**Open Quick Links**

**Current Location**

1. [Intro to Data Science (Summer 2021)](https://blackboard.wm.edu/webapps/blackboard/execute/courseMain?course_id=_20138_1)
2. Syllabus

[Hide Course Menu](https://blackboard.wm.edu/webapps/blackboard/content/listContent.jsp?course_id=_20138_1&content_id=_1563361_1&mode=reset)

**Menu Management Options**

* [Refresh](https://blackboard.wm.edu/webapps/blackboard/content/listContent.jsp?course_id=_20138_1&content_id=_1563361_1&mode=reset)
* [Display Course Menu in a Window](https://blackboard.wm.edu/webapps/blackboard/content/listContent.jsp?course_id=_20138_1&content_id=_1563361_1&mode=reset)

**Course Menu:**

[**Intro to Data Science (Summer 2021)**](https://blackboard.wm.edu/webapps/blackboard/content/listContent.jsp?course_id=_20138_1&content_id=_1563361_1&mode=reset)

* [Home Page](https://blackboard.wm.edu/webapps/blackboard/content/launchLink.jsp?course_id=_20138_1&toc_id=_372202_1&mode=view&mode=reset)
* [Syllabus](https://blackboard.wm.edu/webapps/blackboard/content/listContent.jsp?course_id=_20138_1&content_id=_1563361_1&mode=reset)
* [Instructor Contact Information](https://blackboard.wm.edu/webapps/blackboard/content/launchLink.jsp?course_id=_20138_1&tool_id=_153_1&tool_type=TOOL&mode=view&mode=reset)

**Syllabus**

Top of Form

Bottom of Form

**Content**

**Syllabus**

**ZOOM LINK**

<https://cwm.zoom.us/j/96537949677?pwd=MnJJQ2xhOWVYQkhVT2FKcmxiLy8yZz09>

Meeting ID: 965 3794 9677

Passcode: 089972

**INSTRUCTOR**

Dr. Ronald D. Smith

rdsmith@wm.edu

Office hours available by appointment

**COURSE OVERVIEW**

In this course students will learn the fundamentals of data processing and modeling in the context of Data Science.  Emphasis will be placed on careful planning and deliberate decision making when working with data and building models.  Programming will be done in the Python language and we will be making extensive use of the scikit-learn collection (<https://scikit-learn.org/stable/>).

After a review of the basics of Python programming, students will be introduced to a variety of supervised and unsupervised machine-learning techniques including various methods for regression, classification, and clustering. By the end of the course, students are not expected to be an expert on any particular technique, but should exhibit a solid high-level understanding of the goals of each method, be able to determine when a particular type of model is more or less suitable to a real-world problem and, most importantly, demonstrate a keen attention to detail when working with data.

**COLLABORATION**

Collaboration is both allowed and encouraged!  Most of the time in life we are not working in isolation, and it is both wise and efficient to use all resources available to you, including professors, coworkers, other students, the internet, etc.  The only thing that is off limits is discussion of specific answers to any graded assignment.

Our primary means of communication will be Piazza, so post your questions there, and when you see a question that you can answer, please do!  <https://piazza.com/wm/summer2021/data146/home>

*(The professor will make every attempt to respond to email within 24 hours, however you are strongly encouraged to post your questions on Piazza!)*

**RESOURCES**

Much of the material in this course will be presented in the form of Python notebooks (.ipynb files).

You may choose to use the JupyterHub (hosted by W&M): <https://jupyterhub.wm.edu/>

The first time you log on, it may take a moment to set up your environment (this should only happen the first time).  For any other issues, contact hpc-help@wm.edu

Alternatively, if you would like to run Python locally on your computer, I recommend the Anaconda distribution: <https://www.anaconda.com/products/individual>

Or, if you already have your own solution, that's fine too!  Just make sure you can work with Python notebooks, and that you are running an up-to-date version of Python.

**COURSE SCHEDULE**

**Part 1: Python Review**

Our first goal will be to review some of the basics of programming in Python.  This will include working with data using the pandas and numpy libraries, and some basic plotting using the matplotlib library.

Keywords: *pandas, numpy, matplotlib, DataFrame*

**Part 2: Descriptive Analysis, Intro to Modeling (via regression) and Model Validation**

Before building any model, it is always a good idea to "get to know" your data.  Summary statistics and visualizations are often a good way to go about this.  After being introduced to some common and useful ways to summarize your data, we will build our first model and learn about concepts such as model validation, and what it means for a model to be overfit (or underfit).  Other topics explored will include dimensionality reduction techniques such as Principal Component Analysis and tSNE.

Keywords: *scatter plot, box plot, histogram, central tendency, variability, percentile, quantile, summary statistic, dimensionality reduction, Principal Component Analysis (PCA), tSNE, linear regression, ordinary least squares, internal validity, external validity, K-fold cross validation*

**Part 3: More on Regression, Regularization, and Hyperparameters**

We will continue our discussion on regression, and learn about regularization techniques such as Ridge and Lasso regression, which aim to prevent model overfitting.  Along the way, we will learn about the concept of a "hyperparameter".  The week will conclude with examples of Decision Trees and Random Forests (applied to regression), which we will see again in Week 4 when we discuss classification.

Keywords: *regularization, hyperparameter, Ridge regression, Lasso regression, Elastic Net regression, Decision Tree Regression, Random Forest Regression*

**Part 4: Classification Techniques**

We will be introduced to several classification methods, which are all types of supervised learning.  While the goal of each method is similar, the methods and results can be quite different.  We will explore the differences between these methods when applied to the same data sets and discuss the pros and cons of each.  Additionally, we will be introduced to a classic type of neural network known as the Multilayer Perceptron, and see how it can be used to classify handwritten digits.

Keywords: *Logistic Regression, K-nearest neighbors, Decision Tree Classification, Random Forest Classification, Multilayer Perceptron, supervised learning*

**Part 5: Clustering Techniques**

Our semester will conclude with an introduction to clustering techniques, which are types of unsupervised learning.  Similar to the unit on classification, we will compare these methods when applied to the same data sets, and discuss some pros and cons of each.

Keywords: *K-means, DBSCAN, Agglomerative Hierarchical Clustering, unsupervised learning*

**Part 6: Optional Topics**

If time allows, we will explore some basic concepts related to neural networks and/or natural language processing

**COURSE GRADING POLICY**

Your assignments each week will consist of one checkpoint, and one "mini-project" (2 submissions per week, for a total of 10 submissions).  All submissions will be made through the Assignments link on Blackboard.  You should expect a mixture of conceptual questions related to the techniques we are learning as well as technical questions related to code you have written.  ***I will not be asking you anything that will require you to have memorized code we have written, and all assignments are open-book/open-note.  The emphasis in this class is on understanding why we are doing what we are doing.***

Checkpoints: Brief quizzes due every Weds by 11:59 PM.  ***8% each x5 = 40%***

Mini-projects: You will be reporting the output of code you have written, as well as answering some conceptual questions about the methods being used.  These will be due each Sunday by 11:59 PM.  ***10% each x5 = 50%.***

The ***remaining 10% of your grade is based on participation***, and can be earned by coming to office hours (on Zoom), participating in online discussion (Piazza), or otherwise demonstrating active engagement in the course.

Extra credit: Attend a Data Science Alumni meetup (career talks).  More details coming soon (or ask me!).  **+2%**

Lateness Policy: Excepting documented illness or other extenuating circumstances, late assignments will not be accepted and will earn you a zero.

Final Grades:  Below is a Python function to compute your final letter grade, based on your numerical grade:

def GetLetterGrade(pct\_score):

if pct\_score > 0.91:

return "A"

elif pct\_score > 0.9:

return "A-"

elif pct\_score > 0.89:

return "B+"

elif pct\_score > 0.81:

return "B"

elif pct\_score > 0.8:

return "B-"

elif pct\_score > 0.79:

return "C+"

elif pct\_score > 0.71:

return "C"

elif pct\_score > 0.7:

return "C-"

elif pct\_score > 0.69:

return "D+"

elif pct\_score > 0.61:

return "D"

elif pct\_score > 0.6:

return "D-"

else:

return "F"

**ACCOMODATIONS**

It is the policy of William & Mary to accommodate students with disabilities and qualifying diagnosed conditions in accordance with federal and state laws. Any student who feels s/he may need an accommodation based on the impact of a learning, psychiatric, physical or chronic health diagnosis should be referred to [Student Accessibility Services](http://www.wm.edu/offices/deanofstudents/services/studentaccessibilityservices/) staff at 757-221-2509 or at [sas@wm.edu](mailto:sas@wm.edu). SAS staff will work with you to determine if accommodations are warranted, and if so, to help you obtain an official letter of accommodation.