

TABLA DE CONTENIDOS

	página
Dedicatoria	I
Agradecimientos	II
Tabla de Contenidos	IV
Índice de Figuras	VII
Índice de Tablas	VIII
Resumen	X
1. Introducción	11
1.1. Descripción de la propuesta	11
1.2. Hipótesis	13
1.3. Objetivos	13
1.4. Alcances	14
2. Revisión Literaria	15
2.1. Definiciones fundamentales de minería de datos	15
2.2. Tipos de aprendizaje automático	18
2.3. Medidas de desempeño y validación de datos	21
2.4. Reducción de dimensionalidad	24
2.5. Algoritmos de clasificación para predicción	25
3. Metodología experimental	30
3.1. Origen de datos	30
3.2. Limpieza de datos	30
3.3. Transformación de atributos	30
3.4. Medida de desempeño	31
3.5. Método de validación	31

4. Conjunto de datos	32
4.1. Conjunto de datos iniciales	32
4.2. Conjunto de datos final	37
5. Análisis del Problema	42
5.1. Alternativas de solución	42
5.2. Clasificador Naive Bayes	42
5.3. Análisis de componentes principales	46
5.4. Diseño de la solución	51
6. Resultados experimentales	54
6.1. Introducción	54
6.2. Predicción usando todas las superficies	54
6.3. Predicción usando superficie de arcilla	56
6.4. Predicción usando superficie de pasto	58
6.5. Predicción usando superficie de cemento	59
6.6. Predicción Roland Garros 2015 con datos de entrenamiento 2012 . . .	61
6.7. Análisis de Resultados	62
7. Conclusiones y trabajo futuro	64
7.1. Conclusiones	64
7.2. Trabajo Futuro	65
Bibliografía	67
Glosario	69
Anexos	
A: Metodología de Desarrollo	71
A.1. Implementación	71
A.2. Control de Versiones	72
A.3. Scrum	74
B: Arquitectura de la aplicación	76
C: Medidas de desempeño clasificador Naive Bayes	79

D: Implementación de Análisis de Componentes Principales en R	81
E: Screenshots sistema de predicción para encuentros de tenis	83

ÍNDICE DE FIGURAS

	página
2.1. Los procesos de minería de datos.	16
2.2. Representación de una Red Neuronal Artificial.	26
4.1. Desempeño jugadores ATP temporada 2012 desde el sitio web oficial.	32
4.2. Ejemplo de salida del script de limpieza y transformación de datos de desempeño jugadores 2012.	33
4.3. Ejemplo de enfrentamientos de Novak Djokovic temporada 2012. . . .	36
4.4. Ejemplo de salida del script de limpieza y transformación de datos de enfrentamientos temporada 2012.	36
4.5. Ejemplo del formato para el conjunto final de datos temporada 2012.	41
5.1. Componentes principales para el conjunto de datos (todas las super- ficies) de los mejores 200 jugadores del ATP durante el año 2012. . . .	46
5.2. Dispersión de los componentes principales para enfrentamientos du- rante la temporada 2012.	48
5.3. Dispersión de los componentes principales en 3D para enfrentamientos durante la temporada 2012.	50
5.4. Diseño solución del sistema.	52
B.1. Representación de la arquitectura de la aplicación.	77
E.1. Vista principal del sistema de predicción para encuentros de tenis profesional.	84
E.2. Ingreso de jugadores y tipo de superficie al sistema de predicción para encuentros de tenis profesional.	84
E.3. Vista de resultados sistema de predicción para encuentros de tenis profesional.	85

ÍNDICE DE TABLAS

	página
2.1. Matriz de confusión para el cálculo de medidas de desempeño.	21
2.2. Medidas de desempeño para clasificación binaria.	22
2.3. Medidas de desempeño para clasificación multi-clase.	23
2.4. Cuadro comparativo de algoritmos de clasificación.	29
4.1. Agrupación del ranking de jugadores en categorías normalizadas. . . .	38
5.1. Varianzas acumuladas PCA para cada componente principal del conjunto de enfrentamientos en todas las superficies para la temporada 2012.	47
6.1. Validación cruzada enfrentamientos 2012 en todas las superficies (10 folds).	54
6.2. Validación con enfrentamientos de 2012 para conjunto de entrenamiento y enfrentamientos de 2013 para conjunto de prueba en todas las superficies.	55
6.3. Validación con enfrentamientos de 2012 para conjunto de entrenamiento y enfrentamientos de 2014 para conjunto de prueba en todas las superficies.	55
6.4. Validación cruzada enfrentamientos de 2012 en arcilla (10 folds). . . .	56
6.5. Validación con enfrentamientos de 2012 para conjunto de entrenamiento y enfrentamientos de 2013 para conjunto de prueba en arcilla.	56
6.6. Validación con enfrentamientos de 2012 para conjunto de entrenamiento y enfrentamientos de 2014 para conjunto de prueba en arcilla.	57
6.7. Validación cruzada enfrentamientos de 2012 en pasto (10 folds). . . .	58
6.8. Validación con enfrentamientos de 2012 para conjunto de entrenamiento y enfrentamientos de 2013 para conjunto de prueba en pasto.	58
6.9. Validación con enfrentamientos de 2012 para conjunto de entrenamiento y enfrentamientos de 2014 para conjunto de prueba en pasto.	59
6.10. Validación cruzada enfrentamientos de 2012 en cemento (10 folds). . . .	59
6.11. Validación con enfrentamientos de 2012 para conjunto de entrenamiento y enfrentamientos de 2013 para conjunto de prueba en cemento.	60

6.12. Validación con enfrentamientos de 2012 para conjunto de entrenamiento y enfrentamientos de 2014 para conjunto de prueba en cemento.	60
6.13. Resultados predicción de encuentros campeonato Roland Garros 2015.	61
C.1. Matriz de confusión validación cruzada con el conjunto de 17 atributos durante la temporada 2012.	79
C.2. Medidas de desempeño con el conjunto de 17 atributos durante la temporada 2012.	79