

# Identificación de tráfico de red basado en Aprendizaje Automático

Santiago Egea Gómez

([santiago.egea@alumnos.uva.es](mailto:santiago.egea@alumnos.uva.es))

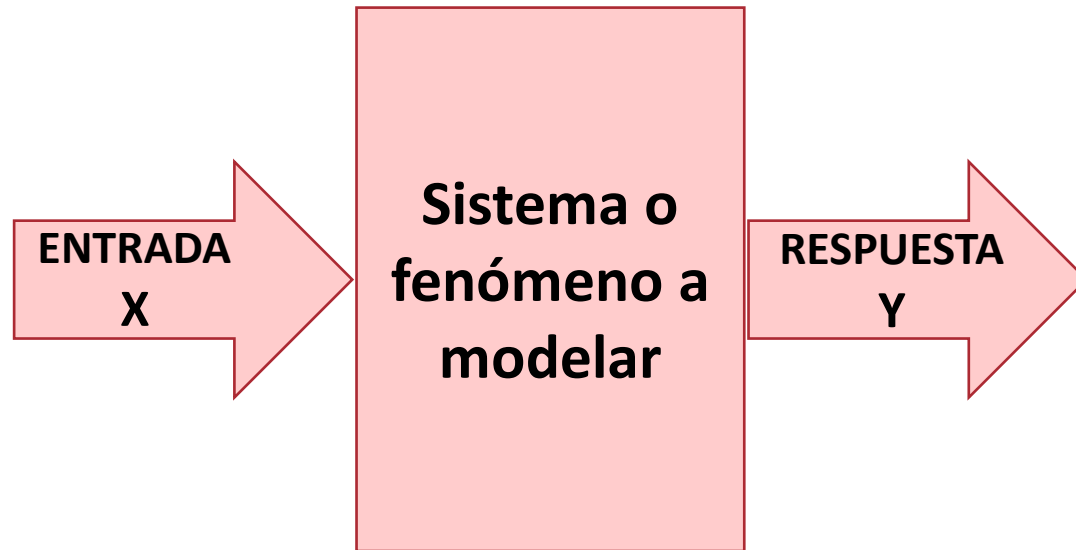
Universidad de Valladolid

Jornadas Técnicas RedIRIS – 15 de Junio 2017

# ¿Qué es el Aprendizaje Automático?

- Concepto y definición
- Árboles de decisión
- Algoritmos ensamblados

# Concepto y definición



- Rama de la IA
- Objetivo: Otorgar a las máquinas la capacidad de aprender por sí solas

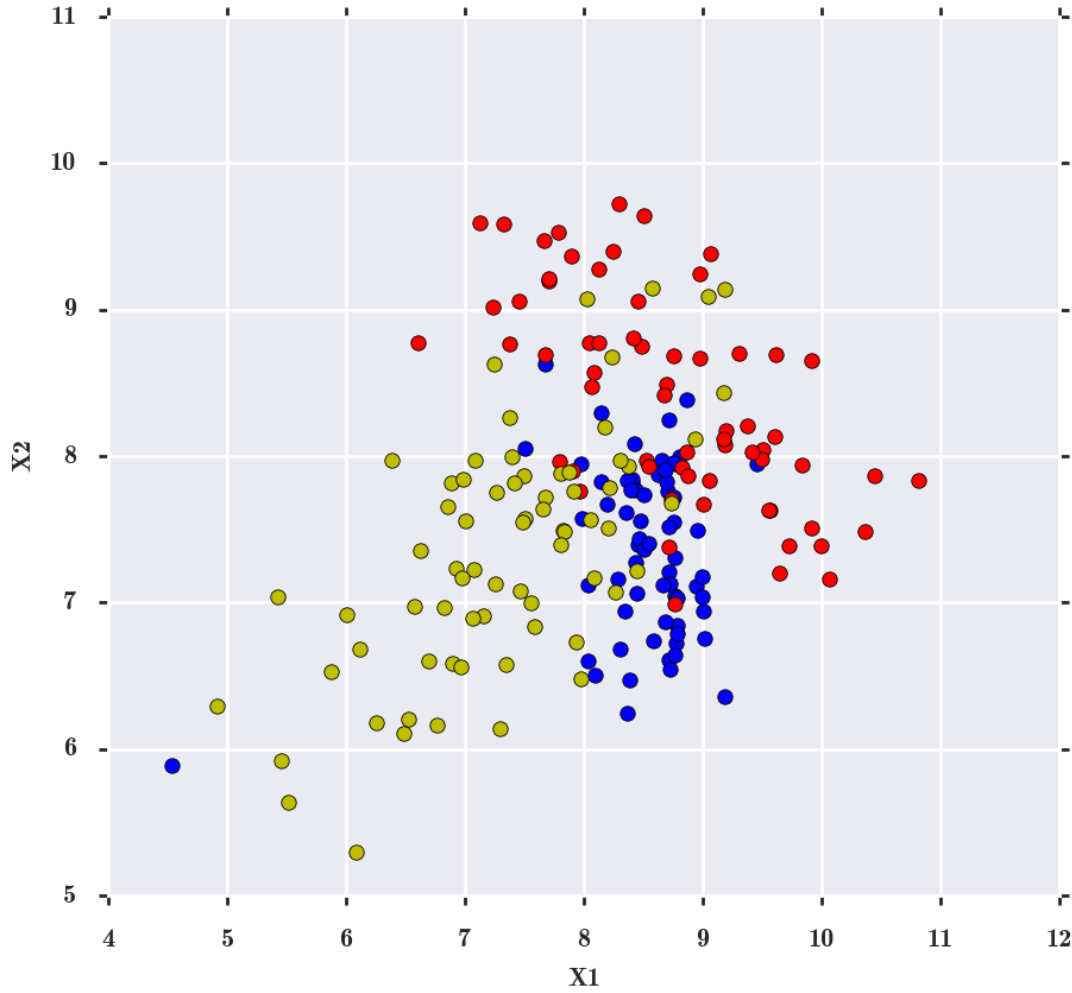
# Concepto y definición

- Tipos de problemas de Aprendizaje Automático:

Aprendizaje Supervisado: Se conoce "Y"		Aprendizaje No Supervisado No se conoce "Y"
<b>Clasificación</b> Y son valores discretos o categoría	<b>Regresión</b> Y puede tomar valores continuos	<b>Clustering</b> Identificación de grupos (Y) en función de semejanzas
Identificación de tráfico de red	Predicción de precio de una vivienda	Sistemas de búsqueda de pajeras

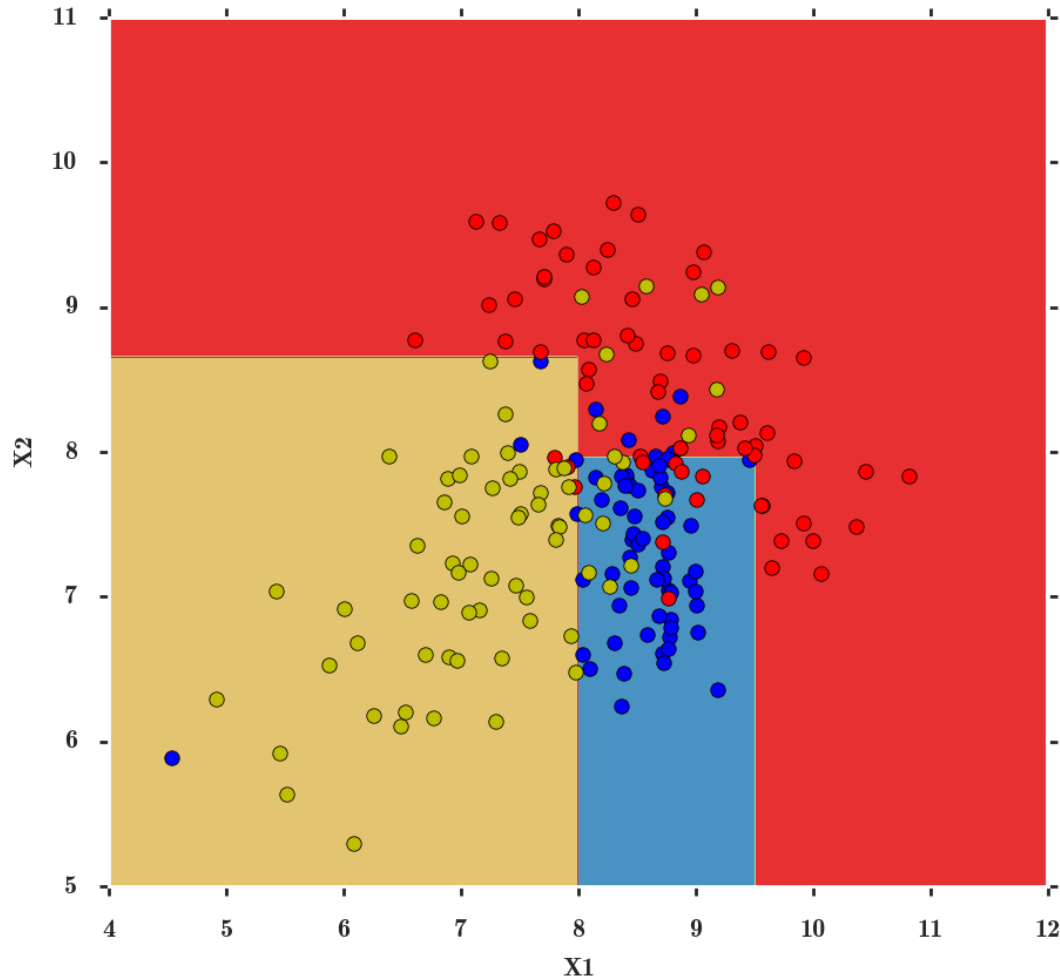
# Árboles de Decisión

PROBLEMA DE  
CLASIFICACIÓN  
SUPERVISADO

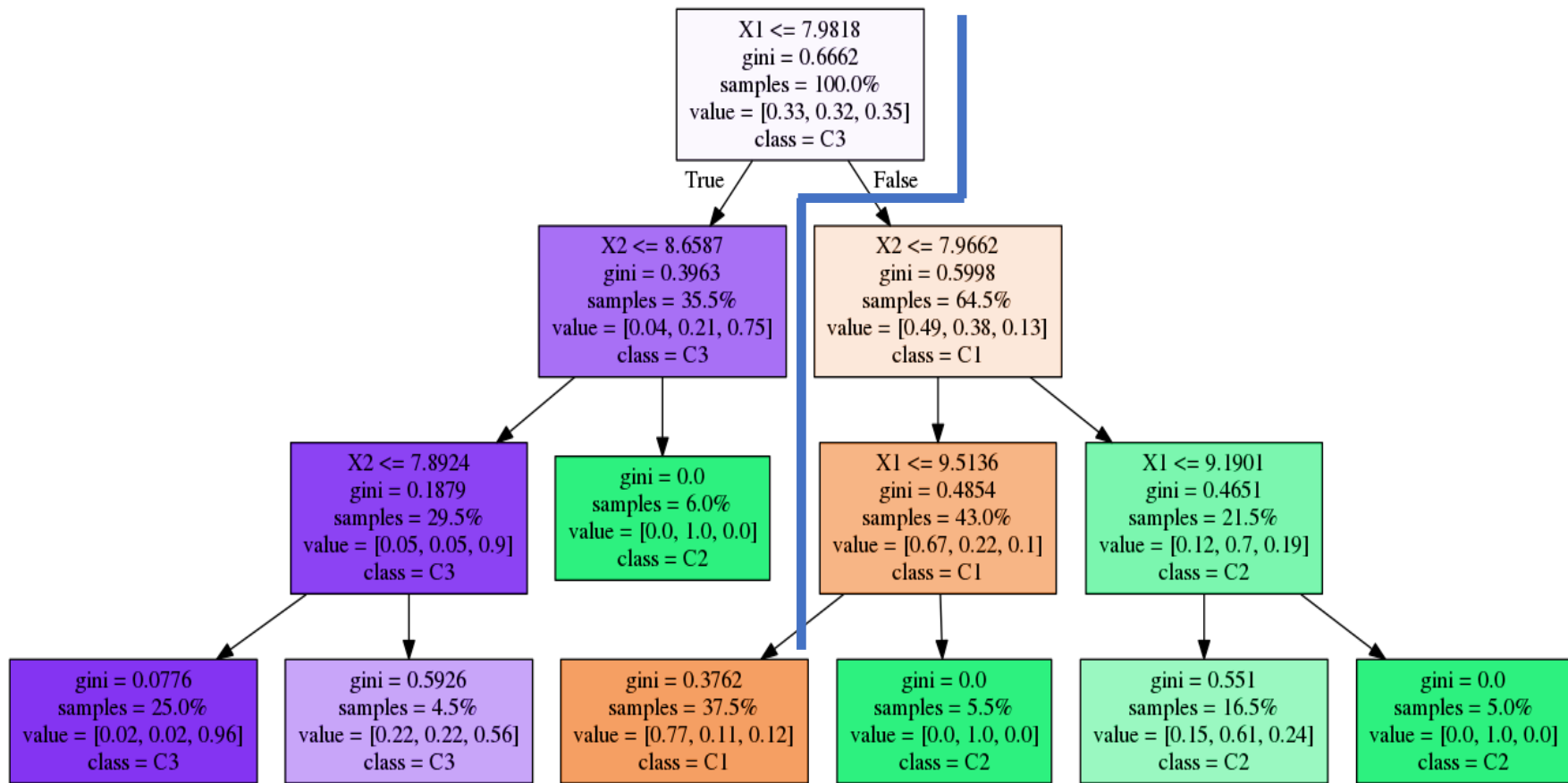


# Árboles de Decisión

POSIBLE  
SOLUCIÓN

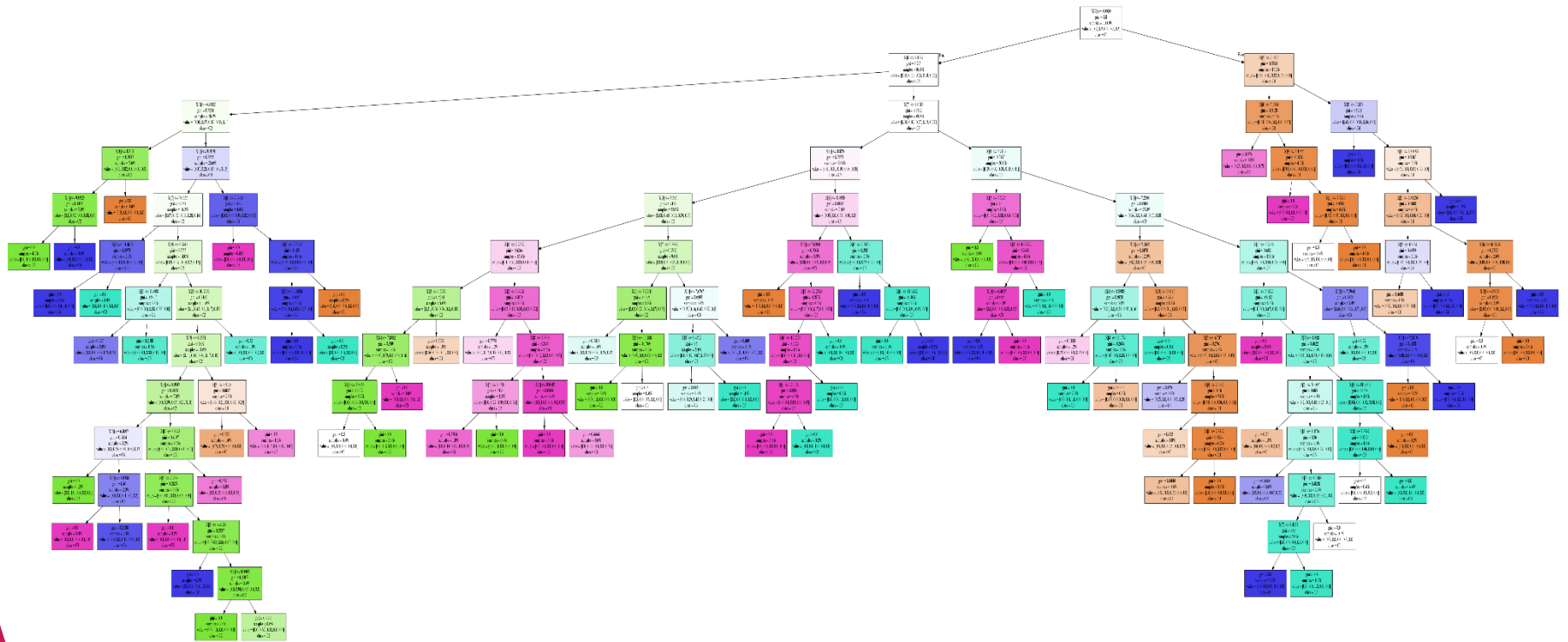


# Árboles de Decisión



Objetivo: minimizar  $GINI\ INDEX = \sum_K p_{mk} (1 - p_{mk})$

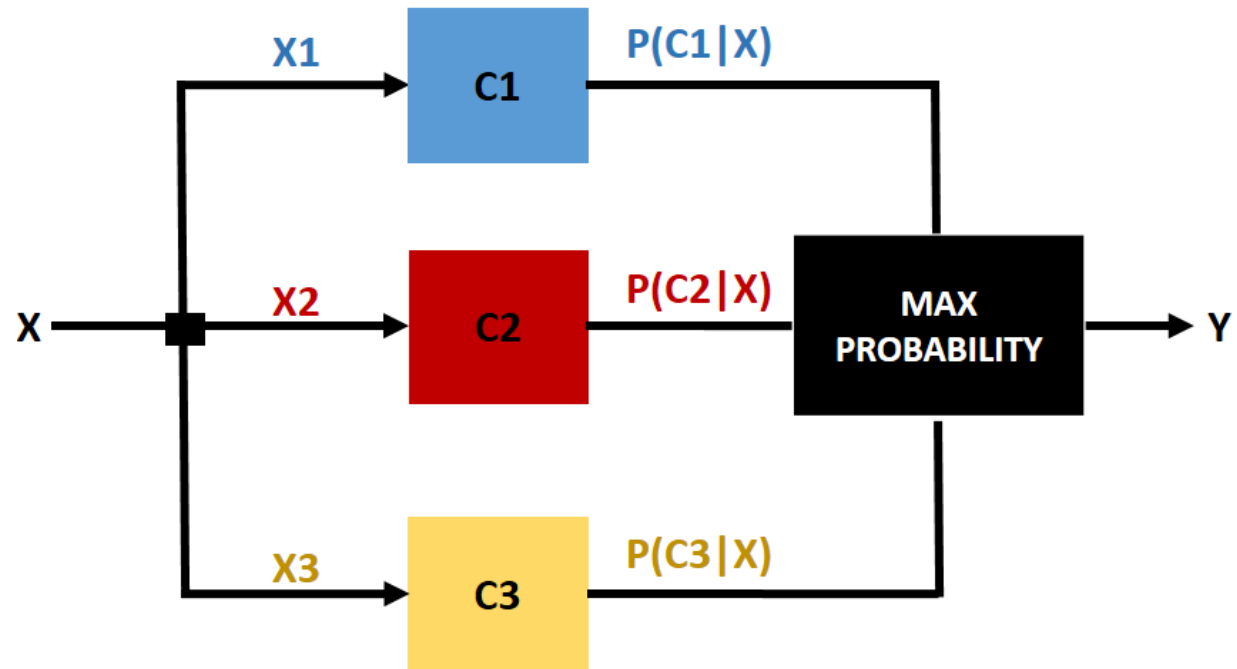
# Árboles de Decisión





# Algoritmos Ensamblados



- Algoritmos complejos formados por algoritmos base que cooperan en base a estrategias de entrenamiento y clasificación.
- Ejemplo:
  - OnevsRest



# Identificación de tráfico en redes

- Herramientas
- Ventajas del Aprendizaje Automático

# Herramientas

- Identificación basada en puertos:
  - Well-Known ports (IANA): FTP Data (20), Telnet (23), HTTP (80), ...
  - Técnica muy simple 
  - Precisiones reportadas: 50% - 70% 
  - Aplicaciones que usan puertos dinámicos: P2P...
  - Ej: [Coral Reef](#)

# Herramientas

- Deep Packet Inspection (DPI):
  - Identificación basada en patrones extraídos de la capa de aplicación.



# Herramientas

## ■ Fortalezas y Debilidades de DPI:

- Alta precisión
- Granularidad fina
- Baja eficiencia computacional
- Mantenimiento tedioso
- Tráfico encriptado
- Vulnera la privacidad del cliente



## ■ Aplicaciones DPI:

- [L7-Filter](#)
- [PACE](#)
- [nDPI](#)

# Herramientas

- Otras soluciones:
  - Híbridas: Puertos-DPI
  - Basadas en el comportamiento del usuario (BLINC)
  - Basada en inspección de IPs conocidas e interacciones DNS

# Ventajas del Aprendizaje Automático

- Buen ratio entre precisión y complejidad computacional
- Respeto a la privacidad del usuario
- Capaz de identificar tráfico encriptado
  
- Aún quedan muchas líneas por cerrar

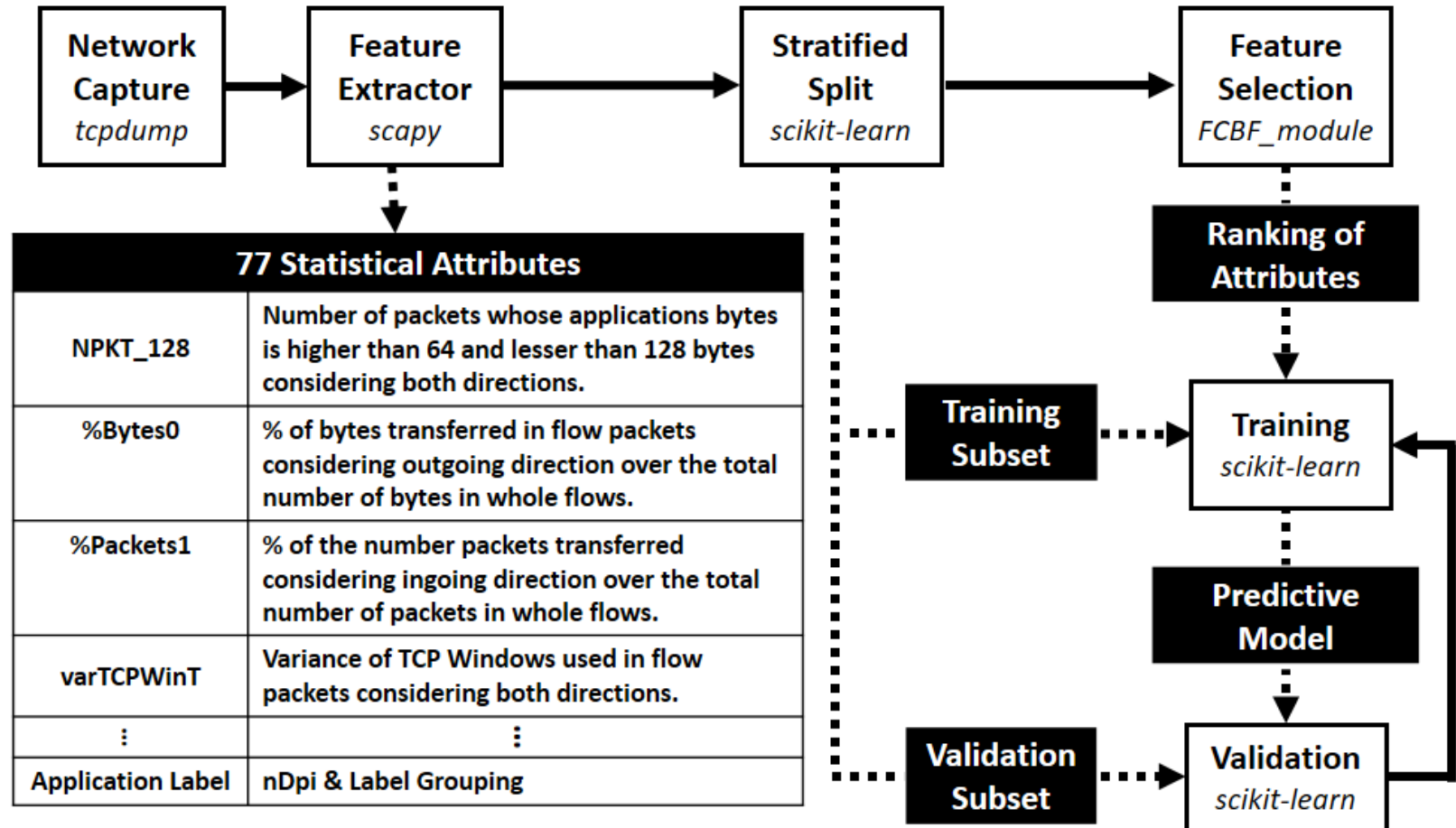


# Identificación de tráfico basado en Aprendizaje Automático

- Concepto “Early Traffic Classification”
- Metodología
- Comparación de algoritmos ensamblados



# Metodología



# Resultados

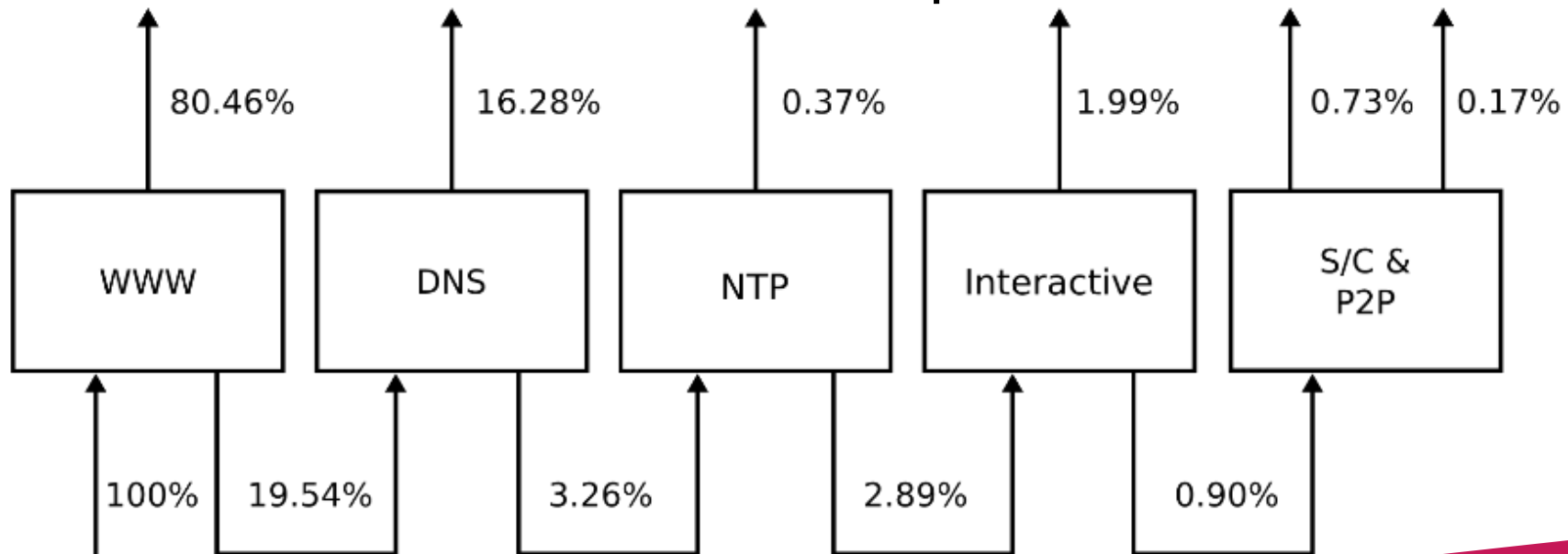
	<b>Overall Accuracy</b>	<b>Byte Accuracy</b>	<b>Training Time</b>	<b>Classification time</b>
<b>DT</b>	0.95310	0.93077	4.33370	0.03834
<b>OneVsRest</b>	0.96681	0.95995	10.55774	0.11443
<b>OneVsOne</b>	0.95384	0.95169	12.78923	0.52626
<b>RF</b>	0.95962	0.95390	7.81902	0.49980
<b>Bagging</b>	0.95467	0.95002	14.49463	0.42822
<b>ExtraTrees</b>	0.95492	0.94860	4.26417	0.61417
<b>OuputCode</b>	0.97396	0.96047	23.63508	0.20806
<b>ADA</b>	0.95587	0.94488	4,11500	0,41584

# Resultados

	Overall Accuracy	Byte Accuracy	Training Time	Classification time
DT	0.95310	0.93077	4.33370	0.03834
OneVsRest	0.96681	0.95995	10.55774	0.11443
OneVsOne	0.95384	0.95169	12.78923	0.52626
RF	0.95962	0.95390	7.81902	0.49980
Bagging	0.95467	0.95002	14.49463	0.42822
ExtraTrees	0.95492	0.94860	4.26417	0.61417
OuputCode	0.97396	0.96047	23.63508	0.20806
ADA	0.95587	0.94488	4,11500	0,41584
TDTC	0,97604	0,97317	3.87237	0.06602

# Tailored Decision Tree Chain - TDTC

- Características del problema:
  - Clases altamente desbalanceadas
  - Clases más fácil de identificar que otras



# Conclusiones

- Se confirman las ventajas en términos de precisión de los algoritmos de ensamblados
- Se confirma la penalización en términos de latencia de estos algoritmos
- TDTC. Se consigue retener las ventajas en precisión mientras que se mantiene una latencia aceptable



# Futuras Líneas

- Colección de atributos robustos
- Identificación de tráfico temprana
- Experimentación Online
- Auto-entrenamiento

# Muchas Gracias