

# Machine Learning Mastery With Weka

Analyze Data, Develop Models  
and Work Through Projects

---

Jason Brownlee

**MACHINE  
LEARNING  
MASTERY**



Jason Brownlee

# **Machine Learning Mastery With Weka**

**Analyze Data, Develop Models and Work Through Projects**

**Machine Learning Mastery With Weka**

© Copyright 2017 Jason Brownlee. All Rights Reserved.

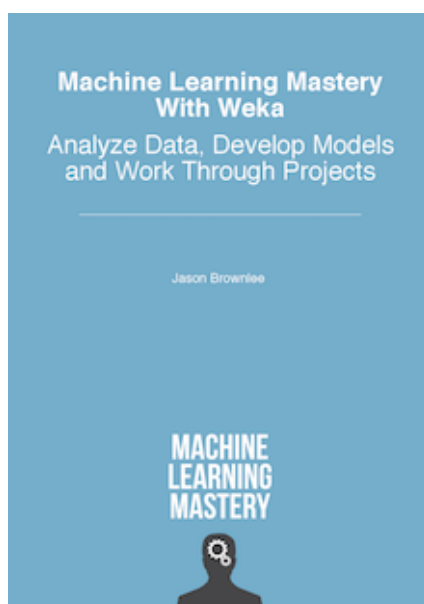
Edition: v1.2

# This is Just a Sample

Thank-you for your interest in **Machine Learning Mastery With Weka**.

This is just a sample of the full text. You can purchase the complete book online from:

<https://machinelearningmastery.com/machine-learning-mastery-weka>



# Contents

<b>1</b>	<b>Welcome</b>	<b>1</b>
1.1	Applied Machine Learning the Wrong Way . . . . .	1
1.2	Applied Machine Learning with Weka . . . . .	1
1.3	Book Overview . . . . .	2
1.4	Your Outcomes From This Process . . . . .	4
1.5	What This Book is Not . . . . .	4
1.6	Summary . . . . .	5
<b>2</b>	<b>Rapidly Accelerate Your Progress in Applied Machine Learning With Weka</b>	<b>6</b>
2.1	Starting in Applied Machine Learning is Hard . . . . .	6
2.2	Focus on Learning Just One Thing . . . . .	7
2.3	Learn the Process of Applied Machine Learning . . . . .	7
2.4	How to Best Use Weka . . . . .	7
2.5	Summary . . . . .	8
<b>3</b>	<b>How to Normalize and Standardize Your Machine Learning Data</b>	<b>9</b>
3.1	About Data Filters in Weka . . . . .	9
3.2	Normalize Your Numeric Attributes . . . . .	11
3.3	Standardize Your Numeric Attributes . . . . .	14
3.4	Summary . . . . .	15

# Chapter 1

## Welcome

*Welcome to Machine Learning Mastery With Weka.* This book is your guide to applied machine learning. You will discover the step-by-step process that you can use to get started and become good at machine learning for predictive modeling using the Weka platform.

### 1.1 Applied Machine Learning the Wrong Way

Here is what you should not do when you start in applied machine learning:

- Get really good at the math that underlies machine learning theory.
- Deeply study the underlying theory and parameters for machine learning algorithms.
- Avoid or lightly touch on all of the other tasks needed to complete a real project.

This approach can work for some people, but it is a really slow and a roundabout way of getting to your goal. It teaches you that you need to spend all your time learning how to use individual machine learning algorithms. It also does not teach you the process of building predictive machine learning models that you can actually use to make predictions. Sadly, this is the approach used to teach machine learning that I see in almost all books and online courses on the topic.

### 1.2 Applied Machine Learning with Weka

This book focuses on a specific sub-field of machine learning called predictive modeling. This is the field of machine learning that is the most useful in industry and the type of machine learning that the Weka platform excels at facilitating.

Unlike statistics, where models are used to understand data, predictive modeling is laser focused on developing models that make the most accurate predictions at the expense of explaining why predictions are made. Unlike the broader field of machine learning that could feasibly be used with data in any format, predictive modeling is primarily focused on tabular data (e.g. tables of numbers like a spreadsheet). This book was written around three themes designed to get you started and practicing applied machine learning effectively and quickly. These three parts are as follows:

- **Weka:** Weka is the very best platform for beginners getting started and practicing applied machine learning.
- **Lessons:** Learn how the subtasks of a machine learning project map onto Weka and the best practice way of working through each task.
- **Projects:** Tie together all of the knowledge from the lessons by working through case study predictive modeling problems.

These are the three pillars of this book that will quickly and effectively take you from where you are now to your goal of confidently working through and delivering results on your own applied machine learning projects.

## 1.3 Book Overview

This book was carefully designed to quickly and effectively take you from beginner to confident machine learning practitioner capable of working through your own projects end-to-end. As such, this book is divided into 4 parts:

- Part 1: Introduction
- Part 2: Lessons
- Part 3: Projects
- Part 4: Conclusions

### 1.3.1 Part 1: Introduction

The introduction makes the case that Weka is the best platform for beginners getting started in applied machine learning. It covers:

- Why applied machine learning is so hard and how Weka makes it easy.
- What the Weka machine learning workbench provides.
- How to make best use of Weka by developing a portfolio of completed projects.

After completing this part you will be ready to actually get started learning applied machine learning using the Weka workbench.

### 1.3.2 Part 2: Lessons

This part provides the meat of the book, providing you with specific instruction on how to use Weka for applied machine learning. Each tutorial is standalone. The benefit of this is that you can dip in to specific lessons if and when you need them, or work through them sequentially one-by-one until you have all the knowledge you need to work through a problem. Each lesson will teach you one key skill in using Weka for applied machine learning. The full list of the 18 lessons provided are as follows:

- **Lesson 01:** How to Download and Install the Weka Machine Learning Workbench.
- **Lesson 02:** A Tour of the Weka Machine Learning Workbench.
- **Lesson 03:** How To Load CSV Machine Learning Data.
- **Lesson 04:** How to Load Standard Machine Learning Datasets.
- **Lesson 05:** How to Better Understand Your Machine Learning Data.
- **Lesson 06:** How to Normalize and Standardize Your Machine Learning Data.
- **Lesson 07:** How to Transform Your Machine Learning Data.
- **Lesson 08:** How To Handle Missing Values In Machine Learning Data.
- **Lesson 09:** How to Perform Feature Selection With Machine Learning Data.
- **Lesson 10:** How to Use Machine Learning Algorithms.
- **Lesson 11:** How To Estimate The Performance of Machine Learning Algorithms.
- **Lesson 12:** How To Estimate A Baseline Performance For Your Models.
- **Lesson 13:** How To Use Top Classification Machine Learning Algorithms.
- **Lesson 14:** How To Use Top Regression Machine Learning Algorithms.
- **Lesson 15:** How to Use Top Ensemble Machine Learning Algorithms.
- **Lesson 16:** How To Compare the Performance of Machine Learning Algorithms.
- **Lesson 17:** How to Tune the Parameters of Machine Learning Algorithms.
- **Lesson 18:** How to Save Your Machine Learning Model and Make Predictions.

After completing all of the lessons, you will be ready to work through standalone projects, end-to-end.

### 1.3.3 Part 3: Projects

This part contains three end-to-end projects that tie together the lessons from the previous part. Each project focuses on a different type of problem. The projects increase in complexity, starting off easy and straightforward and finish by using many advanced techniques you have learned. The projects you will work through in this part include:

- **Project 01:** Multiclass classification project to predict iris flower species from flower measurements.
- **Project 02:** Binary class classification project to predict the onset of diabetes from patient medical details.
- **Project 03:** Regression project to predict suburban house price from suburb details.

After completing this part, you will have solidified your knowledge of working through applied machine learning projects end-to-end and be ready to take on your own projects.



### 1.3.4 Part 4: Conclusions

Now that you are ready to take on your own projects, this part takes a moment to look back at how far you have come. The skills of applied machine learning are in great demand and it is important to appreciate exactly what you have learned and how you can bring those skills to your own projects. This part also lists valuable resources that you can consult to get more information and get answers to the inevitable technical questions that will come up.

## 1.4 Your Outcomes From This Process

This book will lead you from being a developer who is interested in applied machine learning to a developer who has the resources and capability to work through a new dataset end-to-end using Weka and develop accurate predictive models. Specifically, you will know:

- How to work through a small to medium sized dataset end-to-end.
- How to deliver a model that can make accurate predictions on new unseen data.
- How to complete all subtasks of a predictive modeling problem with Weka.
- How to learn new and different techniques in Weka.
- How to get help with Weka.

From here you can start to dive into the specifics of the techniques and algorithms used with the goal of learning how to use them better in order to deliver more accurate predictive models, more reliably in less time.

## 1.5 What This Book is Not

This book was written for professional developers who want to know how to build reliable and accurate machine learning models.

- **This is not a machine learning textbook.** We will not be getting into the basic theory of machine learning (e.g. induction, bias-variance trade-off, etc.). You are expected to have some familiarity with machine learning basics, or be able to pick them up yourself.
- **This is not an algorithm book.** We will not be working through the details of how specific machine learning algorithms work (e.g. random forest). You are expected to have some basic knowledge of machine learning algorithms or how to pick up this knowledge yourself.
- **This is not a programming book.** We will not be writing any code at all. Weka provides a Java API, but this API will not be covered in this book. We will focus exclusively on developing models using the Weka graphical user interface.

The beauty of Weka is that you can learn the process of applied machine learning and get good at delivering results without a strong background in algorithms or machine learning theory. The details and theory can come later, as you work to get better at the process of applied machine learning and delivering robust predictions and predictive models.

## 1.6 Summary

I hope you are as excited as me to get started. In this introduction chapter you learned that this book is unconventional. Unlike other books and courses that focus heavily on machine learning algorithms and theory and focus on little else, this book will walk you through each step of a predictive modeling machine learning project.

### 1.6.1 Next

Let's dive in. The next section will make the case as to why Weka is the best platform for beginners in applied machine learning.

## Chapter 2

# Rapidly Accelerate Your Progress in Applied Machine Learning With Weka

Why start with Weka over another tool like the R environment or Python for applied machine learning? In this chapter you will discover why Weka is the perfect platform for beginners interested in rapidly getting good at applied machine learning. After reading this chapter you will know:

- Why getting started in applied machine learning is hard.
- The one most important thing to focus on when getting started in applied machine learning.
- How to make best use of Weka when getting started in applied machine learning.

Let's get started.

### 2.1 Starting in Applied Machine Learning is Hard

When you start out in applied machine learning, there is so much to learn. For example:

- There are the algorithms.
- There is the data.
- There is the specific problem you are working on.
- There is the mathematics behind it all.
- There is the tool that you plan to use.

Often you are convinced that you need to learn a new programming language before you can get started in applied machine learning, like Python or more esoteric languages like Matlab or R. This does not have to be the case. It is so much easier to learn one thing well rather than try, and possibly fail to learn a host of new things.

## 2.2 Focus on Learning Just One Thing

The one thing to learn when you are starting in machine learning is how to deliver a result. That is, given a problem, how to work through it and deliver a set of predictions or how to deliver a model that can generate predictions. Not just predictions, but accurate predictions that can be delivered robustly and reliably, that you can put your name or your company's name against and in which you can feel confident. This is the most important skill to learn. It often involves steps like:

1. Defining your problem.
2. Preparing your data.
3. Evaluating a suite of algorithms.
4. Improving your results with tuning and ensembles.
5. Finalizing your model and present results.

This is the process of applied machine learning.

## 2.3 Learn the Process of Applied Machine Learning

The best tool to learn this process is the Weka machine learning workbench. There are 3 main reasons why this is the case:

- **Speed:** you can work your problem fast, giving you more time to try lots of ideas.
- **Focus:** it is just you and your problem, the tool gets out of your way.
- **Coverage:** it provides lots of state-of-the-art algorithms to choose from.

It saves you from the cruft that you can encounter with other platforms. You do not need to spend weeks learning a new language or API, and can focus on learning how to work through problems efficiently and effectively. You can focus on the one valuable thing you need to learn: the process of applied machine learning and delivering a result. Later, you can learn how to use more and different tools.

## 2.4 How to Best Use Weka

There is a specific way that you can use Weka to best aid you on your machine learning journey.

- **Practice on small in-memory datasets.** These are datasets with hundreds or thousands of instances so they are fast to work with and are standard datasets in the field, so that they are well understood.
- **Practice on different problem types.** Select standard datasets from a range of problem domains, such as biology, physics and advertising, and a range of problem types, such as binary and multiclass classification, regression, unbalanced datasets, and more.

- **Practice by exercising different parts of the tool.** Use a range of different techniques on different problems, including filtering methods, machine learning algorithms and even unsupervised methods like clustering and association rules.

These three simple principles will help you greatly accelerate your progress in developing skills in applied machine learning. Your learning will be focused on working through a problem and delivering a result in the form a set of accurate and reliable predictions or a model that can make ongoing predictions. We will go into more detail on how to make the best use of Weka in Chapter ???. The benefits of this approach will mean that you can greatly outpace others starting out in the field that are:

- Still figuring out how to implement an algorithm from scratch in code.
- Still figuring out how to use an esoteric programming language or API.
- Still figuring out how to setup their environment.

In applied machine learning, fast, reliable and systematic turnaround of results is more important than most other things. For this and more, Weka is your way forward.

## 2.5 Summary

In this chapter you discovered the importance of the Weka machine learning workbench for beginners in applied machine learning. You learned:

- That getting started in applied machine learning is hard because there is so much to learn.
- That the one most important thing to focus on in applied machine learning is delivering a reliable and robust result.
- That Weka can best be used by practicing on a suite of standard machine learning datasets.

### 2.5.1 Next

In the next section we will take a closer look at the Weka workbench and the features and benefits it provides to beginners in applied machine learning.

# Chapter 3

## How to Normalize and Standardize Your Machine Learning Data

Machine learning algorithms make assumptions about the dataset you are modeling. Often, raw data is comprised of attributes with varying scales. For example, one attribute may be in kilograms and another may be a count. Although not required, you can often get a boost in performance by carefully choosing methods to rescale your data. In this lesson you will discover how you can rescale your data so that all of the data has the same scale. After reading this lesson you will know:

- How to normalize your numeric attributes between the range of 0 and 1.
- How to standardize your numeric attributes to have a zero mean and unit variance.
- When to choose normalization or standardization.

Let's get started.

### 3.1 About Data Filters in Weka

Weka provides filters for transforming your dataset. The best way to see what filters are supported and to play with them on your dataset is to use the *Weka Explorer*. The *Filter* pane allows you to choose a filter.

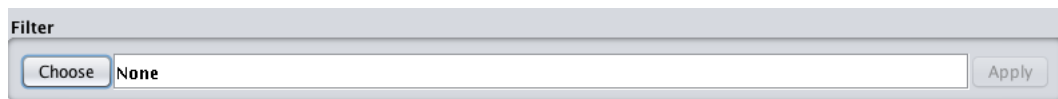


Figure 3.1: Weka Filter Pane for Choosing Data Filters.

Filters are divided into two types:

- **Supervised Filters:** That can be applied but require user control in some way. Such as rebalancing instances for a class.
- **Unsupervised Filters:** That can be applied in an undirected manner. For example, rescale all values to the range 0-to-1.

Personally, I think the distinction between these two types of filters is a little arbitrary and confusing. Nevertheless, that is how they are laid out. Within these two groups, filters are further divided into filters for Attributes and Instances:

- **Attribute Filters:** Apply an operation on attributes or one attribute at a time.
- **Instance Filters:** Apply an operation on instance or one instance at a time.

This distinction makes a lot more sense. After you have selected a filter, its name will appear in the box next to the *Choose* button. You can configure a filter by clicking its name which will open the configuration window. You can change the parameters of the filter and even save or load the configuration of the filter itself. This is great for reproducibility.

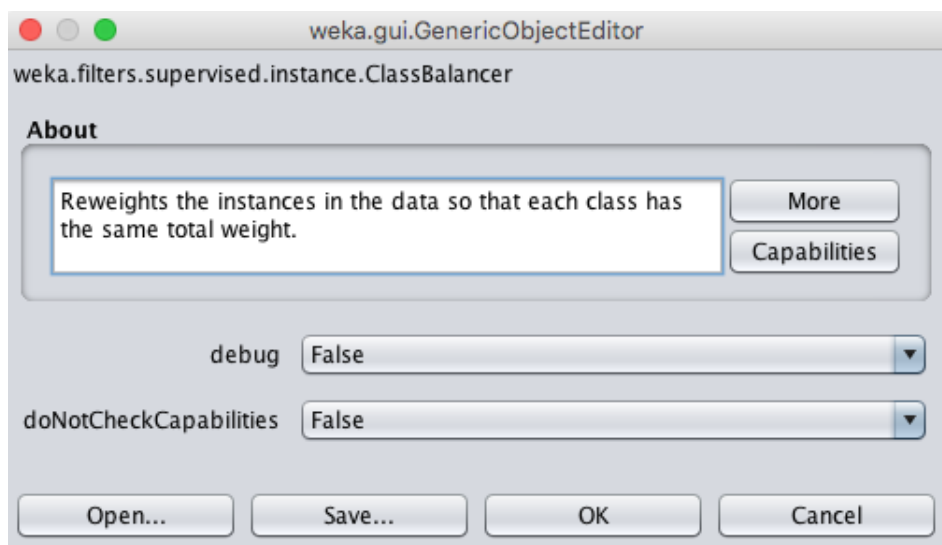


Figure 3.2: Weka Data Filter Configuration.

You can learn more about each configuration option by hovering over it and reading the tooltip. You can also read all of the details about the filter including the configuration, papers and books for further reading and more information about the filter works by clicking the *More* button.

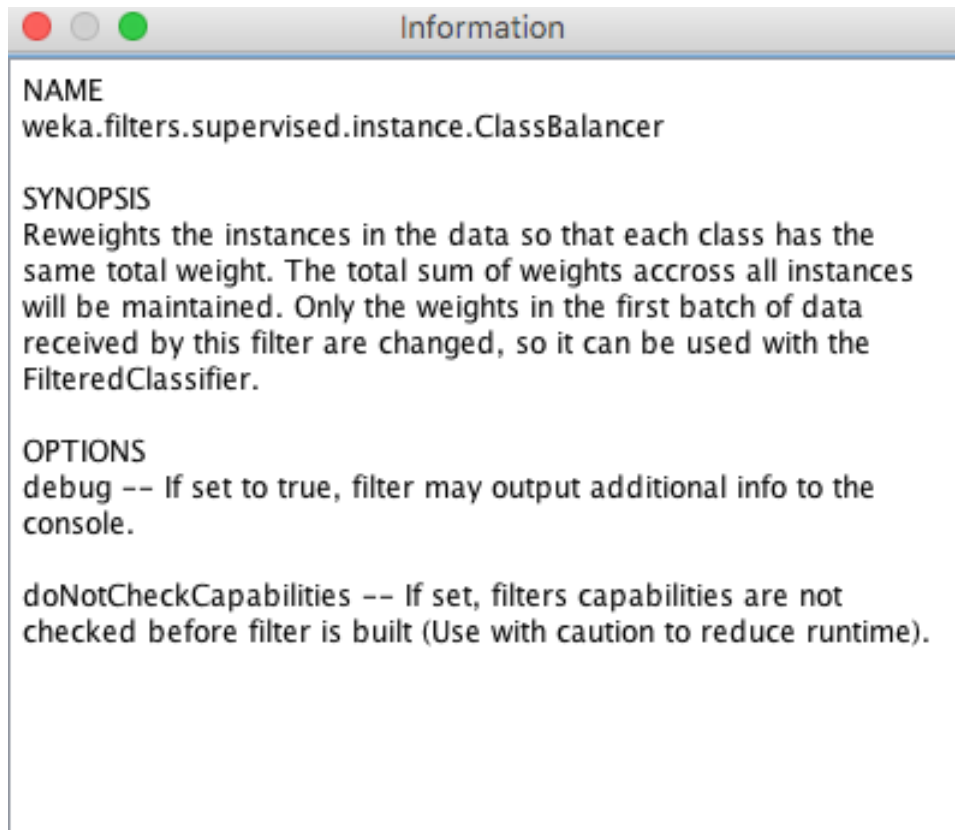


Figure 3.3: Weka Data Filter More Information.

You can close the help and apply the configuration by clicking the *OK* button. You can apply a filter to your loaded dataset by clicking the *Apply* button next to the filter name.

## 3.2 Normalize Your Numeric Attributes

Data normalization is the process of rescaling one or more attributes to the range of 0 to 1. This means that the largest value for each attribute is 1 and the smallest value is 0. Normalization is a good technique to use when you do not know the distribution of your data or when you know the distribution is not Gaussian (a bell curve).

The dataset used for this example is the Pima Indians onset of diabetes dataset. You can learn more about this dataset in Section ???. You can normalize all of the attributes in your dataset with Weka by choosing the *Normalize* filter and applying it to your dataset. You can use the following recipe to normalize your dataset:

- 1. Open the *Weka Explorer*.
- 2. Load the `data/diabetes.arff` dataset.



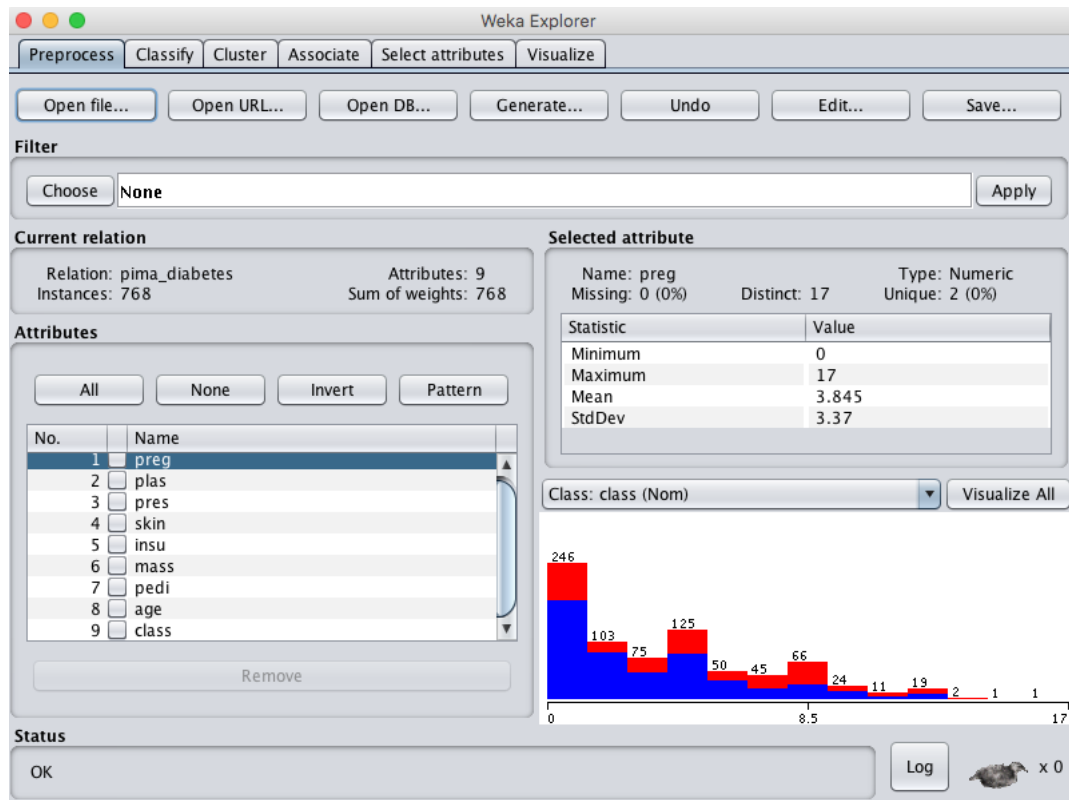


Figure 3.4: Weka Explorer Loaded Diabetes Dataset.

- 3. Click the *Choose* button and select the *unsupervised.attribute.Normalize* filter.

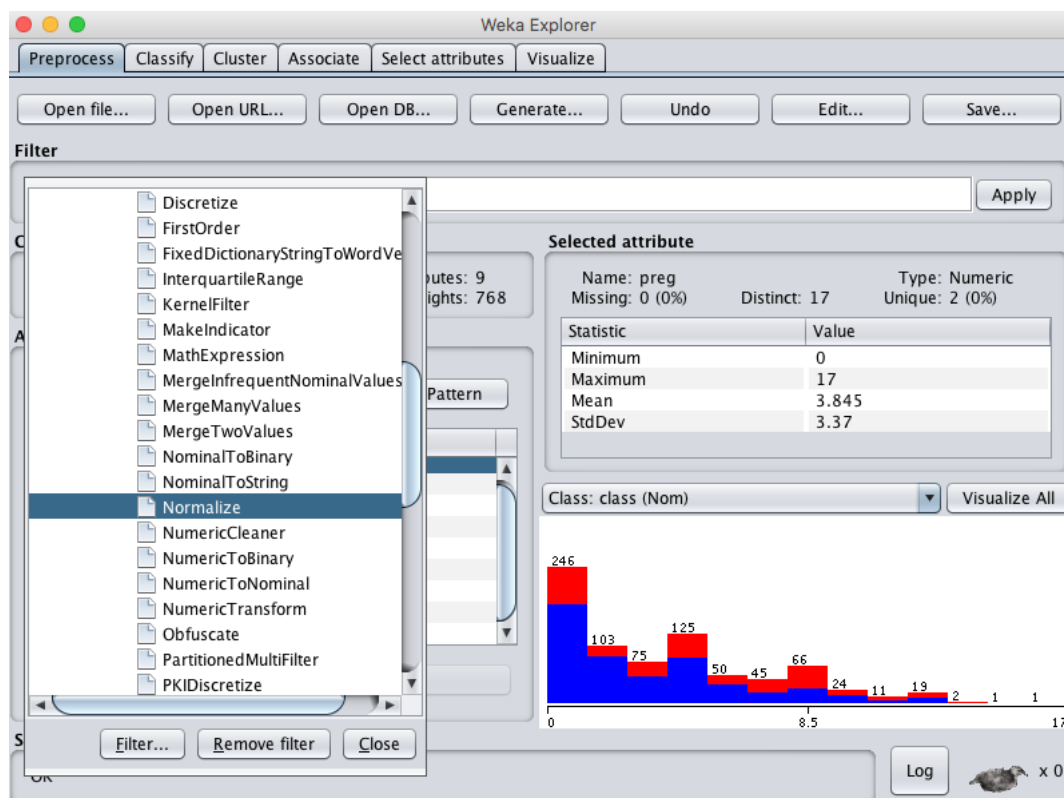


Figure 3.5: Weka Select Normalize Data Filter.

- 4. Click the *Apply* button to normalize your dataset.
- 5. Click the *Save* button and type a filename to save the normalized copy of your dataset.

Reviewing the details of each attribute in the *Selected attribute* window will give you confidence that the filter was successful and that each attribute was rescaled to the range of 0 to 1.

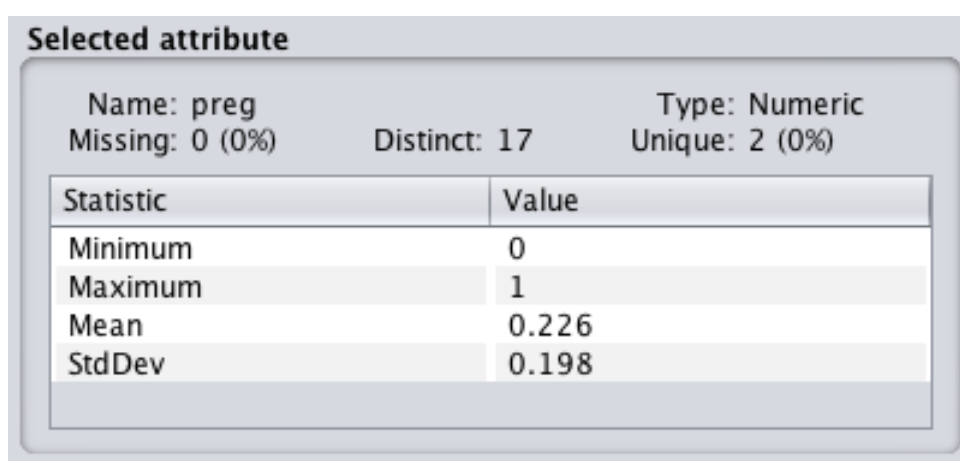


Figure 3.6: Weka Normalized Data Distribution.

You can use other scales such as -1 to 1, which is useful when using Support Vector Machines and AdaBoost. Normalization is useful when your data has varying scales and the algorithm

you are using does not make assumptions about the distribution of your data, such as  $k$ -Nearest Neighbors and Artificial Neural Networks.

### 3.3 Standardize Your Numeric Attributes

Data standardization is the process of rescaling one or more attributes so that they have a mean value of 0 and a standard deviation of 1. Standardization assumes that your data has a Gaussian (bell curve) distribution. This does not strictly have to be true, but the technique is more effective if your attribute distribution is Gaussian. You can standardize all of the attributes in your dataset with Weka by choosing the *Standardize* filter and applying it your dataset. You can use the following recipe to standardize your dataset:

- 1. Open the *Weka Explorer*.
- 2. Load the `data/diabetes.arff` dataset.
- 3. Click the *Choose* button to and select the *unsupervised.attribute.Standardize* filter.

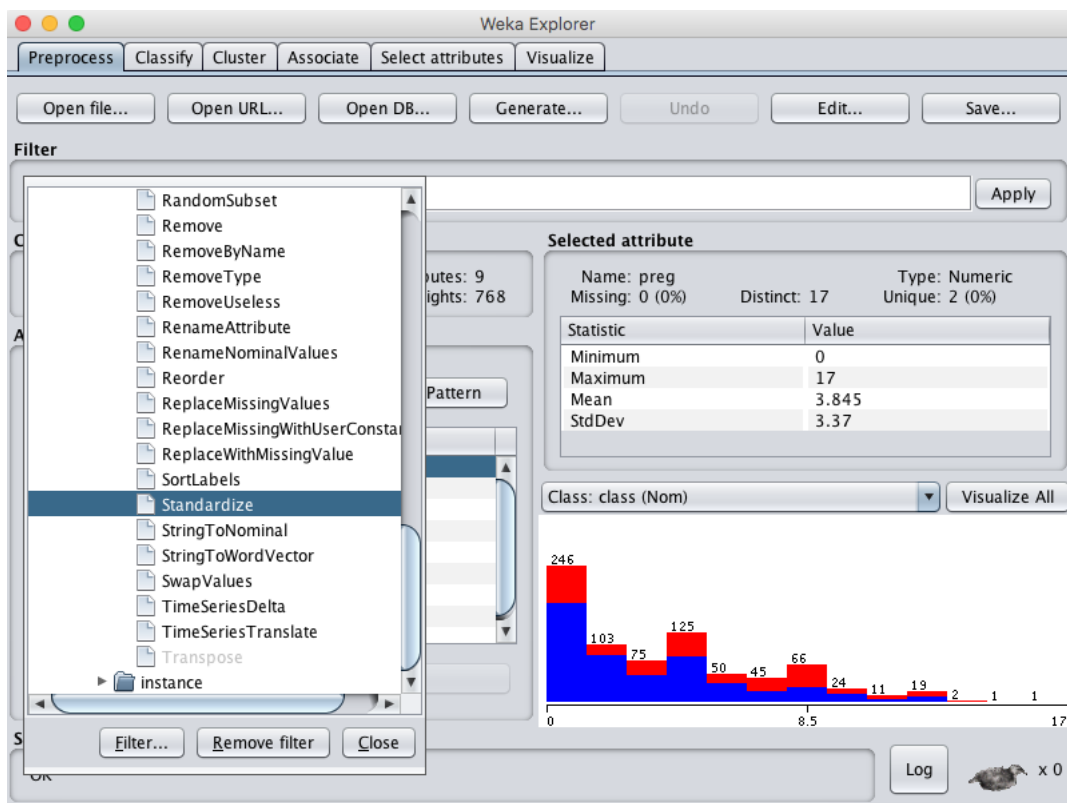
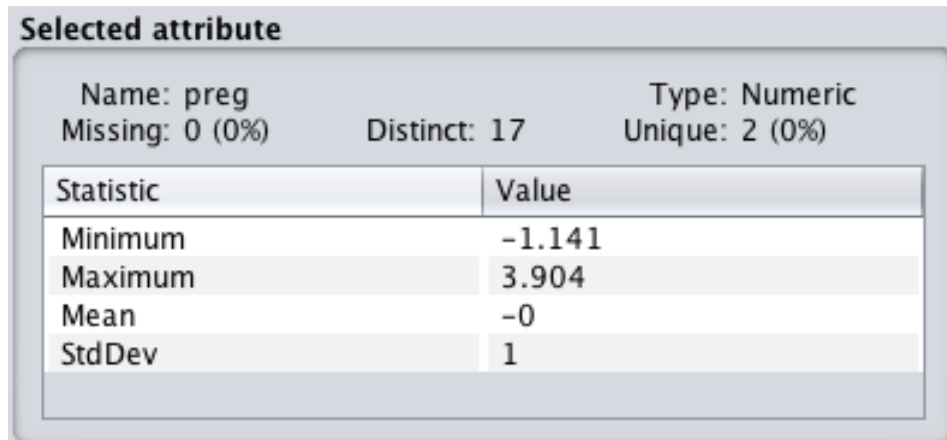


Figure 3.7: Weka Select Standardize Data Filter.

- 4. Click the *Apply* button to normalize your dataset.
- 5. Click the *Save* button and type a filename to save the standardized copy of your dataset.

Reviewing the details of each attribute in the *Selected attribute* window will give you confidence that the filter was successful and that each attribute has a mean of 0 and a standard deviation of 1.



Selected attribute	
Name: preg	Type: Numeric
Missing: 0 (0%)	Distinct: 17
	Unique: 2 (0%)
Statistic	Value
Minimum	-1.141
Maximum	3.904
Mean	-0
StdDev	1

Figure 3.8: Weka Standardized Data Distribution.

Standardization is useful when your data has varying scales and the algorithm you are using does make assumptions about your data having a Gaussian distribution, such as linear regression, logistic regression and linear discriminant analysis.

## 3.4 Summary

In this lesson you discovered how to rescale your dataset in Weka. Specifically, you learned:

- How to normalize your dataset to the range 0 to 1.
- How to standardize your data to have a mean of 0 and a standard deviation of 1.
- When to use normalization and standardization.

### 3.4.1 Next

Weka provides a large assortment of data filters. In the next lesson you will learn how you can transform attributes using more advanced data filters.

# This is Just a Sample

Thank-you for your interest in **Machine Learning Mastery With Weka**.  
This is just a sample of the full text. You can purchase the complete book online from:  
<https://machinelearningmastery.com/machine-learning-mastery-weka>

