DEPARTMENT OF BIOMEDICAL ENGINEERING

24AI671-FUNDAMENTALS OF DATA SCIENCE

UNIT V- DATA SCIENCE FOR REAL-WORLD ENGINEERING SOLUTIONS

1. PREDICTIVE ANALYTICS

Predictive analytics is the use of **past (historical) data, statistical methods and machine learning algorithms** to forecast what is likely to happen in the future. It does not tell us what will happen with certainty; it gives the **probability or best estimate** of a future outcome. It answers the question **"What will happen?"**

In engineering, huge amounts of data are produced every second by sensors, machines, control systems and software. Predictive analytics learns the patterns hidden in this data and uses them to predict events such as machine breakdown, electricity demand, product defects or traffic jams, so that action can be taken **before** the problem occurs.

Four Stages of Analytics

Predictive analytics is the third stage in the analytics journey. Each stage adds more value but needs more data and skill.

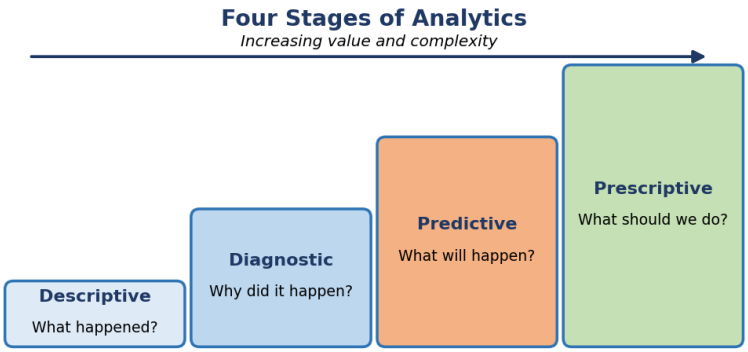

Fig 1.1 Four stages of analytics

- **Descriptive analytics** – summarises past data using reports and charts. Example: last month's production report.
- **Diagnostic analytics** – finds the reason for an event. Example: why a machine stopped.
- **Predictive analytics** – forecasts future events. Example: the machine will stop within 10 days.
- **Prescriptive analytics** – suggests the best action. Example: repair the machine on day 7 to reduce cost.

Steps in Predictive Analytics

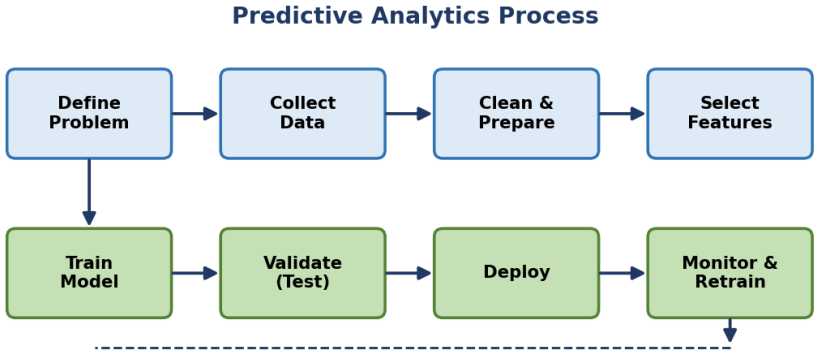

Fig 1.2 Predictive analytics process

- **Define the problem** – clearly decide what has to be predicted and how success will be measured. For example, "predict whether a motor will fail in the next 7 days".

- **Collect data** – gather relevant data from sensors, databases, log files, SCADA systems and IoT devices. More good-quality data normally gives a better model.
- **Clean and prepare data** – real data has missing values, errors and outliers. These are corrected or removed, and values are scaled or encoded so that the algorithm can use them.
- **Select features** – choose the input variables that really influence the output (for example temperature, vibration, current). Useless features are dropped to make the model simpler.
- **Train the model** – the data is split into training and testing sets; the algorithm learns the relationship between inputs and output from the training set.
- **Validate the model** – the model is tested on data it has not seen before. Techniques such as cross-validation make sure it works well in general and does not just memorise the training data (overfitting).
- **Deploy** – the final model is connected to an application, dashboard or control system so that it gives predictions in real time.
- **Monitor and retrain** – with time, the real data changes and accuracy falls (model drift). The model is checked regularly and retrained with new data.

Techniques Used

The choice of technique depends on what is to be predicted. If the output is a number, regression is used; if the output is a category, classification is used.

Technique	Algorithms	Used For
Regression	Linear, Multiple, Polynomial regression	Predicting numeric values such as load, cost, temperature

Technique	Algorithms	Used For
Classification	Logistic regression, Decision tree, Random forest, SVM, Naive Bayes, k-NN	Predicting categories such as fail / not fail, defective / good
Time series	ARIMA, Exponential smoothing, LSTM	Forecasting values that change over time
Neural networks	ANN, CNN, RNN, Deep learning	Complex, non-linear data such as images, speech and signals

Linear Regression

Linear regression is the simplest predictive model. It fits a straight line through the data so that the difference between actual and predicted values is as small as possible. With one input it is called simple linear regression; with many inputs it is multiple regression.

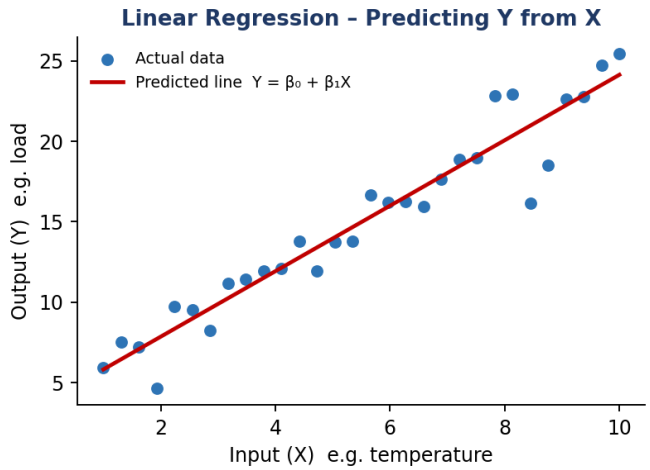

Fig 1.3 Linear regression ($Y = \beta_0 + \beta_1 X + \varepsilon$)

Here **Y** is the value to be predicted, **X** is the input, β_0 is the intercept (value of Y when X = 0), β_1 is the slope (change in Y for a one-unit change in X) and ε is the error term.

Decision Tree and Random Forest

A **decision tree** splits the data using a series of yes/no questions (for example, "Is temperature > 80 °C?") until it reaches a decision. A **random forest** builds many decision trees on random parts of the data and combines their votes, which gives higher accuracy and reduces overfitting.

Measuring Model Performance

- **For regression:** MAE (mean absolute error), MSE (mean squared error), RMSE (root mean squared error) and R^2 (how well the model explains the data). Lower error and R^2 close to 1 are better.
- **For classification:** Accuracy (overall correct predictions), Precision (how many predicted positives are really positive), Recall (how many real positives were found), F1-score (balance of precision and recall) and ROC-AUC.
- **Confusion matrix:** a table of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) from which the above measures are calculated.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \qquad \text{RMSE} = \sqrt{[\ (1/n)\ \Sigma\ (y_i - \hat{y}_i)^2\ ]}$$

Engineering Applications

- **Predictive maintenance** – sensor data of motors, turbines and bearings is used to estimate the remaining useful life, so repairs are planned before breakdown.
- **Energy sector** – forecasting electricity demand and solar or wind power output helps the grid balance supply and demand.
- **Civil engineering** – predicting crack growth, settlement and strength of structures from monitoring data.
- **Manufacturing** – predicting defects and quality problems, which reduces waste and rework.

- **Transport** – predicting traffic congestion, travel time and vehicle fuel consumption.
- **Healthcare** – predicting disease risk and patient condition from medical records.

Advantages and Limitations

Advantages	Limitations
Reduces downtime and maintenance cost	Needs a large amount of good-quality historical data
Detects risks and faults early	Model accuracy falls over time (model drift)
Improves planning and use of resources	Overfitting; complex models are hard to explain
Supports data-driven decisions	Gives probabilities, not certainty

OBSERVE OPTIMIZE OUTSPREAD