Milestone 2 instructions: For Milestone 2, you will write a 2-4 page report on learning models written by other developers for one dataset you would like to work with. Your report should include a description of your dataset, along the same lines as was requested for Milestone 1. Some ideas on what to put in the report: \* Some learning tasks that people have worked on, that impress you. \* Comments on how they accomplish these tasks. \* References to their code, with links. \* Things that you found in the code that we have not covered in the class. You can also add your own ideas and observations. While you are working on creating this report, you can also start thinking/discussing your own variants and extensions of the existing solutions, and you can discuss such extensions in your report. Keep in mind that for your Mini-Project to be viable, you will need to propose some new contribution beyond anything that has been done in the past for your set. \* After you are done, please submit your response as a .pdf file upload. \* For maximum credit, your responses should be well-organized, neat, relevant and thorough. \* Looking at online ML tutorials can provide additional ideas as to what a makes for a well-written presentation (but beware, some are better than others!). Feedback: What information do your classmates need to know about your proposed project so that they can decide whether to form a group with you? They need to understand the characteristics of the dataset, as well as at least one learning task that you are considering for it. They also need to know what new contributions the group would expect to make with respect to the best work that has already been done on Kaggle, and for that, thorough descriptions must first be given for the relevant Kaggle projects. At a minimum, we need to know about the learning tasks they have considered, what learning methods they used, what feature was targeted for prediction, what features were used as predictors, any issues arising from data preparation or exploratory data analysis, the measures of prediction performance considered, and the prediction performance itself. Any proposals should clearly state what new contributions are to be made with respect to one or more of these issues (e.g. higher accuracy, or different learning methods, or the effects of different feature choices, or different data preparation strategies). When appropriate, some consideration should also be given to how a fair comparison can be made with existing Kaggle models - will any of their learning models or data preparation methods be used or re-implemented by you? Responses that are "relevant and thorough" should address all of these questions. (Keep in mind that we will need to see all this information in your Milestone 4 report as well.) Extra information is of course welcome. Dataset: Apple Quality \* General description provided (e.g. source of data, uses, class information): yes. \* Number of data records: 4000. \* Number of data features: 7. \* Types of features (e.g. real-valued? ordinal? categorical? text? time & location?): standardized real values. \* Possible learning task: classification. \* Target feature for prediction: Quality. \* Likely features for use as predictors: all. \* Kaggle usability rating: not stated. (10.0) \* Data preparation issues: "little preprocessing needed". Other learning tasks: \* Highest-ranked Kaggle contributions surveyed: yes, with a rationale given for choosing one. \* Relevant learning tasks: binary classification. \* Target feature for prediction: Quality. \* Features used as predictors: all. \* Data preparation and exploratory data analysis: correlation map, feature value histograms. \* Model descriptions: SVC, RF, XGBoost, KNN, etc. \* Measures of prediction performance: accuracy. \* Prediction performance outcomes: >90%. \* Other relevant issues: classes are balanced. Proposed project: \* Anticipated new contributions: not stated. \* Reuse of existing models: not stated. \* Reuse of previous data preparation: not stated. \* Basis for a fair comparison with Kaggle projects: not stated. (accuracy?) Overall: \* Description of dataset and learning task: (2 out of 2 marks) (Excellent / Good / Fair / Problematic / Missing): Excellent \* Investigation of impressive Kaggle contributions: (2 out of 2 marks) (Excellent / Good / Fair / Problematic / Missing): Excellent \* Proposed contributions relative to Kaggle competitors: (0 out of 1 mark) (Excellent / Good / Fair / Problematic / Missing): Missing \* Well-organized and neat? (0.75 out of 1 mark) (Excellent / Good / Fair / Problematic / Missing): Good Other comments: Overall, this is a good report, with an excellent survey of the dataset and the models developed for it on Kaggle. However, you don't give any indication of what improvements you have in mind relative to the Kaggle contributions. If you try to achieve better accuracy, you might have a difficult time, since their models have accuracies above 90%. Please consider what you can do that will guarantee some sort of new contribution even if your models are less accurate. You could do this by using learning methods that they haven't considered, or by preparing the data in different ways. Keep in mind that you will need to ensure that your comparison with the Kaggle models is fair. Your report is a bit long. For Milestone 4, please consider beginning your report with a compact description of the dataset and of the Kaggle models (in that order), before discussing your own contributions. Compact descriptions allow the readers to better understand your points. The feature values are generally bell-shaped... we cannot say that they are "distributed evenly" (which would indicate that they are approximately uniform). Informally speaking, principal component analysis involves the transformation of the data to a new representation in which the first axis (principal component) captures the greatest possible spread (variance) in the data, the second axis captures the second greatest spread, etc. We will see its uses in dimensionality reduction soon. For Milestone 3, please form a new group around your proposed project, or join 1 or 2 other students to work on their proposal. Be sure follow the instructions for Milestone 3 very carefully, and remember that Milestone 3 marks are awarded only to groups for which all members have satisfied all the conditions of M3 perfectly. There is one other student in Section 852 who proposed the Apple Quality dataset to work on.