

# Learn data science by doing data science (and football analytics)

Hans Samson, ANWB

March 7, 2024

# How to teach yourself data science from scratch

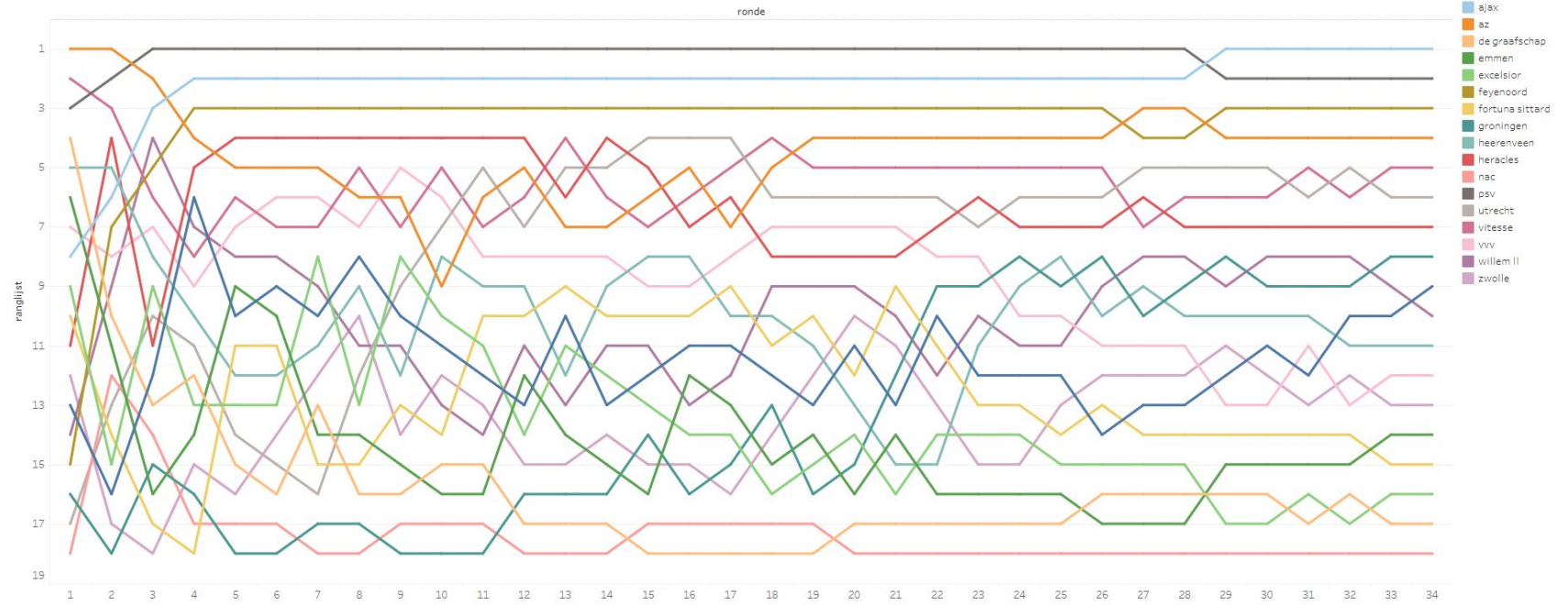
- Define a real-world use case
- Get started, Just Do It
- Choose or create a familiar dataset
- Take small, manageable steps, try the simplest things first
- When stuck; don't panic
- Reflect on your progress and outcomes
- Stay motivated and curious, keep on learning

# Define a real-world use case.

Started 4 years ago recreating this visualisation in Tableau.



Positie op de ranglijst

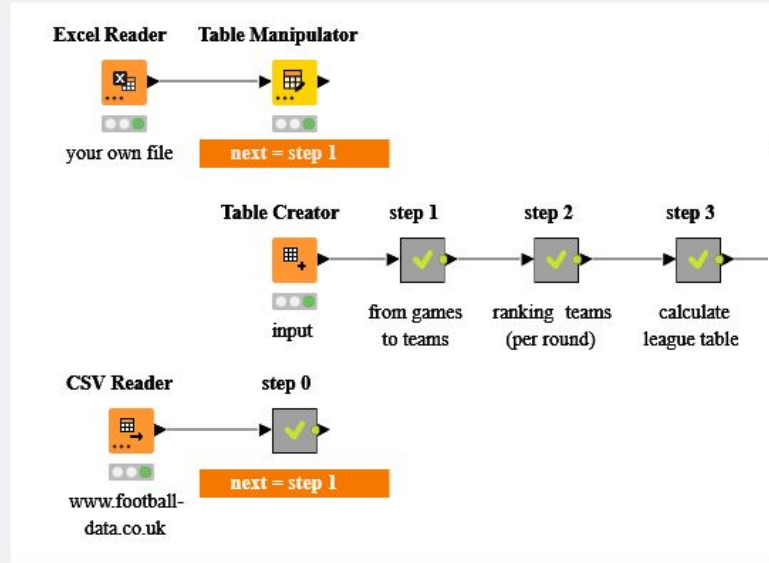


# Choose or create a familiar dataset

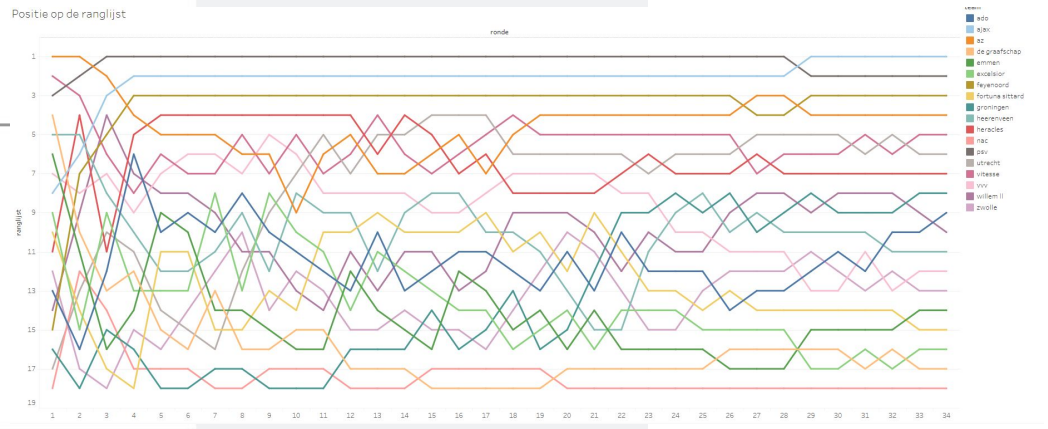
## 02 Creating Your Sports Analytics Dataset

Sport analytics | Insights | Football | Soccer | Dataset

Draft - Latest edits on Nov 16, 2023 7:48 AM



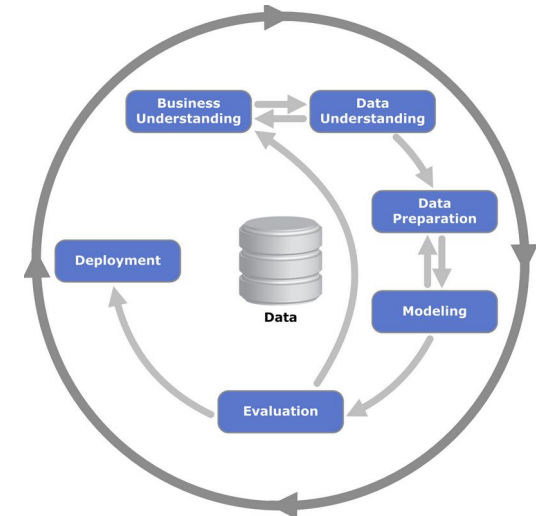
S	Date	S	Time	S	HomeTeam	S	AwayTeam	I	FTHG	I	FTAG
	29/10/2023		14:00		Aston Villa		Luton		3		1
	29/10/2023		14:00		Brighton		Fulham		1		1
	29/10/2023		14:00		Liverpool		Nott'm Forest		3		0
	29/10/2023		15:30		Man United		Man City		0		3
	04/11/2023		12:30		Fulham		Man United		0		1
	04/11/2023		15:00		Brentford		West Ham		3		2
	04/11/2023		15:00		Burnley		Crystal Palace		0		2
	04/11/2023		15:00		Everton		Brighton		1		1
	04/11/2023		15:00		Man City		Bournemouth		6		1
	04/11/2023		15:00		Sheffield United		Wolves		2		1
	04/11/2023		17:30		Newcastle		Arsenal		1		0
	05/11/2023		14:00		Nott'm Forest		Aston Villa		2		0



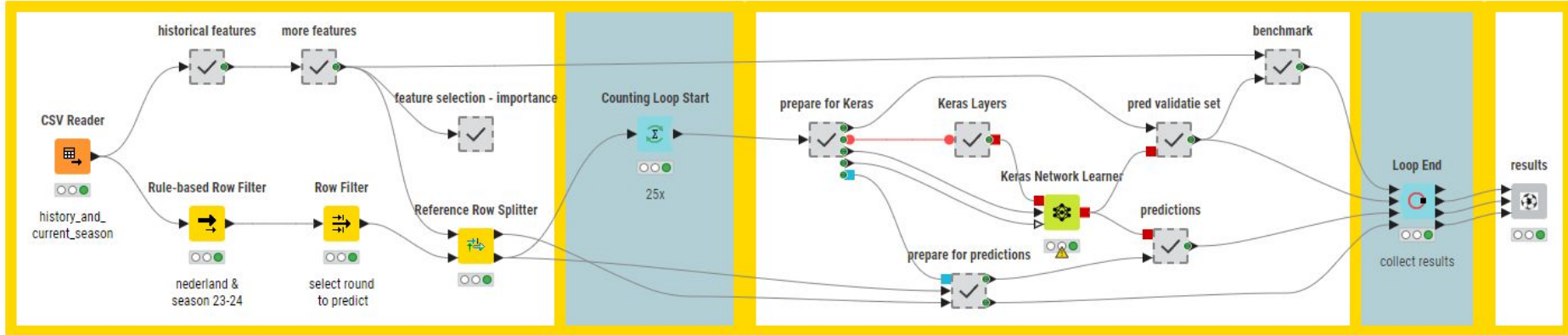
# From input to output - take small, manageable steps

Predicting the outcome of a football match

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
<p><b>Determine Business Objectives</b> Background Business Objectives Business Success Criteria</p> <p><b>Assess Situation</b> Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits</p> <p><b>Determine Data Mining Goals</b> Data Mining Goals Data Mining Success Criteria</p> <p><b>Produce Project Plan</b> Project Plan Initial Assessment of Tools and Techniques</p>	<p><b>Collect Initial Data</b> Initial Data Collection Report</p> <p><b>Describe Data</b> Data Description Report</p> <p><b>Explore Data</b> Data Exploration Report</p> <p><b>Verify Data Quality</b> Data Quality Report</p>	<p><b>Select Data</b> Rationale for Inclusion/Exclusion</p> <p><b>Clean Data</b> Data Cleaning Report</p> <p><b>Construct Data</b> Derived Attributes Generated Records</p> <p><b>Integrate Data</b> Merged Data</p> <p><b>Format Data</b> Reformatted Data  Dataset Dataset Description</p>	<p><b>Select Modeling Techniques</b> Modeling Technique Modeling Assumptions</p> <p><b>Generate Test Design</b> Test Design</p> <p><b>Build Model</b> Parameter Settings Models Model Descriptions</p> <p><b>Assess Model</b> Model Assessment Revised Parameter Settings</p>	<p><b>Evaluate Results</b> Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models</p> <p><b>Review Process</b> Review of Process</p> <p><b>Determine Next Steps</b> List of Possible Actions Decision</p>	<p><b>Plan Deployment</b> Deployment Plan</p> <p><b>Plan Monitoring and Maintenance</b> Monitoring and Maintenance Plan</p> <p><b>Produce Final Report</b> Final Report Final Presentation</p> <p><b>Review Project</b> Experience Documentation</p>

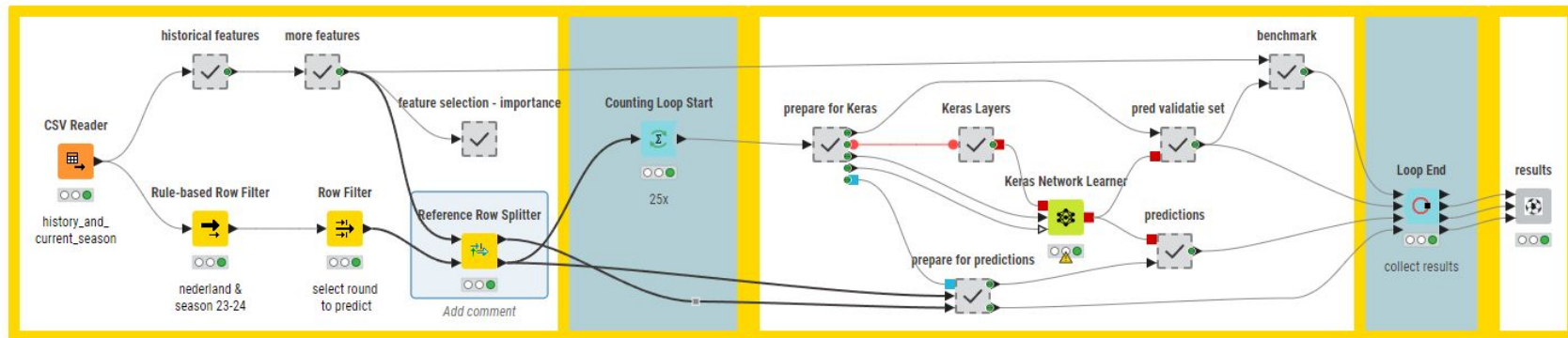


# From input to output





# From input to output



► 1: Splitted table comprising reference rows. ► 2: Splitted table comprising all other rows. 📄 Flow Variables

Rows: 9 | Columns: 121

Table 📄 Statistics 📄

#	RowID	team	tegensta...	wedstrijd...	ronde	score	thuis_uit	datum	seizoen	source	thuisstea...	thuisstea...	thuisstea...	thuisstea...	thuisstea...
		String	String	String	Number (inte...	String	String	Local Date	String	String	String	String	String	String	Number
1	Row...	ajax	nec	ned-23-24-196	22	gelijk	thuis	2024-02-18	23-24	ned	ajax	ned-23-24-196	thuis	23-24	8
2	Row...	feyenoord	rkc	ned-23-24-198	22	gelijk	thuis	2024-02-18	23-24	ned	feyenoord	ned-23-24-198	thuis	23-24	0
3	Row...	fortuna sittard	az	ned-23-24-191	22	gelijk	thuis	2024-02-17	23-24	ned	fortuna sittard	ned-23-24-191	thuis	23-24	1
4	Row...	heerenveen	ga eagles	ned-23-24-194	22	gelijk	thuis	2024-02-17	23-24	ned	heerenveen	ned-23-24-194	thuis	23-24	0
5	Row...	psv	heracles	ned-23-24-190	22	gelijk	thuis	2024-02-16	23-24	ned	psv	ned-23-24-190	thuis	23-24	0
6	Row...	sparta	excelsior	ned-23-24-192	22	gelijk	thuis	2024-02-17	23-24	ned	sparta	ned-23-24-192	thuis	23-24	1

# From input to output

The screenshot displays a KNIME workflow and its results. On the left, a workflow diagram shows a 'CSV Reader' feeding into a 'Rule-based Row Filter' and a 'historical fe...' node. The 'Rule-based Row Filter' is connected to a 'nederlan...' node. The main window shows a table of match results with columns for match number, team, opponent, wins, draws, losses, and prediction. Below the table, a 'Flow Variables' section shows a list of match pairs, with 'heerenveen - ga eagles' selected. To the right, a box plot titled 'heerenveen - ga eagles' compares the distribution of 'verlies' (loss) and 'winst' (win) predictions. The 'verlies' box is blue and the 'winst' box is orange. The plot shows the median, quartiles, and range for both categories.

wedstrijdnummer	team	tegenstander	winst	gelijk	verlies	prediction
ned-23-24-190	psv	heracles	25			winst
ned-23-24-191	fortuna sittard	az	2	1	22	verlies
ned-23-24-192	sparta	excelsior	24	1		winst
ned-23-24-193	zwolle	almere	25			winst
ned-23-24-194	heerenveen	ga eagles	12		13	verlies
ned-23-24-195	twente	utrecht	25			winst
ned-23-24-196	ajax	nec	25			winst
ned-23-24-197	vitesse	volendam	25			winst
ned-23-24-198	feyenoord	rkc	25			winst

Showing 1 to 9 of 9 entries

Flow Variables

Count: 2

Owner ID

6:1204

Refresh

heerenveen - ga eagles

psv - heracles  
fortuna sittard - az  
sparta - excelsior  
zwolle - almere  
**heerenveen - ga eagles**  
twente - utrecht  
ajax - nec  
vitesse - volendam  
feyenoord - rkc

heerenveen - ga eagles

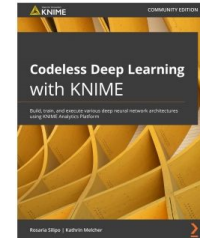
Box Plot Data:

Category	Min	Q1	Median	Q3	Max
verlies	0.35	0.38	0.40	0.43	0.46
winst	0.34	0.37	0.39	0.45	0.49



# When stuck; don't panic

The screenshot shows the KNIME software interface. At the top, there are navigation tabs for 'All', 'Workflows', 'Nodes', 'Components', 'Extensions', and 'Collections'. A search bar is on the right. Below the navigation, there's a 'Filter by tag' section with buttons for 'Codeless Deep Learning with KNIME', 'Deep learning', 'Neural networks', and 'Keras'. The main content area displays a workflow card for 'Simple Example for Binary Classification with Keras'. The card includes a 'Workflow' icon, the title, a description 'This workflow trains a fully connected, feedforward network using the adult dataset.', the author 'kathrin', and a 'Draft only' status. There are also tags for 'Neural networks', 'Deep learning', and 'Codeless Deep Learning with KNIME'.



 Machine Learning Mastery  
<https://machinelearningmastery.com> › Blog

## Multi-Class Classification Tutorial with the Keras Deep ...

7 Aug 2022 — In this tutorial, you will discover how to use **Keras** to develop and evaluate neural network models for **multi-class classification** problems. ...

 Stack Overflow  
<https://stackoverflow.com> › questions › how-to-do-multi- ...

## How to do Multiclass classification with Keras?

5 Nov 2020 — You can play with the layers, add more neurons, increase or reduce the number of layers. If you notice that the accuracy is good on training but ...

2 answers · Top answer: You need to convert your string categories to integers, there is a me...

**Keras LSTM Multiclass Classification** structure - Stack Overflow 31 Jan 2021  
How to improve accuracy with **keras multi class classification**? 29 Apr 2020  
How to do **multi-class** image **classification** in **keras**? 11 Oct 2017  
**Multiclass classification LSTM keras** - python - Stack Overflow 7 Jul 2021  
More results from [stackoverflow.com](https://stackoverflow.com)



 HackerNoon  
<https://hackernoon.com> › multiclass-classification-with-k... ›

## Multiclass Classification with Keras

19 Sept 2022 — In this article, we will be focusing on a **multiclass classification** problem with practical code examples written with **Keras**. We will use cross- ...

  **KNIME 4.1: Keras error ( Selected Keras backend 'Keras(Tensorflow)' is not available anymore**  
KNIME Extensions deep-learning  
Jan '20 - Hi, I am trying to configure the **Keras** Network Learner and I am getting an error. Screenshot below. Error%20Screenshot I have set up the Python and Deep Learning libraries as instructed. W...

  **KNIME 3.6.0 error: "Selected Keras back end 'Keras (TensorFlow)' is not available anymore"**  
KNIME Extensions deep-learning  
Jul '18 - Hi, I tested the **Keras**+Tensorflow capabilities of KNIME 3.5 last week and found them very good. After updating to KNIME 3.6 I try to use (again) the example like "01\_Classi...

  **Keras Network Learner ERROR - Failed to save Keras deep learning network | Ubuntu 22.04**  
KNIME Extensions  
Apr '23 - Hi, I use Knime on Ubuntu 22.04. I'm a beginner and I received the following error when executing a workflow ERROR **Keras** Network Learner 3:16 Execute failed: An error occurred while creating the **Keras** network from its layer specifications. Details: Failed to save **Keras**...

  **Changing Keras deep learning network configuration within loop stops loop flow: Keras Network Learner**  
KNIME Analytics Platform  
Feb '20 - I have a **Keras** network learner within a loop. The loop cycles through various previously saved **Keras** networks (saved as .h5 files with different network configurati...

# Reflect on your progress and outcomes

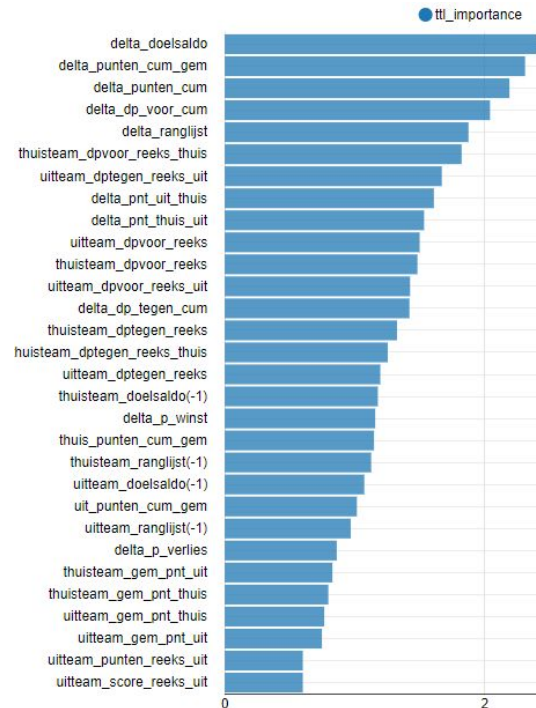
Does my input and output makes sense?

Cross Tabulation of dp\_voor by dp\_tegen

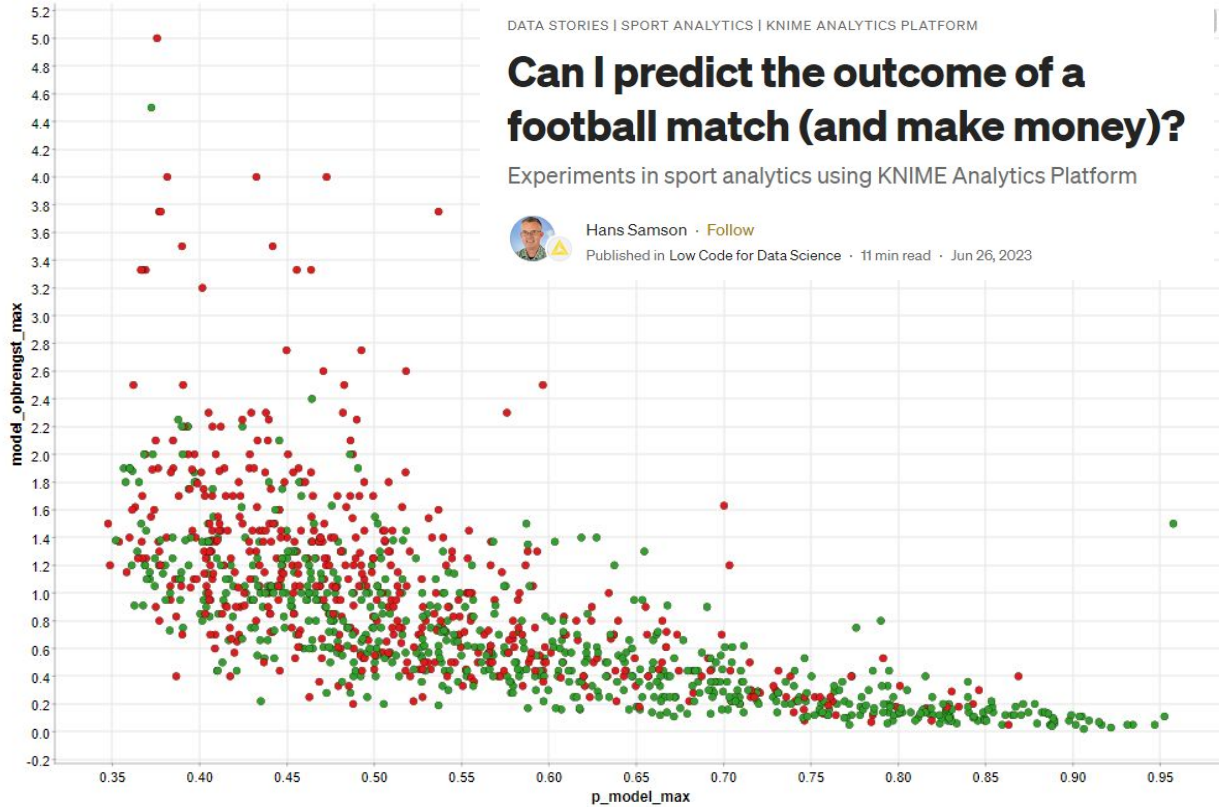
Frequency	0	1	2	3	4	5	6	7	8	9	13	Total
0	1.731	1.855	1.183	591	247	81	24	14	4	2	1	5.733
1	2.336	2.998	1.714	752	306	89	21	8	1			8.225
2	1.883	2.225	1.269	514	167	50	16	3	1			6.128
3	1.115	1.126	605	266	70	17	5					3.204
4	499	434	231	92	23	7						1.286
5	232	198	90	27	6	1						554
6	78	60	29	7	1							175
7	15	15	2	2								34
8	11	1	2									14
9	4	1										5
10	1		1									2
Total	7.905	8.913	5.126	2.251	820	245	66	25	6	2	1	25,360

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa ( $\kappa$ )	Correctly Classified	Incorrectly Classified
52.01%	47.99%	0.196	28046	25879



# Stay motivated and curious



As the chance of success increases, the potential profit decreases.

# My 7 tips to teach yourself data science from scratch

1. Define a real-world use case
2. Get Started, Just Do It
3. Choose or create a familiar dataset
4. Take small, manageable steps
5. When stuck; don't panic
  - a. Seek additional information
  - b. Engage in structured learning
  - c. Don't hesitate to ask for help
6. Reflect on your progress and outcomes
7. Stay motivated and curious, keep on learning