**Título:** APPLICATION OF ADVANCED MACHINE LEARNING TECHNIQUES TO EARLY NETWORK TRAFFIC CLASSIFICATION

**Nombre:** Egea Gómez, Santiago

**Universidad:** Universidad de Valladolid

**Departamento:** Teoría de la señal y comunicaciones e ingeniería telemática

**Fecha de lectura:** 24/02/2020

**Mención a doctor europeo:** concedido

**Programa de doctorado:** Programa de Doctorado en Tecnologías de la Información y las Telecomunicaciones por la Universidad de Valladolid

**Dirección:**

> **Director:** BELEN CARRO MARTINEZ

> **Codirector:** Luis Hernández Callejo

> **Codirector:** ANTONIO JAVIER SÁNCHEZ ESGUEVILLAS

**Tribunal:**

> **presidente:** RAQUEL BARCO MORENO

> **secretario:** JUAN IGNACIO ARRIBAS SANCHEZ

> **vocal:** FERNANDO PÉREZ CRUZ

**Descriptores:**

> INTELIGENCIA ARTIFICIAL

> INTERNET E INTRANET

> INGENIERIA DE CONTROL

**El fichero de tesis** ya ha sido incorporado al sistema

**Localización:** BIBLIOTECA UNIVERSITARIA REINA SOFIA

**Resumen:** The fast-paced evolution of the Internet is drawing a complex context which imposes demanding requirements to assure end-to-end Quality of Service. The development of advanced intelligent approaches in networking is envisioning features that include autonomous resource allocation, fast reaction against unexpected network events and so on. Internet Network Traffic Classification constitutes a crucial source of information for Network Management, being decisive in assisting the emerging network control paradigms. Monitoring traffic flowing through network devices support tasks such as: network orchestration, traffic prioritization, network arbitration and cyberthreats detection, amongst others.
The traditional traffic classifiers became obsolete owing to the rapid Internet evolution. Port-based classifiers suffer from significant accuracy losses due to port masking, meanwhile Deep Packet Inspection approaches have severe user-privacy limitations. The advent of Machine Learning has propelled the application of advanced algorithms in diverse research areas, and some learning approaches have proved as an interesting alternative to

the classic traffic classification approaches. Addressing Network Traffic Classification from a Machine Learning perspective implies numerous challenges demanding research efforts to achieve feasible classifiers. In this dissertation, we endeavor to formulate and solve important research questions in Machine-Learning-based Network Traffic Classification. As a result of numerous experiments, the knowledge provided in this research constitutes an engaging case of study in which network traffic data from two different environments are successfully collected, processed and modeled.

Firstly, we approached the Feature Extraction and Selection processes providing our own contributions. A Feature Extractor was designed to create Machine-Learning ready datasets from real traffic data, and a Feature Selection Filter based on fast correlation is proposed and tested in several classification datasets. Then, the original Network Traffic Classification datasets are reduced using our Selection Filter to provide efficient classification models. Many classification models based on CART Decision Trees were analyzed exhibiting excellent outcomes in identifying various Internet applications. The experiments presented in this research comprise a comparison amongst ensemble learning schemes, an exploratory study on Class Imbalance and solutions; and an analysis of IP-header predictors for early traffic classification. This thesis is presented in the form of compendium of JCR-indexed scientific manuscripts and, furthermore, one conference paper is included.

In the present work we study a wide number of learning approaches employing the most advance methodology in Machine Learning. As a result, we identify the strengths and weaknesses of these algorithms, providing our own solutions to overcome the observed limitations. Shortly, this thesis proves that Machine Learning offers interesting advanced techniques that open prominent prospects in Internet Network Traffic Classification.