Project 3

Deadline:	Submit by midnight of 23 September 2018.
Evaluation:	25% of your final course grade.
Late Submission:	See Course Guide and refer to message from Rachel sent on Wednesday, 22 August.
Work	This assignment may be done in **pairs**. No more than two people per group are allowed. Should you choose to work in pairs, upon submission of your assignment, you will need to fill out and submit a form (to be provided) indicating your contribution to the project.
Purpose:	Learning outcomes 1 - 5 from the course outline.

Project outline:

Kaggle (https://www.kaggle.com/) is a crowdsourcing, online platform for machine learning competitions, where companies and researchers submit problems and datasets, and the machine learning community compete to produce the best solutions. This is a perfect training ground for real-world problems. It is an opportunity for data scientists to develop their portfolio which they can advertise to their prospective employers, and it is also an opportunity to win prizes.

For this project, you are going to work on a real-world Kaggle dataset in the 'playground competition' section, which tries to solve the "Costa Rican Household Poverty Level Prediction" problem, found here https://www.kaggle.com/c/costa-rican-household-poverty-prediction

Kaggle describes the problem as follows: "*The Inter-American Development Bank is asking the Kaggle community for help with income qualification for some of the world's poorest families. Are you up for the challenge?*

Here's the backstory: Many social programs have a hard time making sure the right people are given enough aid. It's especially tricky when a program focuses on the poorest segment of the population. The world's poorest typically can't provide the necessary income and expense records to prove that they qualify.

In Latin America, one popular method uses an algorithm to verify income qualification. It's called the Proxy Means Test (or PMT). With PMT, agencies use a model that considers a family's observable household attributes like the material of their walls and ceiling, or the assets found in the home to classify them and predict their level of need.

While this is an improvement, accuracy remains a problem as the region's population grows and poverty declines.

To improve on PMT, the IDB (the largest source of development financing for Latin America and the Caribbean) has turned to the Kaggle community. They believe that new methods beyond traditional econometrics, based on a dataset of Costa Rican household characteristics, might help improve PMT's performance.

Beyond Costa Rica, many countries face this same problem of inaccurately assessing social need. If Kagglers can generate an improvement, the new algorithm could be implemented in other countries around the world"

The playground competition is live and submissions will be accepted to the leader-board until September 20, 2018.

Task:

Your work is to be done using the Jupyter Notebook, which you will submit as the primary component of your work.

Your tasks are as follows:

1. You will first need to create an account with Kaggle.
2. Then familiarise yourself with the Kaggle platform.
3. Familiarise yourself with the submission process.
4. Download the datasets, then explore and perform thorough EDA.
5. Devise an experimental plan for how you intend to empirically arrive at the most accurate solution.
6. Explore the accuracy of kNN for solving the problem.
7. Explore scikit-learn (or other libraries) and employ a suite of different machine learning algorithms not yet covered in class.
8. Investigate which subsets of features are effective, and retain and build solutions based on this analysis and reasoning.
9. Devise solutions to these machine learning problems that are creative, innovative and effective. Since much of machine learning is trial and error, you are asked to continue refine and incrementally improve your solution. Keep track of all the different strategies you have used, how they have performed, and how your accuracy has

improved/deteriorated with different strategies. Provide also your reasoning for trying strategies and approaches. Remember, you can submit up to four solutions to Kaggle per day. Keep track of your performance and consider even graphing them.

10. Take a screenshot of your final and best submission score and standing on the Kaggle leader-board for both competitions and save that as a jpg file. Then embed this jpg screenshots into your Notebooks, and record your submission scores on the class Google Doc (found on Stream) where the class leader-boards will be kept.

The Kaggle platforms and the community of data scientists provide considerable help in the form of ‘kernels’, which are often Python Notebooks and can help you with getting started. There are also discussion fora which can offer help and ideas on how to go about in solving problems. Copying code from this resource is not acceptable for this assignment. Doing so can be regarded as plagiarism, and can be followed with disciplinary action.

Marking criteria:

Marks will be awarded for different components of the project using the following rubric:

Component	**Marks**	**Requirements and expectations**
EDA	5	Variety of exploratory research and inquiry into different aspects of the dataset, use of broad and appropriate range of visualisations and their effective communication. Thoroughness in data preparation.
Cluster analysis	5	Use of cluster analysis for exploring the dataset.
Classification modelling using kNN	30	Experimentation with kNN. Considering different values of *k* and effects of different distance metrics.
Classification modelling using a variety of algorithms	20	It is unlikely that kNN will produce the best accuracy on this kind of problem. Therefore, you are asked to explore and use a variety of algorithms either from scikit-learn, or elsewhere, in order to arrive at your best solution for the competition.
Analysis	20	The manner in which you have devised your experiments, evaluated your classifiers, interpreted your findings, as well as conducted feature analysis and feature selection.
Kaggle submission score	20	Successful submission of predictions to Kaggle, listing of the score on the class leader-boards and position on the class leader-boards. The winning submission with the best accuracy on the class leader-board will receive full marks. The next best submission will receive 15 marks, and every subsequent placing will receive one less point. If a submission is made **by September 12 and the class leader board document is updated, then 5 marks will be guaranteed for the submission irrespective of placing on the leader-board.**
BONUS MARKS		
Additional feature extraction	5	Bonus marks will be awarded for extracting additional features from dataset and incorporating them into the training set, together with the comparative analysis showing whether or not they have increased predictive accuracy.

Hand-in: **Zip**-up all your **notebooks**, any other .py files you might have written as well as jpgs, of your screenshots into a single file and submit through Stream. **If, and only if** Stream is down, then email the solution to the lecturer.

If you have any questions or concerns about this assignment, please ask the lecturer sooner rather than closer to the submission deadline.