The overall goal of the project is to design, implement, and evaluate a classifier (model) for a domain of interest to you along with an executable jupyter notebook describing your results. Your project must consist of the following.

- **Data Discovery.** You will need to find relevant datasets that can be used to train and evaluate your classifier based on types of predictions you plan to perform. See below for a list of possible sites where you can find data (however, you are not limited to using only these sites).

- **Exploratory Analysis and Visualization.** You will need to perform exploratory analysis with corresponding visualizations to provide a better understanding of the datasets chosen (both for yourself and as part of the report).

- **Data Preprocessing/Cleaning.** You will need to preprocess and clean the datasets as needed, which will likely include data merging, dealing with missing values, feature selection, and discretization.

- **Classifier Development.** You must use the basic classification approaches described in class (k-nn, naive bayes, decision trees) and evaluate their effectiveness for your data. As part of this work you will need to then develop a tailored classification approach for your data based on the evaluation results. The tailored approach could be an ensemble of existing approaches (including, e.g., random forests), a variant of an algorithm we covered, or an algorithm we did not cover in class, with the goal of providing better predictive performance than the three basic approaches. Note that your work must be driven on evaluation results and predictive performance.

Additional requirements are provided below. The project will be broken into a series of steps to be done over the remainder of the semester. Each step is described further below. Note that this is an individual project and each student project must be different.

**Step 1: Proposal (due Tues, 10/10).** For this step, your job is to describe the domain you are interested in exploring, the information you want to predict via a classifier, and initial datasets you will use for your project. The datasets you choose should be interesting (at least to you) and be non trivial (for classification). While difficult to give exact parameters, the primary dataset should have at least 6 attributes and on the order of 1000s (or more) instances. Note that these are not hard and fast rules—please check with me if you have a dataset that doesn’t meet these criteria. In addition, you must include at least one additional dataset that can be combined with the primary dataset to provide richer results. Examples of areas you may want to explore include data related to gaming, sports information, census
or government statistics, scientific observations, social media data (movies, music, etc.),
consumer behaviors, socio-economic data, political trends, and security information (e.g.,
intrusion detection and/or vulnerability data). The following links are to sites that contain
(or link to) public datasets. You may want to use these to get started or if you are having
trouble finding data for your project.

- [https://data.gov](https://data.gov)
- [https://www.kaggle.com/datasets](https://www.kaggle.com/datasets)
- [https://registry.opendata.aws/](https://registry.opendata.aws/)
- [https://data.worldbank.org](https://data.worldbank.org)
- [https://www.google.com/publicdata/directory](https://www.google.com/publicdata/directory)
- [https://www.yelp.com/dataset](https://www.yelp.com/dataset)
- [https://www.wikidata.org](https://www.wikidata.org)
- [https://www.pewresearch.org/internet/datasets/](https://www.pewresearch.org/internet/datasets/)
- [http://labrosa.ee.columbia.edu/millionsong](http://labrosa.ee.columbia.edu/millionsong)
- [https://www.sports-reference.com](https://www.sports-reference.com)
- [https://www.wunderground.com/history](https://www.wunderground.com/history)
- [https://developer.bestbuy.com/apis](https://developer.bestbuy.com/apis)
- [https://github.com/shramos/Awesome-Cybersecurity-Datasets](https://github.com/shramos/Awesome-Cybersecurity-Datasets)

For this step, turn in a typed description of the domain you will be exploring, what informa-
tion you will try to predict via a classifier, and the initial datasets you plan to use for
your project. For each dataset, include a brief description of its contents, where it is located,
and its general format. Note that you should start working on exploratory analysis, data
preprocessing/cleaning, and feature selection for the next step.

**Step 2: Exploratory Analysis and Feature Selection (due Thurs, 11/9).** For this
step, you must:

(a). Finalize the datasets you plan to use for your classifier.

(b). Provide information about your datasets including the attributes and attribute types,
the number of instances, and the attribute(s) you will be using as class labels for
predictions.
(c). Provide initial data visualizations highlighting important and/or interesting aspects of your datasets related to your classification problem. Visualizations may include frequency distributions, comparisons of attributes (scatterplot, multiple frequency diagrams), and so on. The goal is not to include all possible diagrams, but instead to select and highlight diagrams that provide insight about the dataset and your classification itself. Each visualization must be clearly and accurately labeled.

(d). A list of the features (attributes) you will be using to train your classifiers and a justification for their use.

(e). The types of cleaning operations you will be performing, including handling of missing values and discretization.

For this step, you will need to create a jupyter notebook with the above information. Submit the code you have written along with your jupyter notebook to your GitHub repository for the project (details will be posted to Piazza). Note that once you finish this step, you should start working on Step 3, which is due only one week later.

**Step 3: Initial Classifiers and Evaluation Results (due Thurs, 11/16).** For this step, you must:

(a). Implement the data cleaning operations and feature selection tasks necessary for developing your classifiers.

(b). Implement the three basic classifiers for your datasets: k-nn, naive bayes, and decision trees.

(c). Provide evaluation results for the three classifiers.

As in Step 2, you will need to create a jupyter notebook with the above information. Submit the code you have written along with your jupyter notebook to your GitHub repository for the project (details will be posted to Piazza). Once you finish this step, you should start working on your custom classifier for Step 4.

**Step 4: Project Implementation and Report (due by Fri, 12/15).** This is the last step of the project, and involves finishing your tailored classifier, evaluating your results, and writing your final report as an executable jupyter notebook. You must submit all of your code, your tests, and your notebook to your GitHub project repository by the due date (details will be posted to Piazza). Your final report must be structured as follows. Note that your notebook must be self-contained, professionally written, and organized such that it could be used to help demonstrate to future employers your ability to think critically about, develop, and solve “real-world” data analysis problems.
Section 1: Introduction. This section must briefly describe the datasets you used and the classification tasks you implemented (e.g., what were you trying to classify in the dataset). You must also briefly summarize your findings in terms of classifier performance. This portion should be text (not executable code).

Section 2: Data Analysis. This section must provide details about the datasets used for your project. You must describe the datasets used, statistical summaries and visualizations relevant to the classification tasks, the steps involved in cleaning the dataset, and the features selected for classification. This section must be in paragraph form that includes (well written) text interspersed with executable code (for all statistical summaries and visualizations, data cleaning steps, and feature selection). Additionally, each figure included must have a figure caption (Figure number and textual description) that is referenced from the text (e.g., “Figure 2 shows a frequency diagram for ...”).

Section 3: Classification Results. This section must describe your approach for generating your training and test sets, the classifiers used, parameter settings, your evaluation approach, and the evaluation results of the classifiers. You will need to spend some time explaining your tailored classifier (which should outperform the three simple approaches). Similar to Section 2, this section must be in paragraph form, with text interspersed with executable code (for generating training and testing sets, executing the classifiers, and generating evaluation metrics), and results presented as labeled figures. Finally, this section should provide a summary of the results along with your observations regarding the results and the performance of the classifiers you used.

Section 4: Conclusion. Provide a brief conclusion of your project focused on any of its inherent challenges for classification, ideas you have on ways to improve classification performance, and next steps you would perform if you had more time.

Grading. The project is worth a total of 70 points. Steps 1–3 will be worth 4 points each. The remaining points will be based on your final implementation and report (jupyter notebook) described in Step 4. Note that your project will be graded on the following criteria.

- Quality. The overall quality of your code and report.

- Completeness. Whether your work is complete, you correctly followed all instructions, and there are no bugs in your code and report.

- Difficulty. The overall scope of the project, the amount of effort put into developing a tailored classifier, the amount of effort put into the report, the amount of testing performed, and the strength of your results.

- Report. Covers each aspect asked for, is clear and professional, and is well organized.