Introduction to Data Science

Teacher guidance

General guidance

The MEI Introduction to Data Science course is a free, self-study course that is suitable for both Core Maths and A Level Mathematics students. It is designed to introduce students to some key ideas in data science and provide them with some valuable insights into careers and opportunities for further study in this exciting field.

Use of real data

This course uses the A Level Mathematics large data sets for the AQA, Edexcel, MEI and OCR specifications as examples throughout the course, but it doesn't matter if you haven't used them before.

Interactive coding activities

The lessons involve coding activities that can be accessed in Kaggle, a web-based programming environment. No previous coding experience is needed to access the course as all the commands used are given. Students will interact with activities by copying, pasting and editing existing code.

Lessons

1. Introduction to Data Science
2. Pre-processing and Cleaning Data
3. Data Presentation and Visualisation
4. The Data Science Cycle
5. Introduction to Machine Learning
6. Machine Learning, AI and Bias

Lesson 1: Introduction to Data Science

Lesson objectives

- To introduce students to the key themes of data science.
- To give students an opportunity to work with some data in a coding environment.

General guidance for the lesson

The three main themes of data science are:

- Statistical understanding,
- Technology skills,
- Domain knowledge (making context-based decisions about working with data).

The key focus of this lesson is for students to be comfortable in working with data in a coding environment. The coding environment used in the course is Kaggle, a web-based platform, where students can access pages featuring text and code, known as *notebooks*. Students should ensure that they can use Kaggle before moving on to the rest of the lessons in the course.

Presentation notes

Slide	Title	Notes
1	Title	
2	Data Science: gaining insight from data	What is data science? Data science is the intersection of three main ideas: - Statistical understanding e.g. using averages and charts to make sense of data - Technology skills – processing the data at scale needs technology – you'll see how code such as Python is the fastest way to do this - Domain knowledge – you will need to make lots of decisions in context. Understanding the context is a key skill for a data scientist e.g. with weather data negative temperature is OK but negative rainfall isn't
3	Why is there a need for Data Science?	• In the last 20 years there has been a massive shift in the quantity of data – there is more data recorded every day than in the whole of human history until the year 2000. • There is also a lot of live data being processed. • Data Science is using technology to spot patterns in these data. In practice, code is the most efficient way to work with large quantities of data.
4	Working with real data	A couple of examples • Amazon: customer data, lots of it, data comes in live. • Met office: data from a large number of different sensors around the world. You wouldn't want to analyse these data with a spreadsheet!

Slide	Title	Notes
5	The solution: code	• Coding is scalable: if it works for a data set of 100 items then we can use it on one with a million or more – the only limit is the processing power. • Python has been used in this course as it used by real data scientists, and you might have met it before at school or college.
6	Programming in Python using Kaggle	Time to get going on a notebook in Kaggle – an online platform where you can use Python to work with data and save your *notebooks* (the name for documents mixing code and text) The course should work for both new and experienced coders: • If you have not used Python much **don't worry**: all the code you need is given. • If you are confident with Python **don't worry**: there are plenty of opportunities to explore further. **You will need to create an account at kaggle.com to save your notebooks.** If you have a Google account it is straightforward to use this to generate a Kaggle account; if not you can create one linked to any email address.
7	Lesson 1 activity	In this activity you will explore a data set using two main sets of commands in Python – these are known as *libraries.* The libraries you will use are *Pandas* and *Seaborn*. To start using the activity click **Copy and edit**. There is more help in the "before you start" and "guidance" sections. See below for guidance on the activity.
8	Data Science: gaining insight from data	In the activity you used all three of these skills as a data scientist: - You demonstrated **domain knowledge** when you interpreted the results based on what you know about weather and the geography of the UK. - You used **technology skills** in working with the code. - You used **statistical understanding** to interpret different statistics and charts. For example, you understood the difference between a mean and a median.

9	Key strands of data science	As you work through this course you will meet 4 key strands of data science: • Data need **pre-processing**. This is an umbrella term that covers everything between getting the data and being able to calculate some statistics or create some charts. This includes things such as formatting numbers/text, cleaning data by removing obvious errors, removing characters, moving data from one source to another, … and much more! It's sometimes called *data wrangling.* If you tried the extension for the activity in this lesson you will have seen this for how to deal with rainfall recorded as "trace". The next lesson in the course focusses on this in more detail. • **Exploratory data analysis** – in data science you are often given a data set and your first task is to look for any patterns in it. This is different to collecting data to solve a problem as you need to decide what questions to ask about the data. • **Visualisations** help people understand data - it's a very creative area beyond the histograms and box plots you meet in school maths e.g. Gapminder. • **Predictive models** can be built from data using machine learning. A predictive model will often be a formula or process for predicting an unknown quantity, for example, how much a customer will spend next year or whether someone will like a TV program on a streaming platform. This is different to an explanatory model which attempts to find if there is a link between things, for example people with healthier diets tend to suffer less illness. You'll see this in one of the later lessons.
10	Lessons	Well done on completing lesson 1. Describe lessons 2-6.
11	*Meet a data scientist* videos	One of the additional features of this course is that you will get to hear from practising data scientists. In each section there is a "Meet a data scientist" video, These are about 5 minutes long and give you an insight into the many and varied uses of data science.
12	Further support with Python and using Kaggle	There is more help in the "before you start" and "guidance" sections.

Activity notes

The main purpose of this activity is for students to become familiar with opening, running and editing a notebook on Kaggle.

To get started students should:

- Click **Copy and edit**
  This will save the notebook to their own Kaggle area (if they've created an account and logged-in). The notebooks auto-save whilst they are being worked on but you can also use the Save button top-right to save a version.
- Click **Run All**
  The code in the notebook needs to run in the order it is on the page. For example, the libraries need to be imported before the commands can be used. It is essential that Run All is pressed every time the page is accessed. Occasionally the page can lose access to the Kaggle server – if this happens then pressing Run All again will reconnect it.
- Try changing some code in one of the code boxes and then pressing Run for that block.
- Try copying, pasting and editing some code into a blank code block.

An example of a completed notebook, including suggested answers to the checkpoints, is given at: *https://mei.org.uk/introduction-to-data-science/further-support/*

Lesson 2: Pre-processing data

Lesson objectives

- To introduce students to the importance of pre-processing data.
- To give students an opportunity to work with some data that requires pre-processing.

General guidance for the lesson

Pre-processing data is an umbrella term that encompasses the processes involved in preparing data so that statistics can be calculated, and charts and other visualisations can be produced. Students might previously have 'cleaned' data by removing errors; however, there are many other aspects to pre-processing data, such as storing and moving data, creating new features, and filtering the data. The pre-processing decisions will usually depend on the context so it is difficult to give a comprehensive list of all the processes that could be needed.

When analysing large quantities of data using code this is a much more significant part of working with data than in traditional school statistics. Some estimates suggest that this can be 60-80% of a data scientist's time. Pre-processing data is often referred to as *data wrangling*.

Presentation notes

Slide	Title	Notes
1	Title	
2	Pre-processing and cleaning data	Working with large quantities of data using code creates new challenges. These include: • moving data between systems, • checking and cleaning the data, • formatting the data types so they are usable. This is a major aspect of data science. It is often known as *data wrangling*. • Data might be stored in a database or coming from a live feed. This will then need to be imported into the package being used for the analysis. • You will need to know something about the context to decide if there are errors in the data. • You've already seen the importance of knowing if data is stored as numbers or text. You will increasingly see that this is a major part of working with data. Some estimates suggest that 60-80% of a data scientist time is on data wrangling.
3	Example: extension activity in lesson 1	If you tried the extension in the lesson 1 activity you will have seen an example of this already. In the extension you looked at the column for rainfall which was stored as text as some of the values were tr (short for trace). The code you were shown replaced the tr values with 0.025 and then converted the number to a decimal (or *float*). You'll see some more examples of pre-processing in the activities for this lesson.

Slide	Title	Notes

4	Cleaning, formatting and filtering data	Some of the things you will meet in the activities are:
		
		• Cleaning:
		  ○ Removing incorrect data e.g. you'll see some cars where the mass is 0
		  ○ Removing characters: e.g. you'll see some number that have a comma separator for the 1000s
		• Formatting data types:
		  ○ e.g. once you've removed the comma the text feature needs to be converted to a number so pandas can analyse it
		• Derived fields (this is often called *feature engineering*)
		  ○ Percentages: e.g. percentage of people in a local authority that travel by bike
		  ○ Creating text from numeric codes: the cars data uses a code of 1 for petrol and 2 for diesel – it is useful to have a text feature for this
		• Filtering: You can then look at just the petrol or just the diesel cars by filtering.
		
		These are just some examples of pre-processing data. The important aspect to focus on in each case is that it makes sense in the context you are working in. This is you using your domain knowledge about cars or how people travel to work.

5	Activities	• Activity 2-1: Are people more likely to cycle to work or walk to work in different parts of the country? • Activity 2-2: Are petrol or diesel cars heavier and which have higher CO2 emissions? Are there differences for cars of different makes? See below for guidance on the activities.
6	Cleaning, formatting and filtering data	You've met some different examples of pre-processing data in the activity. For example: • You removed some cars data where the mass was 0. • You created a column for the percentage of people in a local authority that travel by bike. • You filtered for petrol cars. This is not an exhaustive list of everything you might need to do to pre-process data. Every data set is different and understanding the context is important. For example, it is likely that a value of 0 is an error for the mass of a car but a value of 0 for the number of people who travel by train in the Scilly Isles could be realistic.

Slide	Title	Notes
7	Coding: automating processes	One of the advantages of using code is that the solutions are scalable. You changed the value for 347 rows, but it would have been no more difficult to change it for 347,000 rows, or even more, and, if you know a little more coding, you could adapt this so it did all the columns at once. If you want to know more about pre-processing, or data wrangling, there are some links in the further reading section of this lesson. In the next lesson you'll see some examples of how you can enhance your data analysis with some engaging visualisations.

Activity notes

- **Activity 2-1: Are people more likely to cycle to work or walk to work in different parts of the country?**

  This activity is designed to demonstrate two key aspects of pre-processing data: changing the data type and creating a derived feature. It is common for numerical data to have been recorded in a way that it is stored as text (such as because of a comma separator). These need changing to a numerical feature before it can be processed.

  There are also cases when it useful to create a new feature from existing features, such as by finding the total of some columns or dividing one by the other to calculate a proportion.

- **Activity 2-2: Are petrol or diesel cars heavier and which have higher CO2 emissions? Are there differences for cars of different makes?**

  This activity is designed to demonstrate two further aspects of pre-processing data: cleaning the data and changing the data type to text. Students should be aware that data is often recorded or stored with errors. One common error that they meet in this activity is the use of '0' when there is no data. This can occur because of the way the data has been recorded or how it has been transferred electronically. It is important that students should be able to give a valid context-based reason why data points should be removed or not. In this example a car of mass 0kg is not possible; however, there are other cases where a value of 0 is possible, for example, rainfall in a day. When removing data it is good practice to take a new copy of the data table with the values removed so that the original table is still available.

  The other pre-processing strategy covered in this activity is conversion from a numerical to text feature. Numbers are often used as a key to store options for a text feature. This can be because it is easier to record and can take less space to store it. Many of the commands built into both the Pandas and Seaborn libraries will automatically interpret numerical features as representing numerical quantities and so this can result in unwanted analysis. For example, in this case it wouldn't make sense to calculate the mean Propulsion Type. Creating an additional text feature with the words for the categories avoids the potential for confusion.

An example of completed notebooks, including suggested answers to the checkpoints, is given at: *https://mei.org.uk/introduction-to-data-science/further-support/*

Lesson 3: Data presentation and visualisation

Lesson objectives

- To explore how visualisations can be used to observe and communicate patterns in data.
- To give students an opportunity to explore the charts available using the *Seaborn* data visualisation library in Python.

General guidance for the lesson

Visualisations are an important part of data science. They can be used for two purposes:

- As an exploratory tool to find patterns in data.
- As a communication tool to display patterns in the data.

The phrase "*A picture paints a thousand words*" is commonly used. This applies to data visualisation: it is often easier to understand patterns in data when they are represented visually as opposed to when the data is presented in a table or the analysis of data is presented using only numerical statistics.

In this lesson students will meet some charts they are familiar with; however, they will also meet some less familiar ones. Generating these in code means that students do not need to focus on how to draw them and, instead, they can focus on how useful the visualisations are for displaying any patterns.

Presentation notes

Slide	Title	Notes
1	Title	
2	Visualisation	Charts and diagrams represent data graphically to: • Emphasise the patterns in the data • Make the patterns easier to understand Visualisations can help you see patterns in the data and explain those patterns to other people. In this lesson you will mainly focus on the first of these: how you can observe patterns in the data using visualisations. You might be able to think of some examples where you've seen visualisations to communicate trends in data, such as in the news. For some good examples of this see Beautiful News (informationisbeautiful.net).
3	Standard data presentation charts in Seaborn	• Box plot • Histogram • Scatter diagram You should have met these at GCSE and in A level/Core Maths. These visualisations that have been used extensively for many years as they are very effective for showing the distribution of a set of numbers of links between pairs of numbers.

Slide	Title	Notes
4	Additional charts in Seaborn	You'll also meet some new visualisations as well as ways to enhance the ones you already know. In all these cases you should be thinking about what this is telling you about the context and how the layout of the diagram is conveying information.
5	Activities	• Activity 3-1: How does the weather differ at the five different UK stations? • Activity 3-2: Which features of countries are associated with longer life expectancy? • Activity 3-3 (extension): How have income and house prices changed in different areas over time? In these activities you will see one-dimensional plots for exploring a single column from a data set and two-dimensional plots like scatter diagrams that show the relationship between two numerical columns. There is also an extension activity if you are interested in exploring time series.
6	Visualisation: telling a story with data	In the activities you explored lots of visualisations for representing the patterns in the data. One of the main aims of a visualisation is to be able to communicate or tell a story about the data. For example, these box plots clearly show the difference in average temperature between the stations. Thinking about all the visualisations you met in the activities – which were the best for communicating information? This could be very subjective!
7	Association	Two variables are associated when their values are linked in some way. In the activity you saw that there is an association between GDP and life expectancy, though not a linear one. In statistics the word correlation is used to describe specifically linear association. But association can be non-linear, like this one. However, in every day English people often say 'correlation' when they mean 'association'.
8	Adding more variables to a scatter diagram	In the second activity you saw how using colour or shape to represent a third feature on a scatter diagram can reveal important details that could otherwise be missed.
9	Visualisation	For examples of well-designed visualisations, try: • Information is beautiful: *https://informationisbeautiful.net/* • Gapminder: *https://www.gapminder.org/*

Activity notes

Seaborn is a very powerful library of data visualisation techniques that can be imported into Python. In the activities in this lesson students should focus on how the visualisations are helping them observe patterns in the data. More detail on Seaborn is given at *https://seaborn.pydata.org/*

- **Activity 3-1: How does the weather differ at the five different UK stations?**

  The focus for this activity is for students to create one dimensional plots for numerical features in the data set and to group these using different categorical features. The version of a histogram used by Seaborn is to have equal interval widths and frequency on the vertical axis: this is different to what students will have met but is suitable to observe the distribution.

- **Activity 3-2: Which features of countries are associated with longer life expectancy?**

  The focus for this activity is for students to create two dimensional plots for pairs of numerical features. These can then be further developed by adding a third feature to the plot, either as the colour or size of the data points.

- **Activity 3-3 (extension): How have income and house prices changed in different areas over time?**

  Seaborn can create time series using a similar technique to how it creates scatter plots. The expectation is that the user has defined the horizontal axis as a feature that is appropriate to use.

An example of completed notebooks, including suggested answers to the checkpoints, is given at: *https://mei.org.uk/introduction-to-data-science/further-support/*

Lesson 4: The data science cycle

Lesson objectives

- To understand how working with data is an iterative process that can be viewed as a cycle.
- To give students an opportunity to work on an investigation using the data science cycle.

General guidance for the lesson

In practice the process of working with data doesn't usually follow a linear process of steps that are completed once. There is an iterative aspect to working with data: often when working on one step of the process you will want to go back to a previous step. For example, once you have started your data analysis you might realise that you need to do some more pre-processing of the data. The *data science cycle* is a way to represent this non-linear and iterative approach to working with data.

In this lesson students have an opportunity to work on an investigative problem. This allows them to gain the experience of the *data science cycle* and, for A Level Mathematics students, to become more familiar with the Large Data Set for their specification. Students of Core Maths can choose whichever of the four activities appeals to them most.

Presentation notes

Slide	Title	Notes
1	Title	
2	The data science cycle	In this lesson you will meet what is known as "the data science cycle". This is a way of describing the steps involve in working on a problem: • Start by defining what your problem is. • Then get the data you need. • Once you have the data, you need to explore it. This is sometimes referred to as being a *data detective*: you will explore the data to see what patterns there are by trying lots of charts and statistics. • When you have found some patterns, you can complete your more formal analysis of the data to confirm these. • You will then be able to communicate the results to your audience and this will often involve using visualisations. The cycle shows that this is a non-linear, iterative process: you might go round the cycle a few times and you might want to jump back a step or two. For example, when you start exploring the data you might realise you need more data.

3	Activities	In the activity for this lesson you should choose one of the following as the starting point for an investigation. You will use the data science cycle as a checklist for how to progress with the investigation. • Activity 4a is based on the Cars data from the AQA large data set • Activity 4b is based on the Weather data from the Edexcel large data set • Activity 4c is based on the Countries data from one of the MEI large data sets • Activity 4d is based on the Travel by local authority data from the OCR large data set If you're studying A Level Maths you should just do the activity relevant to your A Level; if you're studying Core Maths you can choose whichever of these interests you the most.
4	The data science cycle	In the activity you worked through the data science cycle in context. It is worth reflecting now on the iterative nature of the cycle: as you worked through it did you: • Go round the cycle more than once? • Jump back a stage in the cycle? You might have noticed that when you started communicating your results, your findings suggested other ways you could explore further.
5	PPDAC	An alternative that is often used to describe the data science cycle is the shorthand: PPDAC. This stands for Problem, Plan, Data, Analysis, Conclusion. This is useful as it emphasises that: • **Problem:** You will often start with a "real world" problem described in words. • **Plan:** You will then need to plan how you are going to get data to help you solve this problem. • **Data:** Once you have the data it is likely that it will need some pre-processing. • **Analysis:** The analysis is where you calculate the statistics – this is where school statistics often focusses, but it is only one aspect of solving problems with data. • **Conclusion:** Your conclusion then needs to be communicated effectively – good domain knowledge and skill with visualisation are useful here. You can think about how this process differs from the data science cycle given in the earlier slide. How would your work on the investigation have been different if you'd used this cycle?

6	CRISP-DM	An alternative description of the stages in data science is describe by the shorthand CRISP-DM or the Cross Industry Standard Process for Data Mining. This is used by many organisations and businesses for describing how they work with data. It has more of a focus on how the data can be used by businesses to make improvements. CRISP DM has the following steps: • Start by understanding the business needs. • Once you understand the business needs you should try to understand the data you have available to you. Sometimes you will be able to choose what data is collected but at other times the data will be given to you. • The data will need pre-processing. As discussed in lesson 2 this will often be a time-consuming stage in the process. • CRISP-DM emphasises the importance of building models with the data. For example, you might wish to build a *model* that predicts which customers will return to a site based on their previous behaviour. • The evaluation is an attempt to measure how well your predictions form your models match the real world. • You can then deploy the models, if they appear to be useful, or go back to the start of the process if they are not. If you want to know more about how data science is applied in practice, please see the further reading in this section.

Activity notes

This activity is an opportunity for students to practise the techniques they have learned on an investigation. A Level Maths students can investigate the activity for their specification's large data set; Core Maths students can choose whichever of these interests them the most. There is the potential for any students, who are interested, to use this as the start of a longer project exploring one of the A Level large data sets, for example, as a summer project.

Lesson 5: Introduction to machine learning

Lesson objectives

- To understand that machine learning is the process by which computers find models from data using iteration.
- To be able to build and measure a linear model in Python.

General guidance for the lesson

The term *machine learning* is used to describe a family of algorithms that build models from data. The machine *learns* from the data by using an iterative, 'trial and improvement' style method. For example, in finding a linear model of the form $y = mx + c$, a computer can try lots of different values for m and c and continually tweak them until it has the best line for the data points.

A standard technique used when building and measuring the effectiveness of a model is to perform a *training/testing split* on the data. Typically, this will be in a ratio of 80:20 or 75:25. The training data is used to build the model. This is the data given to the algorithm to find the best possible values of the parameters, e.g. m and c for a linear model. The testing data is then used to measure the model. This is effectively the same as measuring how the model would perform on unseen data.

Presentation notes

Slide	Title	Notes
1	Title	
2	What is machine learning?	So far you've been looking for patterns directly using statistics and visualisation. But for large data sets and complex patterns this can be too time-consuming or too difficult for humans. Machine learning is the process of using a computer algorithm to find patterns in the data. Once you've found a pattern, it can help you make predictions or decisions based on data: e.g. websites such as Amazon develop recommendation algorithms to try and predict what products you might be interested in. Another example is medical science which uses health data to track the spread of a disease or assess whether drugs or other treatments are effective.
3	Predicting	The word prediction here is being used in a very precise sense: this isn't fortune telling. Data science won't tell you who will win the next world cup or the exact price of a share on the stock market in a month's time. In general, predictions aren't necessarily about the future at all: e.g., one common predictive model is spam detection for email. Your email provider will try to predict whether an incoming email is spam or not based on other data, such as the sender, subject and content of the email. It's important to realise that the predictions are expected values, and there will be some level of variation built in. You might have seen this in graphs like this one for climate change that give a sense of the range of possible values predicted by the model. Similarly, if Amazon predict the expected amount you will spend on their site next year this is unlikely to be perfect but will give an average spend for customers with a similar profile to you.

Slide	Title	Notes
4	A simple model: line of best fit	One model you should be familiar with is a simple linear model that predicts an expected value based on a straight line drawn on a scatter diagram. In GCSE Mathematics you will have seen this as a line of best fit. In this example a linear model has been created that will predict weight for a given height. The spread of the data shows that this won't be a perfect prediction but will give an expected weight for a given height value: for example, according to this model, the expected weight of people who are 170cm tall is 80kg.
5	Activity	Activity 5-1: Can engine size and mass be used to predict a car's CO2 emissions? In the activity for this lesson, you will look at creating a model to predict the CO2 emissions of cars based on their engine size and mass. You'll try fitting a simple linear model in one variable and then extend this to building a model based on two input variables.
6	More complex models	In this activity you used linear functions to model the relationship between mass, engine size and emissions. In practice there's a whole range of mathematical functions that might fit data better, depending on the context. These could involve products or powers for example, or more complicated functions that are harder to write down. Don't worry, the methods you've learnt in this lesson can easily be extended to non-linear models too. However, in practice linear models are still one of the most widely used types of model. This is because they're relatively easy to understand and interpret.
7	Machine learning: an iterative process	Machine learning is an iterative process based on making small changes to a model and then seeing if the fit is better or not. By repeating these iterations thousands or millions of times the machine learns the best model. It's a lot like how you might learn a sports skill such as a tennis shot. You would practise it thousands of times and each time you would be making small adjustments. If a small change results in a better shot you are more likely to repeat that in future.
8	Other types of machine learning	The activity in this lesson focussed on building a model to predict an average numerical value. There are other kinds of machine learning that try to build different types of models. For example, identifying whether an email is spam requires a model that outputs true or false; this is an example of a binary classifier. You'll build binary classification models in lesson 6. Models for more complex tasks like facial recognition require more complex machine learning techniques. If you are interested in machine learning and the associated area of artificial intelligence there are some links in the further reading section for this lesson.

Activity notes

In this activity students use a machine learning algorithm to build a linear model to predict the CO_2 emissions of cars. Students should be familiar with the idea of a line of best fit for a scatter graph and they might also have found regression lines for data before. The data science approach covered here differs in the following ways:

- The focus is on creating *predictive models* as opposed to *explanatory models.* For example, you are trying to create a model that would predict the amount of CO_2 emitted. This is different to finding out if there is a link between mass and CO_2.
- A training-testing split is performed on the data. In this example 75% of the data is used for 'training' the model, i.e., finding the best m and c for the data, and 25% is reserved for 'testing' the model. This allows you to build a variety of different models and then make relative judgements about which is best. The testing data is used so that the models can be compared using 'unseen' data.
- The model is built using a machine learning algorithm. The line of code: `linear_model = LinearRegression().fit(input_train, target_train)` instructs the machine to use an iterative process to find the best m and c for the training data. This is similar to a trial and improvement method; however, the exact details of the algorithm are beyond the scope of this course.
- Models can be evaluated using the measure R^2. This is a value on a scale of 0-1, where a higher value implies a better fit to the data. For a linear model this is the square of the product moment correlation coefficient, r. However, R^2 is often preferred in data science contexts as it represents the proportion of the variation in the output feature that can be attributed to variation in the input features.
- Models with multiple inputs can also be generated. These are harder to visualise graphically but can be understood algebraically, e.g. $y = m_1x_1 + m_2x_2 + c$.

Lesson 6: Machine learning, AI and Bias

Lesson objectives

- To give students an opportunity to build a binary classification model in Python.
- To understand how most recent applications of artificial intelligence are based on models created using machine learning.
- To develop an awareness of how models built using machine learning can be biased.

General guidance for the lesson

In lesson 5 students met one type of machine learning algorithm: linear regression. In this lesson they will get an opportunity to build another type of model: binary classification. This is a model that predicts whether something will fall into one of two categories using the values of other variables. These categories could be *True* and *False*, *A* and *B*, or some other pair of values.

Once students have built two different types of models this can be used to introduce the ideas of *Artificial Intelligence*. One definition of Artificial Intelligence is that it is any automated decision-making process. For example, deciding which products to show a customer on a shopping website or deciding whether to tag an email as 'spam' are both cases where the decision will be based on a predictive model derived from data.

Students should also consider the ways in which models can be biased. In particular, if the model is based on data that does not represent the population the model will be applied to then it is likely to underperform. This can have serious consequences as shown in the news stories in one of the slides.

The data for this lesson is not one of the A Level large data sets.

Presentation notes

Slide	Title	Notes
1	Title	
2	What is a model?	A model is a simplification of the real world. For example, you can think of the mean as being a model for the data that says all the values are the same. This model is not an accurate representation of all the values in the data set but it could help you make decisions. A simple model in a shop could be that customers spend £20 (on average). This could be useful to predict how much extra money you would take if you had 10 extra customers. Models can help you make decisions. In the shop example it could be that you are deciding whether to place advertising in a local magazine and the decision could be made by comparing the cost of the advert to the amount of extra income you expect to receive. In this course you meet examples of models built from data using machine learning algorithms. You've met a line of best fit, which has two numbers, or *parameters*. Increasingly complex models will have more parameters, for example, ChatGPT is a model for generating text that has over 1 trillion parameters.

Slide	Title	Notes
3	Binary classification	Last lesson, you built models that aimed to predict a number – regression models. Another common task for machine learning is to predict a category. This is called classification. The simplest version of classification is binary classification, where the target is a yes/no answer. For example, spam detection algorithms try to predict whether an email is spam or not. This is based on the results of a machine learning algorithm that has been built using lots of email messages.
4	Activity	Activity 6-1: Can a penguin's features be used to predict its species? The data for this activity comes from a study of three species of penguins: Adelie, Gentoo and Chinstrap. You'll build a binary classification model that uses body measurements to try and predict whether a penguin is a Gentoo penguin.
5	What about Artificial Intelligence?	Artificial Intelligence (AI) describes systems that are designed to make decisions automatically. The success of recent AI systems is due to improvements in machine learning technology. A good example of this is language translation, such as French to English. Early models that tried to translate language using rules were not very successful. Recent models built by spotting patterns in massive quantities of texts in both languages are much more effective. Developing your skills in data analysis will give you a solid foundation for understanding machine learning and this is the key to most modern applications of AI.
6	Building AI models from data	As models become more complex it is more difficult to devise a rule for finding the parameters. In the case of a mean it is straightforward: find the total of values and divide by how many there are. For models with large numbers of parameters the only viable approach is to learn these iteratively from data. Some models can sit at the overlap between the two processes: lines of best fit and spam detection can either be created using rules or learned iteratively from the data. More complicated models will be learned from the data. This covers the examples here: recommender systems (like on Spotify or Netflix), Deep Blue (the first chess program to defeat a grandmaster) and ChatGPT.
7	Bias, Ethics and Artificial Intelligence	Your classification model for penguins was very successful at distinguishing Gentoo penguins from Adelie or Chinstrap penguins. But there are a total of five different species of penguin across the Antarctic. If you tried to use your model somewhere where there were also Emperor penguins, for example, it might not perform as well, as the data that the model was trained on would not be representative of the population you are attempting to use it on. This is an example of *bias*. This might not be a disaster in this context, but sometimes bias can have serious consequences as shown in the example news stories.

Slide	Title	Notes
8	Varied uses of data science	Throughout the course you have seen the varied uses of data science in the *Meet a data scientist* videos. Data science impacts on many areas of life, study and work. If you are interested in data science you can probably find an example of it in a context that is of interest to you.
9	Your next steps?	There are a variety of routes into data science: • You could study or work in an area that interests you and then develop your data skills in that area. It is common for people to undertake postgraduate courses in data science after studying or working in other areas. • You could study for a degree in data science or a related area such as mathematics, statistics or computing. • You could do a degree apprenticeship in data analytics. There are interesting sites in the further reading … … and you can always have fun practising on the huge amounts of free data that is available on sites like Kaggle.

Activity notes

In this activity students build a binary classification model for penguins. The model attempts to predict whether a penguin is a Gentoo penguin or not based on various numerical features.

The structure of the process is similar to the one used for building a linear regression model in lesson 5:

- Perform a training-testing split.
- Define the input and output features.
- Use a machine learning algorithm to find the best values of the parameters for the training data. In this case it is the value of the input variable on which to partition the data.
- Measure the model on the testing data using a suitable metric. In this case students use the *accuracy*, i.e., the proportion of penguins that the model has correctly identified in the testing set.