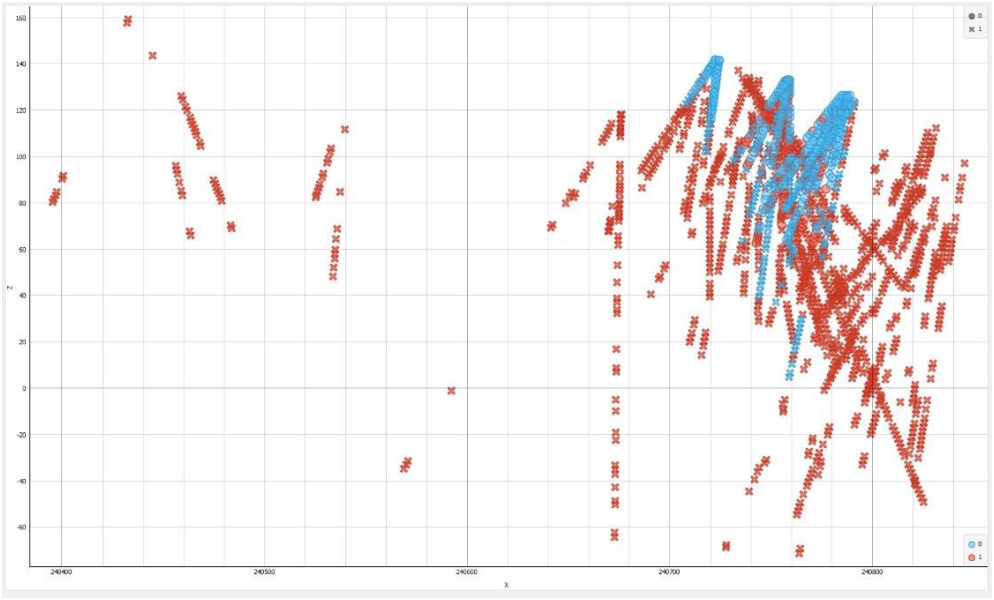

Ilustración 42: Plano XZ redes neuronales datos de entrenamiento para litología AND (elaborado en Orange Canvas).

Ahora bien, en la tabla 17 se puede ver el resultado de la matriz de confusión para el presente caso de estudio, en donde se puede notar la alta cantidad de datos estimados como aciertos positivos, con un valor de 1253 datos de los 1729 datos totales que se tienen para el conjunto de entrenamiento de las redes neuronales artificiales, y en cuanto a los aciertos negativos, tiene un valor bastante moderado con respecto al total, y lo que nos indica que este método resulta hasta el momento ser un buen estimador de variables nominales.

		Predicción		
		0	1	Total
Real	0	254	103	357
	1	119	1253	1372
	Total	373	1356	1729

Tabla 17: Matriz de confusión datos de entrenamiento para litología AND (elaboración propia).

En la tabla 18 se cuenta con los resultados de las medidas de desempeño para los datos de entrenamiento de la litología AND:

Redes Neuronales				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
AND	0.87	0.92	0.91	0.71

Tabla 18: Resultados estadísticos datos de entrenamiento para litología AND (elaboración propia).

Según se observa en la tabla 18 se tiene valores bastante altos y buenos para las medidas de desempeño de los datos de entrenamiento.

En las ilustraciones 43 y 44 se presentan los resultados de la estimación mediante redes neuronales artificiales para los datos de prueba de la litología AND:

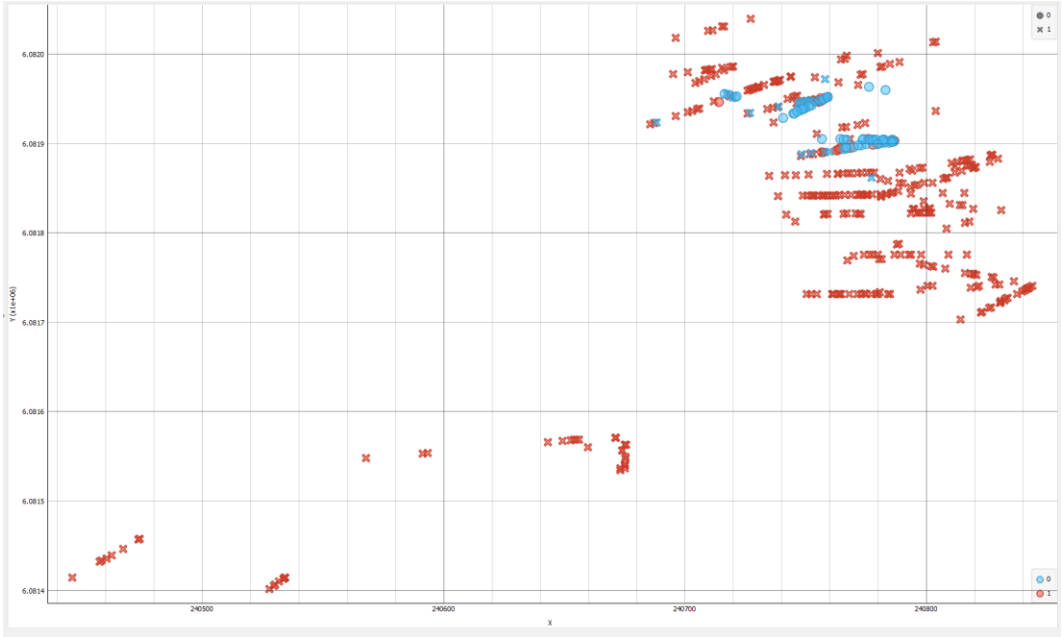

Ilustración 43: Plano XY redes neuronales datos de prueba para litología AND (elaborado en Orange Canvas).

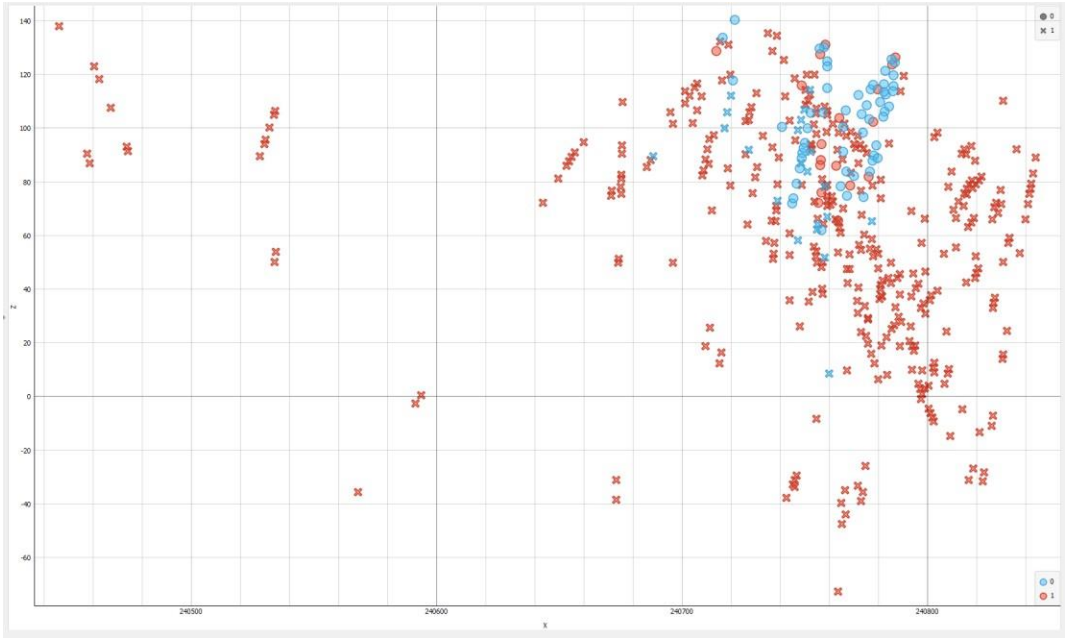

Ilustración 44: Plano XZ redes neuronales datos de prueba para litología AND (elaborado en Orange Canvas).

Analizando las ilustraciones 43 y 44, se nota aún una alta cantidad de datos con color rojo, es decir, con la presencia probable de la litología en estudio, pero cabe destacar que son menos datos a los ya analizados en la parte de entrenamiento, esto es normal, ya que es la distribución de datos que se quiso dar al asignar un 80% de datos para entrenamiento y un 20% de datos para la prueba del modelo.

En la tabla 19 se tiene el resultado de la matriz de confusión para los datos de prueba de la litología AND:

		Predicción		
		0	1	Total
Real	0	54	21	75
	1	18	339	357
	Total	72	360	432

Tabla 19: Matriz de confusión datos de prueba para litología AND (elaboración propia).

De acuerdo a lo observado en la tabla anterior, se puede decir que la matriz de confusión de los datos de prueba de la litología AND, resulta tener buenas predicciones, en otras palabras, el modelo predice 339 datos como aciertos positivos y 54 datos como aciertos negativos y el resto como falsos positivos y falsos negativos, estos valores nos indican que el modelo de prueba esta haciendo buenas estimaciones, y esto se debe al buen desempeño realizado para los datos de entrenamiento, que luego se transfiere dicho modelo al modelo de prueba.

En la tabla 20 se cuenta con los resultados de las medidas de desempeño para los datos de prueba de la litología AND:

Redes Neuronales				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
AND	0.90	0.94	0.95	0.72

Tabla 20: Resultados estadísticos datos de prueba para litología AND (elaboración propia).

Para litología S2:

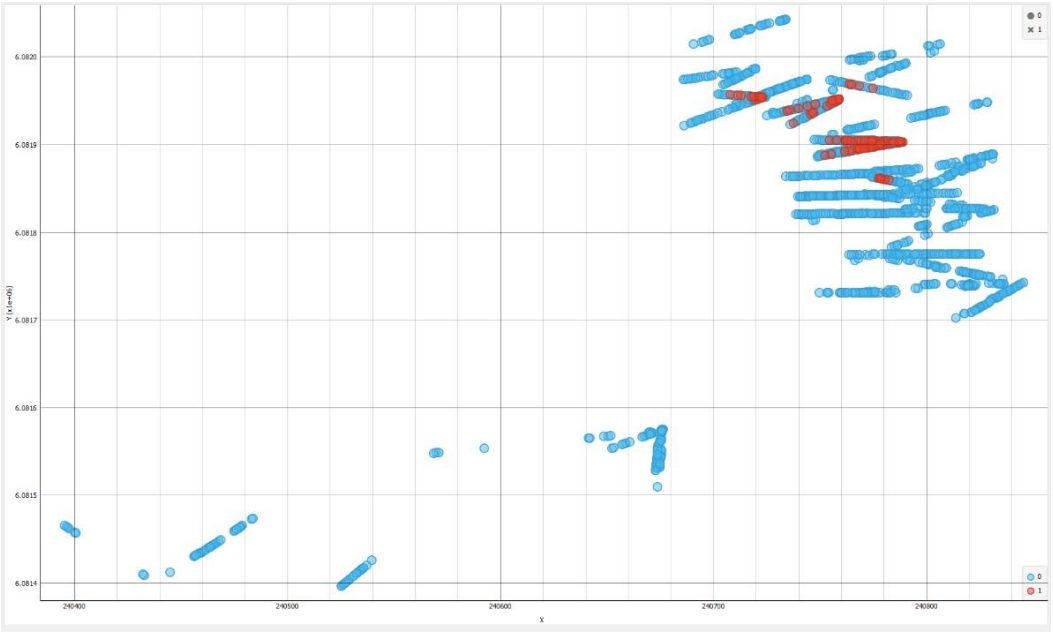

Ilustración 45: Plano XY redes neuronales datos de entrenamiento para litología S2 (elaborado en Orange Canvas).

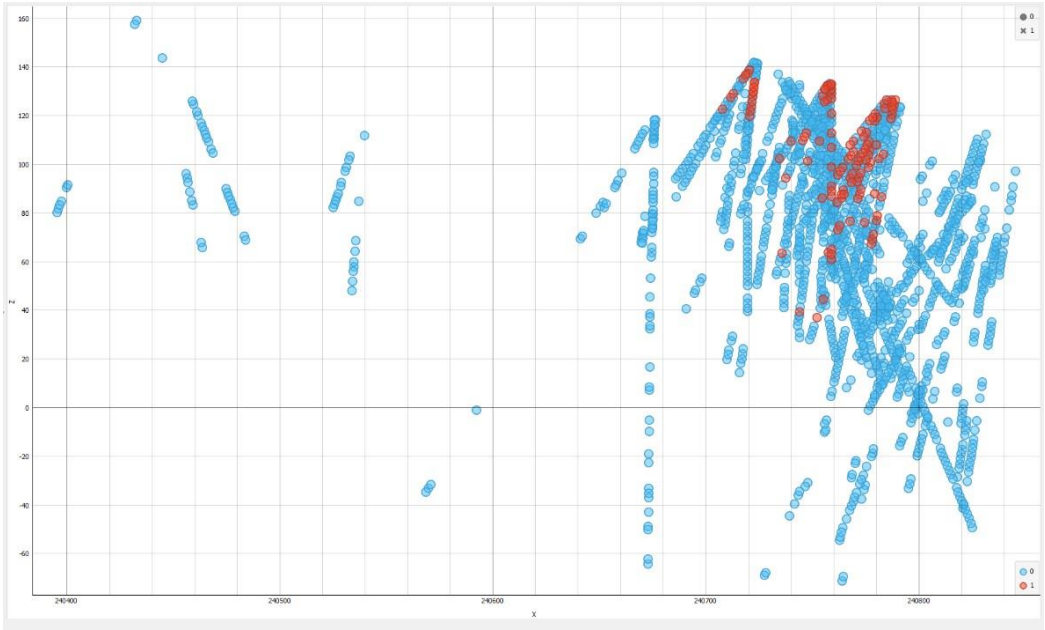

Ilustración 46: Plano XZ redes neuronales datos de entrenamiento para litología S2 (elaborado en Orange Canvas).

En las ilustraciones 45 y 46 se tiene los resultados de las estimaciones mediante el método de redes neuronales artificiales con los datos de entrenamiento para la litología S2. Según lo que se puede observar, se tiene una alta cantidad de datos en color azul y una baja cantidad de datos con color rojo, esto nos indica que el color azul es la no presencia probable de la litología en estudio y el color rojo representa la presencia probable de la litología S2.

A continuación se presentan las tablas de la matriz de confusión y de resultados estadísticos para la litología S2 (en el caso de la litología HBX y Mixto se encuentran en el apéndice D).

		Predicción		
		0	1	Total
Real	0	1560	29	1589
	1	93	47	140
	Total	1653	76	1729

Tabla 21: Matriz de confusión datos de entrenamiento para litología S2 (elaboración propia).

En la tabla 21 se nota la evidente cantidad de datos estimados como aciertos negativos con un valor de 1560 del total de 1729 datos y 47 para el caso de aciertos positivos, esto se produce debido a que S2 cuenta con un conjunto más pequeño de muestras, y por lo tanto el modelo se confunde más para lograr coincidir el valor real del predicho.

En la tabla 22 se encuentran los resultados de las medidas de desempeño para los datos de entrenamiento de la litología S2:

Redes Neuronales				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
S2	0.92	0.61	0.33	0.98

Tabla 22: Resultados estadísticos datos de entrenamiento para litología S2 (elaboración propia).

En las ilustraciones 47 y 48 se tiene una vista en el plazo XY y XZ de las estimaciones realizadas para litología S2. De acuerdo con lo que se observa se nota una disminución de datos en color rojo y azul, esto se debe a que es el conjunto de prueba.

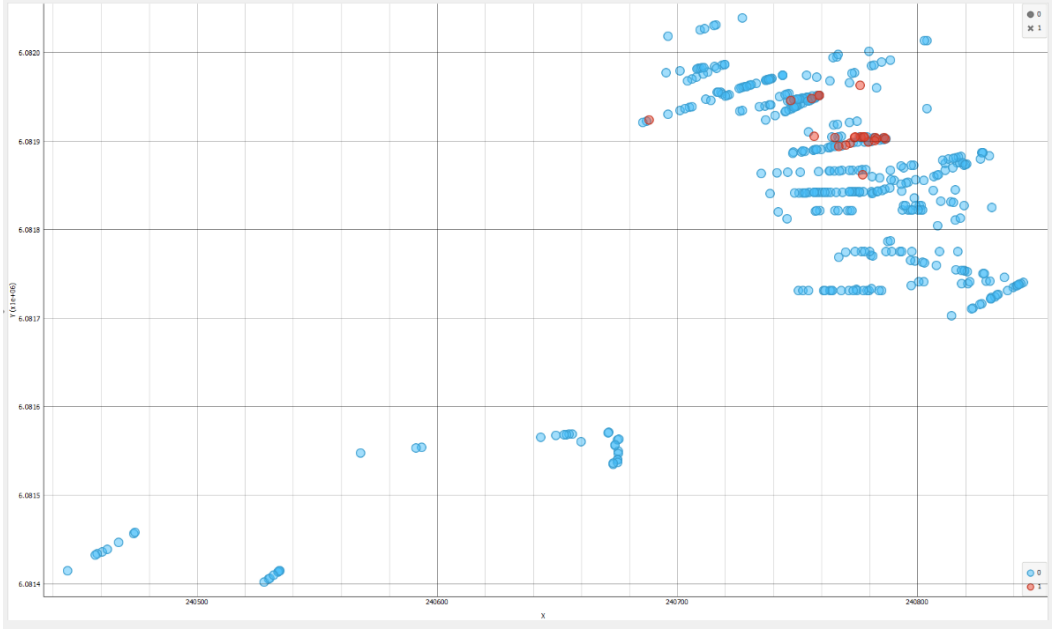

Ilustración 47: Plano XY redes neuronales datos de prueba para litología S2 (elaborado en Orange Canvas).

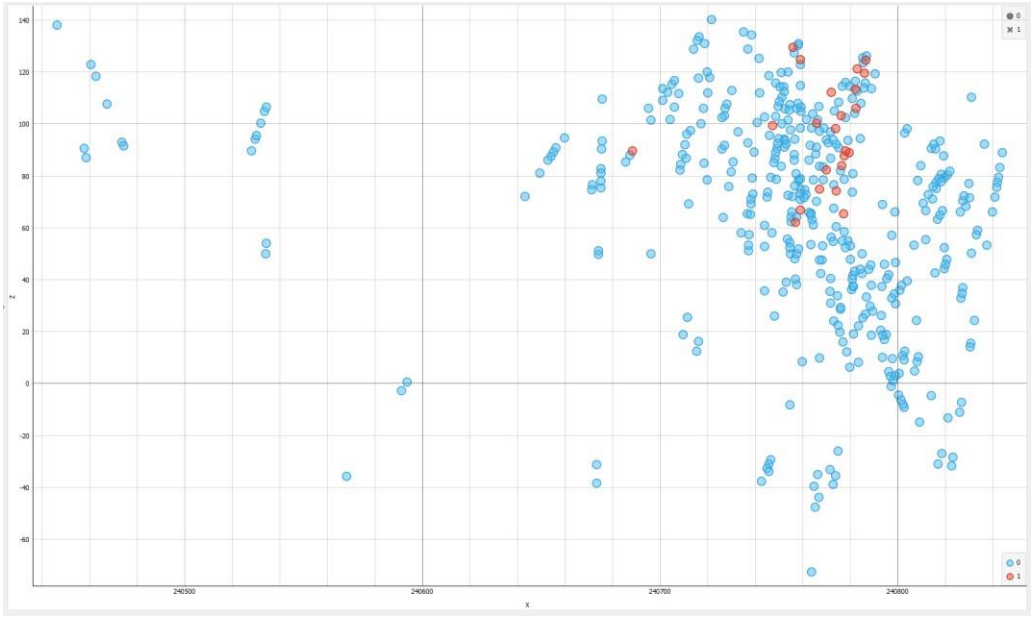

Ilustración 48: Plano XZ redes neuronales datos de prueba para litología S2 (elaborado en Orange Canvas).

A partir de las ilustraciones 47,48 y la tabla 23, se puede analizar que existe una presencia más baja de litología S2 en los datos de prueba con respecto a los datos de entrenamiento, esto sucede ya que en los datos de entrenamiento se cuenta con una mayor cantidad de datos para la litología S2, en cambio, para la prueba del modelo de redes neuronales artificiales se disminuye la cantidad de datos. A pesar de lo anterior, si se analiza la tabla 22 con respecto a la tabla 24, la exactitud como medida principal de desempeño, aumenta de 0.92 a 0.94.

		Predicción		
		0	1	Total
Real	0	394	2	396
	1	21	15	36
	Total	415	17	432

Tabla 23: Matriz de confusión datos de prueba para litología S2 (elaboración propia).

En la tabla 24 se encuentran los resultados de las medidas de desempeño para los datos de prueba de la litología S2:

Redes Neuronales				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
S2	0.94	0.88	0.41	0.99

Tabla 24: Resultados estadísticos datos de prueba para litología S2 (elaboración propia).

5.3. Estimación Regresión Logística

A continuación se verán los resultados obtenidos mediante el método de regresión logística para las litologías AND y S2 (para el caso de las litologías HBX y Mixto, se encuentran en el apéndice C).

Para litología AND:

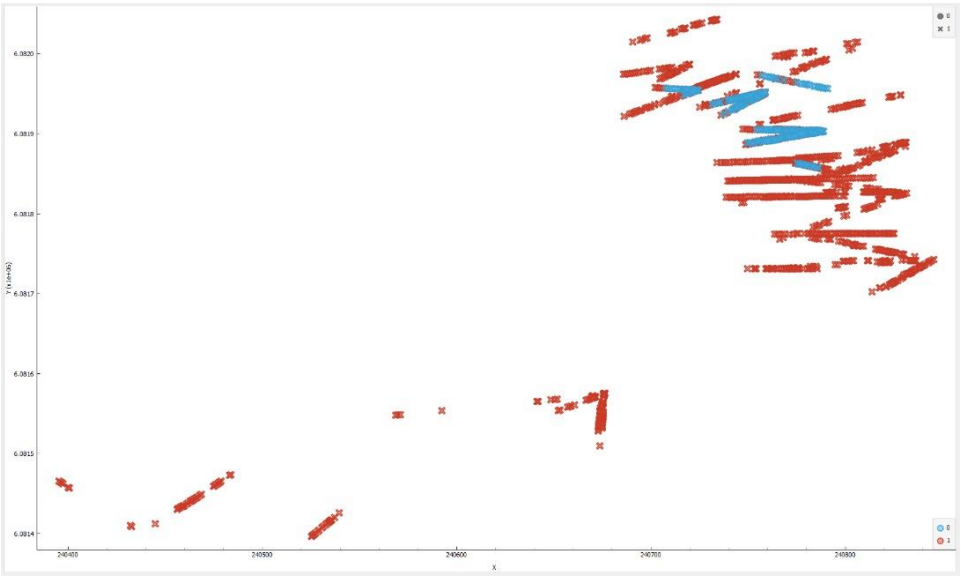

Ilustración 49: Plano XY regresión logística datos de entrenamiento para litología AND (elaborado en Orange Canvas).

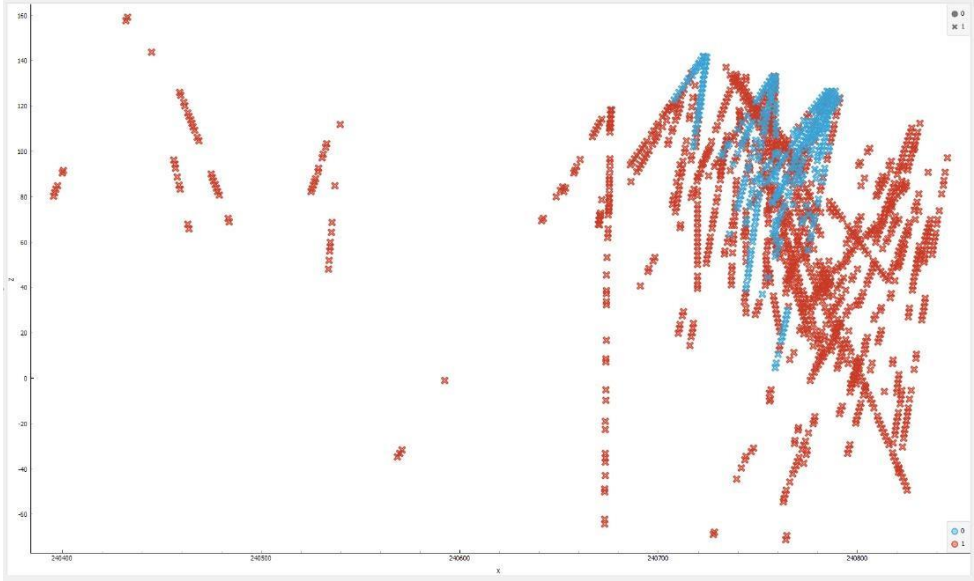

Ilustración 50: Plano XZ regresión logística datos de entrenamiento para litología AND (elaborado en Orange Canvas).

Como se puede ver en las ilustraciones 49 y 50, se tiene un gran conjunto de datos en color rojo y un conjunto más pequeño en el costado derecho con color azul, el rojo nos indica la alta presencia probable de litología AND para el conjunto de entrenamiento, y el color azul indica la baja presencia o la no presencia probable de litología AND.

A continuación se tiene las tablas 25 y 26 que representan la matriz de confusión y sus resultados de desempeño estadísticos:

		Predicción		
		0	1	Total
Real	0	84	335	419
	1	22	1288	1310
	Total	106	1623	1729

Tabla 25: Matriz de confusión datos de entrenamiento para litología AND (elaboración propia).

Regresión Logística				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
AND	0.79	0.79	0.98	0.2

Tabla 26: Resultados estadísticos datos de entrenamiento para litología AND (elaboración propia).

En la tabla 25 se cuenta con 1288 datos que su valor real coincidió con su valor predicho para los aciertos positivos y, 84 datos fueron predichos como aciertos negativos del total de 1729 de la base de datos de entrenamiento. Por lo anterior, el método nos indica que es un buen predictor para los aciertos positivos y esto queda más claro en su tabla de desempeño con un valor de 0,79 para la exactitud,0.79 para la precision,0.98 para la sensibilidad y con un 0.2 para la especificidad, de estos valores el que más llama la atención es la baja especificidad que produce el modelo para los datos de entrenamiento, es decir, no es un buen estimador para datos falsamente predichos, aunque esta métrica no es la más relevante para entrenar el modelo, ya que la que mayor peso tiene para producir mejores resultados es la exactitud y, este modelo es un buen estimador de la exactitud.

En las ilustraciones 51 y 52 se puede ver una disminución de los datos, ya que estamos frente a los resultados obtenidos por el modelo para los datos de prueba, es evidente que para los datos de prueba también estima bien para la litología AND, ya que se encuentran en color rojo varios puntos del espacio tridimensional y esto nos indica la clara presencia probable de la litología AND por sobre las demás litologías.

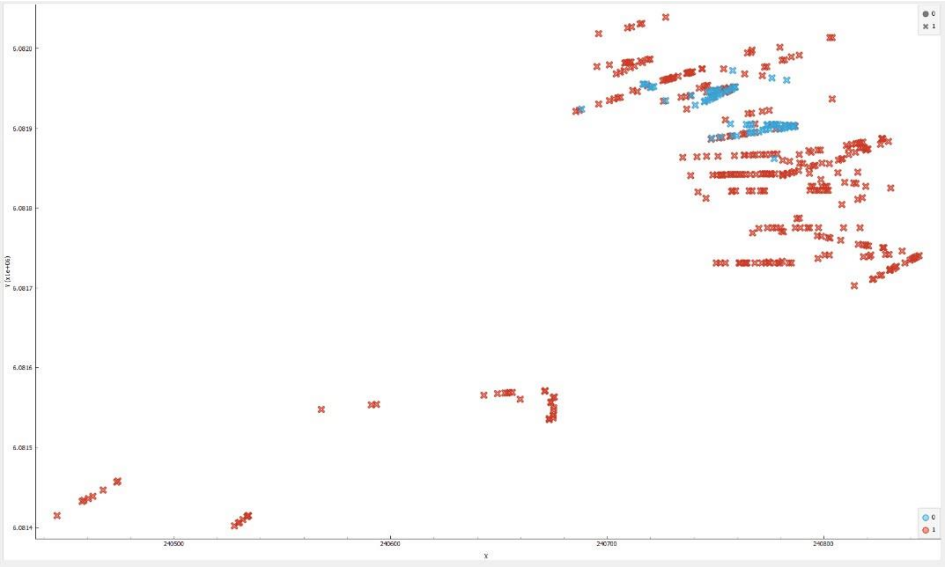

Ilustración 51: Plano XY regresión logística datos de prueba para litología AND (elaborado en Orange Canvas).

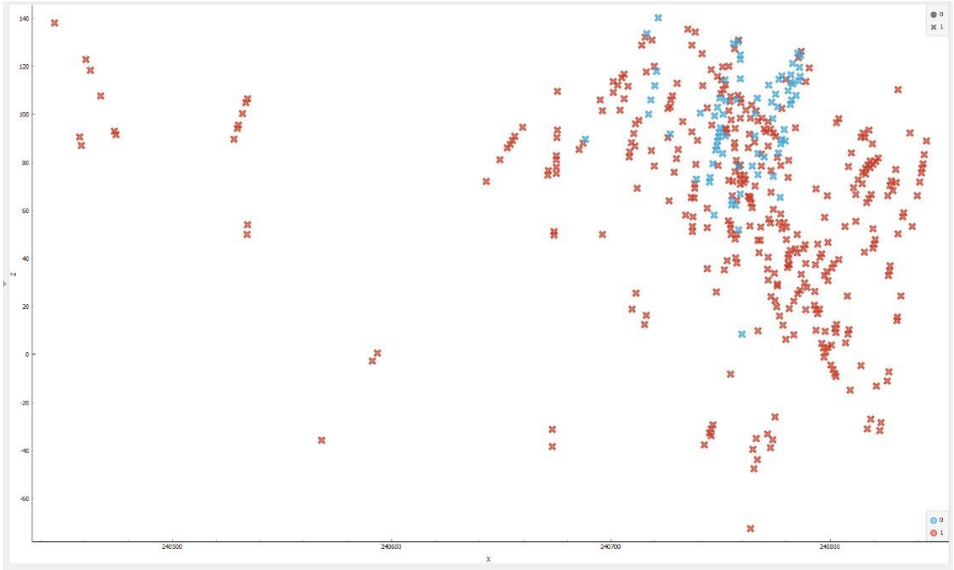

Ilustración 52: Plano XZ regresión logística datos de prueba para litología AND (elaborado en Orange Canvas).

En las tablas 27 y 28 se tiene la matriz de confusión y sus métricas de desempeño para los datos de prueba:

		Predicción		
		0	1	Total
Real	0	42	56	98
	1	19	315	334
	Total	61	371	432

Tabla 27: Matriz de confusión datos de prueba para litología AND (elaboración propia).

Regresión Logística				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
AND	0.82	0.84	0.94	0.42

Tabla 28: Resultados estadísticos datos de prueba para litología AND (elaboración propia).

Para litología S2:

En las ilustraciones 53 y 54 se tiene una vista en el plano XY y XZ para los resultados obtenidos mediante el uso de redes neuronales artificiales con los datos de entrenamiento para la litología S2:

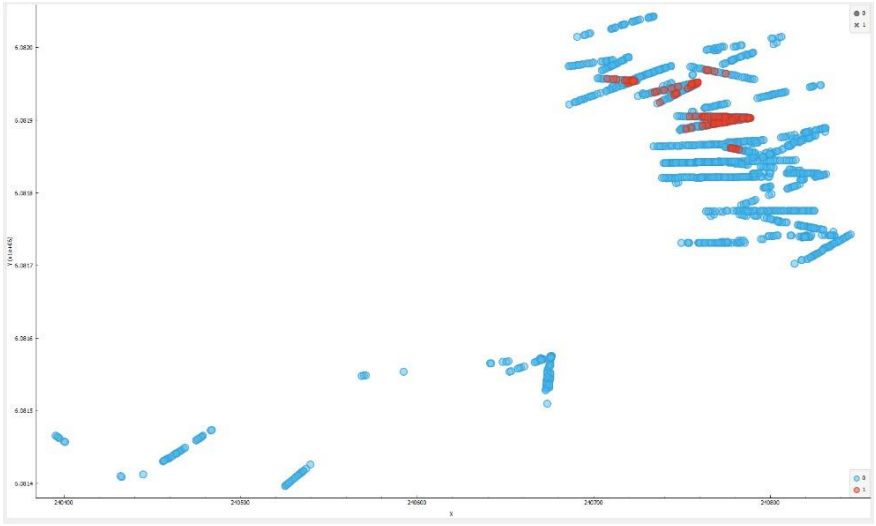

Ilustración 53: Plano XY regresión logística datos de entrenamiento para litología S2 (elaborado en Orange Canvas).

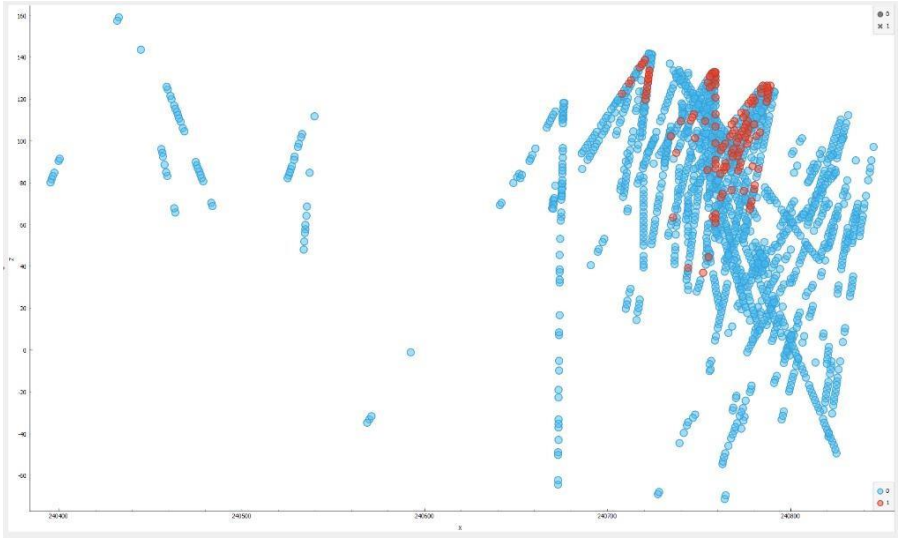

Ilustración 54: Plano XZ regresión logística datos de entrenamiento para litología S2 (elaborado en Orange Canvas).

Según se observa en los resultados anteriores, es claro ver que existe una gran cantidad de datos con color azul y un conjunto más reducido con color azul. Lo anterior nos indica que el modelo predijo como 0 la no presencia de la litología S2 y como 1 la presencia probable de la litología S2.

En las siguientes tablas se presentan los resultados obtenidos para la matriz de confusión y los resultados estadísticos para la litología S2 (en el caso de la litología HBX y Mixto se encuentran en el apéndice D).

		Predicción		
		0	1	Total
Real	0	1576	18	1594
	1	104	31	135
	Total	1680	49	1729

Tabla 29: Matriz de confusión datos de entrenamiento para litología S2 (elaboración propia).

Regresión Logística				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
S2	0.92	0.63	0.22	0.98

Tabla 30: Resultados estadísticos datos de entrenamiento para litología S2 (elaboración propia).

En las tablas anteriores, se tiene una predicción de 1576 datos del total de 1729 datos, que fueron predichos como la no presencia probable de litología S2 y con 31 datos como que existe la presencia probable de litología S2, el resto de los datos que se tiene en la diagonal son los datos mal predichos o falsos positivos o negativos, esto se produce debido a que se tiene otros tipos de litologías y el modelo tiende a confundirse. A pesar de lo anterior, el método resulta ser un buen estimador para las métricas de desempeño, en donde el que mas destaca es la exactitud con un valor de 0,92 o en porcentaje seria de 92%.

Para el caso de las predicciones llevadas a cabo por el modelo de regresión logística con los datos de prueba, ellos se presentan en las siguientes ilustraciones 55 y 56:

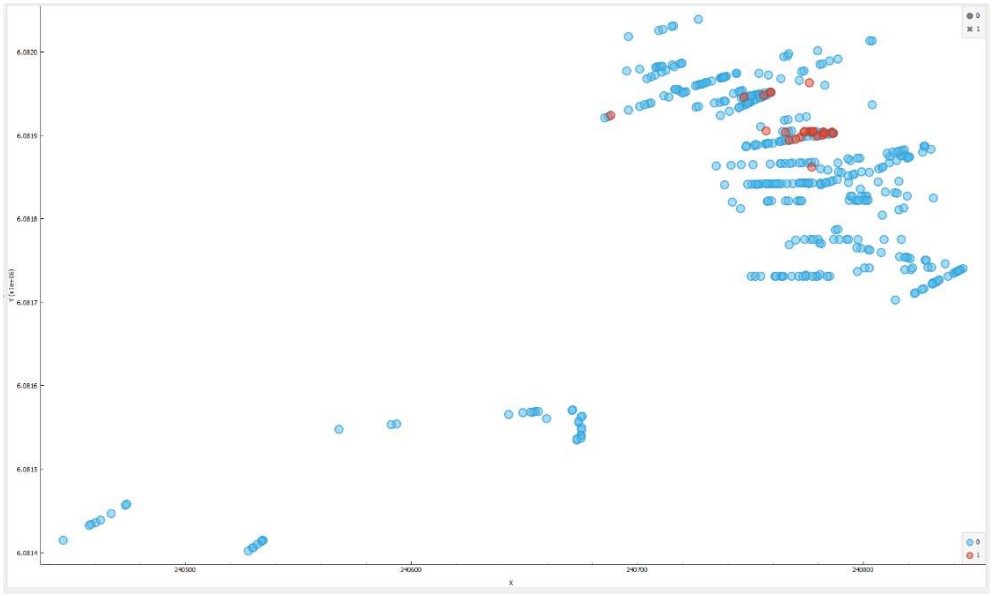

Ilustración 55: Plano XY regresión logística datos de prueba para litología S2 (elaborado en Orange Canvas).

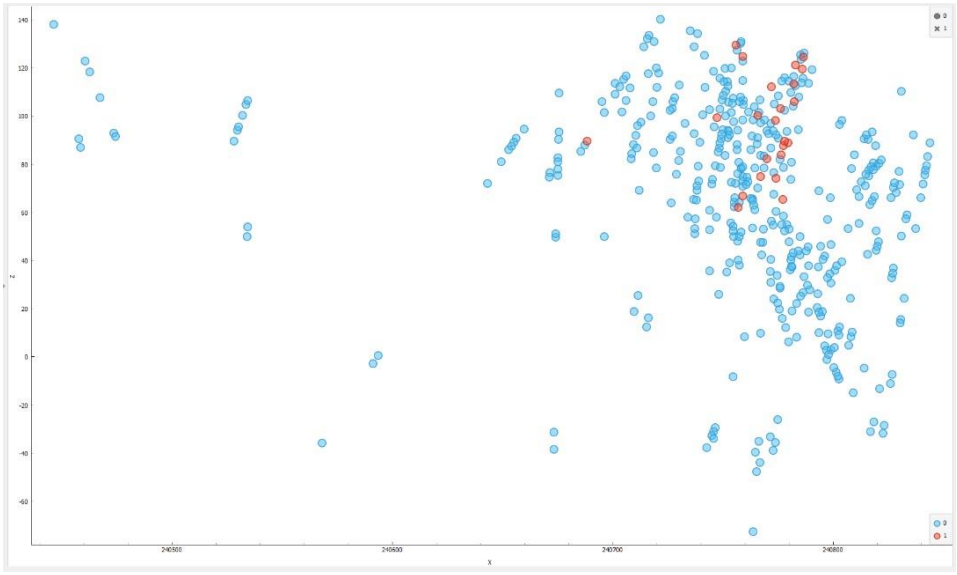

Ilustración 56: Plano XZ regresión logística datos de prueba para litología S2 (elaborado en Orange Canvas).

De los resultados anteriores, se aprecia que también existe una cantidad considerable de datos estimados como 0 o color azul, y una cantidad más pequeña estimada como 1 o en color rojo, esto nos quiere decir que el modelo de regresión logística para los datos de prueba predijo una cantidad disminuida como presencia probable de litología S2, esto es normal, ya que se cuenta con pocos datos para probar el modelo obtenido del conjunto de entrenamiento.

A continuación se presentan las tablas 31 y 32 con la matriz de confusión y las métricas de desempeño para la estimación mediante regresión logística con datos de prueba y con litología del tipo S2:

		Predicción		
		0	1	Total
Real	0	401	5	406
	1	18	8	26
	Total	419	13	432

Tabla 31: Matriz de confusión datos de prueba para litología S2 (elaboración propia).

Regresión Logística				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
S2	0.94	0.61	0.3	0.98

Tabla 32: Resultados estadísticos datos de prueba para litología S2 (elaboración propia).

5.4. Comparación de métodos de estimación

A continuación se verá las distintas comparaciones realizadas a los métodos de estimación desarrollados durante la memoria. Cabe destacar que debido a que los resultados obtenidos por los programas SGeMS y Orange Canvas visualmente son distintos y no sería una comparación valida solo evaluándolos de ese modo, es que se comparan mediante sus medidas de desempeño vistos para cada tipo de litología. Además, debido a que los métodos de machine learning tienen un procedimiento de trabajo distinto al método de Kriging de indicadores, ya que consideran un 80% de datos para entrenamiento y un 20% para prueba, es necesario tener la misma cantidad de datos para poder comparar los 3 métodos. Por lo anterior, lo que se hace es dejar el 100% de los datos de Kriging de indicadores, es decir, 2161 datos y para la parte de machine learning, se crea un nuevo modelo como resultado de los datos de entrenamiento y prueba, en otras palabras, el modelo estima para el 100% de los datos considerando el aprendizaje obtenido de los modelos de entrenamiento y prueba.

Primeramente se comenzará el análisis comparativo para los 3 métodos estudiados con litología de tipo AND:

Para litología AND:

		Predicción					
		Kriging Indicador		Redes Neuronales		Regresión Logística	
		0	1	0	1	0	1
Real	0	323	160	308	124	126	391
	1	175	1503	137	1592	41	1603
	Total	498	1663	445	1716	167	1994
		2161		2161		2161	

Tabla 33: Matriz de confusión comparación métodos de estimación para litología AND (elaboración propia).

En la tabla 33 se tiene un resumen de los resultados de la matriz de confusión obtenidos para la litología de tipo AND, para el Kriging de Indicadores, Redes Neuronales y Regresión Logística.

A continuación en la ilustración 57, se tiene un gráfico de barras de la litología de tipo AND, para los 3 tipos de métodos o técnicas propuestos:

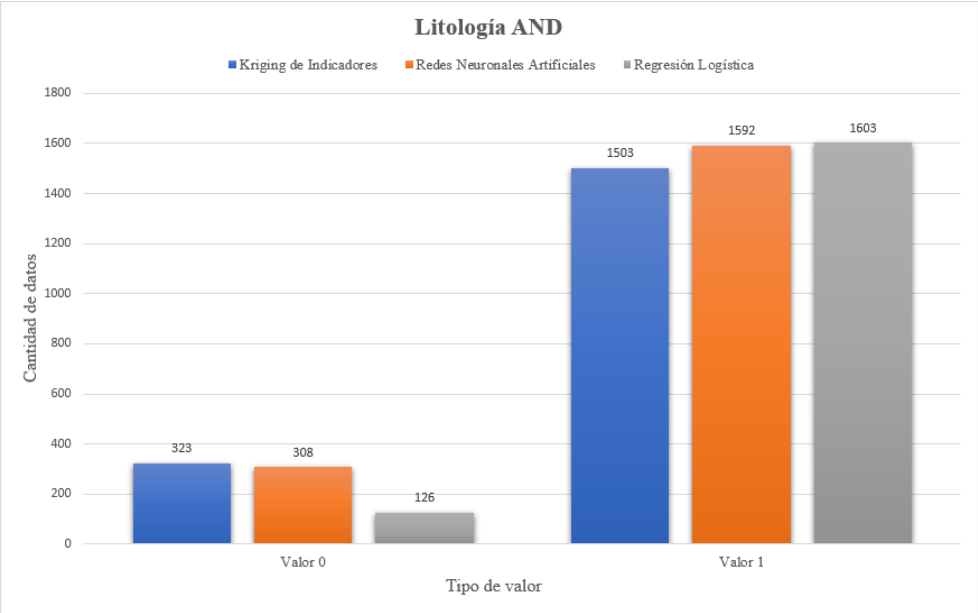

Ilustración 57: Gráfico de barras de la litología de tipo AND, para los 3 tipos de métodos (elaboración propia).

Se puede observar en la barra gris del método de Regresión Logística con un valor de 1603, como el valor que más alto coincide con respecto a el valor real del predicho, es decir, estima de mejor manera los aciertos positivos, en comparación a las Redes Neuronales y el Kriging de indicadores.

En la barra gris para el tipo de valor 0, es decir, para el método de Regresión Logística se tiene el valor más bajo de aciertos negativos, lo que nos indica que este método predice menos valores como aciertos negativos. Hasta este punto se podría pensar que la Regresión Logística es el mejor método para estimar litologías de tipo AND, el problema que tiene este método en comparación a los otros 2 propuestos, es que estima demasiados valores como falsos positivos y por lo tanto se pierden muchos valores que podrían haber sido considerados como aciertos positivos o negativos.

En cuanto al Kriging de indicadores y las Redes Neuronales resultan ser métodos bastante homogéneos en sus estimaciones de aciertos negativos con un valor de 323 y 308 respectivamente.

De acuerdo con lo anterior, lo llamativo está en sus predicciones de aciertos positivos, en donde el Kriging de indicadores tiene un valor de 1503 y las Redes Neuronales un valor de 1592, bastante por encima, lo que nos estaría indicando que las Redes Neuronales son mejores estimadores para aciertos positivos de la litología de tipo AND.

Para litología S2

		Predicción					
		Kriging Indicador		Redes Neuronales		Regresión Logística	
		0	1	0	1	0	1
Real	0	1907	76	1954	31	1977	23
	1	125	53	114	62	122	39
	Total	2032	129	2068	93	2099	62
		2161		2161		2161	

Tabla 34: Matriz de confusión comparación métodos de estimación para litología S2 (elaboración propia).

A continuación en la ilustración 58, se tiene el gráfico de barras para la litología de tipo S2, según los 3 tipos de métodos o técnicas propuestos:

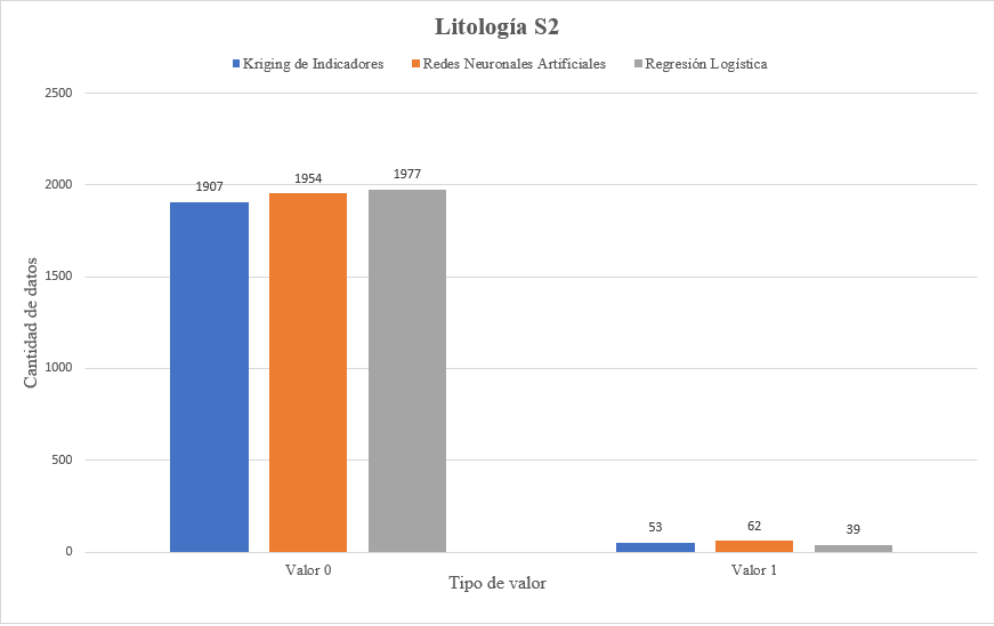

Ilustración 58: Gráfico de barras de la litología de tipo S2, para los 3 tipos de métodos (elaboración propia).

En la ilustración 58 se observa que las Redes Neuronales son las que mejor estiman aciertos positivos, con un valor de 62, y en segundo lugar el Kriging de indicadores con 53 datos, y por último la Regresión Logística con 39 datos. Además, si se analiza comparativamente la cantidad de datos con aciertos negativos, es decir, predichos con un valor 0, como la no presencia de litología S2, el que mayor predice de esta forma es la regresión logística con 1977 datos de los 2161 totales. Cabe destacar la disminución de aciertos positivos en comparación a la litología AND, esto se produce debido a que se tiene menor cantidad de datos para el entrenamiento y prueba del modelo por parte de la base de datos de la litología S2.

Para litología HBX:

		Predicción					
		Kriging Indicador		Redes Neuronales		Regresión Logística	
		0	1	0	1	0	1
Real	0	1908	67	2006	21	2030	23
	1	129	57	65	69	63	45
		2037	124	2071	90	2093	68
	Total	2161		2161		2161	

Tabla 35: Matriz de confusión comparación métodos de estimación para litología HBX (elaboración propia).

De acuerdo con la tabla 35 se obtiene la siguiente ilustración 59 para la litología de tipo HBX:

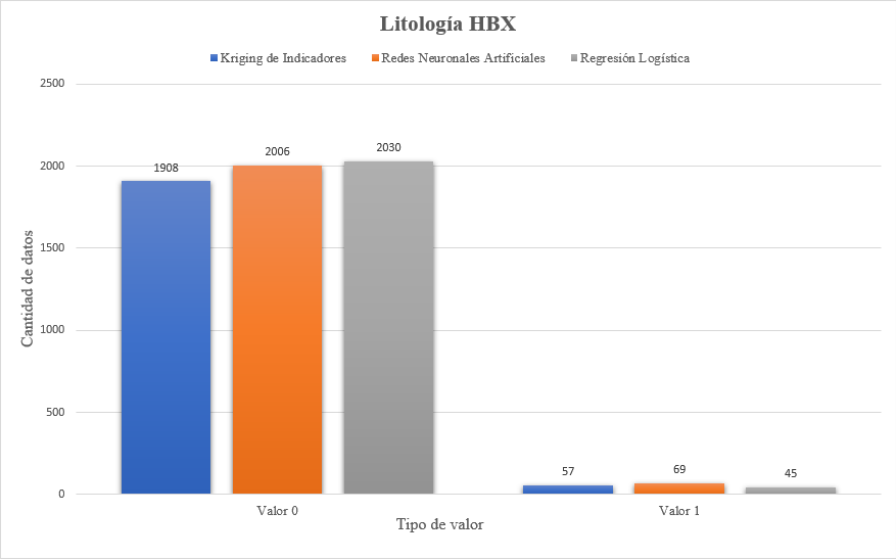

Ilustración 59: Gráfico de barras de la litología de tipo HBX, para los 3 tipos de métodos (elaboración propia).

En la ilustración 59 se tiene el gráfico de barras obtenido según la matriz de confusión de la tabla 35 para la litología HBX. Según se observa las Redes Neuronales son las que mejor predicen la presencia de litología HBX en comparación a los otros 2 métodos, con un valor de 69 y por detrás con 57 el Kriging de indicadores y en último lugar la Regresión Logística con 45 aciertos positivos. Así también, para el caso de los aciertos negativos el que predice mayor cantidad de 0, es decir, la no presencia de litología HBX, es la Regresión Logística con un valor de 2030 datos, cercanamente le sigue las Redes Neuronales con 2006 datos y bastante por detrás el Kriging de indicadores. Debido a lo anterior, nos indica que el Kriging de indicadores es un buen predictor para la no presencia probable de la litología HBX.

Para litología Mixto:

		Predicción					
		Kriging Indicador		Redes Neuronales		Regresión Logística	
		0	1	0	1	0	1
Real	0	1767	48	1853	29	1873	24
	1	151	195	165	114	177	87
	Total	1918	243	2018	143	2050	111
		2161		2161		2161	

Tabla 36: Matriz de confusión comparación métodos de estimación para litología Mixto (elaboración propia).

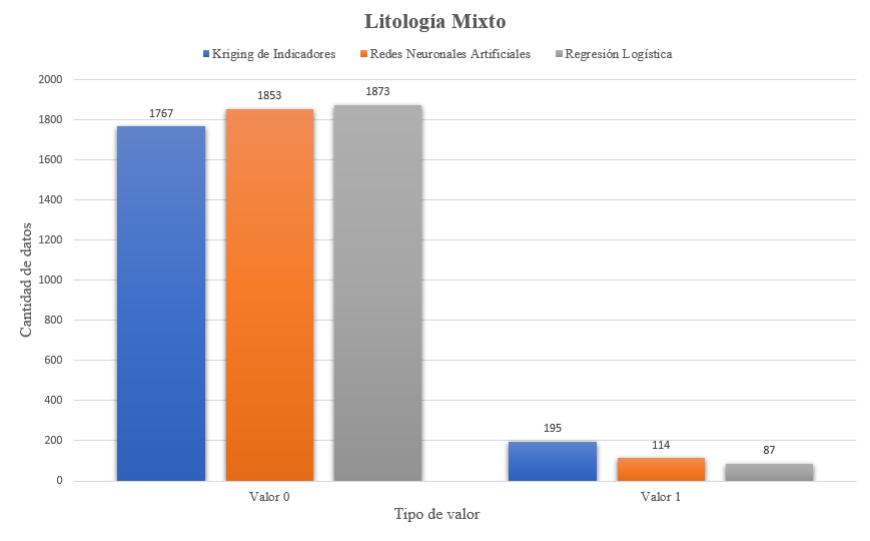

Ilustración 60: Gráfico de barras de la litología de tipo Mixto, para los 3 tipos de métodos (elaboración propia).

Para la litología Mixto, que está conformada por 4 tipos de litologías, que son STW, TBX, MQV y S1, se observa en la ilustración 60, que el Kriging de indicadores es el que mejor detecta la presencia de este tipo de litologías con un valor de 195 datos predichos como aciertos positivos, en cambio las Redes Neuronales predicen 114 datos como aciertos positivos, y la Regresión Logística bastante por detrás con 87 datos predichos como aciertos positivos. Ahora bien, analizando la cantidad de datos predichos como ceros o como la no presencia probable de la litología de tipo Mixto, la Regresión Logística es el que más datos predice como aciertos negativos, con un valor de 1873 datos, y de cerca le sigue las Redes Neuronales con 1853 datos y bastante por detrás el Kriging de indicadores. Por lo anterior, esto nos quiere decir que la Regresión Logística es un buen estimador de aciertos negativos para la litología de tipo Mixto, en comparación a los otros 2 métodos estudiados.

Todas las litologías:

	Kriging de Indicadores				Redes Neuronales				Regresión Logística			
Litología	Exactitud	Precisión	Sensibilidad	Especificidad	Exactitud	Precisión	Sensibilidad	Especificidad	Exactitud	Precisión	Sensibilidad	Especificidad
AND	0.84	0.90	0.89	0.66	0.87	0.92	0.92	0.71	0.80	0.80	0.97	0.24
S2	0.90	0.41	0.29	0.96	0.93	0.66	0.35	0.98	0.93	0.62	0.24	0.98
HBX	0.89	0.45	0.30	0.96	0.96	0.76	0.51	0.98	0.96	0.66	0.41	0.98
Mixto	0.90	0.80	0.56	0.97	0.91	0.79	0.40	0.98	0.90	0.78	0.32	0.98
Promedio	0.88	0.64	0.51	0.89	0.92	0.78	0.55	0.91	0.90	0.72	0.49	0.80

Tabla 37: Resultados estadísticos comparación métodos de estimación para todas las litologías (elaboración propia).

En la tabla 37 se cuenta con un resumen de las 4 métricas de desempeño para las 4 litologías estudiadas según los 3 métodos de estimación propuestos.

Primeramente, se comparará las métricas de desempeño para litología AND, y luego sucesivamente para las otras restantes, para lo cual se analizarán los gráficos de barras obtenidos según las 4 métricas de desempeño que son exactitud, precisión, sensibilidad y especificidad.

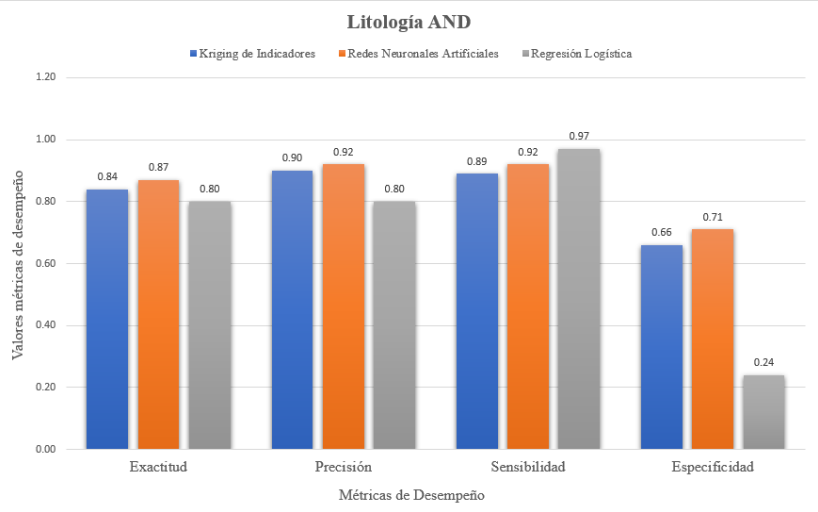

Ilustración 61: Gráfico de barras de Métricas de Desempeño para la litología de tipo AND, para los 3 tipos de métodos (elaboración propia).

En la ilustración 61 se tiene el análisis para la litología AND, en donde el método que mejor estima de acuerdo con sus respectivas métricas de desempeño, son las Redes Neuronales, ya que en general tiene mejores resultados para la exactitud y precision y en cuanto a la sensibilidad son valores más similares entre los 3 métodos, pero en cuanto a la especificidad, se nota un evidente aumento de 0.77 con respecto a la Regresión Logística que posee una especificidad de apenas 0.24.

A continuación en la Ilustración 62 se presenta el análisis realizado al grafico de barras de litología de tipo S2:

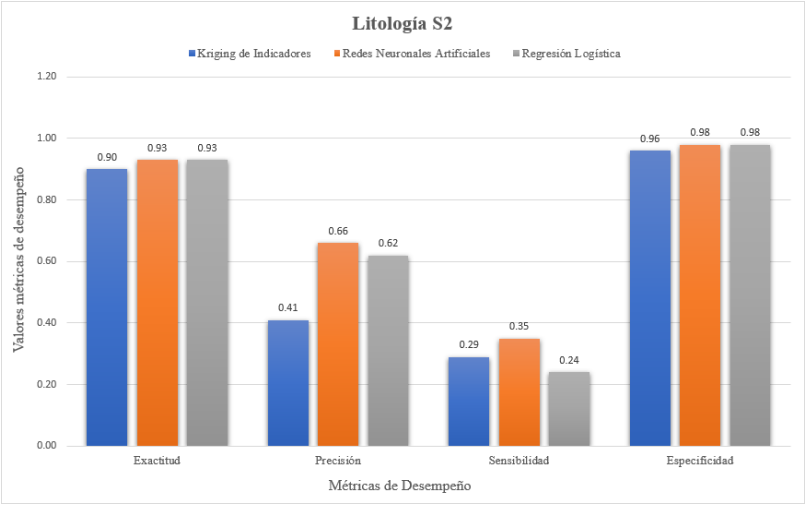

Ilustración 62: Gráfico de barras de Métricas de Desempeño para la litología de tipo S2, para los 3 tipos de métodos (elaboración propia).

Para la litología S2, el método que mejores resultados en general se observan en la ilustración 62, es el de Redes Neuronales, con una exactitud de 0.93 y precision de 0.66, en cambio para la sensibilidad y especificidad se observan valores más similares entre los 3 métodos por lo tanto no se puede asegurar cuál de los 3 métodos tendría una mejor estimación en otros casos de estudio.

A continuación en la Ilustración 63 se tiene el análisis realizado al grafico de barras de la litología de tipo HBX:

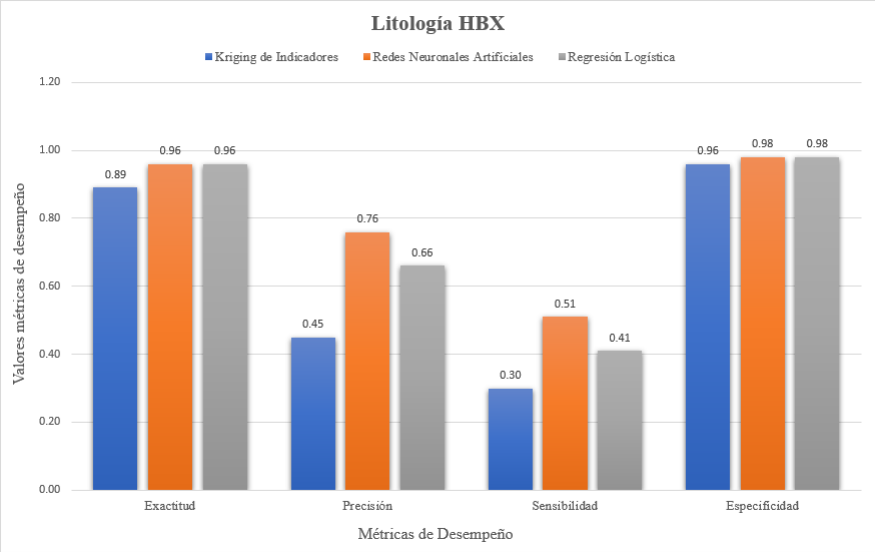

Ilustración 63: Gráfico de barras de Métricas de Desempeño para la litología de tipo HBX, para los 3 tipos de métodos (elaboración propia).

Para la litología HBX, una vez más las redes neuronales se imponen con respecto a los otros métodos, al menos en cuanto a exactitud y precision, por su parte la sensibilidad de las Redes Neuronales parece ser significativamente más alta, con un valor de 0.51, en comparación a los 0.30 del Kriging de indicadores y el 0.41 de la Regresión Logística, por su parte la especificidad tiene valores más similares entre los 3 métodos estudiados con valores por sobre los 0.90.

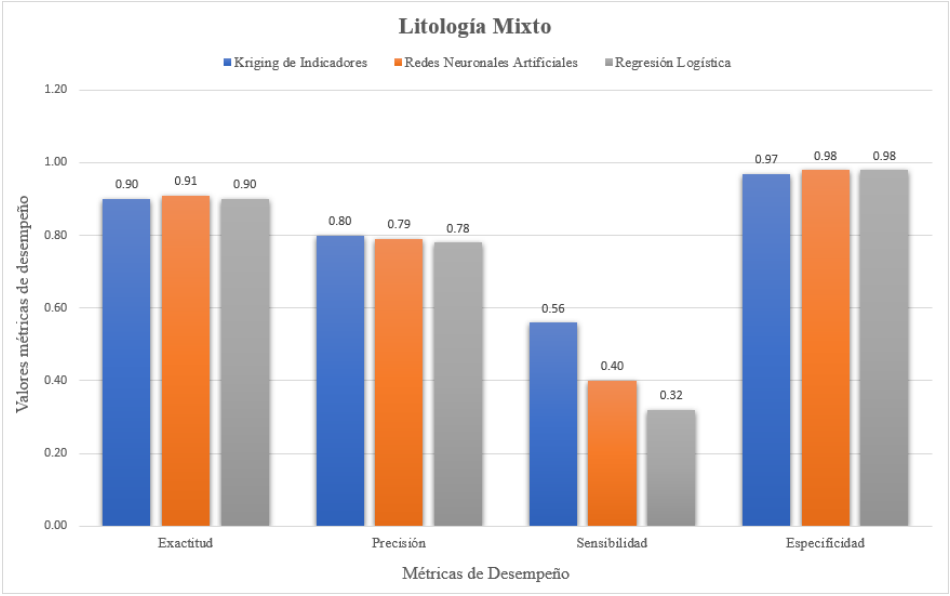

Ilustración 64: Gráfico de barras de Métricas de Desempeño para la litología de tipo Mixto, para los 3 tipos de métodos (elaboración propia).

Finalmente, para el caso en estudio de la litología de tipo Mixto, el que mejor resultados obtiene son las Redes Neuronales, con una exactitud de 0.91,se presenta una diferencia en la precisión en comparación a los otros casos ya estudiados, ya que el Kriging de indicadores obtiene mejores resultado para la métrica de desempeño precision, con un valor de 0.80,y de cerca le sigue las Redes Neuronales con un valor de 0.79, y para la Regresión Logística con 0,78.De lo anterior, se obtienen valores bastante similares entre las precisiones y por lo tanto no se puede asegurar, cual es mejor que otra.

Ahora bien con respecto, a la sensibilidad el Kriging de indicadores tiene mejores resultados con un valor de 0.56, en comparación a los 0.40 y 0.32 de las Redes Neuronales y Regresión Logística respectivamente. Para la restante métrica de desempeño, se tiene para los 3 métodos valores que superan los 0.97, y por lo tanto cualquiera de los 3 métodos da buenos resultados para la especificidad.

Hasta ahora de acuerdo con las métricas de desempeño analizadas para cada tipo de litología, y según su método de estimación propuesto, las Redes Neuronales aparecen a grandes rasgos como el que mejores resultados obtiene según el tipo de litología que se quiera estudiar. Por lo anterior, es que se quiso hacer una tabla final, con el promedio para todas las litologías según su métrica en estudio, en otras palabras, si se tiene la exactitud, será el promedio de las 4 litologías estudiadas.

En la tabla 38 se muestran los resultados finales para todos los métodos de estimación, según el promedio para métrica de desempeño.

Método	Exactitud	Precisión	Sensibilidad	Especificidad
Kriging de Indicadores	0.88	0.64	0.51	0.89
Redes Neuronales	0.92	0.78	0.55	0.91
Regresión Logística	0.90	0.72	0.49	0.80

Tabla 38: Resultados estadísticos comparación métodos de estimación para todas las litologías (elaboración propia).

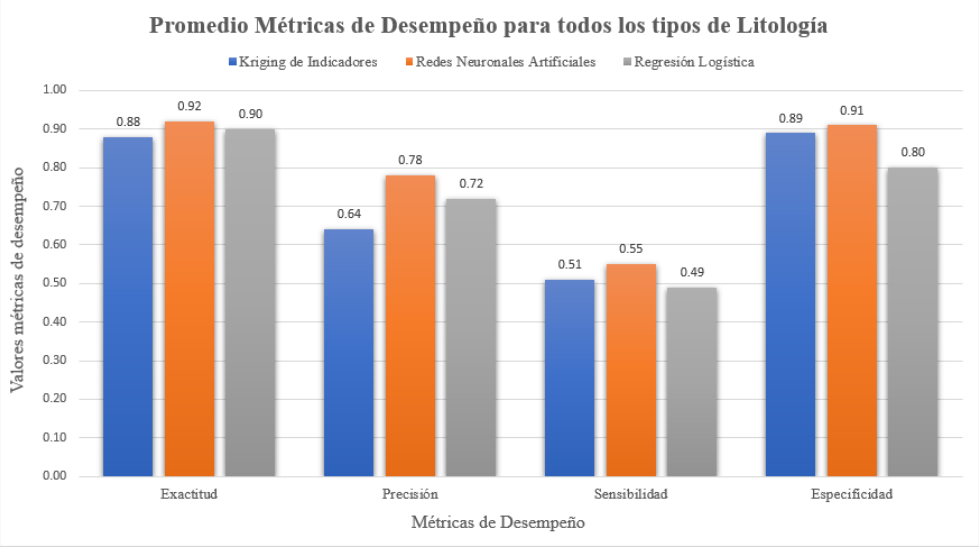

Ilustración 65: Gráfico de barras de Métricas de Desempeño para el promedio de todos los tipos de litología, para los 3 tipos de métodos (elaboración propia).

De acuerdo con la tabla final y su respectivo gráfico de barras, es que en la ilustración 65 se puede notar que las Redes Neuronales son las que mejor estiman la exactitud con un valor de 0.92, luego se tiene la Regresión Logística con un valor de 0.90, y por último el Kriging de indicadores con un valor de 0,88.

Cabe destacar que tanto los resultados obtenidos para cada tipo de litología, como así también el promedio final de la tabla 38, se tienen valores bastantes altos, si se toma como métrica de desempeño principal la exactitud, con valores cercanos al 90%, esto quiere decir, que los 3 métodos estiman bien tanto los aciertos positivos como aciertos negativos, en otras palabras, predicen de gran manera la presencia probable o no presencia probable de las litologías en estudio.

Estos valores por encimas de 90%, posiblemente sea por la naturaleza de los datos que se utilizaron para desarrollar la memoria, puesto que era una base de datos con ceros y unos, y en general los métodos predicen de mejor forma las variables nominales, en comparación a tener valores continuos, como son las leyes de minerales, que harían que los modelos tengan que trabajar más para obtener mejores resultados.

Conclusión

En primer lugar con respecto a los 3 métodos empleados durante el desarrollo de la presente memoria, el Kriging de indicadores demuestra ser un método bastante positivo y correcto en sus resultados, y debido a que sus algoritmos están respaldados por la información provenientes de los variogramas, resulta ser bastante confiable en comparación a los métodos de machine learning. La parte negativa del Kriging de indicadores, es que ante la posible entrada de nuevos datos al modelo, se requiere nuevamente realizar el estudio variográfico y se convierte en un proceso lento y tedioso.

En el caso de la regresión logística, tiende a construir modelos más simplificados y con un menor costo de tiempo y procesamiento, en comparación al Kriging de indicadores y las redes neuronales, por lo tanto puede ser una opción viable en la ayuda de la interpretación geológica sin tener que ejercer tanto tiempo para llegar a resultados preliminares.

Para el caso de las redes neuronales artificiales, son las que demuestran en promedio los mejores resultados para los estadísticos considerados, los cuales fueron la exactitud, precisión, sensibilidad y especificidad, en donde particularmente son las que mejor estiman la exactitud con un valor de 0.92, es decir, que el valor real más se acerca al valor estimado. La parte negativa de las redes neuronales artificiales, es que según sus características tienden al overfitting, en otras palabras, aprenden tan bien a estimar un modelo para los datos de entrenamiento ,que ante la presencia de los datos de prueba, también los estima de una manera demasiado correcta, el problema se causa cuando se está ante la presencia de nuevos datos que nunca han pasado por el modelo de entrenamiento y menos por el modelo de prueba, por lo tanto se pueden dar el caso de overfitting o underfitting, y seguir arrastrando errores. Además, agregar que las redes neuronales en general toman como una fuente valiosa el tener nuevos datos para el modelo, lo complicado está en que al tener mayores cantidades de atributos o características, hacen que los tiempos de entrenamiento y prueba, crezcan exponencialmente y provocando un mayor requerimiento computacional.

Ahora bien, con respecto a los objetivos de la memoria, se puede decir que los métodos de machine learning resultan una buena alternativa para la estimación de variables geológicas, y en nuestro caso, para variables categóricas o nominales, pero debido a la cercanía de los resultados estadísticos no se podría asegurar con total confianza que un método supera con creces a otro, ya que los resultados son muy cercanos y bastaría con seguir mejorando las iteraciones para producir una alteración de dichos resultados, y por lo tanto podría cambiar quien estima mejor que otro. Debido a lo anterior, es que el utilizar un método ya sea geoestadístico o de machine learning, queda en manos del experto del área.

Entre los problemas generales que se tiene al estimar mediante los 3 métodos planteados, el que resulta más como fuente de implicancias, es la cantidad y calidad de los datos. Por lo anterior, es que dentro de las mejoras futuras que se plantean para la investigación, es disponer de una mayor cantidad de datos, y sobre todo de datos más variados y representativos, ya que como se revisó en la memoria, había una representación bastante considerable de la litología AND, en comparación a los otros tipos de litologías.

Otras de las consideraciones futuras que se pueden tener, es plantear otros métodos tanto para la parte geoestadística, como así también, para la parte de machine learning o aprendizaje de máquinas. Para la parte geoestadística se podría considerar utilizar el CoKriging de indicadores, con la intención de encontrar mayores correlaciones entre las variables a utilizar, y en cuanto al machine learning, se podría considerar utilizar árboles de decisión o support vector machine, los cuales son métodos bastante utilizados en estos tiempos para clasificar valores dicotómicos o nominales.

Dentro de las mejoras que se plantean también, están el considerar usar paquetes de programación de Python o programas más avanzados, como por ejemplo Keras, TensorFlow, Numpy, etc. Con el objetivo de tener un mayor control en cada proceso que se lleva a cabo en las estimaciones.

Además, se podría considerar realizar un estudio sobre la categorización de recursos y su respectivo análisis económico, con los resultados obtenidos para los métodos que se quieran estudiar, ya sea geoestadístico o de machine learning, con la finalidad de estudiar el comportamiento de las variables nominales y su relación con las variables continuas.

En conclusión, las técnicas provenientes del machine learning tienen un enorme potencial de aplicación, y por lo tanto se recomienda seguir estudiando otros casos en los que se pueda aplicar machine learning, ya sea para variables nominales o variables continuas, o también aplicarlo a otras áreas del proceso minero y compararlos con herramientas clásicas de la geoestadística.

Referencias bibliográficas

Alfaro, M. (2007). Estimación de recursos mineros. Universidad de Chile, Santiago, Chile

Devore, J. (2012). Probabilidad y estadística para ingeniería y ciencias (Séptima Ed).

Covarrubias, G. (2012). Construcción y validación de una metodología de seguimiento para modelos de regresión logística. Universidad de Chile, Santiago, Chile.

Caparrini, F. (2018). Redes Neuronales: una visión superficial.

Emery, X. y González, K. (2007). Probabilistic modelling of lithological domains and its application to resourse evaluation.Universidad de Chile,Santiago,Chile.

Emery, X. (2013). Geoestadística. Universidad de Chile,Santiago,Chile.

Espinoza, B. (2019). Evaluación de uso de redes neuronales para la estimación de leyes como alternativa a la geoestadística. Universidad de Talca, Curicó, Chile.

Flórez, R. y Fernández, J. (2008). Las Redes Neuronales Artificiales: Fundamentos teóricos y aplicaciones prácticas.

Frez, T. (2014). Kriging y simulación secuencial de indicadores con proporciones localmente variables. Universidad de Chile,Santiago,Chile.

Giraldo, R. (2002). Introducción a la Geoestadística, Teoría y Aplicación. Universidad Nacional de Colombia, Bogotá, Colombia.

Hosmer, D. y Lemeshow, S. (2000). Applied Logistic Regression.2ª ed. United States, Wiley.

Journel, A. (1983). Nonparametric estimation of spatial distributions. Journal of the International Association for Mathematical Geology.

Kohavi, R. y Provost, F. (1998). Glossary of terms. Machine Learning-Special Issue on Applications of Machine Learning and the Knowledge Discovery Process.

Matheron, G. (1971). The Theory of Regionalized Variables and its Applications, Ecole Nationale Supérieure des Mines de Paris, Fontainebleau,France.

Matich, D. (2001). Redes Neuronales: Conceptos Básicos y Aplicaciones.

Murphy, K. (2012). Machine learning a probabilistic perspective. Massachusetts Institute of Technology.

Mery, N. (2016). Modelamiento geoestadístico de la incertidumbre en leyes y tipos de roca en un yacimiento ferrífero. Universidad de Chile, Santiago, Chile.

Rossi, Mario E. y Deutsch, Clayton V. (2013). Mineral resource estimation. Springer Science & Business Media.

Rodríguez-Sahagún, P. (2018). Aplicación de redes neuronales convolucionales y recurrentes al diagnóstico de autismo a partir de resonancias magnéticas funcionales. Memoria para optar al Título de Ingeniero Industrial. Universidad Politécnica de Madrid, Escuela Técnica Superior de Ingenieros Industriales, Madrid, España.

Rodríguez, E. (2019). Estimación de recursos en presencia de incertidumbre geológica. Universidad de Concepción, Chile.

Retamal, J. (2020). Predicción y diagnóstico en suelos contaminados por DAM usando Machine Learning. Universidad de Talca, Curicó, Chile.

Sucapuca, L. (2017). Vínculos entre relaciones de contacto y variogramas de indicadores para el modelamiento de variables categóricas. Universidad de Chile, Santiago, Chile.

Salas, M. (2019). Predicción geoestadística de unidades geológicas o geometalúrgicas utilizando información de variables cuantitativas. Universidad de Chile, Santiago, Chile.

Vapnik, Vladimir N. (1995). The Nature of Statistical Learning Theory. New York: Springer.

Apéndice A: Análisis exploratorio de datos

En los siguientes apéndices se presentan todas aquellas tablas o gráficos que no pertenecen al cuerpo principal de la memoria.

Desde la tabla **A.1** hasta la tabla **A.2** se encuentran las estadísticas básicas para las leyes de plata y oro, además de sus histogramas y Boxplot, desde la ilustración **A.1** hasta ilustración **A.4**.

Ley de Plata (ppm)	
Media	5.87
Error Típico	0.60
Mediana	2.19
Moda	1.00
Desviación estándar	40.18
Varianza de la muestra	1614.16
Curtosis	2119.06
Coeficiente de asimetría	44.45
Rango	2079.99
Mínimo	0.01
Máximo	2080
Suma	26373.21
Datos	4495

Tabla A.1: Estadísticas Básicas Ley de Plata (Elaboración propia)

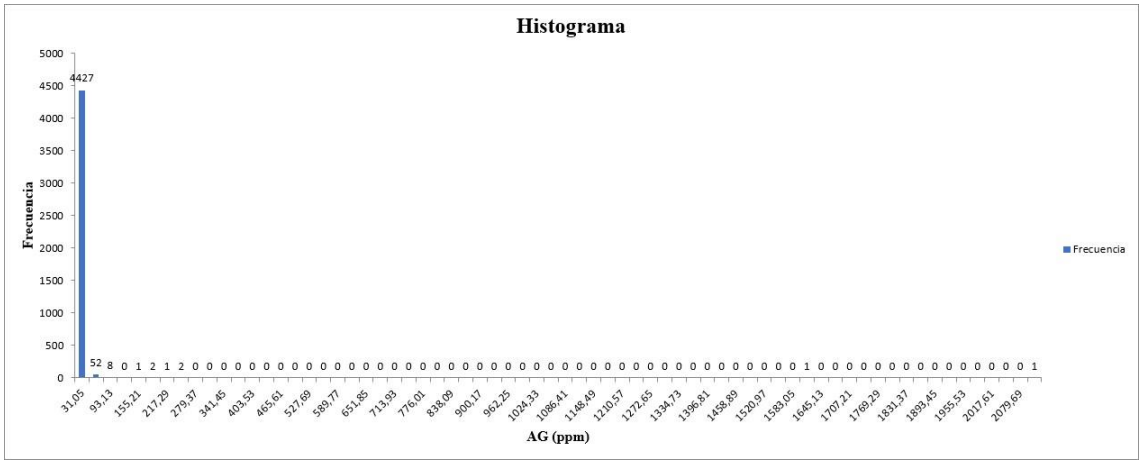

Ilustración A.1: Histograma Ley de Plata (Elaboración propia)

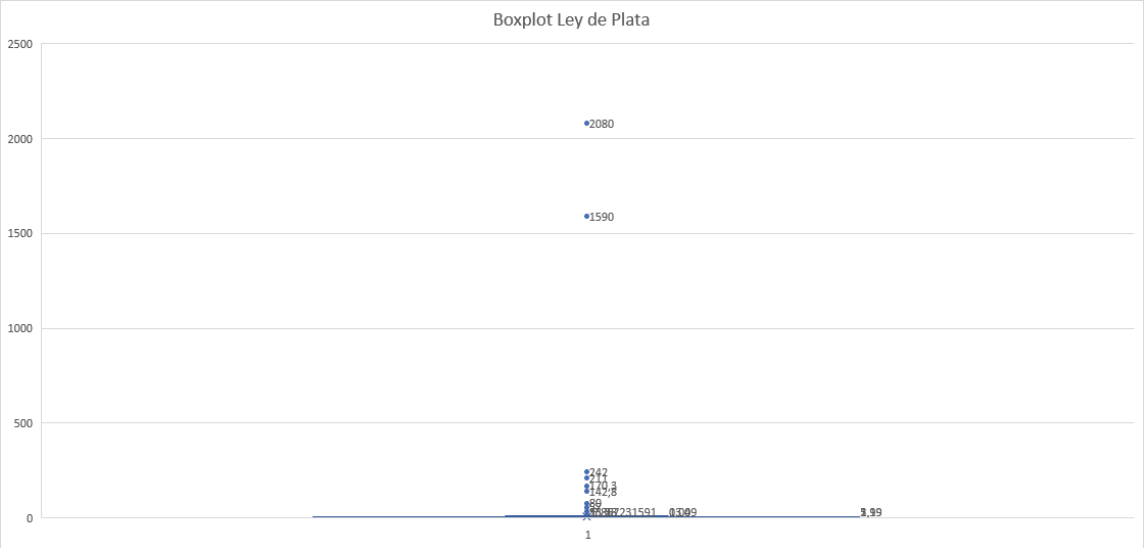

Ilustración A.2: Boxplot Ley de Plata (Elaboración propia)

Ley de Oro (%)	
Media	0.72
Error Típico	0.06
Mediana	0.08
Moda	0.03
Desviación estándar	4.07
Varianza de la muestra	16.54
Curtosis	1201.87
Coeficiente de asimetría	29
Rango	192.8
Mínimo	0.00
Máximo	192.8
Suma	3224.79
Datos	4495

Tabla A.2: Estadísticas Básicas Ley de Oro (Elaboración propia)

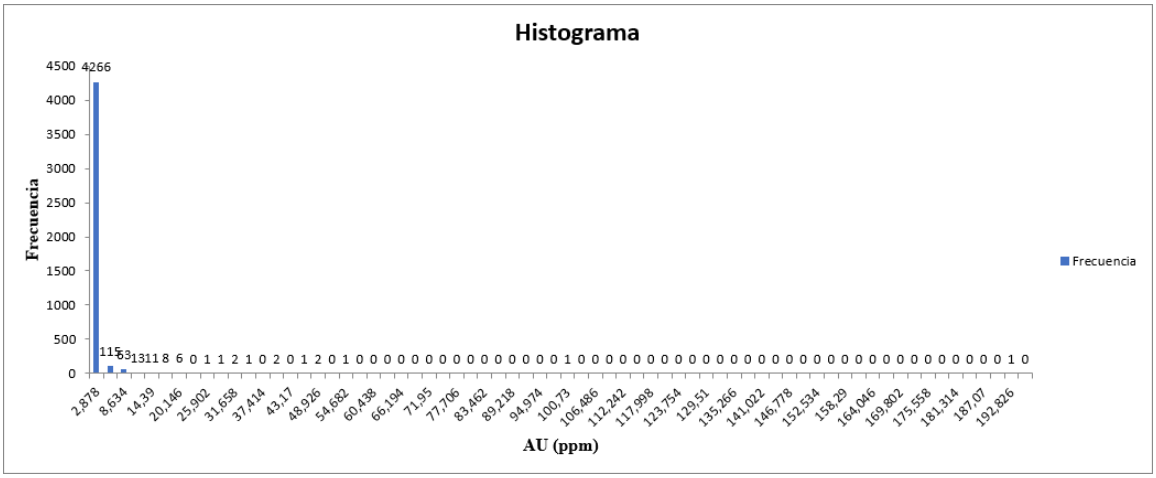

Ilustración A.3: Histograma Ley de Oro (Elaboración propia)

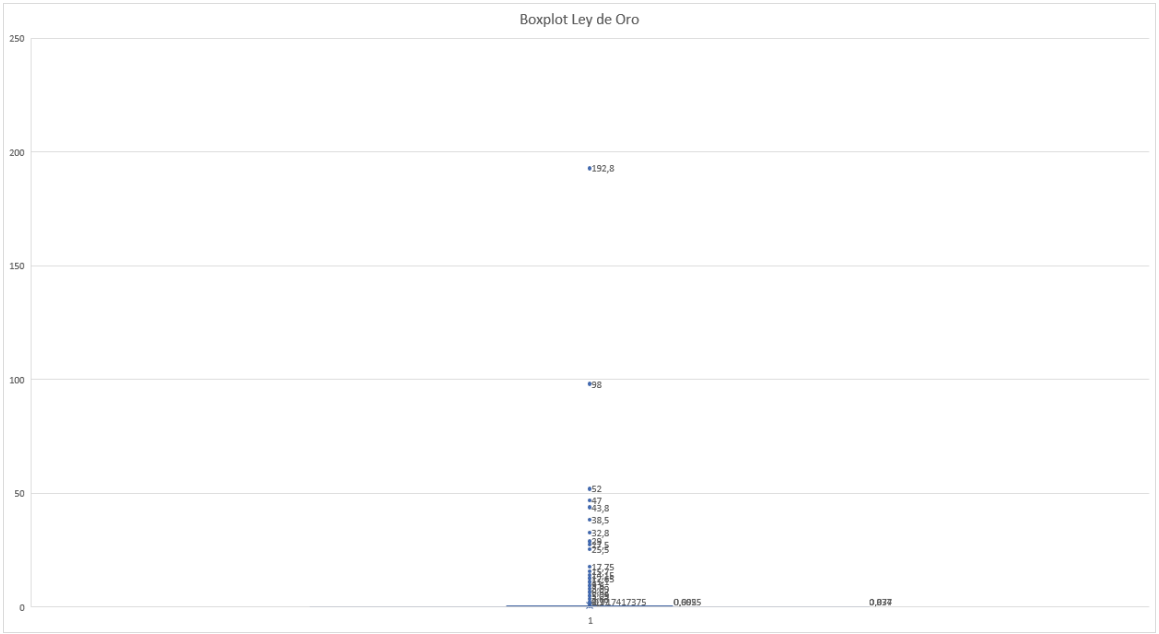

Ilustración A.4: Boxplot Ley de Oro (Elaboración propia)

Apéndice B: Variogramas experimentales y modelados

En el presente apéndice se encuentran desde la ilustración **B.1** hasta la ilustración **B.18** los variogramas experimentales y modelados para todos los ángulos estudiados de las litologías S2, HBX y Mixto.

Para litología S2:

Desde la ilustración **B.1** hasta **B.6** se encuentran los variogramas experimentales y modelado de la litología S2.

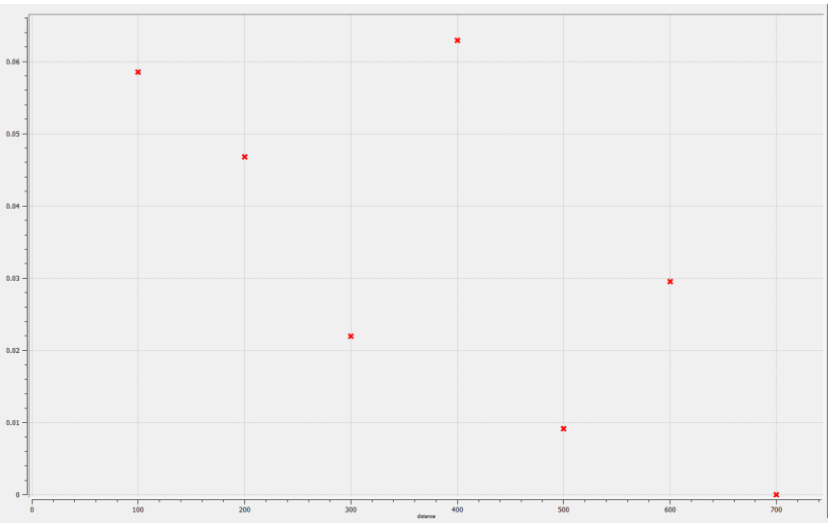

Ilustración B.1: Variograma experimental 0° para litología S2 (elaborado en SGeMS).

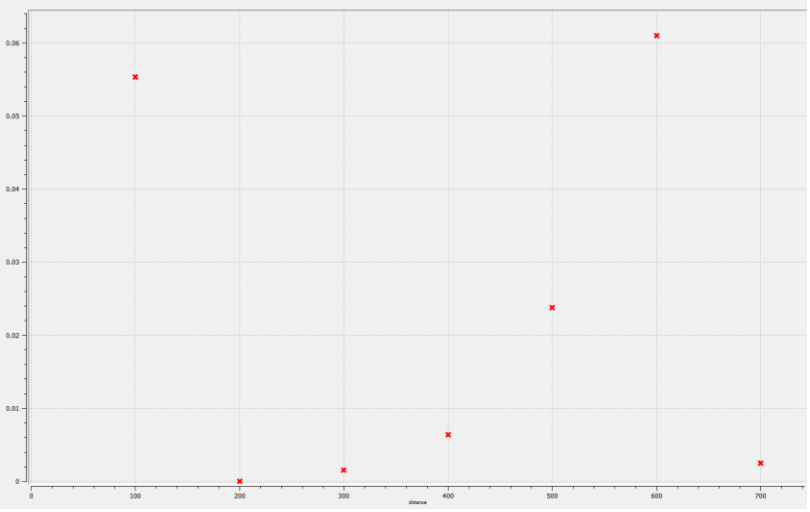

Ilustración B.2: Variograma experimental 45° para litología S2 (elaborado en SGeMS).

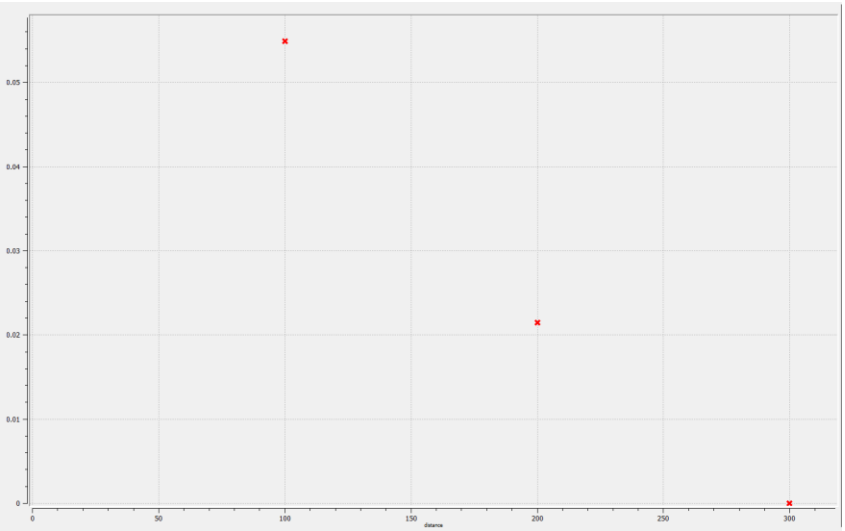

Ilustración B.3: Variograma experimental 90° para litología S2 (elaborado en SGeMS).

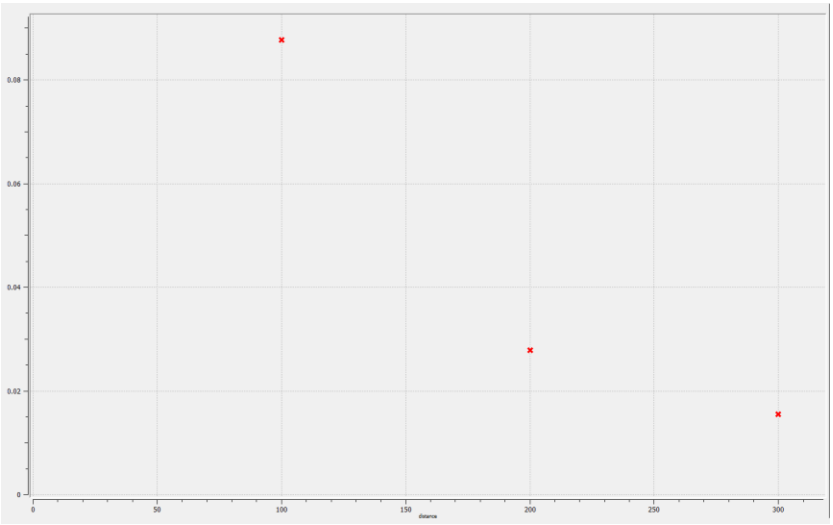

Ilustración B.4: Variograma experimental 135° para litología S2 (elaborado en SGeMS).

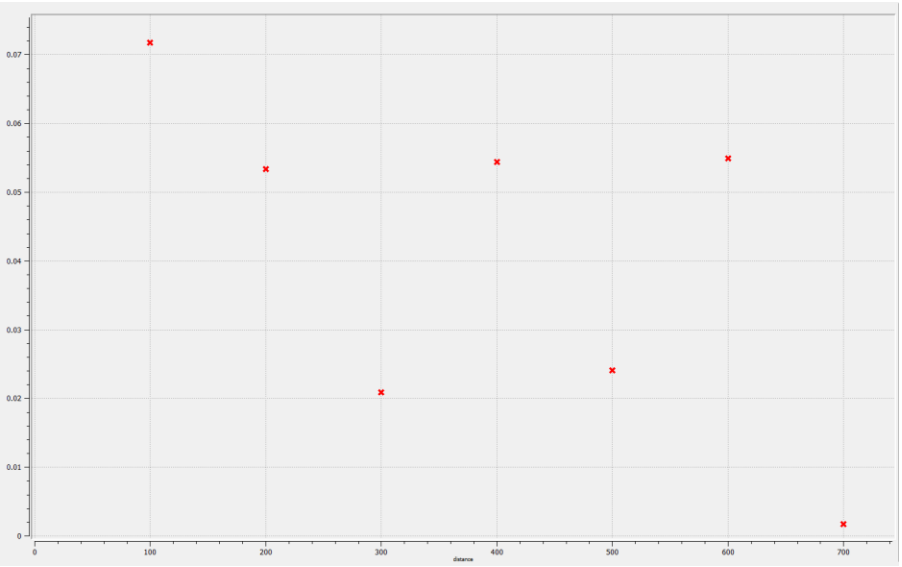

Ilustración B.5: Variograma experimental omnidireccional para litología S2 (elaborado en SGeMS).

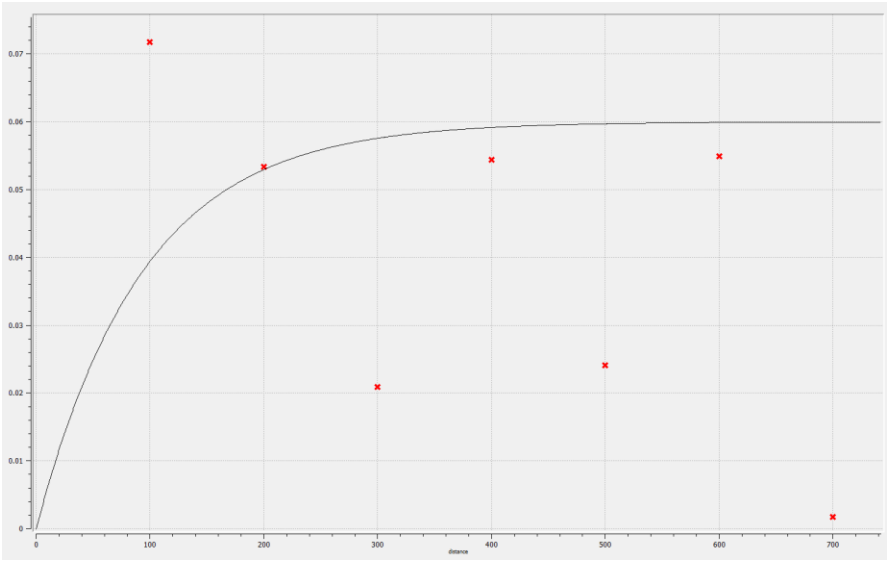

Ilustración B.6: Variograma modelado para litología S2 (elaborado en SGeMS).

Para litología HBX:

Desde la ilustración **B.7** hasta **B.12** se encuentran los variogramas experimentales y modelado de la litología HBX.

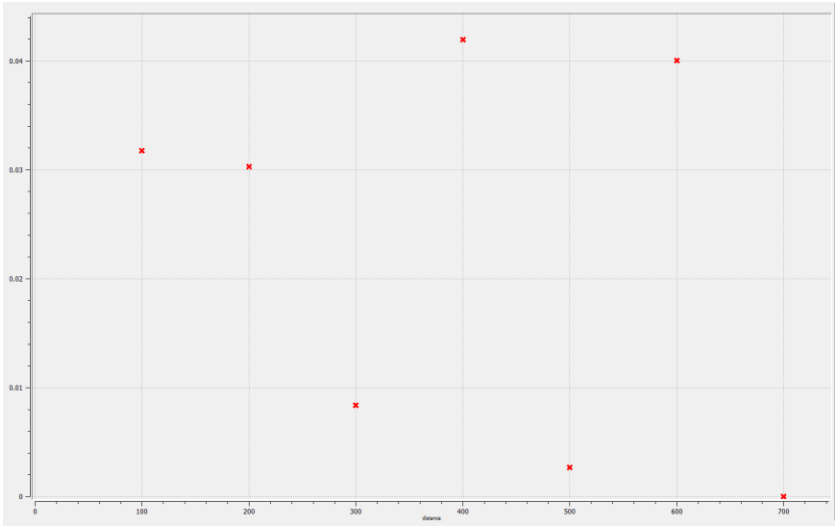

Ilustración B.7: Variograma experimental 0° para litología HBX (elaborado en SGeMS).

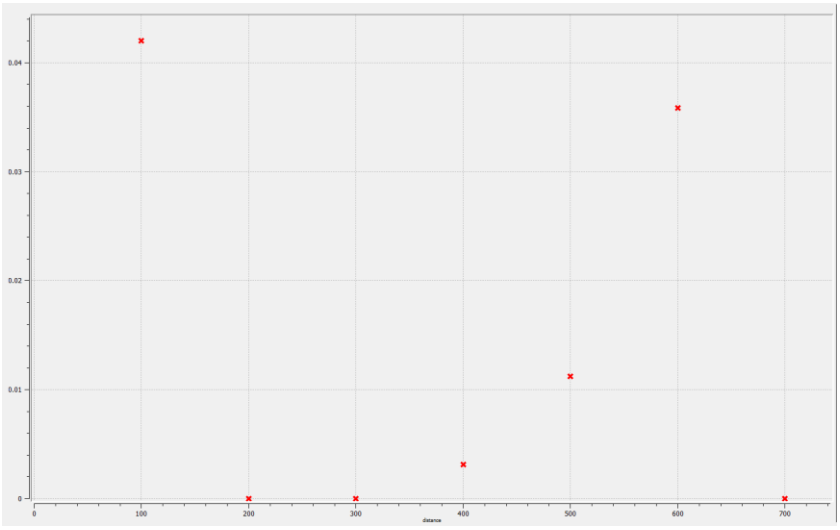

Ilustración B.8: Variograma experimental 45° para litología HBX (elaborado en SGeMS).

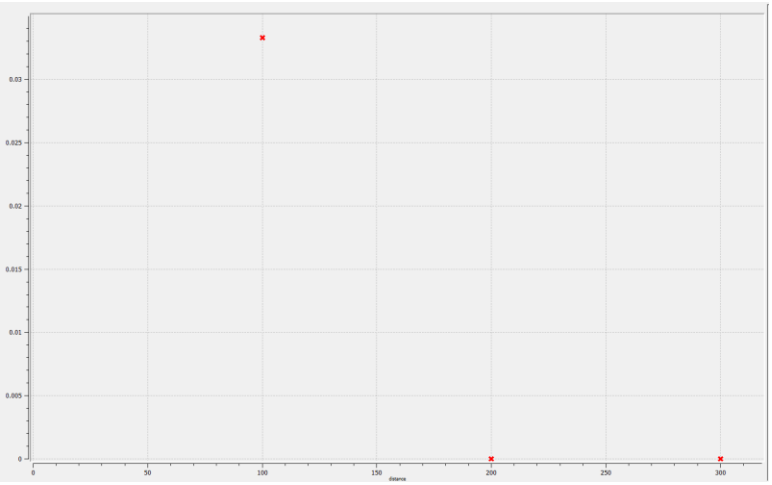

Ilustración B.9: Variograma experimental 90° para litología HBX (elaborado en SGeMS).

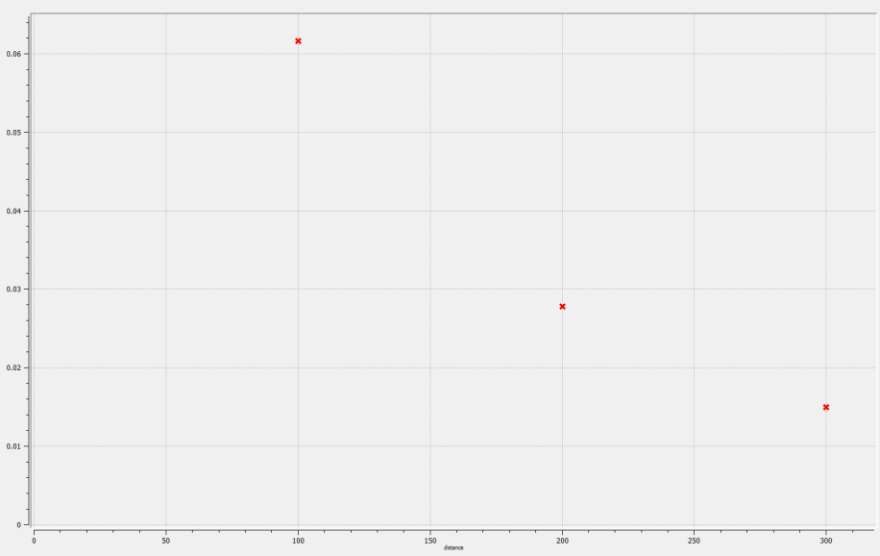

Ilustración B.10: Variograma experimental 135° para litología HBX (elaborado en SGeMS).

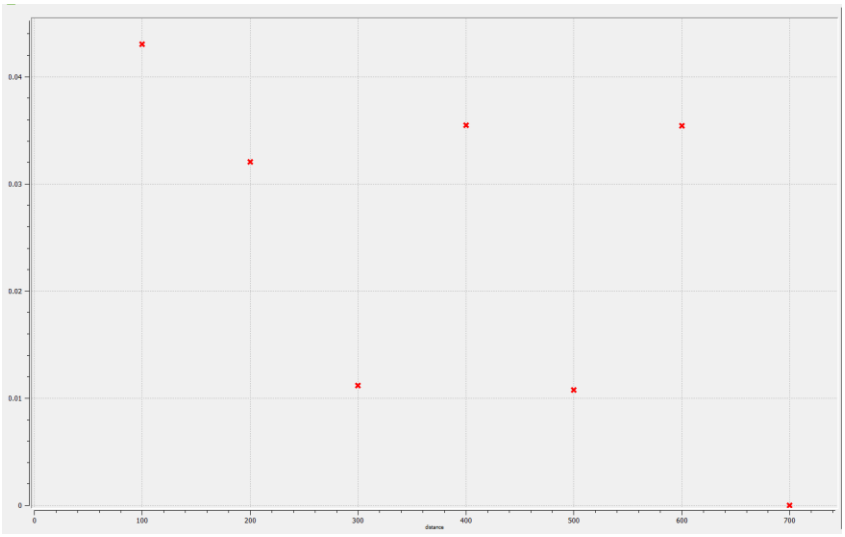

Ilustración B.11: Variograma experimental omnidireccional para litología HBX (elaborado en SGeMS).

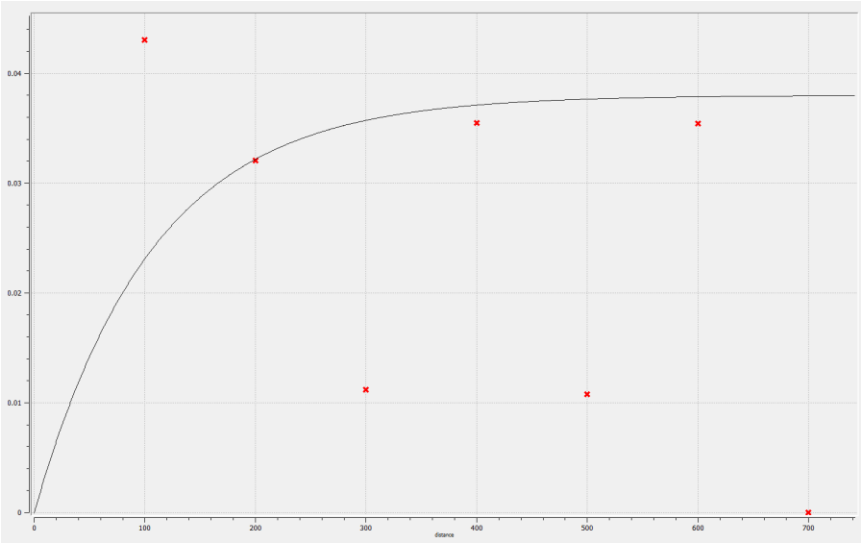

Ilustración B.12: Variograma modelado para litología HBX (elaborado en SGeMS).

Para litología Mixto:

Desde la ilustración **B.13** hasta **B.18** se encuentran los variogramas experimentales y modelado de la litología Mixto.

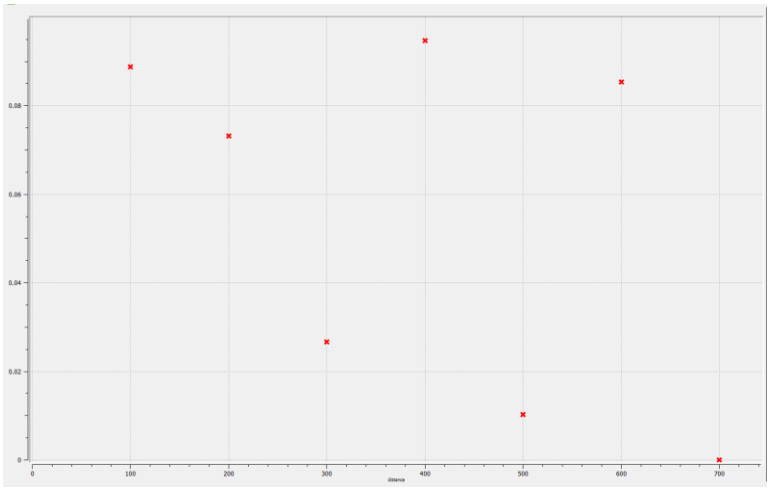

Ilustración B.13: Variograma experimental 0° para litología Mixto (elaborado en SGeMS).

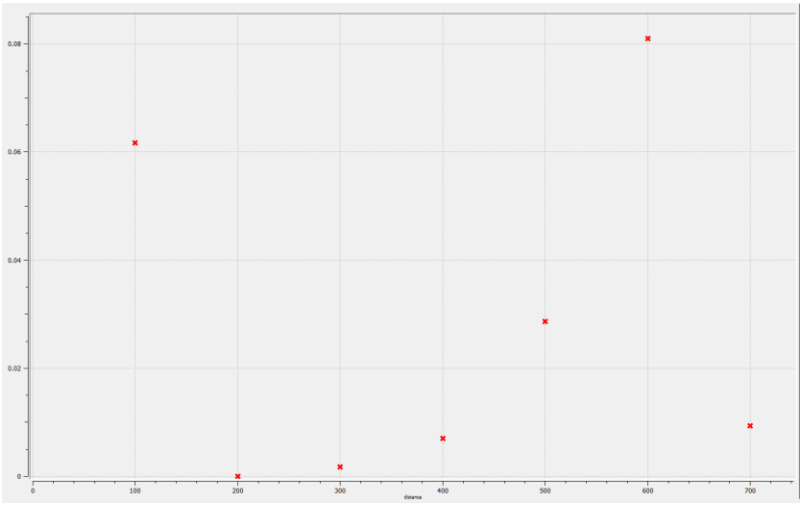

Ilustración B.14: Variograma experimental 45° para litología Mixto (elaborado en SGeMS).

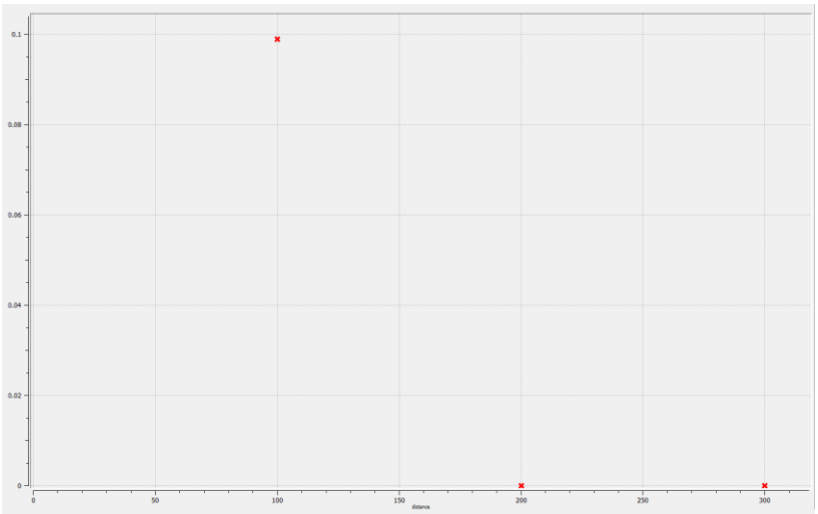

Ilustración B.15: Variograma experimental 90° para litología Mixto (elaborado en SGeMS).

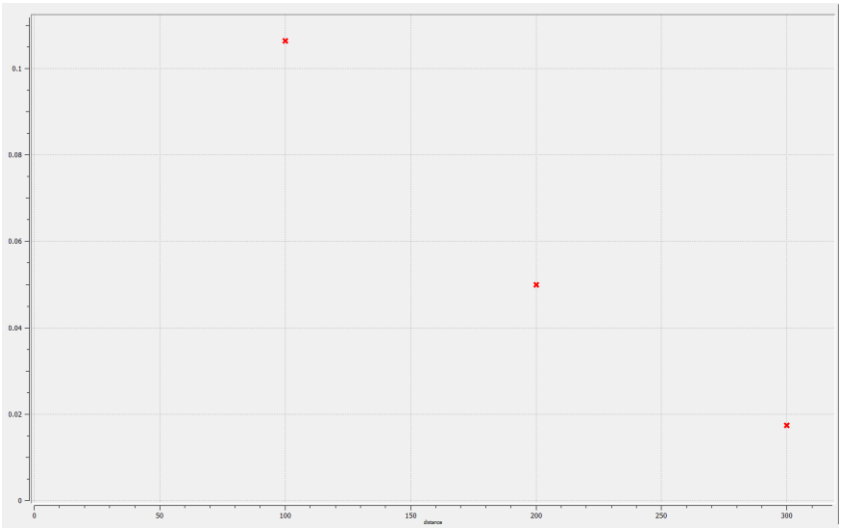

Ilustración B.16: Variograma experimental 135° para litología Mixto (elaborado en SGeMS).

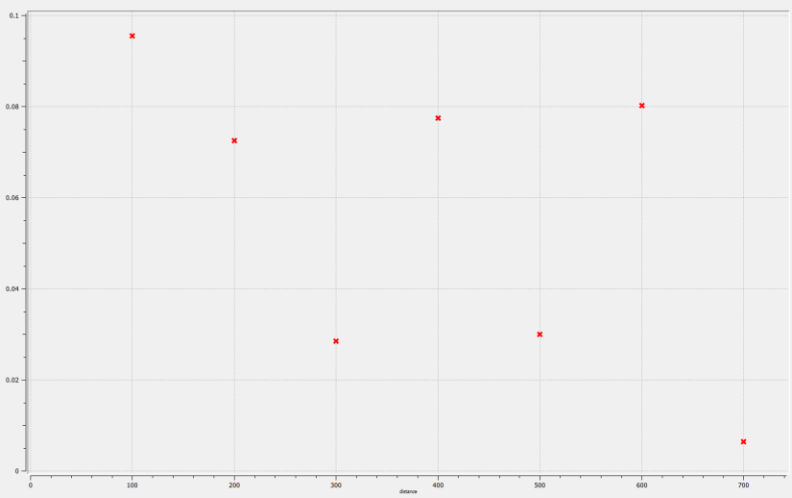

Ilustración B.17: Variograma experimental omnidireccional para litología Mixto (elaborado en SGeMS).

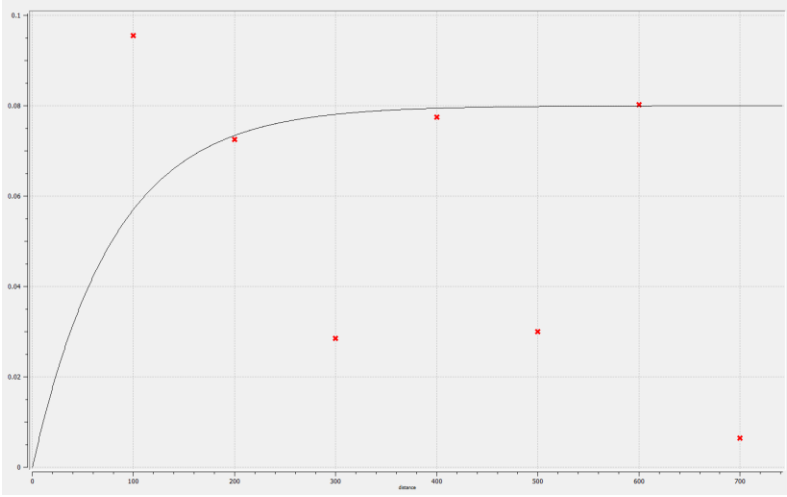

Ilustración B.18: Variograma modelado para litología Mixto (elaborado en SGeMS).

Apéndice C: Resultados de estimaciones

En el presente apéndice C se tiene los resultados de las estimaciones llevadas a cabo con los 3 métodos propuestos, los cuales son Kriging de indicadores, redes neuronales artificiales y regresión logística para las litologías HBX y de tipo Mixto.

Desde la ilustración **C.1** hasta la ilustración **C.6** se tiene los resultados del Kriging de indicadores.

Para Kriging de indicadores:

Ilustración C.1: Vista de frente Kriging de indicadores para litología HBX (elaborado en SGeMS).

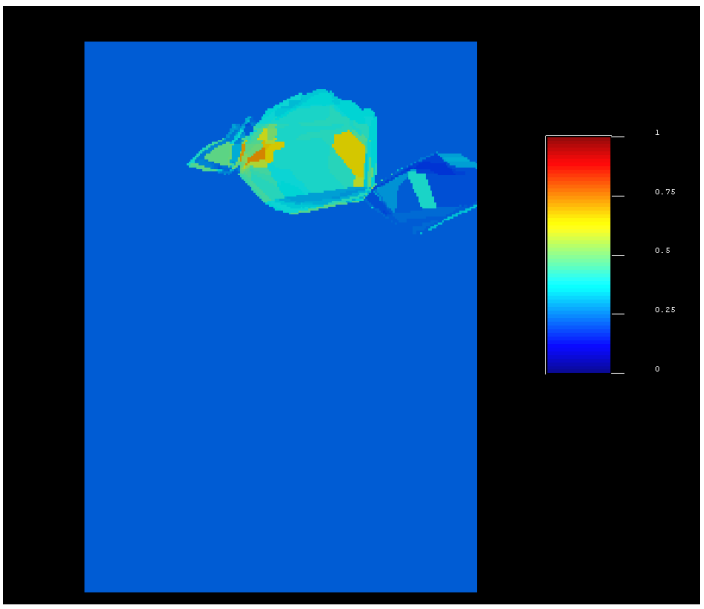

Ilustración C.2: Vista en planta Kriging de indicadores para litología HBX (elaborado en SGeMS).

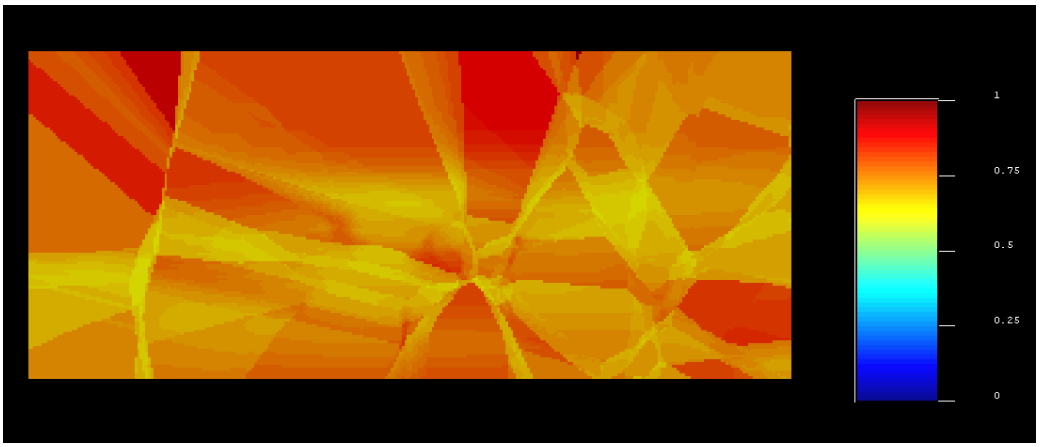

Ilustración C.3: Varianza Kriging de indicadores para litología HBX (elaborado en SGeMS).

Ilustración C.4: Vista de frente Kriging de indicadores para litología Mixto (elaborado en SGeMS).

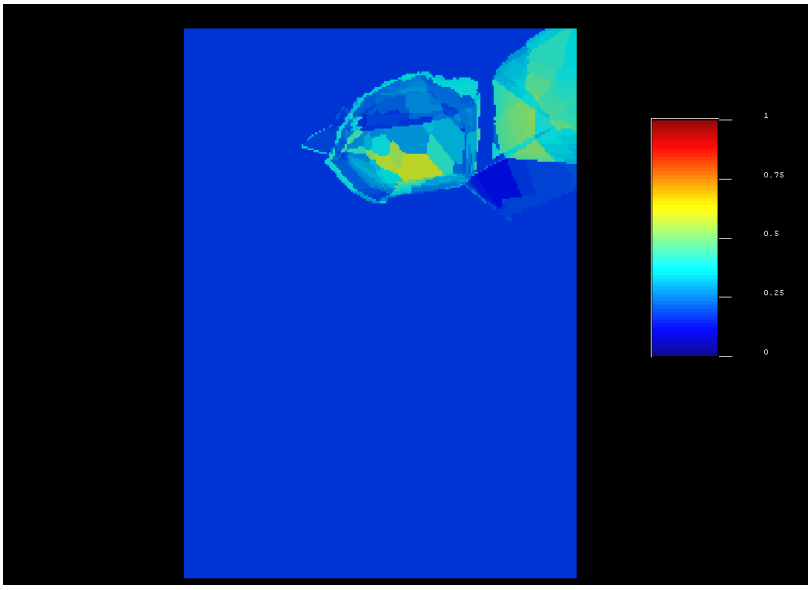

Ilustración C.5: Vista en planta Kriging de indicadores para litología Mixto (elaborado en SGeMS).

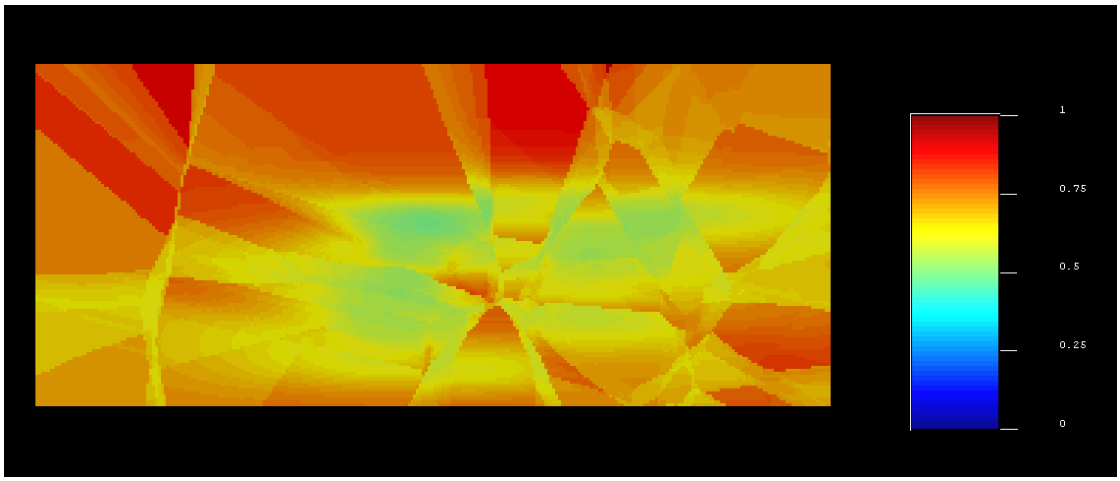

Ilustración C.6: Varianza Kriging de indicadores para litología Mixto (elaborado en SGeMS).

Desde la ilustración **C.7** hasta la ilustración **C.14** se tiene los resultados mediante redes neuronales artificiales con los conjuntos de datos de entrenamiento y prueba.

Para Redes Neuronales Artificiales y litología HBX:

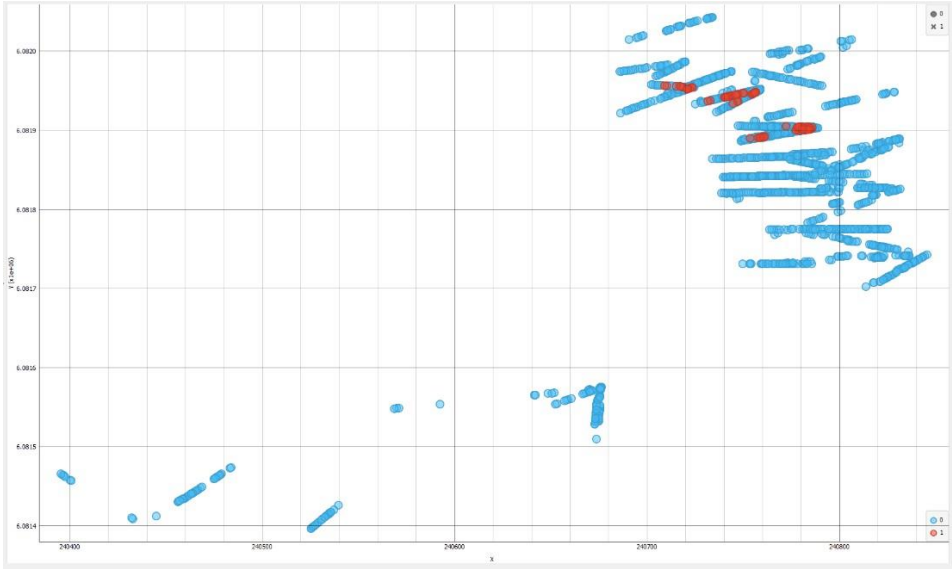

Ilustración C.7: Plano XY redes neuronales datos de entrenamiento para litología HBX (elaborado en Orange Canvas).

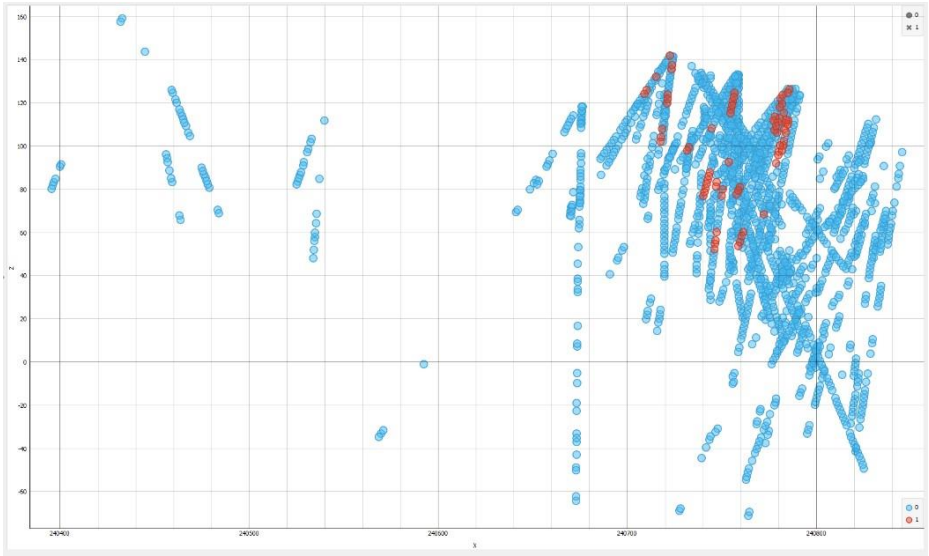

Ilustración C.8: Plano XZ redes neuronales datos de entrenamiento para litología HBX (elaborado en Orange Canvas).

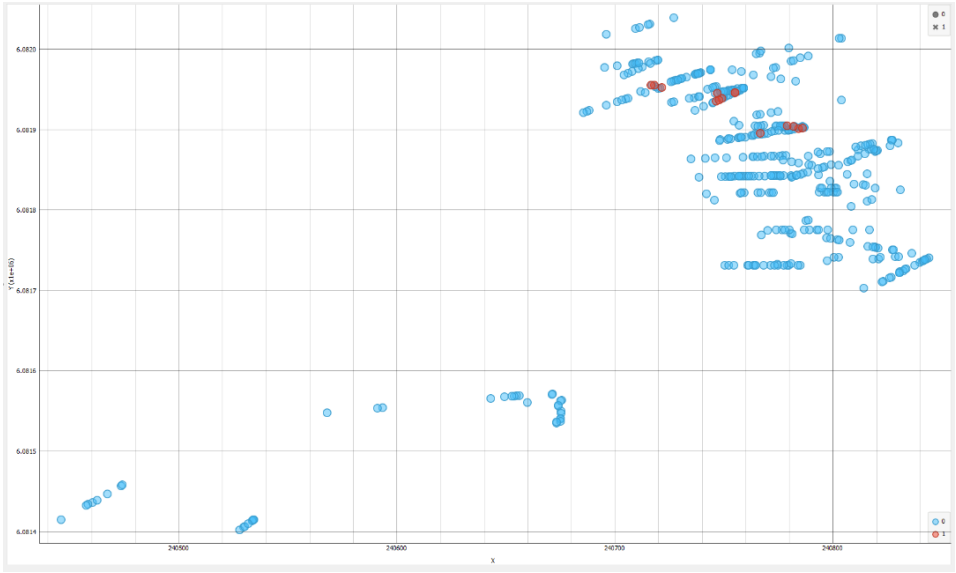

Ilustración C.9: Plano XY redes neuronales datos de prueba para litología HBX (elaborado en Orange Canvas).

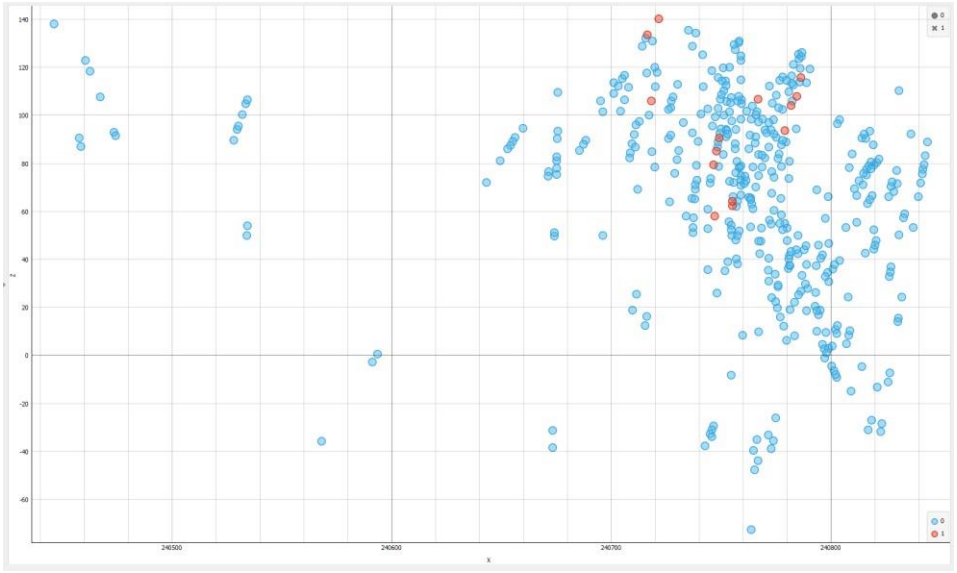

Ilustración C.10: Plano XZ redes neuronales datos de prueba para litología HBX (elaborado en Orange Canvas).

Para Redes Neuronales Artificiales y litología Mixto:

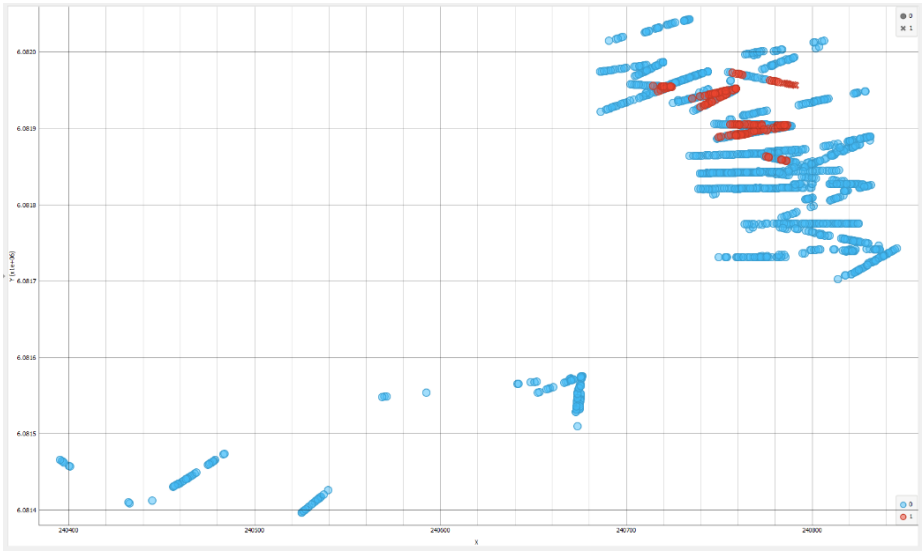

Ilustración C.11: Plano XY redes neuronales datos de entrenamiento para litología Mixto (elaborado en Orange Canvas).

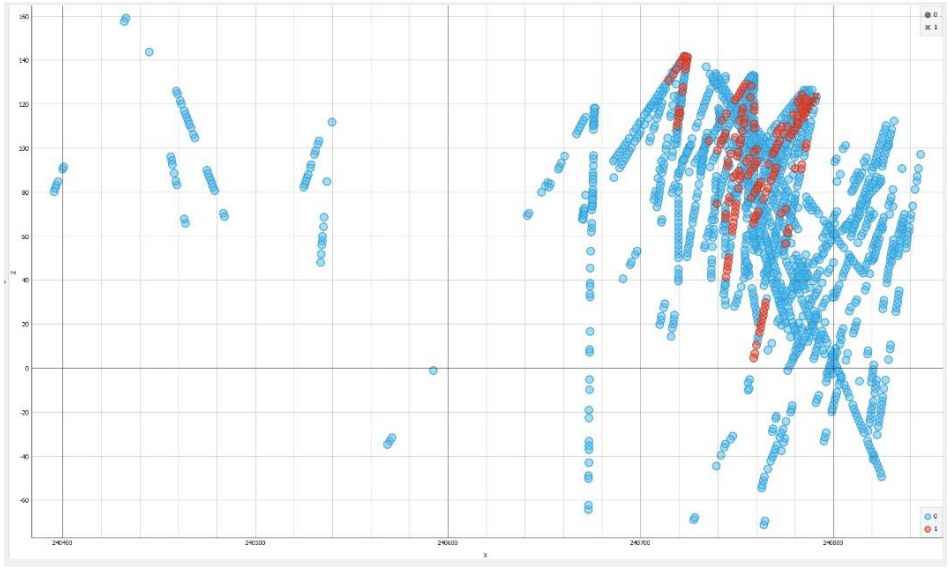

Ilustración C.12: Plano XZ redes neuronales datos de entrenamiento para litología Mixto (elaborado en Orange Canvas).

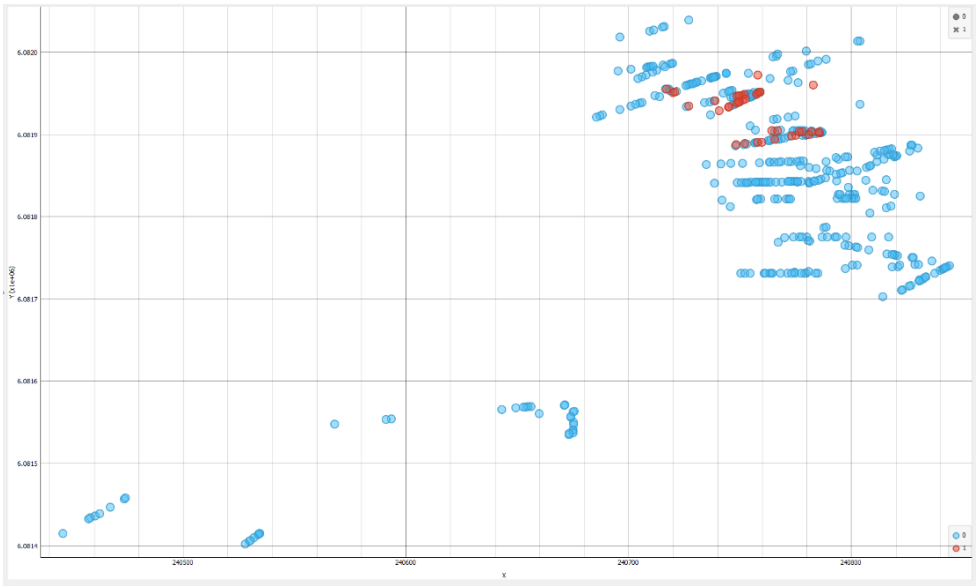

Ilustración C.13: Plano XY redes neuronales datos de prueba para litología Mixto (elaborado en Orange Canvas).

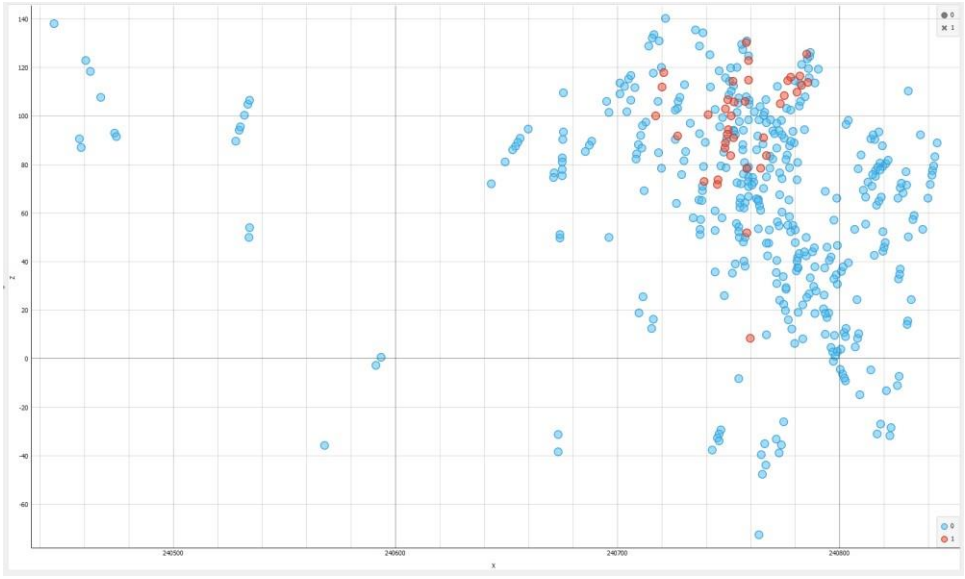

Ilustración C.14: Plano XZ redes neuronales datos de prueba para litología Mixto (elaborado en Orange Canvas).

Desde la ilustración **C.15** hasta la ilustración **C.22** se tiene los resultados mediante regresión logística para el conjunto de datos de entrenamiento y prueba.

Para Regresión Logística y litología HBX:

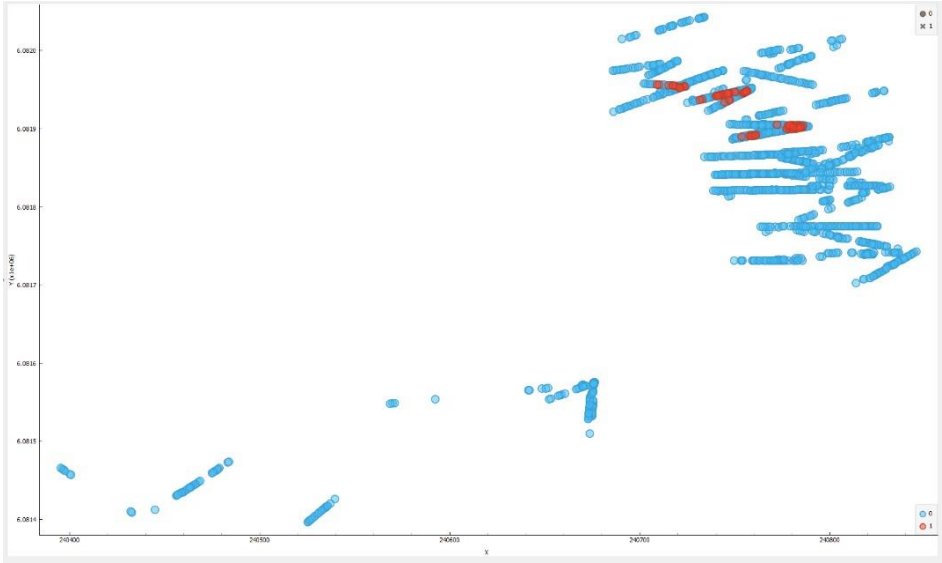

Ilustración C.15: Plano XY regresión logística datos de entrenamiento para litología HBX (elaborado en Orange Canvas).

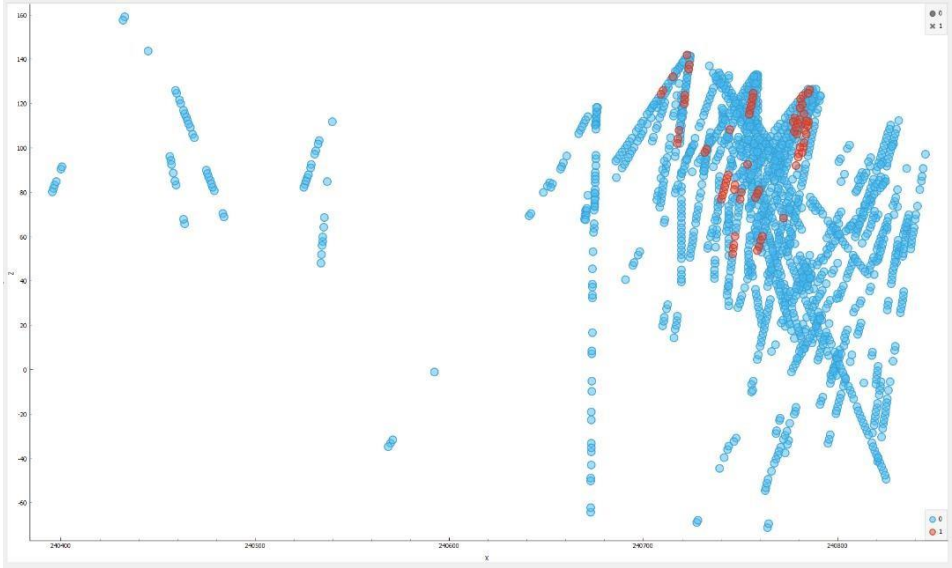

Ilustración C.16: Plano XZ regresión logística datos de entrenamiento para litología HBX (elaborado en Orange Canvas).

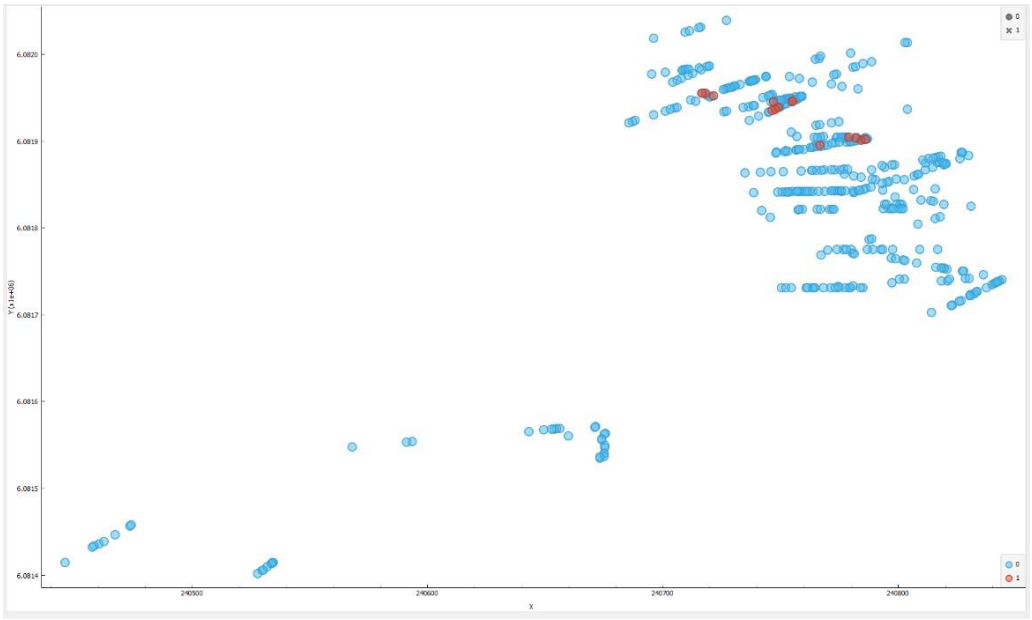

Ilustración C.17: Plano XY regresión logística datos de prueba para litología HBX (elaborado en Orange Canvas).

Ilustración C.18: Plano XZ regresión logística datos de prueba para litología HBX (elaborado en Orange Canvas).

Para Regresión Logística y litología Mixto:

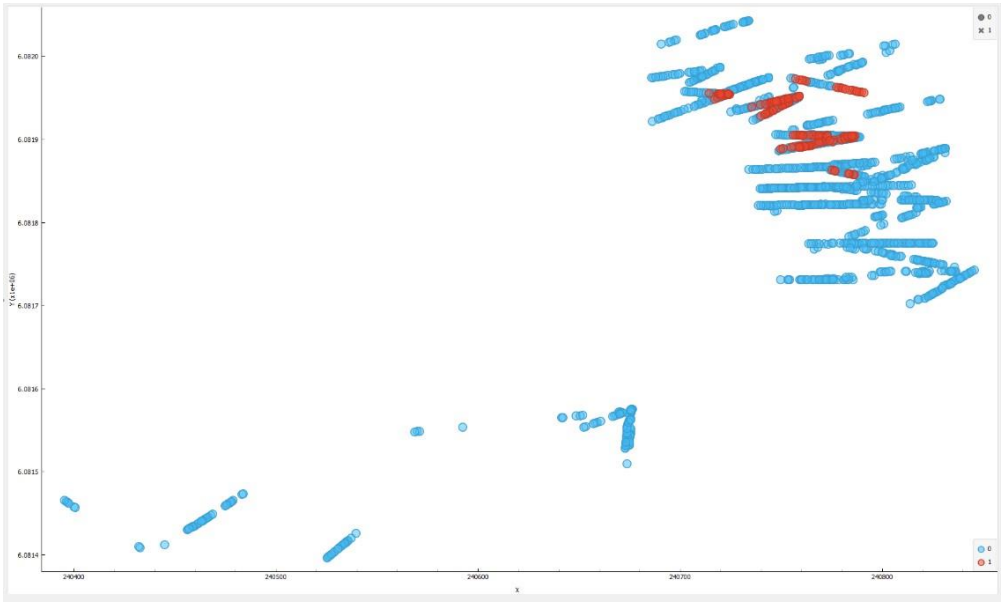

Ilustración C.19: Plano XY regresión logística datos de entrenamiento para litología Mixto (elaborado en Orange Canvas).

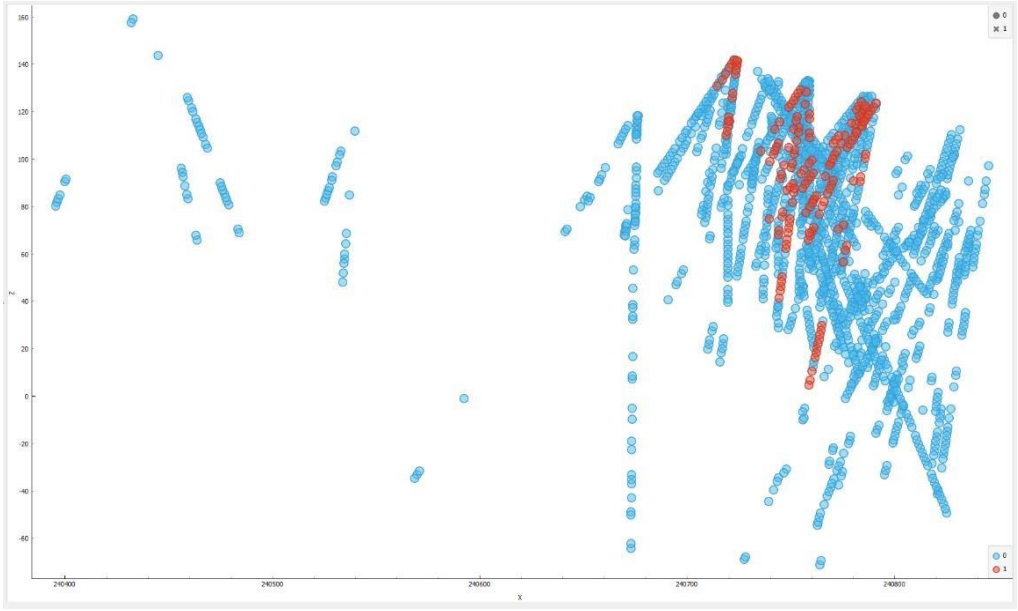

Ilustración C.20: Plano XZ regresión logística datos de entrenamiento para litología Mixto (elaborado en Orange Canvas).

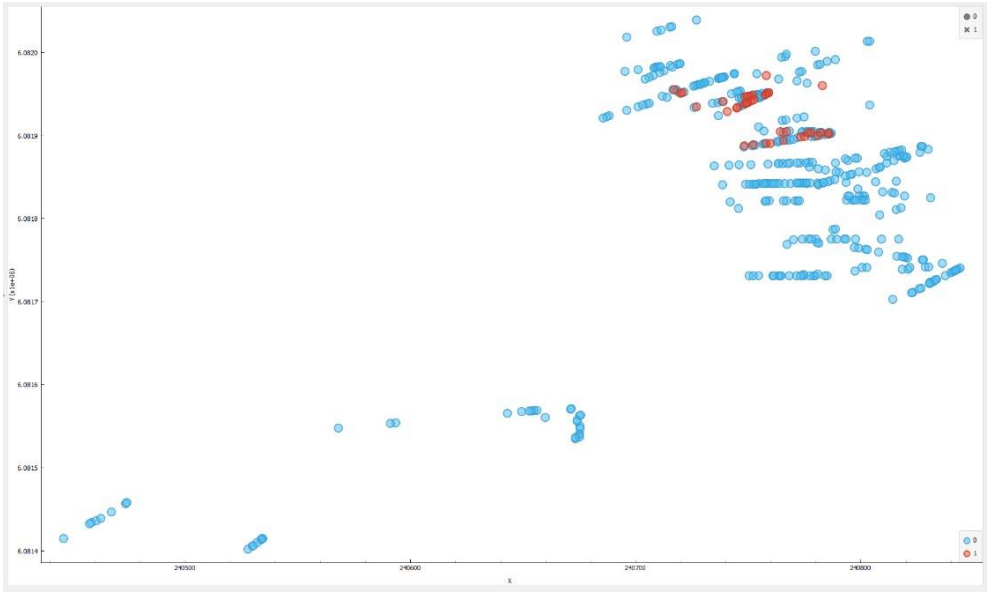

Ilustración C.21: Plano XY regresión logística datos de prueba para litología Mixto (elaborado en Orange Canvas).

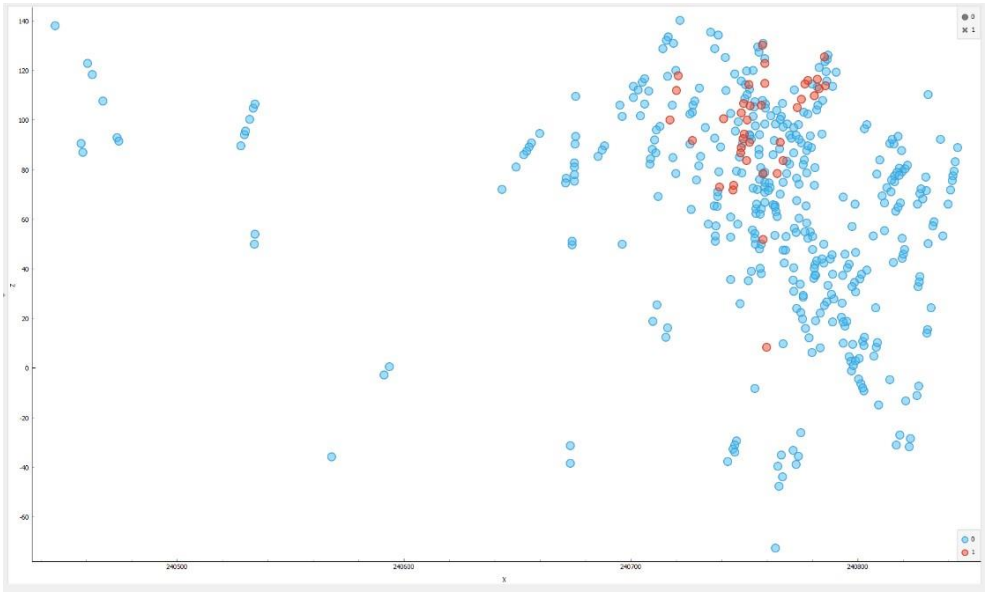

Ilustración C.22: Plano XZ regresión logística datos de prueba para litología Mixto (elaborado en Orange Canvas).

Apéndice D: Matriz de confusión y resultados estadísticos litologías

En el presente apéndice se verán los resultados obtenidos mediante la matriz de confusión y métricas de desempeños para los 3 métodos propuestos y las litologías HBX y Mixto.

Para Kriging de Indicadores:

Desde la tabla **D.1** hasta la tabla **D.4** se presentan la matriz de confusión y métricas de desempeño mediante el método Kriging de indicadores.

		Predicción		
		0	1	Total
Real	0	1908	67	1975
	1	129	57	186
	Total	2037	124	2161

Tabla D.1: Matriz de confusión Kriging de indicadores para litología HBX (elaboración propia).

Kriging de Indicadores				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
HBX	0.90	0.45	0.3	0.96

Tabla D.2: Resultados estadísticos Kriging de indicadores para litología HBX (elaboración propia).

		Predicción		
		0	1	Total
Real	0	1767	48	1815
	1	151	195	346
	Total	1918	243	2161

Tabla D.3: Matriz de confusión Kriging de indicadores para litología Mixto (elaboración propia).

Kriging de Indicadores				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
Mixto	0.90	0.8	0.56	0.97

Tabla D.4: Resultados estadísticos Kriging de indicadores para litología Mixto (elaboración propia).

Para Redes Neuronales Artificiales:

Desde la tabla **D.5** hasta la tabla **D.12** se presentan la matriz de confusión y métricas de desempeño mediante el método redes neuronales artificiales para el conjunto de datos de entrenamiento y prueba.

Para litología HBX:

		Predicción		
		0	1	Total
Real	0	1598	15	1613
	1	57	59	116
	Total	1655	74	1729

Tabla D.5: Matriz de confusión datos de entrenamiento para litología HBX (elaboración propia).

Redes Neuronales				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
HBX	0.95	0.79	0.5	0.99

Tabla D.6: Resultados estadísticos datos de entrenamiento para litología HBX (elaboración propia).

		Predicción		
		0	1	Total
Real	0	408	6	414
	1	8	10	18
	Total	416	16	432

Tabla D.7: Matriz de confusión datos de prueba para litología HBX (elaboración propia).

Redes Neuronales				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
HBX	0.96	0.62	0.55	0.98

Tabla D.8: Resultados estadísticos datos de prueba para litología HBX (elaboración propia).

Para litología Mixto:

		Predicción		
		0	1	Total
Real	0	1514	21	1535
	1	135	59	194
	Total	1649	80	1729

Tabla D.9: Matriz de confusión datos de entrenamiento para litología Mixto (elaboración propia).

Redes Neuronales				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
Mixto	0.90	0.73	0.3	0.98

Tabla D.10: Resultados estadísticos datos de entrenamiento para litología Mixto (elaboración propia).

		Predicción		
		0	1	Total
Real	0	339	8	347
	1	30	55	85
	Total	369	63	432

Tabla D.11: Matriz de confusión datos de prueba para litología Mixto (elaboración propia).

Redes Neuronales				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
Mixto	0.91	0.87	0.64	0.97

Tabla D.12: Resultados estadísticos datos de prueba para litología Mixto (elaboración propia).

Para Regresión Logística:

Desde la tabla **D.13** hasta la tabla **D.20** se presentan la matriz de confusión y métricas de desempeño mediante el método regresión logística para el conjunto de datos de entrenamiento y prueba.

Para litología HBX:

		Predicción		
		0	1	Total
Real	0	1619	21	1640
	1	51	38	89
	Total	1670	59	1729

Tabla D.13: Matriz de confusión datos de entrenamiento para litología HBX (elaboración propia).

Regresión Logística				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
HBX	0.95	0.64	0.42	0.98

Tabla D.14: Resultados estadísticos datos de entrenamiento para litología HBX (elaboración propia).

		Predicción		
		0	1	Total
Real	0	411	2	413
	1	12	7	19
	Total	423	9	432

Tabla D.15: Matriz de confusión datos de prueba para litología HBX (elaboración propia).

Regresión Logística				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
HBX	0.96	0.77	0.36	0.99

Tabla D.16: Resultados estadísticos datos de prueba para litología HBX (elaboración propia).

Para litología Mixto:

		Predicción		
		0	1	Total
Real	0	1522	17	1539
	1	146	44	190
	Total	1668	61	1729

Tabla D.17: Matriz de confusión datos de entrenamiento para litología Mixto (elaboración propia).

Regresión Logística				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
Mixto	0.90	0.72	0.23	0.98

Tabla D.18: Resultados estadísticos datos de entrenamiento para litología Mixto (elaboración propia).

		Predicción		
		0	1	Total
Real	0	351	7	358
	1	31	43	74
	Total	382	50	432

Tabla D.19: Matriz de confusión datos de prueba para litología Mixto (elaboración propia).

Regresión Logística				
Litología	Exactitud	Precisión	Sensibilidad	Especificidad
Mixto	0.91	0.86	0.58	0.98

Tabla D.20: Resultados estadísticos datos de prueba para litología Mixto (elaboración propia).