﻿**IDS Final Exam**

**Marks: 20 Time: 3 hours**

You are given an email classified data in to spam and non-spam. In the column labelled spam, a 0 indicates the not spam and a 1 indicates spam

**Your task:** build a model to predict if a given email is spam. Your goal is to catch as many spam emails as possible without misclassifying too many not spam as spam. False positives (marking good email as spam) are very undesirable.

Note: Try to build all classification models you learnt in IDS and choose the best one. (KNN, Logistic Regression and Decision Tree Classifier)

**Split randomly** the given dataset after necessary pre-process into x\_train, y\_train, x\_test and y\_test. (you can use the library ‘﻿from sklearn.model\_selection import train\_test\_split’ to split the data set)

**The deliverable:** A function called spam\_predict which satisfies:

* input: x\_test, a set of email spam predictors for a group of emails
* output: y\_pred, a set of labels, one for each email; 0 for non-spam and 1 for spam

We provide you with some benchmarks for comparison.

**Baseline Model:**

~90% expected accuracy on non-spam emails in observed data

~90% expected accuracy on spam emails in observed data

~85% expected accuracy on non-spam emails in future data

~90% expected accuracy on spam emails in future data

time to build: 5 min

**Reasonable Model:**

~93% expected accuracy on non-spam emails in observed data

~95% expected accuracy on spam emails in observed data

~91% expected accuracy on non-spam emails in future data

~93% expected accuracy on spam emails in future data

time to build: 5 min

Grading: Your grade will be based on:

your model's ability to out-perform benchmarks

your ability to carefully and thoroughly follow the data science pipeline to the extent to which all choices are reasonable and defensible by methods you have learned in IDS

Submit all the code and mention the best model you built.