



---

## SpamDetector

Project for Data Mining and Machine Learning

---

Francesco Campilongo

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Datasets</b>	<b>3</b>
2.0.1	Data Pre-Processing . . . . .	3
<b>3</b>	<b>Design</b>	<b>4</b>
<b>4</b>	<b>Use Case Diagram</b>	<b>5</b>
<b>5</b>	<b>UML Class Analysis</b>	<b>6</b>
<b>6</b>	<b>Implementation and Tests</b>	<b>7</b>
<b>7</b>	<b>Results</b>	<b>8</b>

# Chapter 1

## Introduction

The purpose of this documentation is to prove the machine learning algorithms utility for the detection of spam, useless messages, into dataset of emails.

This is a basic text classification problem, and so the algorithm used are very known in the lecture.

The project is carried out on python.

# Chapter 2

## Datasets

The datasets used for this project are two:

- E-mail dataset
- SMS dataset

Those two datasets were found on Kaggle.com both of them in a .csv format.

### 2.0.1 Data Pre-Processing

The datasets found were already pretty good for the type of result this project wants to achieve. The E-mail dataset is characterized from three columns: "label", "text" and "label\_num", the second column contains the actual message, the first and third columns are the same, but in the first one there are the actual words "ham", for the messages which are not classified as spam, and "spam", instead in the third column there are "0" for the ham messages and "1" for the spam messages;

The first column has been discharged since not useful for the execution of the classification algorithm.

The SMS file had need a little more work to obtain a dataset usable, because it has five columns named "v1", "v2", "", "" and "", the first contains the words "ham", for the messages which are not classified as spam and "spam"; the second column contains the actual messages and the last three columns are basically blank, so in order to obtain a good dataset to work with the last three columns have been discharged the first two are taken in consideration, but in the first one the "ham" word is changed with a "0" and the "spam" word with a "1", in order to have the same kind of datasets between the E-mail and the SMSs.

## Chapter 3

# Design

## Chapter 4

# Use Case Diagram

## Chapter 5

# UML Class Analysis

## Chapter 6

# Implementation and Tests



## Chapter 7

## Results