AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

*CSCE 581: Introduction to Trusted AI*

# Lectures 5 and 6: AI, ML and Supervised ML

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

27TH AND 29TH JAN 2026

**Carolinian Creed: "I will practice personal and academic integrity."**

**Credits**: **Copyrights of all material reused acknowledged**

# Organization of Lectures 5, 6

- Introduction Section
  - Recap from Week 2 (Lectures 3 and 4)
- Main Section
  - L5: Common ML methods
  - L5: Case Study on Data, ML and Fairness – Face Recognition
  - L6: Supervised ML
  - L6: Measuring Goodness of Supervised ML
  - Project Discussion
- Concluding Section
  - About next week – Lectures 7, 8
  - Ask me anything

# Introduction Section

# Recap from Week 2 (Lectures 3, 4)

Week 1: Introduction to Trusted AI

- Week 2
  - Trusted decisions
  - Data and characteristics
  - Common ways to prepare data
  - How to organize content for inferencing / reasoning

# AI News Anniv. (2025) - One Year Later

- DeepSeek R1
- Technical interest, business interest

# #1 NEWS - Genetic Data From Over 20,000 U.S. Children Misused for 'Race Science'

Link - https://www.nytimes.com/2026/01/24/us/children-genetics-race-science.html

- **Context**
  - Adolescent Brain Cognitive Development Study - In the nationwide ABCD Study, more than 11,000 children underwent regular M.R.I.s and clinical tests and had DNA samples taken from their blood or saliva. In a separate study, known as the Philadelphia Neurodevelopmental Cohort, researchers at the University of Pennsylvania collected similar information from about 9,500 children.
- **Findings**
  - At least 63 times since 2007, data from some of the 28 human genomic repositories that the N.I.H. controls was improperly released to researchers, used for unapproved purposes or made vulnerable to theft, according to government records reviewed by The New York Times
- **Insight**
  - Good faith declarations are meaningless
  - Poor research, poorer researcher-public collaboration

# #2 NEWS - 100+ citations in 50+ papers at Neurips with Hallucinations

Link - https://fortune.com/2026/01/21/neurips-ai-conferences-research-papers-hallucinations/ ,
Details - https://gptzero.me/news/neurips/

- **Context**
  - Neurips is a top AI conference
  - 20k+ papers submitted, 4k+ papers accept
- **Findings**
  - Paper has citations will hallucinations
  - These are grounds for papers rejections as per Neurips policy
- **Insight**
  - LLM used to detect hallucination, which can have errors; Manually verified
  - Conferences (organizers, reviewers) and authors need to be more vigilant

| Real Citation | Flawed Citation | Hallucinated Citation |
|---|---|---|
| Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. nature, 521:436-444, 2015. | Y. LeCun, Y. Bengio, and Geoff Hinton. Deep leaning. nature, 521(7553):436-444, 2015. | Samuel LeCun Jackson. Deep learning. *Science & Nature*: 23-45, 2021. |
| A. Yang, B. Zhang, B. Hui, B. Gao, B. Yu, C. Li, D. Liu, J. Tu, J. Zhou, J. Lin, et al. Qwen2.5–math technical report: Toward mathematical expert model via self-improvement. arXiv:2409.12122, 2024. | A. Yang, (missing author), B. Hui, B. Gao, B. Yu, C. Li/, D. Liu, J. Tu, J. Zhou, J. Lin, et al. Qwen 2. 5–math technical report: Toward mathematical expert model via self-improvement. arXiv preprint arXiv:2409.12122, 2024. | A. Yang, B. Yang, C. Yang, et al. Qwen3.5–mathematical report for iterative model self-improvement. arXiv:2909.12233, 2024. |

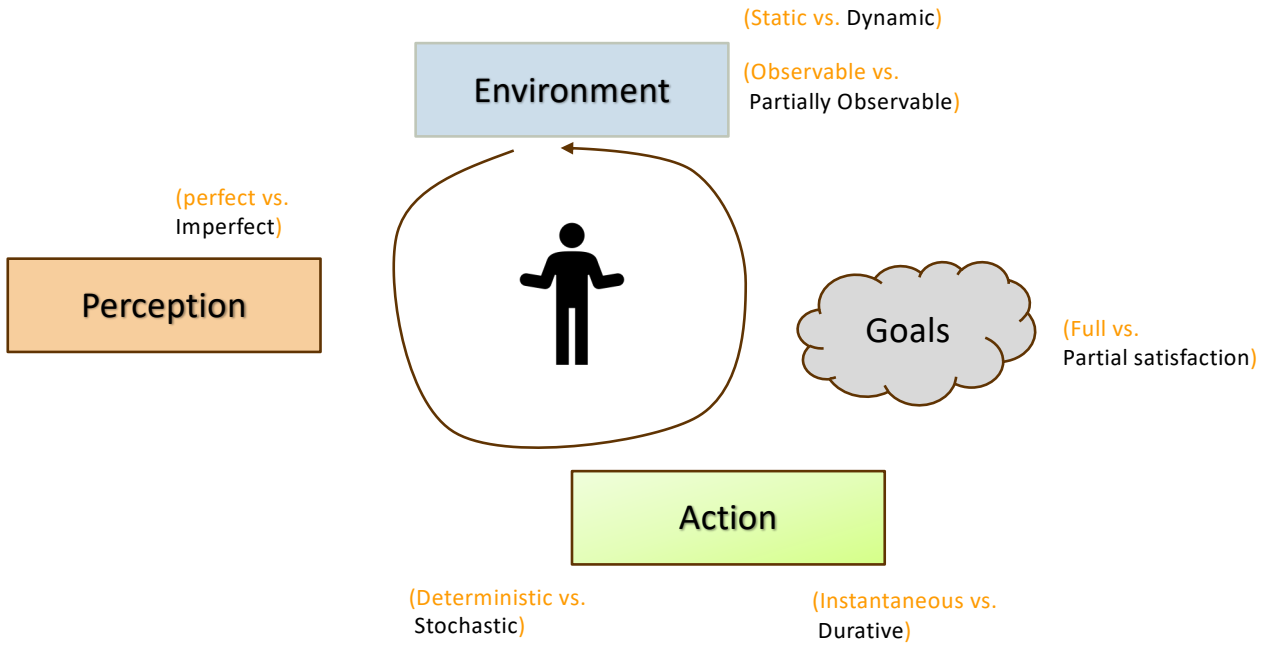
# #3 NEWS - Misleading text in the physical world can hijack AI-enabled robots

Link - https://news.ucsc.edu/2026/01/misleading-text-can-hijack-ai-enabled-robots/
Paper with details - https://arxiv.org/pdf/2510.00181

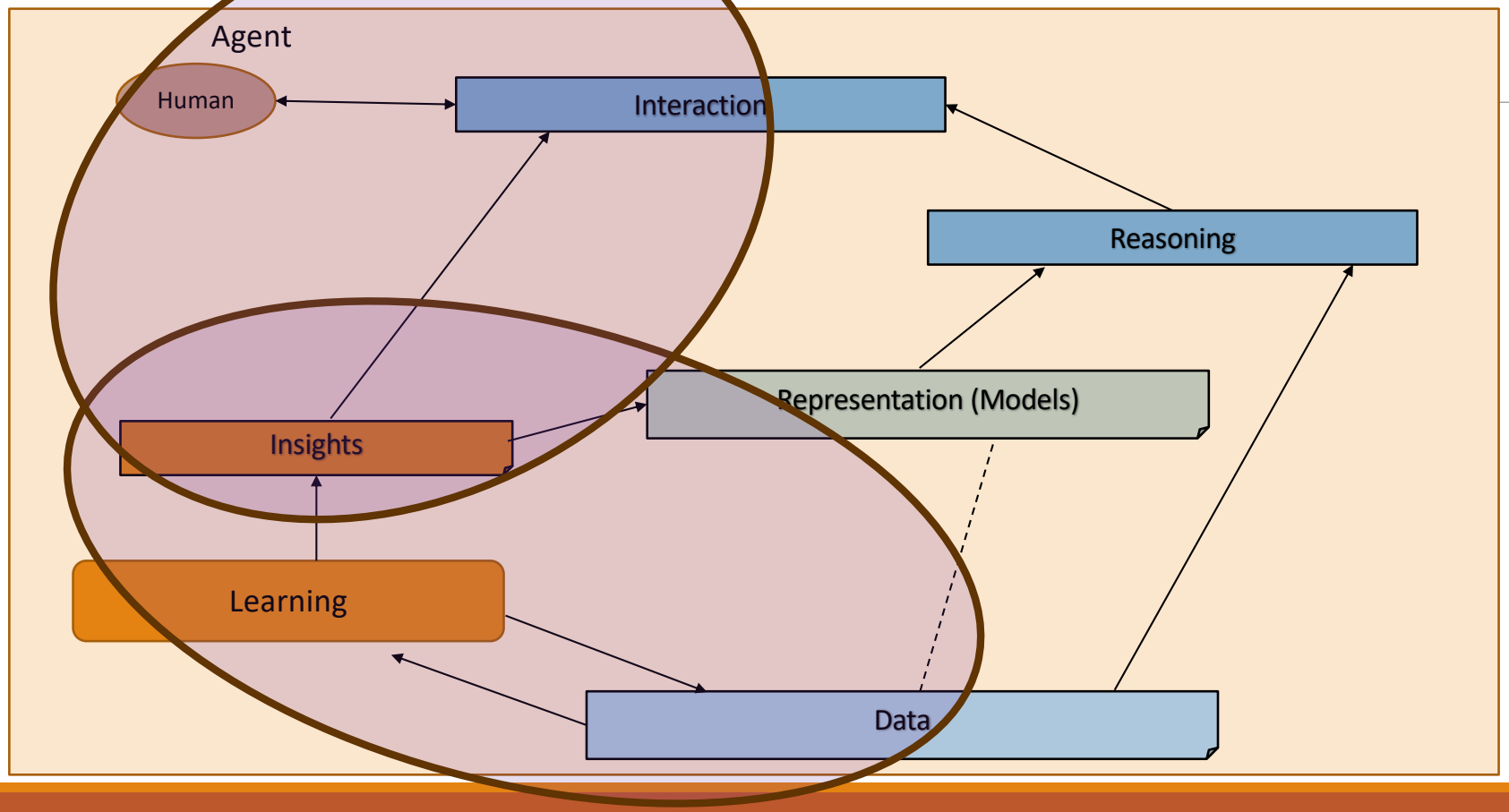
- **Context**
  - Embodied AI systems are around
- **Insight**
  - Self-driving cars, autonomous robots and drones, and other AI systems that use cameras may be vulnerable to these attacks
- **Findings**
  - CHAI - Command Hijacking against embodied AI - a new class of prompt-based attacks that exploit the multimodal language interpretation abilities of Large Visual-Language Models (LVLMs)
  - Prompt-injection attacks happen when code (instructions) and data are mixed.
  - CHAI embeds deceptive natural language instructions, such as misleading signs, in visual input, systematically searches the token space, builds a dictionary of prompts, and guides an attacker model to generate Visual Attack Prompts.



# Intelligent Agent Model



# Relationship Between Main AI Topics (Covered in Course)



# High Level Semester Plan (Adapted, Approximate)

**CSCE 581 –**

- Week 1: Introduction
- Week 2: Background: AI - Common Methods
- Week 3: The Trust Problem
- Week 4: Machine Learning (Structured data) - Classification
- Week 5: Machine Learning (Structured data) - Classification – Trust Issues
- Week 6: Machine Learning (Structured data) – Classification – Mitigation Methods
- Week 7: Machine Learning (Structured data) – Classification – Explanation Methods
- Week 8: Machine Learning (Text data, **vision**) – Classification, **Large Language Models**
- Week 9: Machine Learning (Text data) - Classification – Trust Issues, LLMs
- Week 10: Machine Learning (Text data) – Classification – Mitigation Methods
- Week 11: Machine Learning (Text data) – Classification – Explanation Methods
- Week 12: Emerging Standards and Laws, **Real world applications**
- Week 13: Project presentations
- Week 14: Project presentations, Conclusion

AI/ ML topics and with a focus on fairness, explanation, Data privacy, reliability

# Main Section

# Machine Learning


**Credit**: Retrieved from internet

# Machine Learning – Insights from Data

- Descriptive analysis
  - Describe a past phenomenon
  - **Methods**: classification (feedback from label), clustering, dimensionality reduction, anomaly detection, neural methods, reinforcement learning (feedback from hint/ reward)
- Predictive analysis
  - Predict about a new situation
  - **Methods**: time-series, neural networks
- Prescriptive analysis
  - What an agent should do
  - **Methods**: simulation, reinforcement learning, reasoning

- New areas
  - Counterfactual analysis
  - Causal Inferencing
  - Scenario planning

# Nomenclature

Column, Attribute, Feature

Row, Item

| | PID | ST_NUM | ST_NAME | OWN_OCCUPIED | NUM_BEDROOMS | NUM_BATH | SQ_FT |
|---|---|---|---|---|---|---|---|
| 2 | 100001000 | 104 | PUTNAM | Y | 3 | 1 | 1000 |
| 3 | 100002000 | 197 | LEXINGTON | N | 3 | 1.5 | -- |
| 4 | 100003000 | | LEXINGTON | N | n/a | 1 | 850 |
| 5 | 100004000 | 201 | BERKELEY | 12 | 1 | NaN | 700 |
| 6 | | 203 | BERKELEY | Y | 3 | 2 | 1600 |
| 7 | 100006000 | 207 | BERKELEY | Y | NA | 1 | 800 |
| 8 | 100007000 | NA | WASHINGTON | | 2 | HURLEY | 950 |
| 9 | 100008000 | 213 | TREMONT | Y | 1 | 1 | |
| 10 | 100009000 | 215 | TREMONT | Y | na | 2 | 1800 |

# Types of Attributes/ Columns

- Numeric: has number as value in computational sense; all mathematical functions are valid.
  - Example: SQ_FT
- Categorical: has distinct values
  - Nominal: each value is incomparable with other
    - Example: OWN_OCCUPIED, ST_NAME
  - Ordinal: the values can be ordered
    - Example: ST_NUM, NUM_BEDS
- Comment:
  - Q: what type is a binary variable?
  - A: depends on the semantics – nominal (gender), ordinal (number basements).

| | PID | ST_NUM | ST_NAME | OWN_OCCUPIED | NUM_BEDROOMS | NUM_BATH | SQ_FT |
|---|---|---|---|---|---|---|---|
| 2 | 100001000 | 104 | PUTNAM | Y | 3 | 1 | 1000 |
| 3 | 100002000 | 197 | LEXINGTON | N | 3 | 1.5 | -- |
| 4 | 100003000 | | LEXINGTON | N | n/a | 1 | 850 |
| 5 | 100004000 | 201 | BERKELEY | 12 | 1 | NaN | 700 |
| 6 | | 203 | BERKELEY | Y | 3 | 2 | 1600 |
| 7 | 100006000 | 207 | BERKELEY | Y | NA | 1 | 800 |
| 8 | 100007000 | NA | WASHINGTON | | 2 | HURLEY | 950 |
| 9 | 100008000 | 213 | TREMONT | Y | 1 | 1 | |
| 10 | 100009000 | 215 | TREMONT | Y | na | 2 | 1800 |

# Why is Type of Variable Important

- Handling of missing values
- Distance between
  - Values
  - Data items
- Used for measuring accuracy, error
- Guiding the learning process
  - Selection of algorithms

# Concepts

- **Input data**: data available
  - **Training data**: used for training a learning algorithm and get a model
    - [Optional] **Validation data**: used to tune parameters
  - **Test data**: used to test a learning model

- **Classification problem**
  - Separating data into classes (also called labels, categorical types)
  - One of the attributes is the class label we are trying to learn
  - Class label is the **supervision**
- **Clustering problem**
  - We are trying to learn grouping of data
  - There is no attribute indicating membership in the groups (hence, **unsupervised**)
- **Prediction problem**
  - Learning value of a continuous variable

Reference: https://machinelearningmastery.com/difference-test-validation-datasets/
https://www2.seas.gwu.edu/~bell/csci243/lectures/classification.pdf

# Sample Learning Task

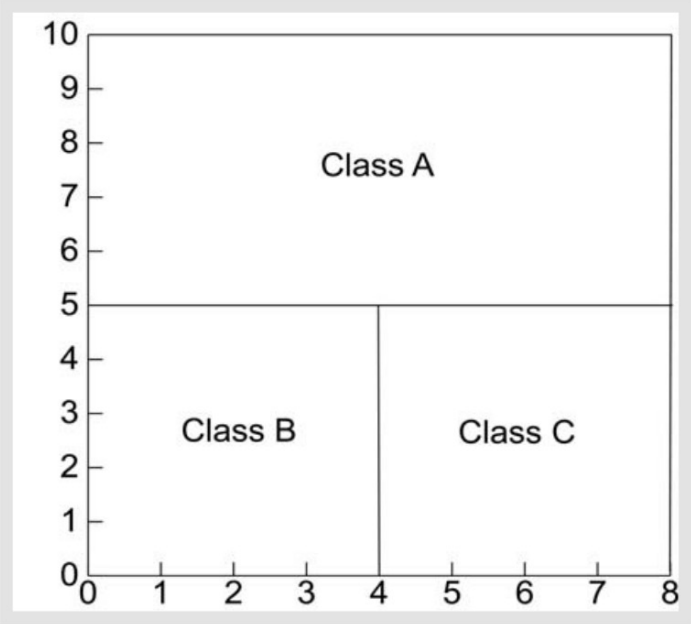
- COVID-19 data

Notebook: https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l6-l7-l8-supervised-ml/Supervised-Regression-Classification.ipynb

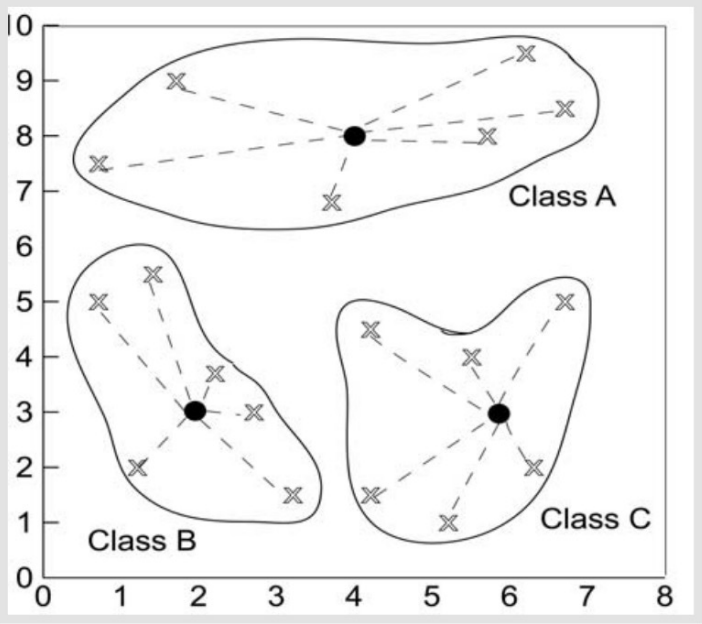
-

# Methods for Classification

**Partitioning Based**




**Distance Based**




Source: https://www2.seas.gwu.edu/~bell/csci243/lectures/classification.pdf

# Data Exercises

# Exercise 1: Weight, Height, BMI

**Objective**: understand BMI of people (humans)

- Understand data and trust issues in the exercise

- **Data**: Collect weight and height information from a population
  - Clean data, as appropriate

- **Model**: Build a model to predict BMI
  - Task 1: To predict weight, given height // prediction
  - Task 2: To classify BMI into 4 categories*, given height // classification

- Understanding:
  - Report performance metrics for the (two) models based on Task
  - Understand how trust issues arise and how they can be resolved/ mitigated

# OngoingActivity

- Do exercise 1

* : see next slide on BMI categories

**: refer to Crawl-Walk-Run approach to scope - https://www.linkedin.com/pulse/crawl-walk-run-approach-ai-based-real-world-problem-biplav-srivastava-pxsre/

# Body Mass Index (BMI)

- BMI Category, based on BMI Range ($kg/m^2$)
  - C1 - Underweight: Less than 18.5
  - C2 - Healthy Weight: 18.5 to less than 25
  - C3 - Overweight: 25 to less than 30
  - C4 - Obesity: 30 or greater
- Details: https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html

# Student Github Setup Instructions

- Github
  - Create a repo called ”CSCE581-Spring2026-<StudentName>
  - Share will “**biplav-s**” (instructor’s Github Id) and “**BharathMuppasani**” (TA’s ID)
  - Enter its name in **“Student-InfoShared ..” sheet, column E**

- All quiz, projects, exercises will go as sub-folders inside it
  - Timestamp of folder will be used to confirm but you can also state explicitly use **“Student-InfoShared ..” sheet’s corresponding** ”Completion” cell. The Github timestamp will overrule in case of mismatch.

# Exercise To Dos (Over Next 2 Weeks)

1. Make a sub-folder in your github repo called “exercise-height-weight”
   1. Create a sub-folder called “**data**” and have all data there. Two specifically are sample .csv as well as cleaned/ prepared .csv file(s)
   2. Create a sub-folder called “**code**”. All code will be below it
      1. Create a sub-folder called “**data-prep**”. Have data preparation and cleaning code there.
      2. Create a sub-folder called “**custom-classifier-model**”. Have classifier training and testing code there
      3. Create a sub-folder called “**custom-regression-model**”. Have regression training and testing code there
   3. Create a sub-folder called “**genai**”. All files related to gpt/chatgpt will be below it
      1. Create a testcase file for classification. (Copy and use the testcase template: https://github.com/biplav-s/book-trustworthy-chatbot/blob/main/ai-testcases/testcase-template.md )
      2. Put transcript/ result of your work there.

**Report results on:**

- 50 cm
- 100 cm
- 150 cm
- 200 cm
- 250 cm

# Gender Shades / Face Recog - Paper


1. Joy Buolamwini, Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA. Volume 81 of Proceedings of Machine Learning Research, pages 77-91, PMLR, 2018.

2. http://gendershades.org/overview.html

# Dataset

a. Megaface, which to date is the largest publicly available set of facial images, was composed utilizing Head Hunter to select one million images from the Yahoo Flicker 100M image dataset
b. LFW, a dataset composed of celebrity faces which has served as a gold standard benchmark for face recognition, was estimated to be 77.5% male and 83.5% White. Performance not broken down by gender or race
c. Intelligence Advanced Research Projects Activity (IARPA) released the IJB-A dataset as the most geographically diverse set of collected face.
d. As of 2017, The National Institute of Standards and Technology is starting another challenge to spur improvement in face gender classification by expanding on the 2014-15 study.

# Pilot Parliament Benchmark

- They also used IJB-A and Adience for comparison.
- One author labeled each image with one of six Fitzpatrick skin types and provided gender annotations for the IJB-A dataset.
- The Adience benchmark was already annotated for gender.
- These preliminary skin type annotations on existing datasets were used to determine if a new benchmark was needed. For PPB, 3 annotators including the authors provided gender and Fitzpatrick labels.
- A board-certified surgical dermatologist provided the definitive labels for the Fitzpatrick skin type.
- Gender labels were determined based on the name of the parliamentarian, gendered title, prefixes such as Mr or Ms, and the appearance of the photo.

# Inequality in Misclassification

a. The gender misclassification rates on the Pilot Parliaments Benchmark replicate this trend across all classifiers. The differences between female and male classification error rates range from 8. 1% to 20. 6%.
b. **Even though darker females make up 21.3% of the PPB benchmark, they constitute between 61.0% to 72.4.1% of the classification error**
c. COTS1 and COTS2 APIs solely output single labels indicating whether the face was classified as female or male. COTS3's API outputs an additional number which indicates the confidence with which the classification was made. The authors note that giving crisp class labels does not give users the ability to analyze true positive (TPR) and false positive (FPR) rates for various subgroups if different thresholds were to be chosen.
d. Errors do not seem to happen because of image quality. They consider South African photos of similar image quality as Europeans.

# Discussion: What is Right (Fair)?

- Equal errors?
- Equal accuracy?
- Errors on individual faces?
- …

# Week 3, Lecture 6


Credit: From Internet

# Lecture 6 - Outline

- News
- Discussion the Heigh—Weight programming exercise
- Metrics
- Comparing Classification Methods
- Project Discussion

# Metric Types

- **Effectiveness**: what the **user** of a system sees, primarily cares about
- **Efficiency**: what the **executor** in a system sees, primarily cares about

# Example: Predicting COVID cases

- **Effectiveness**: what the **user** of a system sees, primarily cares about
  - *How accurate (high) is the prediction?*
  - *How low is the error?*
- **Efficiency**: what the **executor** in a system sees, primarily cares about
  - *How low is the error?*
  - *How fast was prediction made?*
  - *How stable is the prediction to change in data?*

# Example: Detecting Spam in Email

- **Effectiveness**: what the **user** of a system sees, primarily cares about
  - *How many spams identified?*
  - *How many spams missed?*
- **Efficiency**: what the **executor** in a system sees, primarily cares about
  - *How fast were spams detected?*
  - *How much memory was used per million emails processed ?*

# Comparing Classification Methods

- Predictive accuracy
- Interpretability: providing insight
- Robustness: handling noisy data

- Speed
- Scalability: large volume of data

Source: Data Mining: Concepts and Techniques, by Jiawei Han and Micheline Kamber

# Metrics: Accuracy, Precision, Recall

| | Predicted class | | |
|---|---|---|---|
| Actual Class | | Class = Yes | Class = No |
| | Class = Yes | True Positive | False Negative |
| | Class = No | False Positive | True Negative |

**Accuracy** =
(TP+TN)/
(TP+FP+FN+TN)

**Precision** =
( TP)/
(TP+FP)

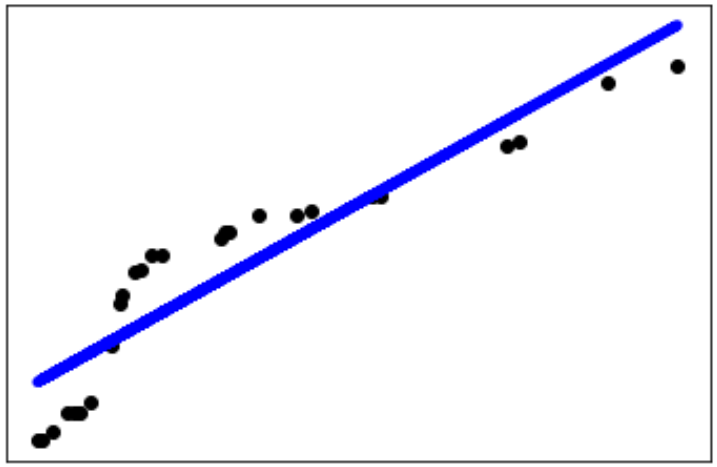
**Recall** =
(TP)/
(TP+FN)

**F1 Score**: *Harmonic Mean*

1/F1 = 1/Precision + 1/Recall

F1 = 2*(Recall * Precision) /
(Recall + Precision)

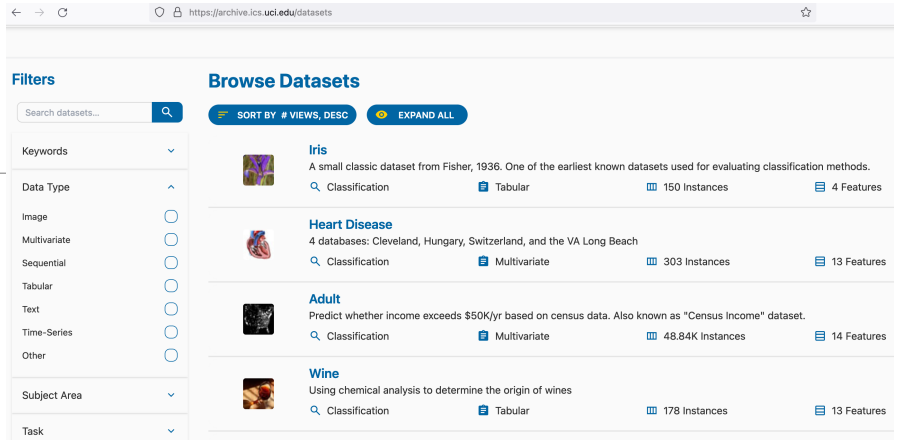
# Linear Regression

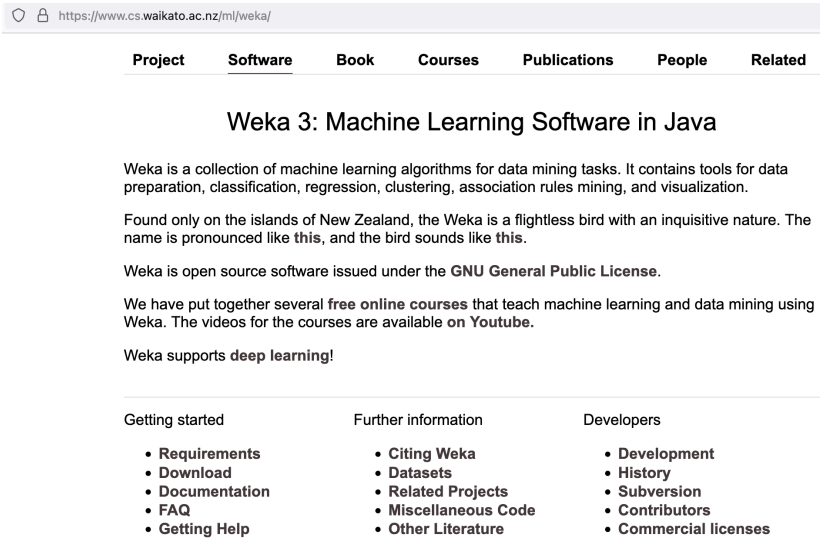


Notebook: https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l6-l7-l8-supervised-ml/Supervised-Regression.ipynb

# Reference and Demo

- Data: UCI Datasets
  - https://archive.ics.uci.edu/datasets
  - Browse or search

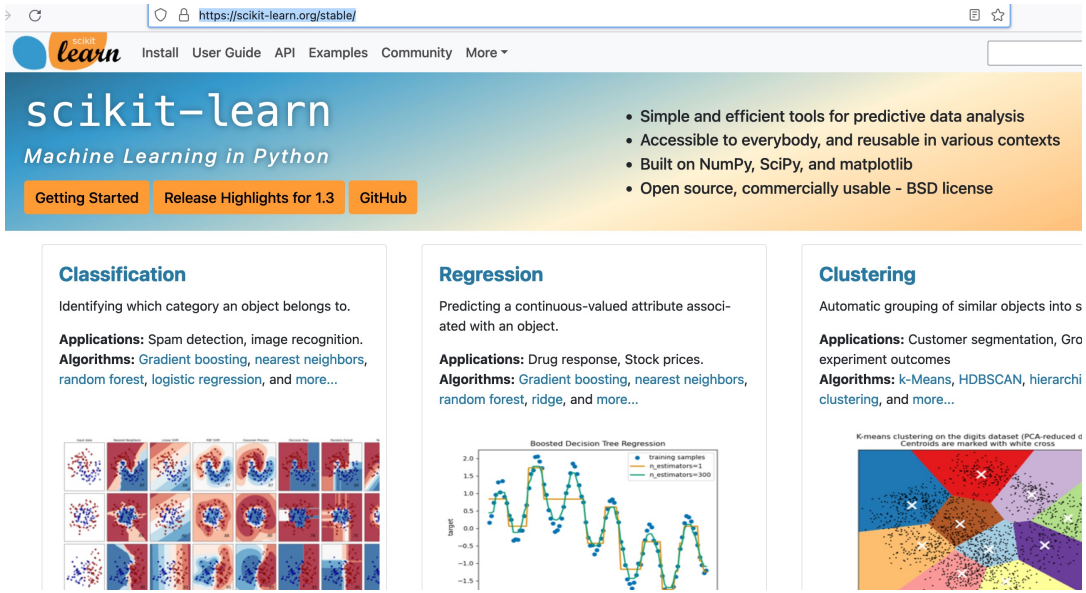



- Tools:
  - Weka - https://www.cs.waikato.ac.nz/ml/weka/
  - Download tool and dataset
- Libraries
  - Scikit - https://scikit-learn.org/stable/

# Reference and Demo

- Data: UCI Datasets
  - https://archive.ics.uci.edu/datasets
  - Browse or search
- Tools:
  - Weka - https://www.cs.waikato.ac.nz/ml/weka/
  - Download tool and dataset
- Libraries
  - Scikit - https://scikit-learn.org/stable/



# Exercise: German Credit

- Check in UCI
- Look at variants
  - https://archive.ics.uci.edu/dataset/573/south+german+credit+update

# Project Discussion

# <Project Title> - <Your Name>

Format for Capturing Project Information

## Project Context

1. Title:
2. Key idea: (2-3 lines)
3. Who will care when done:
4. Data need:
5. AI Methods:
6. Evaluation:
7. Users:
8. Trust issue:

**1 min context, 1 min achievement, 1 min Q/A**

# Course Project

- **Framework**
  1. (Problem) Think of a problem whose solution may benefit people (e.g., health, water, air, traffic, safety)
  2. (User) Consider how the primary user (e.g., patient, traveler) may be solving the problem today
  3. (AI Method) Think of what the solution will do to help the primary user
     1. Solution => ML task (e.g. classification), recommendation, text summarization, …
     2. Use a foundation model (e.g., LLM-based) solution as the baseline
  4. (Data) Explore the data for a solution to work
  5. (Reliability: Testing) Think of the evaluation metric we should employ to establish that the solution will works? (e.g., 20% reduction in patient deaths)
  6. (Holding Human Values) Discuss if there are fairness/bias, privacy issues?
  7. (Human-AI) Finally, elaborate how you will explain the primary user that your solution is trustable to be used by them

# Project Discussion: What to Focus on ?

- Problem: you should care about it
- Data: should be available
- Method: you need to be comfortable with it. Have at least two – one serves as baseline
- Trust issue
  - Due to Users
    - Diverse demographics
    - Diverse abilities
    - Multiple human languages
  - Or other impacts
- What one does to mitigate trust issue

# Rubric for Evaluation of Course Project

**Project**

- Project plan along framework introduced (7 points)
- Challenging nature of project
- Actual achievement
- Report
- Sharing of code

**Presentation**

- Motivation
- Coverage of related work
- Results and significance
- Handling of questions

# Project Discussion

1. Create a private Github repository called "CSCE581-Spring2026-<studentname>-Repo". Share with Instructor (biplav-s)
2. Create a folder called "Project". Inside, create a text file called "ProjectPlan.md" (or "ProjectPlan.txt") and have details by the next class (Jan 30, 2026)

1. Title:
2. Key idea: (2-3 lines)
3. Who will care when done:
4. Data need:
5. Methods:
6. Evaluation:
7. Users:
8. Trust issue:

# Concluding Section

# Week 3 (L5 and L6): Concluding Comments

- We looked at
  - Data and characteristics
  - ML background and methods
- Prepares us for understanding trust issues

# About Next Week – Lectures 7, 8

# Lectures 7, 8: AI / ML Methods and Trust

- Supervised ML
- Trust issues

| Class # | Date | Description | Comments |
|---|---|---|---|
| 1 | Jan 13 (Tu) | Introduction, Trusted AI | W1 |
| 2 | Jan 15 (Th) | Case Studies: Data Analysis for AI, Analysis for Trust [Traffic], Recommendations and Trust [Fairness and ULTRA] | |
| 3 | Jan 20 (Tu) | Review: Trusted Decisions, Expectations, Course Scope; Data | W2 |
| 4 | Jan 22 (Th) | AI: Data Prep, Knowledge Graph | |
| 5 | Jan 27 (Tu) | Common AI methods: ML Landscape | W3 |
| 6 | Jan 29 (Th) | AI - Structured: Analysis – Supervised ML | |
| 7 | Feb 3 (Tu) | AI - Structured: Analysis – Supervised ML | W4 |
| 8 | Feb 5 (Th) | Project discussion (1) | |
| 9 | Feb 10 (Tu) | Quiz 1 | W5<br>Quiz 1 - start |
| 10 | Feb 12 (Th) | AI - Structured: Analysis – Supervised ML – Trust Issues | Quiz 1 - end |