Goal:

* Identifying the fraudulent loan applications using AI and Non-AI techniques
* Predicting the credit worthiness of a non-fraudulent applicant for the loan

Data Gathering: Assume that we have these kind of data

* **KYC (Know Your Customer) Data:**
  + **Personal Identity**:
    - PAN Card
    - Aadhaar Card
    - Passport
    - Voter ID
    - Driver’s License
* **Address Proof**:
  + Utility bills (electricity, water, gas)
  + Rent agreement
  + Bank statement with address
  + Aadhaar, Passport, or Voter ID
* **Personal Details:**
  + Full Name
  + Date of Birth & Age
  + Gender
  + Marital Status
  + Number of Dependents
  + Education
  + Contact Details (Mobile, Email)
* **Employment / Occupational Details:**
  + Salaried Individuals
    - Company name & address
    - Job designation
    - Employee ID
    - Employment duration
    - Latest salary slips (3–6 months)
    - Offer letter / Appointment letter
    - Form 16 / ITRs
  + Self-employed Professionals/Business Owners:
    - Business name and nature
    - Registration certificates (GST, MSME, etc.)
    - Business vintage (how long it's been running)
    - Audited financials (P&L, Balance Sheet)
    - Bank statements (usually 6–12 months)
* **Income and Financial Details:** 
  + Monthly and annual income
  + Other sources of income (rental, investment, etc.)
  + Existing EMIs or loan obligations
  + Credit card usage and dues
  + Bank account statements (typically 6 months)
  + ITR (Income Tax Returns) for 2–3 years
  + Net worth (in some cases)
* **Property or Collateral Details:**
  + Property papers
  + Property tax receipts
* **Credit Information:**
  + Credit report from agencies like CIBIL
  + Credit score
  + Past default or delay records
  + Loan utilization patterns
* **Loan Details (Application-specific)**
  + Loan amount requested
  + Purpose of loan (home, car, education, business, etc.)
  + Preferred tenure
  + Co-applicant or guarantor details
* **Other Discretionary Data (Extra)**
  + Insurance history (if loan includes credit life insurance)
  + Investments (FDs, shares, mutual funds, etc.)
  + Asset declarations (movable and immovable)
  + References (personal or professional)

**Fraudulent application detection**

**Rule-Based Systems/ Non-AI**

1. Predefined rules or logic to flag suspicious behavior:

* PAN mismatch with name/date of birth
* Multiple applications from same IP/device
* Income-to-loan ratio above acceptable limits
* Salary account and credited employer don’t match
* Residence or employment proof issued recently or forged
* Large discrepancy in bank balance and declared income

1. Document Verification:

* **ID & Address Proof Validation**: Cross-checking with government databases (e.g., Aadhaar, PAN, Voter ID)
* **Employment Verification**: Calling employer HR, email verification, or sending physical letters
* **Bank Statement Scrutiny**:
  + Fake transaction patterns (rounded numbers, backdated entries)
  + Sudden inflow of funds before application
* **ITR/Form-16 Verification** with Income Tax department

1. Reputation:

* CIBIL or other credit bureau fraud tags
* Internal fraud database
* Employment or IP origin from known high-risk regions

1. **Field Verification**

* Address verification (ownership, occupancy)
* Employer location & employee confirmation
* Neighbourhood feedback on applicant reputation
* Business existence check

1. **Statistical & Ratio-Based Red Flags:**

* Debt-to-income ratio too high
* Net worth declared but not reflected in bank statements
* Very low or no expenses declared

1. **Pattern Matching:**

* Multiple applications with minor variations in name/address
* Fake employer names used repeatedly
* Applications made just before repayment of an existing loan (possible debt hiding)
* Unusual timing: applying late at night or on holidays

**AI Technique:**

**Supervised Learning (Classification):**

* Assume we have a historic dataset of individual loan applications, capturing demographic information, financial standing, loan specifics, and the crucial fraud\_flag indicating fraudulent applications.
* We also have a transactional data of customer transaction history, including transaction types, amounts, merchant details, and location.
* As we do have a labelled dataset of fraudulent applications we can build a binary classification model by combining loan application characteristics with real-time transactional behaviour.

**Unsupervised Learning** (Anomaly Detection):

* Assume we have limited or no labeled fraud data or fraud patterns are changing rapidly.
* We can Identify data points that **differ significantly** from the majority (normal behavior)
* Most applications will cluster tightly (normal behavior). Some points far away from this cluster could be **anomalies** — potential fraud.

**Creditworthiness Prediction (Credit Scoring)**

This is essentially a **regression or classification problem**, depending on whether you:

* Predict a credit score (regression), or
* Categorize as **Low / Medium / High risk** (classification)

Non AI technique:

* Calculate the **numeric credit score** based on an applicant’s financial and personal information.
* Flag the credit score as Excellent, Good, Average or Poor accordingly

**Demo Application:**

**Fraudulent application prediction:**

* **Build a binary classification model to predict the application is fraud or not.**

Creditworthiness prediction:

* Calculated the **numeric credit score** based on an applicant’s financial and personal information.
* Flagged the credit score as Excellent, Good, Average or Poor accordingly
* CIBIL Score (up to 40 points)
* Monthly Income (up to 20 points)
* Debt-to-Income Ratio (DTI) (up to 15 points)
* EMI to Income Ratio (up to 10 points)
* Loan to Income Ratio (up to 10 points)
* Age (up to 5 points)

| **Credit Score** | **Label** |
| --- | --- |
| 80–100 | Excellent |
| 60–79 | Good |
| 40–59 | Average |
| Below 40 | Poor |