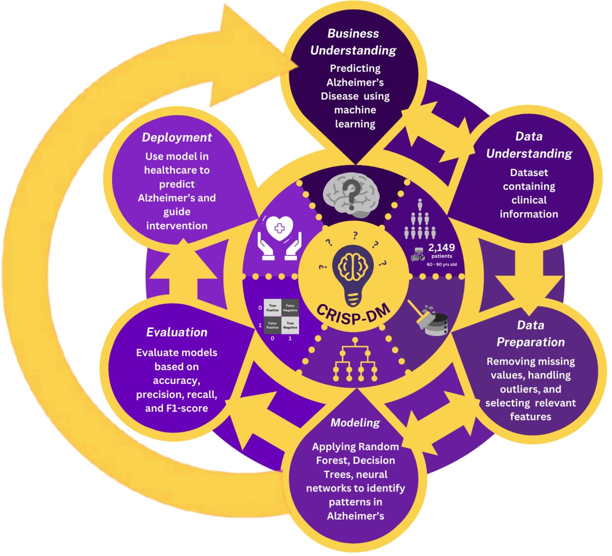
Fariha Kha
Professor Antonio Cerullo
MHC 200
September 4, 2024

Personal Development Project: How AI / ML Aids Research

Project Objective

The intersection of artificial intelligence (AI) and machine learning (ML) with science has the potential to revolutionize our understanding of the world. For this Personal Development Project, I aim to explore how AI/ML can aid research, specifically within the field of computational neuroscience. By reviewing relevant literature and working on simple computational projects, I hope to gain insights into how these technologies can be used to analyze brain activity, model neural networks, and further our understanding of cognitive functions.

Predicting Alzheimer's Disease Using the Cross-Industry Standard Process for Data Mining (CRISP-DM)


Business Understanding
Predicting Alzheimer's Disease using machine learning
Data Understanding
Dataset containing clinical information
Data Preparation
Removing missing values, handling outliers, and selecting relevant features
Modeling
Applying Random Forest, Decision Trees, neural networks to identify patterns in Alzheimer's
Evaluation
Evaluate models based on accuracy, precision, recall, and F1-score
Deployment
Use model in healthcare to predict Alzheimer's and guide intervention
CRISP-DM
2,149
patients
60 - 90 yrs old
True Positive
False Negative
False Positive
True Negative
0
1
0
1

This project will involve a combination of theoretical learning and practical application. Throughout the semester, I will document my progress weekly, summarizing key findings from research articles and reflecting on any challenges or breakthroughs encountered in my projects. The ultimate goal is to understand the impact of AI/ML on neuroscience research and to refine my ability to adopt and apply new knowledge in this rapidly evolving field. I will start from general to specific, so I will go from defining what AI is to implementing what I learned into a project using a public database.

Week 1: 09/04 - 09/10

What is AI?

This week, I came across an article from Google Cloud that dives into the world of artificial intelligence (AI). It offers a comprehensive look at what AI is, how it works, and its various applications. The article not only covers the foundational concepts of AI, such as machine learning and deep learning but also discusses different types of AI and the intricate neural network models that power them. It highlights the transformative benefits of AI, from automating workflows to accelerating research and development.

Summary
The article from Google Cloud defines AI as a technology that allows computers to perform tasks that typically require human intelligence. Key aspects covered include:

- Machine Learning and Deep Learning: AI systems use these methods to learn from data, identify patterns, and make decisions. Machine learning involves training algorithms on data to perform specific tasks, while deep learning employs neural networks to handle more complex data patterns.
- Types of AI: The article outlines various stages of AI development, from reactive machines that follow preprogrammed rules to the more advanced concepts of self-aware AI, which remains theoretical at this stage.
- Neural Network Models: Different neural networks, such as feedforward, recurrent, and convolutional networks, are explained. These models are crucial for tasks like image and speech recognition, which are integral to many AI applications.
- Benefits of AI: The article highlights how AI can enhance efficiency and productivity by automating repetitive tasks, reducing errors, and speeding up research and development processes.

Application to My Project
In the context of my project, which investigates how AI/ML can aid in research, the article provides valuable insights:

- Data Analysis: The ability of AI to process and analyze large volumes of data quickly is crucial for neuroscience research, where datasets can be vast and complex.
- Neural Networks: Understanding different types of neural networks will help me apply the appropriate models to computational neuroscience projects. For instance, convolutional neural networks (CNNs) might be useful for analyzing imaging data, while recurrent neural networks (RNNs) could be applied to time-series data from neural recordings.
- Research Acceleration: The article's discussion on AI's role in speeding up research aligns with my goal to leverage AI for uncovering new insights into brain function. Using

AI tools from platforms like Google Cloud can facilitate efficient data processing and pattern detection, ultimately advancing my research.

This foundational knowledge will inform my approach as I dive deeper into how AI can be applied to computational neuroscience, helping me to select the right techniques and tools for my research objectives.

Ideal Action Steps for Week 2

- ☑ Further Reading: I will continue to explore articles and resources on AI and its applications in research/neuroscience. This will help me build a more comprehensive understanding of how AI can be applied to specific research challenges in the field.
- ☑ Select Tools: Based on the insights from this article, I will evaluate AI tools and platforms that could support my project. Google Cloud's offerings will be considered for its capabilities in handling large datasets and providing advanced neural network models.
- ☑ Project Plan: I will outline a plan for integrating AI techniques into my computational neuroscience project. This includes defining specific research questions, selecting appropriate AI models, and setting up a workflow for data analysis.
- ☑ Explore Datasets: I will visit some free datasets, review their documentation, and download a sample to understand their structure and content. Based on the chosen dataset, I can figure out my project objectives and determine the AI techniques I will need to use.
- ☑ Experimentation: Begin experimenting with AI tools and techniques on smaller datasets to test their effectiveness and understand their potential applications in my research.
- ☑ By the end of Week 2, I aim to have a clearer plan for incorporating AI into my project and to start implementing some of the techniques discussed in the article.

"What Is Artificial Intelligence (AI)? | Google Cloud." Google, Google, cloud.google.com/learn/what-is-artificial-intelligence.

Week 2: 09/11 - 09/17

The Effect of AI on Research

In the article “Speeding up to keep up: exploring the use of AI in the research process,” Jennifer Chubb, Peter Cowling, and Darren Reed explore the increasing role of artificial intelligence (AI) in academic research. They discuss AI’s potential to transform research methodologies, streamline processes, and address the evolving landscape of academic work. Through interviews with leading scholars, the authors analyze AI’s positive and negative impacts on research practice and culture. While AI is seen as a tool that could augment the research process, concerns are raised about how its integration might affect creativity, academic identity, and the traditional values of research institutions.

Summary

- Positive Impacts of AI in Research:
    - Enhanced Efficiency: Chubb, Cowling, and Reed note that AI can assist researchers by streamlining repetitive and mundane tasks, such as data analysis and information gathering, allowing them to focus more on innovative aspects of their work.
    - Support for Interdisciplinarity: AI’s capacity to handle large datasets and facilitate complex analysis can promote interdisciplinary research, enabling collaborations across various fields → I will be handling large datasets!
    - Boost in Information Gathering: The authors highlight that AI’s capability to quickly gather and synthesize vast amounts of information is seen as a significant benefit, aiding researchers in staying up-to-date with the latest developments in their fields.
    - Potential for New Research Methods: AI has the potential to introduce new research methods and processes, leading to novel insights and discoveries that may not be achievable through traditional means.
- Negative Impacts and Concerns:
    - Bureaucratic Pressures: The use of AI to speed up—to “keep up” with metricized academic processes could exacerbate existing issues in the research culture, such as the pressure to publish rapidly and the prioritization of quantity over quality.
    - Threat to Academic Creativity: Chubb, Cowling, and Reed caution that while AI can assist in research, its growing influence might undermine human creativity and the nuanced thought processes that are crucial to groundbreaking research.
    - Risk of Bias and Surveillance: The authors raise concerns about the potential biases that AI might introduce, particularly in peer review processes, and the risks of AI-driven surveillance in research management, which could lead to an overemphasis on measurable outcomes at the expense of scholarly diversity.

    - Ethical and Cultural Implications: They emphasize the need for a more profound exploration of how AI might alter academic identities and the traditional roles of institutions, warning against a purely efficiency-driven approach that could compromise the values and principles that underpin research.
- Future Considerations and Research Directions:
    - Need for Meta-Research: Chubb, Cowling, and Reed advocate for further meta-research into AI's role in research, focusing on how it affects creativity, research integrity, and the values driving academic institutions.
    - Anticipatory Approaches: The authors call for proactive engagement with diverse and critical voices at the policy level to anticipate the broader societal impacts of AI on research, ensuring that its implementation is guided by ethical considerations and a commitment to enhancing the research process without replacing the human element.
    - Responsible AI Integration: They suggest that the future role of AI in research must be carefully managed to support and augment, rather than replace, the creative and intellectual contributions of researchers.

Chubb, J., Cowling, P. & Reed, D. Speeding up to keep up: exploring the use of AI in the research process. AI & Soc 37, 1439–1457 (2022). https://doi.org/10.1007/s00146-021-01259-0

AI / ML Usage in Computational Neuroscience

In the article “Cognitive computational neuroscience,” Nikolaus Kriegeskorte and Pamela K. Douglas explore the integration of cognitive science, computational neuroscience, and artificial intelligence to understand brain information processing. The goal is to build task-performing computational models that mimic brain mechanisms and explain cognition. Historically, cognitive science and neuroscience took separate paths: cognitive science focused on functional models of cognition without brain constraints, while neuroscience emphasized biological plausibility but lacked computational rigor.

Recent advances in AI, brain imaging, and neuroscience offer opportunities to unify these fields into a new discipline called Cognitive Computational Neuroscience. This emerging field seeks to create biologically plausible models capable of performing real-world cognitive tasks while explaining brain activity and behavior. Bridging theory and experiment is central, requiring models that integrate insights from brain connectivity, dynamics, decoding, and representational spaces.
Cognitive Science:

- Focused on decomposing cognition into functional components and building task-performing models
- Achievements include Bayesian cognitive models and symbolic cognitive architectures

- Early limitations: lacked brain data integration and computational complexity

Computational Neuroscience:

- Focused on bottom-up modeling of neuron dynamics and elementary cognitive functions (sensory coding, motor control)
- Challenges: models often fail to scale to complex cognition

Artificial Intelligence:

- Combines component functions for intelligent behavior using deep learning and neural networks
- Strength: ability to handle large datasets and complex learning tasks
- Limitation: current models simplify biological features like action potentials

Challenges in Brain Understanding:

- Need to bridge the gap between cognitive science's task-performing models and neuroscience's biological realism
- Mapping brain activity (via fMRI and imaging) to cognitive functions provides constraints but not mechanisms

Emerging Integration:

- Models must capture both behavior (cognitive fidelity) and neuronal dynamics (biological fidelity)
- Approaches include:
  - Connectivity Models: Analyze brain region interactions and dynamics
  - Decoding Models: Reveal what information is represented in brain activity
  - Representational Models: Characterize representational spaces and adjudicate between computational hypotheses

Kriegeskorte, N., Douglas, P.K. Cognitive computational neuroscience. Nat Neurosci 21, 1148–1160 (2018). https://doi.org/10.1038/s41593-018-0210-5

Machine learning techniques are key to linking brain data with computational models. I am particularly interested in pursuing a deeper understanding of how computational methods can uncover insights into neurological conditions. For the rest of the semester, I plan to focus on an Alzheimer's Disease (AD) dataset, which provides detailed information on biomarkers, cognitive assessments, and genetic factors.

I chose this dataset because Alzheimer's represents one of the most challenging and impactful neurodegenerative disorders, and its complexity demands innovative approaches. By applying machine learning to this data, I hope to identify patterns that could inform early detection or therapeutic strategies. This project aligns with my passion for computational neuroscience, as it bridges the gap between raw data and meaningful interpretations about the brain's function and dysfunction. I aim to contribute to the development of tools that enhance our understanding of the human mind and improve the quality of life for individuals impacted by neurological disorders.

Dataset Choice

Throughout this project, I have chosen to use a synthetic dataset that provides extensive health information on 2,149 patients, ranging from demographic details to cognitive assessments and diagnosis status. This dataset includes key factors such as medical history, lifestyle habits, and clinical measurements, making it ideal for exploring the relationships between these variables and Alzheimer's Disease. The dataset's features include cognitive scores like the Mini-Mental State Examination (MMSE), functional assessments, and symptoms such as confusion and forgetfulness. Additionally, it contains diagnosis information, enabling me to develop predictive models to better understand Alzheimer's onset and progression. By using this dataset, I aim to delve into the complex factors contributing to Alzheimer's Disease, leveraging AI and ML techniques to uncover patterns and potential early indicators. This research aligns with my long-standing interest in Alzheimer's, and I am excited to explore the predictive potential of this data to contribute to advancements in the field. I obtained this dataset from Kaggle, as it allows me to compare my approach to machine learning with those used by other researchers and coders. I will be coding in Python, using Visual Studio Code for development, which is a programming environment I am comfortable with.* In the future, I may also create a Google Colab notebook, as it would make it easier for others to view, run, and experiment with my code.

* Note from Week 5: Using Google Colab, not VS Code - still Python. I will provide links to the Colabs that can be viewed by anyone with the link.

Ideal Action Steps for Week 3

- ☑ Figure out what my next steps are! Plan out an outline of how I would like to conduct this project.
- ☑ Research Alzheimer's Disease.
- ☑ Go through the Kaggle fully. This includes the description of the dataset and what other people have done. - will be done in Week 4
- ☑ Draft a graphical abstract, or at least a plan for it.

Week 3: 09/18 - 09/24

When I first learned about machine learning through a course I took over the summer, I was taught about a process called “CRISP-DM,” which stands for Cross Industry Standard Process for Data Mining. I found the acronym CRISP-DM a bit silly since, technically, it should be CISP-DM. However, CRISP-DM sounds much better and is more memorable…I guess. So,

What is CRISP-DM?

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a standardized framework developed in 1999 to guide data mining and data science projects across industries. It is the most widely used methodology for analytics projects, known for its flexibility to be implemented in either agile or waterfall styles.

Here are the six steps to CRISP-DM:

<u>Business Understanding:</u>

- Establish a clear understanding of project objectives and requirements from both business and technical perspectives
    - Define business objectives and success criteria
    - Assess the situation, resources, risks, and perform cost-benefit analysis
    - Determine data mining goals for technical success
    - Create a detailed project plan

<u>Data Understanding:</u>

- Identify, collect, and analyze relevant datasets
    - Collect initial data and load it into analysis tools
    - Describe the data’s surface properties (format, size)
    - Explore the data using queries and visualization to find relationships
    - Verify data quality and document any issues

<u>Data Preparation: (80% of the project)</u>

- Prepare datasets for modeling; this phase often takes up 80% of the project
    - Select relevant data and justify inclusion/exclusion
    - Clean data by correcting or imputing values
    - Construct new attributes
    - Integrate data from multiple sources
    - Format data for analysis

<u>Modeling:</u>

- Build and test predictive models

    - Select algorithms (regression, neural networks, etc.)
    - Design testing protocols (train-test split)
    - Build models (often simple code execution)
    - Assess models and iterate based on technical success and business goals

Evaluation:

- Determine if the model meets business needs and decide next steps
    - Evaluate model results against business criteria
    - Review the process to ensure thorough execution
    - Decide whether to deploy, iterate further, or start new projects

Deployment:

- Deliver model outputs to stakeholders for real-world use
    - Plan deployment (dashboards, APIs)
    - Plan monitoring and maintenance to ensure long-term success
    - Produce a final report summarizing results
    - Conduct a project retrospective to capture lessons learned

Hotz, Nick. "What Is CRISP DM?" Data Science PM, 9 Dec. 2024, www.datascience-pm.com/crisp-dm-2/.

I have chosen to use the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework for my project because it provides a structured and iterative approach to tackling data-driven problems. In the **Business Understanding** phase, I will define the primary goal of my project: to use machine learning for early detection of Alzheimer's Disease. During **Data Understanding**, I will explore the dataset to uncover key patterns, assess data quality, and identify any potential gaps. The **Data Preparation** phase will involve cleaning, transforming, and engineering features to ensure the dataset is ready for analysis. In the **Modeling** phase, I will develop and fine-tune machine learning models tailored to detect patterns associated with Alzheimer's progression. The **Evaluation** phase will help me assess the effectiveness of my models, ensuring they meet the project's objectives and perform strongly. Finally, the **Deployment** phase will involve creating actionable insights or a practical tool that can assist in early detection, with documentation for stakeholders. This systematic process ensures that my project is both rigorous and aligned with its intended purpose.

Before I look into the Kaggle, I have decided to dedicate this week to learning about Alzheimer's Disease. This will help me get started with the **Business Understanding** phase. Here are some of the notes I have taken that I should know about before going into this project:

General Overview

- Alzheimer’s Disease is a progressive brain disorder that destroys memory, thinking skills, and the ability to perform basic tasks
- Affects over 6 million Americans, primarily aged 65 and older
- Seventh leading cause of death in the U.S. and the most common cause of dementia
- Dementia: Involves loss of cognitive and behavioral functioning, ranging from mild impairment to severe dependency

Key Characteristics

- Main Features:
    - Amyloid plaques (abnormal protein clumps)
    - Neurofibrillary tangles (tau protein fibers)
    - Loss of neuron connections and neuron death
- Progression: Begins in the hippocampus and spreads, causing significant brain shrinkage in later stages

Stages of Alzheimer’s

- Mild Alzheimer’s:
    - Memory loss, wandering, difficulty handling finances, personality changes
    - Often diagnosed at this stage
- Moderate Alzheimer’s:
    - Language issues, confusion, hallucinations, and difficulty with daily tasks like dressing.
- Severe Alzheimer’s:
    - Inability to communicate or care for oneself; significant brain tissue loss.

Risk Factors and Causes

- Unknown Complete Cause: Likely a combination of age, genetics, environment, and lifestyle (scientists don’t fully yet understand)
- Age-Related Brain Changes:
    - Atrophy, inflammation, blood vessel damage, free radicals, and mitochondrial dysfunction

Genetics

- APP, PSEN1, PSEN2 (rare, early-onset Alzheimer’s).
- APOE gene:
    - APOE ε4: Increases risk and leads to earlier onset
    - APOE ε2: May offer some protection
- Down Syndrome: Increased risk due to an extra copy of chromosome 21 (APP gene)

Early Detection and Research

- Biomarkers: Biological indicators (brain scans, cerebrospinal fluid) under study to detect early changes
- Research Focus: Understanding plaques, tangles, and brain changes before symptoms appear

Clinical Trials and Participation

- Vital for discovering causes, treatments, and preventive strategies
- Diverse groups of participants needed by age, sex, and ethnicity
- Join through registries, local trials, or Alzheimer's research centers

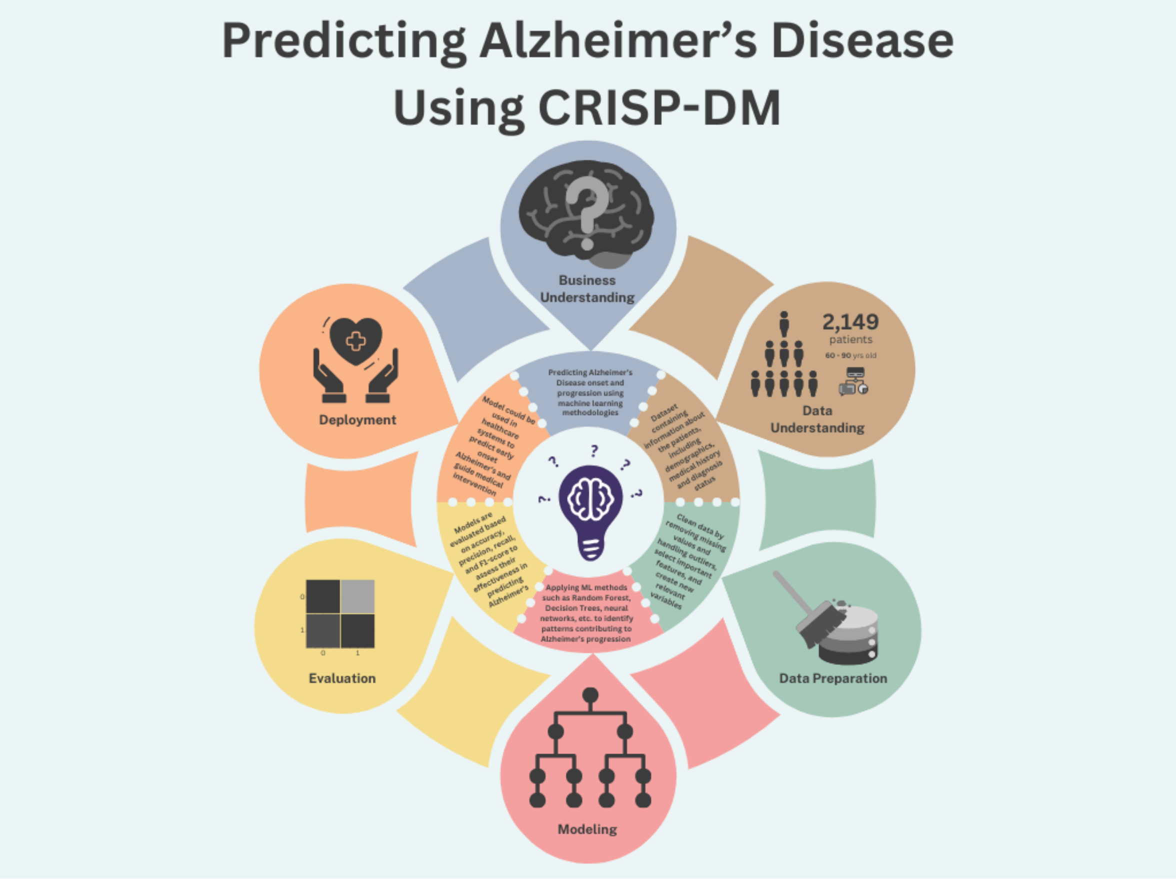
Alzheimer's Disease Fact Sheet | National Institute on Aging, www.nia.nih.gov/health/alzheimers-and-dementia/alzheimers-disease-fact-sheet.

Ideal Next Steps for Week 4

☑ Begin going through Kaggle.
☑ Finalize Graphical Abstract by September 26th @ 11:59 PM
☑ Commence to Step 2 of CRISP-DM: Data Understanding

Week 4: 09/25 - 10/01

Graphical Abstract I:


Predicting Alzheimer's Disease
Using CRISP-DM
Business Understanding
Predicting Alzheimer's Disease onset and progression using machine learning methodologies
2,149
patients
60 - 90 yrs old
Data Understanding
Dataset containing information about the patients, including demographics, medical history and diagnosis status
Clean data by removing missing values and handling outliers, select important features, and create new relevant variables
Data Preparation
Applying ML methods such as Random Forest, Decision Trees, neural networks, etc. to identify patterns contributing to Alzheimer's progression
Modeling
Models are evaluated based on accuracy, precision, recall, and F1-score to assess their effectiveness in predicting Alzheimer's
0
1
0
1
Evaluation
Model could be used in healthcare systems to predict early onset Alzheimer's and guide medical intervention
Deployment

For my first attempt at the graphical abstract, I decided to base the design on the CRISP-DM process, splitting it into six sections to align with its six phases. Using Canva, I created a colorful pie chart (styled like a donut with a hole in the middle) surrounded by bubbles. Each section of the donut represents a phase of CRISP-DM and contains text, while the bubbles touching each section hold related images. At the center, I placed a gray image of a brain inside a light bulb with question marks to symbolize Alzheimer's—a subtle alternative to the overused missing puzzle piece motif. Keeping with a cohesive gray theme for the images, I chose a gray brain with a question mark for **Business Understanding**, figures depicting the 2,149 study participants for **Data Understanding**, and a broom sweeping a data symbol for **Data Preparation** to represent data cleaning. For **Modeling**, I used a chart symbolizing tree-based models like random forests and gradient boosting, while **Evaluation** features a confusion matrix. Finally, for **Deployment**, I included a healthcare-related image to convey the ultimate goal of applying insights to Alzheimer's prevention or treatment. Each section is a different color—blue, brown, green, red, yellow, and orange—against a light blue background. I connected the bubbles to show that the process is tangible—I can go back and forth at certain steps throughout this process.

Notes after feedback: The current design has readability issues: the text is too small, the pictures are disproportionately large, and the general layout needs refinement. The colors may also require adjustment for better aesthetic harmony. I also feel like the "tangible process" is not shown very clearly here, so I will be adding some arrows.

Now it is time to understand the data and see if that helps me go forward with the process.

Dataset Overview

The Alzheimer's Disease dataset contains comprehensive health information for 2,149 patients, uniquely identified by IDs ranging from 4751 to 6900. The dataset includes diverse variables across several categories:

- **Demographic Details:** Age, gender, ethnicity, and education level
- **Lifestyle Factors:** BMI, smoking status, alcohol consumption, physical activity, diet quality, and sleep quality
- **Medical History:** Family history of Alzheimer's, cardiovascular disease, diabetes, depression, head injury, and hypertension
- **Clinical Measurements:** Blood pressure, cholesterol levels (total, LDL, HDL, triglycerides), and more
- **Cognitive and Functional Assessments:** MMSE scores, functional assessments, memory complaints, behavioral problems, and activities of daily living (ADL)
- **Symptoms:** Presence of confusion, disorientation, personality changes, forgetfulness, and difficulty completing tasks
- **Diagnosis Information:** The key outcome—whether the patient has been diagnosed with Alzheimer's Disease (0 = No, 1 = Yes).

Features and Label

<u>Features:</u> Individual measurable properties or characteristics of the data used by a model to make predictions or decisions

Some of the variable that will be used as input to the model:

- Demographics
- Lifestyle Factors
- Medical History
- Cognitive Assessments
- Symptoms

<u>Label:</u> output or target value that a model is trained to predict, often representing the ground truth in supervised learning tasks

- Diagnosis: Whether the patient has Alzheimer's Disease (binary: 0 = No, 1 = Yes).

Visualizing the Data Understanding Phase

- Highlight important trends in the data: distributions of key features like age, MMSE scores, or the prevalence of Alzheimer’s (label)
- Provide summary statistics for numerical features (age, cholesterol levels, MMSE score)
- Visualize key relationships using bar charts (diagnosis vs. gender), histograms, and correlation matrices to show patterns in the dataset.

Based on my understanding of the data, I have determined that the business problem involves performing *binary classification*, where the goal is to predict whether a patient has Alzheimer’s Disease (class 1) or not (class 0) using the provided features.

Binary classification is a supervised learning task where the model predicts one of two possible classes based on the input data. In this case:

- Class 0 (Negative Class): The patient does not have Alzheimer’s Disease.
- Class 1 (Positive Class): The patient has Alzheimer’s Disease.

This is a clear example of how in the CRISP-DM process, it is common to revisit earlier stages as new insights emerge. During the “Data Understanding” phase, I identified that the task involves binary classification. This realization prompted me to revisit the “Business Understanding” phase to refine the project goal, explicitly framing it as predicting Alzheimer’s Disease presence (1) or absence (0). This iterative back-and-forth demonstrates the flexibility of CRISP-DM, ensuring each phase informs and enhances the others for a better project outcome.

Ideal Next Steps for Week 5

☑ Commence to Step 3 of CRISP-DM: Data Preparation

Week 5: 10/02 - 10/08

I have now finally began coding! The first thing I like to do when datasets are involved are analyzing the data through graphs. Visualization is very important, as we learned in the lessons on how to make a graphical abstract. Instead of putting the code and output in this already long Google Doc, I have provided all code and output in this Google Colab. In case someone needs to run something on this, I suggest making a copy of this, downloading the dataset and uploading it to the respective folder (upload the file to Google Drive and then put it in a folder called "Colab Notebooks"), and then running each cell by holding "shift" + "return/enter."

Colab Notebook Link: CO Predicting Alzheimer's Disease - With Model Training

Week 6: 10/09 - 10/15

New updates can be seen in the same Colab Notebook. I decided to look into some more statistical analysis, and if I get to it, I may run some chi-square tests. Unfortunately, I have run into quite a few errors, so I may have to update with data analysis again another week when I figure it out*…Sometimes in computer science, the answer takes time to come to mind…and sometimes you just leave it be.

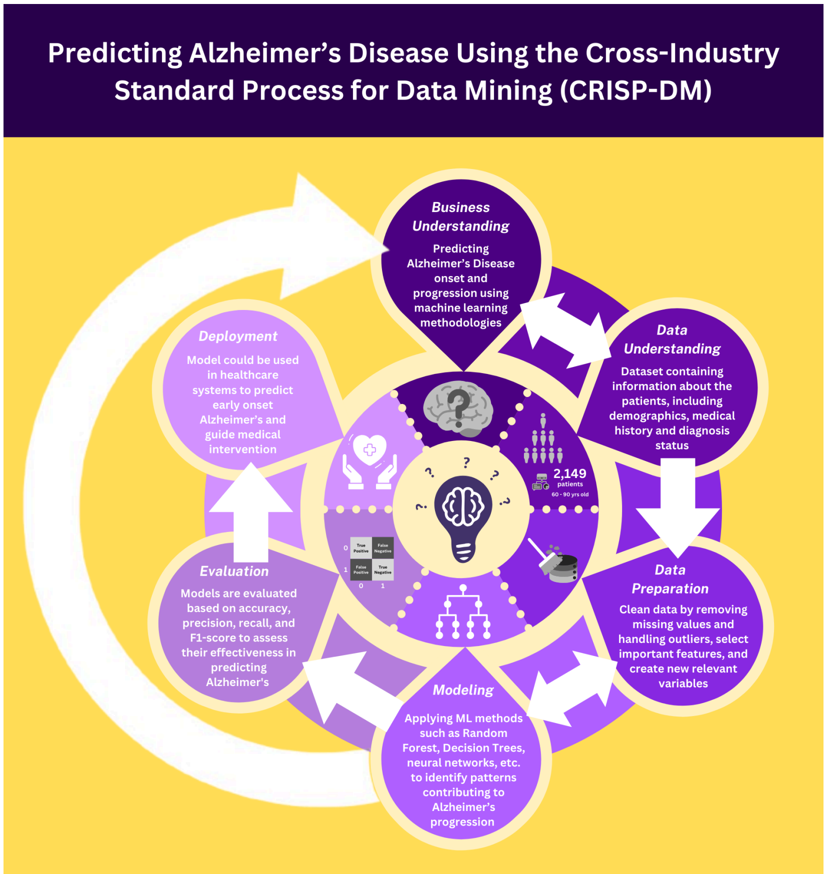
Colab Notebook Link: CO Predicting Alzheimer's Disease - With Model Training

Overall, I have to say I was very lucky to have this dataset. It did not require too much preprocessing. I only had to clean two columns. Everything else that I have done over the past two weeks was just taking a look at the data and analyzing it to find possible correlation/causation relationships with Alzheimer's Disease. I said correlation because no one still knows what "causes" AD.

* Note from Week 15: I never figured out what the problem was with the chi-square tests. It still shows up as "Not a Number" values.

Week 7: 10/16 - 10/22

Graphical Abstract II:



Predicting Alzheimer’s Disease Using the Cross-Industry Standard Process for Data Mining (CRISP-DM)

Business Understanding

Predicting Alzheimer's Disease onset and progression using machine learning methodologies

Data Understanding

Dataset containing information about the patients, including demographics, medical history and diagnosis status

Data Preparation

Clean data by removing missing values and handling outliers, select important features, and create new relevant variables

Modeling

Applying ML methods such as Random Forest, Decision Trees, neural networks, etc. to identify patterns contributing to Alzheimer's progression

Evaluation

Models are evaluated based on accuracy, precision, recall, and F1-score to assess their effectiveness in predicting Alzheimer's

Deployment

Model could be used in healthcare systems to predict early onset Alzheimer's and guide medical intervention

I have now made drastic changes to my graphical abstract. I have a color scheme, involving changing shades of purple from dark purple to light purple as well as a yellow background to make the purple stand out as purple and yellow are complementary colors. I included a more comprehensive title that lets the reader know instantly what “CRISP-DM” is. I should have done that from the beginning because it is always important in writing to inform the reader of what an acronym stands for at first. I also decided to include arrows that point back and forth between each stage to show that CRISP-DM is not a straightforward process. What is great about this process is that people are allowed to go back and change whatever they need to. For example, I knew I wanted to look into Alzheimer’s Disease and predicting it, and I was able to solidify my **Business Understanding** last time by **understanding the data** first, and then going back knowing the features and what my exact label will be. Since the text was too small and the images were too large, I decided to swap their locations—text in bubbles and speech in donut sections.

Week 8: 10/23 - 10/29

Now, it is time to hustle. Each week I will be trying out a different machine learning model to use on the data and analyze the results. Before actually coding, I will find an article that relates that model possibly to a study in Alzheimer's Disease. This week, I decided to focus on Random Forest because when I looked through the Kaggle, many of the competitors used Random Forest and had great results with it.

Random Forest

Notes on "Early Alzheimer's Detection Using Random Forest Algorithm"

- Early detection of Alzheimer's Disease is essential for slowing progression.
- Machine learning, specifically the Random Forest (RF) algorithm, achieved 93.69% accuracy for detecting Alzheimer's
- Researchers used the OASIS-2 longitudinal MRI dataset, which includes 150 individuals aged 60-96 and 373 MRI sessions
- Feature selection was performed using a correlation-based method to identify important variables like age, sex, education, socio-economic status, and MMSE scores
- Data preprocessing involved removing irrelevant features and outliers to improve model performance
- Multiple machine learning models were tested, but Random Forest outperformed others
    - Due to RF's ability to handle high-dimensional data and its strength in identifying important features for classification
    - Effective in using the selected features to distinguish between Alzheimer's Disease and healthy cases with high accuracy.
- A GUI was designed for user-friendly implementation of early detection using the Random Forest algorithm
- Random Forest shows promise for developing effective strategies to detect and slow Alzheimer's progression

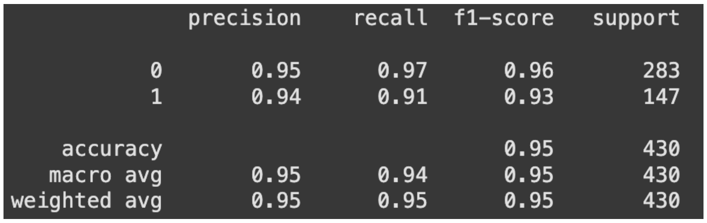
Pros	Cons
● Handles large datasets well ● Works well with both classification and regression ● Less prone to overfitting compared to individual decision trees ● Can handle missing values ● Automatically handles feature	● Can be computationally expensive ● Difficult to interpret ● Model can be slow for real-time predictions ● Requires more memory for large datasets ● Not ideal for small datasets

selection	

Kolte, Pranjlee, et al. “Early Alzheimer’s Detection Using Random Forest Algorithm.” 2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT), 25 May 2023, pp. 1–5, https://doi.org/10.1109/iconscept57958.2023.10170234.

All of my code can be found for this whole project in this Colab. This also includes the data analysis as well: Predicting Alzheimer's Disease - With Model Training

Here is my classification report:



```
              precision    recall  f1-score   support

           0       0.95      0.97      0.96       283
           1       0.94      0.91      0.93       147

    accuracy                           0.95       430
   macro avg       0.95      0.94      0.95       430
weighted avg       0.95      0.95      0.95       430
```

These are stellar results! The best precision, recall, and f1-scores are greater than or equal to 0.8. I may use GridSearchCV to tune hyperparameters, but I might not need to as these scores are ready for deployment. Based on my pros and cons on Random Forest, the results make sense because this dataset was neither large nor small. It was something that Random Forest could definitely handle. The article also mentioned how Random Forest works well with feature selection.

Week 9: 10/30 - 11/05

Gradient Boosting

The next most popular model I saw was Gradient Boosting. Here are my notes from an article I found:

Notes on “A transfer learning approach based on gradient boosting machine for diagnosis of Alzheimer’s disease”

- Gradient Boosting Machines (GBM) are a powerful ensemble learning method used for classification tasks, including Alzheimer’s disease (AD) diagnosis
- GBM builds a model by combining multiple weak learners (typically decision trees) into a strong predictive model through iterative learning
- In the context of Alzheimer’s disease, the TrGB model utilizes GBM for transfer learning, where knowledge from one dataset (source domain) is used to improve predictions on another (target domain)
- The TrGB method adjusts the weights of source domain instances based on their residuals to prevent negative transfer and improve performance on the target dataset
- The model aims to accurately classify different stages of Alzheimer’s disease, particularly in distinguishing between early MCI (EMCI) and late MCI (LMCI)
- The TrGB model leverages the ADNI dataset (source) and Mount Sinai dataset (target) to train and test its classification ability
- By using GBM’s gradient boosting mechanism, TrGB improves classification accuracy by 1.5-4.5% compared to traditional methods
- The model also demonstrates a 5% improvement in early vs. late MCI classification by incorporating knowledge from CN (cognitively normal) vs. AD classification tasks
- GBM’s ability to handle complex data and incorporate transfer learning is key to boosting diagnostic performance in AD-related tasks
- TrGB’s success highlights the effectiveness of GBM in transferring knowledge across datasets to enhance model generalization, particularly in medical domains with limited data

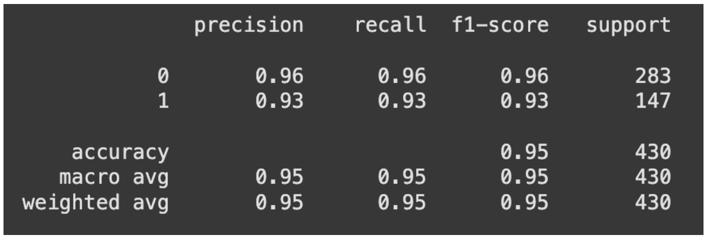
Shojaie, Mehdi, et al. “A transfer learning approach based on gradient boosting machine for diagnosis of alzheimer’s disease.” Frontiers in Aging Neuroscience, vol. 14, 5 Oct. 2022, https://doi.org/10.3389/fnagi.2022.966883.

Pros	Cons
● High predictive accuracy ● Works well with various types of data ● Handles missing data efficiently	● Longer training time ● Prone to overfitting if not tuned properly

● Flexible model (can work with regression and classification tasks) ● Can capture complex relationships in data ● Handles unstructured data well (images and text) ● Built-in feature selection ● Can be parallelized for faster performance	● Sensitive to noisy data and outliers ● Computationally expensive ● Difficult to interpret ● Requires careful hyperparameter tuning ● Slow inference time during prediction ● Memory intensive, especially with large datasets

Colab Notebook: Predicting Alzheimer's Disease - With Model Training

Here is my classification report:



```
              precision    recall  f1-score   support

           0       0.96      0.96      0.96       283
           1       0.93      0.93      0.93       147

    accuracy                           0.95       430
   macro avg       0.95      0.95      0.95       430
weighted avg       0.95      0.95      0.95       430
```

These turned out to be great results as well as all scores are greater than or equal to 0.8.

Week 10: 11/06 - 11/12

Next model:
XGBoost

Notes on “XGBoost-SHAP-based Interpretable Diagnostic Framework For Alzheimer’s Disease”

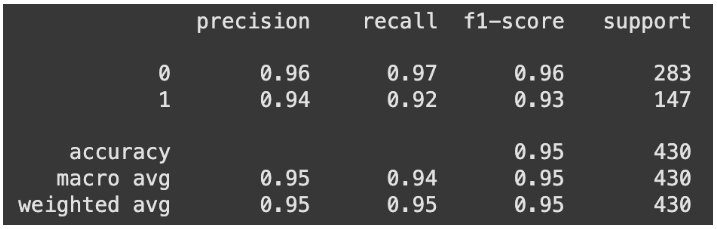
- XGBoost was used to handle imbalanced data in Alzheimer’s disease classification.
- It aimed to classify normal cognition (NC), mild cognitive impairment (MCI), and Alzheimer’s disease (AD)
- Patient data from the ADNI database was used, including clinical info, test results, and neuroimaging
- Three feature selection methods were applied before using XGBoost
- Sample weights were adjusted to deal with class imbalance in the dataset
- SHAP values were linked with XGBoost to explain model predictions
- The framework improved classification accuracy over other methods
- The top 10 features impacting AD diagnosis were identified using SHAP
- The framework achieved high accuracy, sensitivity, and specificity
- It helped in better decision-making by simplifying the classification process for clinicians

Yi, Fuliang, et al. “XGBoost-Shap-based interpretable diagnostic framework for alzheimer’s disease.” *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, 25 July 2023, https://doi.org/10.1186/s12911-023-02238-9.

Pros	Cons
● High accuracy - Often outperforms other models in terms of accuracy ● Built-in options to deal with class imbalance ● Uses L1 and L2 regularization to prevent overfitting ● Efficient with large datasets and high performance ● Works for both regression and classification problems ● Provides insights into the importance of each feature ● Built-in cross-validation helps with hyperparameter tuning	● Requires expertise to tune hyperparameters effectively ● May consume a lot of memory for large datasets ● Can be slower than simpler models on smaller datasets ● Difficult to interpret ● Prone to overfitting if not tuned properly ● Requires preprocessing of categorical variables

I have noticed that in some of the articles I found, some of these models are paired with each other. For instance, I found an article that used AdaBoost to help with Support Vector Machine (SVM). This is not something I know how to do, but I am open to learning more about how to do this if I find the right resources…

Here is my classification report:



```
              precision    recall  f1-score   support

           0       0.96      0.97      0.96       283
           1       0.94      0.92      0.93       147

    accuracy                           0.95       430
   macro avg       0.95      0.94      0.95       430
weighted avg       0.95      0.95      0.95       430
```

Once again, XGBoost is good at minimizing error as it shares a very similar approach to Gradient Bosting. XGBoost has scores greater than or equal to 0.8, so these are accurate scores.

Week 11: 11/13 - 11/19

Logistic Regression

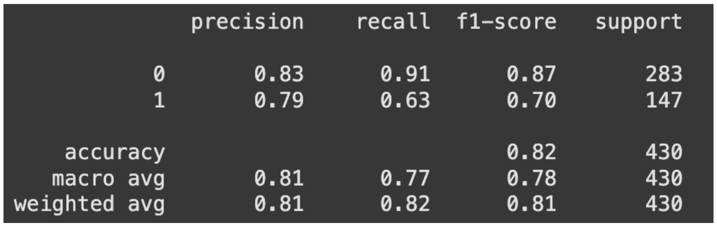
Notes on “Detection of Alzheimer’s Disease Using Logistic Regression and Clock Drawing Errors”

- The study uses logistic regression to build a model for detecting Alzheimer’s disease by analyzing clock drawing errors and other factors
- Logistic regression is applied to predict the likelihood of Alzheimer’s disease based on clock drawing test (CDT) errors, cognitive scores, and genetic and cardiovascular features
- The study proposed four logistic regression models with varying sets of variables, including
  - Clock drawing errors
  - Cognitive features like verbal fluency scores
  - Cardiovascular features
  - Genetic factors (APOE4 status)
- Only three of the ten potential CDT errors were useful in the logistic regression model, demonstrating the importance of feature selection for predictive accuracy
- The base logistic regression model (using clock drawing errors and basic control variables) achieved an AUC of 0.825, indicating decent classification accuracy
- Adding verbal fluency scores significantly improved the model’s AUC to 0.91, showing the value of integrating cognitive features in logistic regression for Alzheimer’s detection
- Logistic regression models were calibrated well and demonstrated good clinical utility, outperforming the base model in Alzheimer’s disease detection
- Adding cardiovascular and genetic data to the logistic regression models did not provide a substantial improvement in the model’s discriminatory power
- The logistic regression model incorporating clock drawing errors and verbal fluency scores offers a potentially effective screening tool for Alzheimer’s disease, showing strength of logistic regression in analyzing complex, high-dimensional data
- Further testing of logistic regression models on larger, more diverse datasets to validate the findings and explore the broader application of this approach in Alzheimer’s disease screening.

Lazarova, Sophia, et al. “Detection of alzheimer’s disease using logistic regression and clock drawing errors.” *Brain Sciences*, vol. 13, no. 8, 29 July 2023, p. 1139, https://doi.org/10.3390/brainsci13081139.

Pros	Cons
● Simple and easy to interpret ● Fast and computationally efficient ● Provides probabilistic outputs ● Works well with small datasets ● Regularization can prevent overfitting	● Assumes linear relationships ● Struggles with non-linear data ● Sensitive to outliers ● Assumes feature independence ● Limited to binary or multinomial outcomes

Here is my classification report:

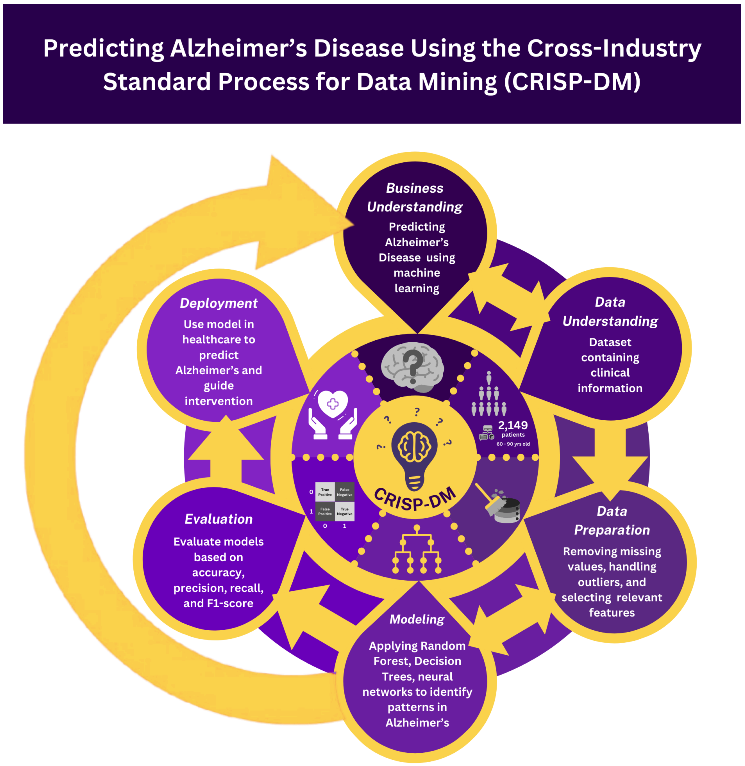


```
              precision    recall  f1-score   support

           0       0.83      0.91      0.87       283
           1       0.79      0.63      0.70       147

    accuracy                           0.82       430
   macro avg       0.81      0.77      0.78       430
weighted avg       0.81      0.82      0.81       430
```

This classification report shows that Logistic Regression may not be as well-performing as the models I have done so far. These are still valid, accurate scores asthey are typically around 0.8 or higher, but this shows that Logistic Regression is not the ideal model for this dataset.

Week 12: 11/13 - 11/19

Graphical Abstract III



Predicting Alzheimer's Disease Using the Cross-Industry Standard Process for Data Mining (CRISP-DM)

Business Understanding
Predicting Alzheimer's Disease using machine learning

Data Understanding
Dataset containing clinical information

Data Preparation
Removing missing values, handling outliers, and selecting relevant features

Modeling
Applying Random Forest, Decision Trees, neural networks to identify patterns in Alzheimer's

Evaluation
Evaluate models based on accuracy, precision, recall, and F1-score

Deployment
Use model in healthcare to predict Alzheimer's and guide intervention

I changed the purple shades to colors that are a bit darker and easier for readers to read the writing on. I made the arrows match the yellow in between the space on the pie chart. This makes the design much cleaner and aesthetically pleasing. I took out the last connecting part between deployment and business understanding because I wanted to show that deployment is truly the last step in the process. This will be my final draft of the graphical abstract.

Next model:
Naive Bayes

Notes on "The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data"

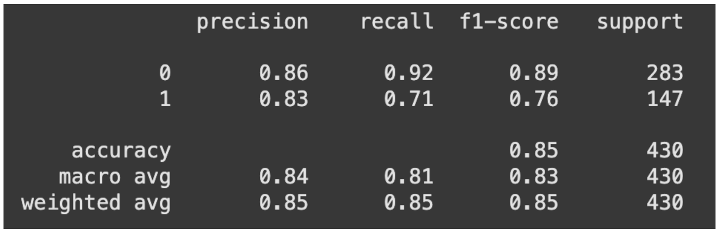
- Naive Bayes (NB) is a machine learning algorithm based on Bayes' theorem and assumes feature independence given the target variable
- It is computationally efficient, allowing quick model construction and prediction
- NB works well in many domains, including bioinformatics, due to its simplicity and speed, even when compared to more complex methods

- The model is miscalibrated, particularly when handling large numbers of features, with predictions often biased toward extreme probabilities (close to 0 or 1)
- Feature selection in NB (FSNB) aims to find the most predictive features, improving performance but increasing computation time
- Naive Bayes has been used in predicting patient outcomes from genome-wide data, showing competitive performance with other algorithms
- The Model-Averaged Naive Bayes (MANB) method averages over different NB models, incorporating various subsets of features, leading to better calibration and performance
- MANB showed a better AUC (0.72) compared to standard NB (0.59) and similar performance to FSNB, with faster training times
- Naive Bayes assumes that features are independent, but this assumption might not hold in real-world data, affecting accuracy
- Despite its simplicity, NB remains effective, especially for large, high-dimensional datasets, with computational efficiency and ease of implementation being its primary strengths

Wei, Wei, et al. “The application of naive Bayes model averaging to predict alzheimer’s disease from genome-wide data.” *Journal of the American Medical Informatics Association*, vol. 18, no. 4, July 2011, pp. 370–375, https://doi.org/10.1136/amiajnl-2011-000101.

Pros	Cons
● Simple and easy to implement ● Fast training and prediction ● Works well with small datasets ● Scalable and efficient for high-dimensional data ● Good for probabilistic predictions	● Assumes feature independence ● Poor performance with highly correlated features ● Requires large amounts of data to perform well ● Assumes the same probability distribution for all features ● Limited expressiveness for complex relationships

Here is my classification report for Naive Bayes:



```
              precision    recall  f1-score   support

           0       0.86      0.92      0.89       283
           1       0.83      0.71      0.76       147

    accuracy                           0.85       430
   macro avg       0.84      0.81      0.83       430
weighted avg       0.85      0.85      0.85       430
```

These scores are technically more accurate than Logistic Regression, but still not as great as the original scores I saw with the first few models. These are accurate as they are greater than or equal to 0.8.

Week 13: 11/20 - 11/26

Support Vector Machine (SVM)

Notes on "Accuracy of Support-Vector Machines for Diagnosis of Alzheimer's Disease, Using Volume of Brain Obtained by Structural MRI at Siriraj Hospital"

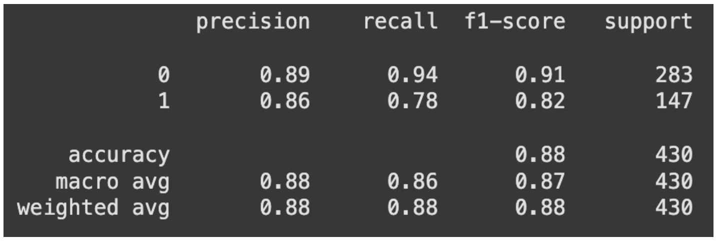
- The study explored the use of Support Vector Machines (SVM) to classify Alzheimer's Disease (AD) based on brain volume and clinical data
- Brain volumes from structural MRI scans were used, focusing on regions such as the hippocampus, caudate, thalamus, and white/gray matter volumes
- 201 subjects were randomly divided into training (92 subjects) and testing groups (91 subjects) for SVM modeling
- The highest classification accuracy (62.64%) was achieved using hippocampus volume as a single feature
- Clinical data, such as cognitive tests (e.g., TMSE, COWA, Clock-drawing test), provided better accuracy (83-90%) for AD classification
- Combining brain volumetry and clinical data did not improve accuracy beyond clinical data alone
- A leave-one-out cross-validation technique was applied to account for the small dataset size
- Radial basis function was used in SVM modeling, with adjustments to parameters like C and gamma for optimal accuracy
- The AD group had an average age of 73.09 years, while the normal control group averaged 69.72 years
- SVM models using clinical data were more accurate for AD classification, while brain volumetry alone provided lower accuracy

Vichianin, Yudthaphon, et al. "Accuracy of support-vector machines for diagnosis of alzheimer's disease, using volume of brain obtained by Structural MRI at Siriraj Hospital." *Frontiers in Neurology,* vol. 12, 10 May 2021, https://doi.org/10.3389/fneur.2021.640696.

Pros	Cons
● Effective in high-dimensional spaces ● Works well for both linear and non-linear problems ● Less prone to overfitting (with appropriate kernel) ● Effective in cases with clear margin of separation	● Memory-intensive, especially with large datasets ● Sensitive to noise, especially in imbalanced datasets ● Longer training times on large datasets ● Choosing the right kernel can be difficult

- Works well for smaller to medium-sized datasets	- Not suitable for very large datasets

Here is my classification report for SVM:



```
              precision    recall  f1-score   support

           0       0.89      0.94      0.91       283
           1       0.86      0.78      0.82       147

    accuracy                           0.88       430
   macro avg       0.88      0.86      0.87       430
weighted avg       0.88      0.88      0.88       430
```

These scores are significantly better than Logistic Regression and Naive Bayes. However, I expected better results from SVM based on the research I did. I thought that this was the perfect dataset with a sufficient amount of complexity. These are valuable scores as they are greater than or equal to 0.8 and are not suspiciously high. When results are too high, that shows that there may not have been enough variety in the dataset. Overall, even though SVM did not perform as well as I thought it would, it is still one of the best-performing model here.

Week 14: 11/27 - 12/03

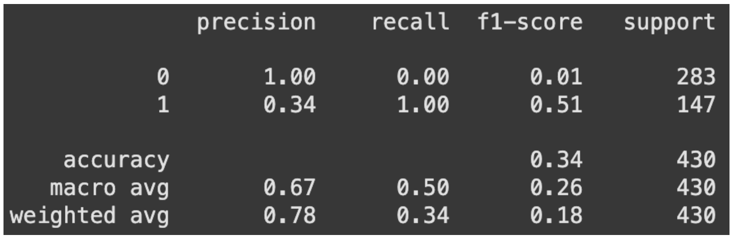
Neural networks

“Characteristics of Neural Network Changes in Normal Aging and Early Dementia”
https://www.frontiersin.org/journals/aging-neuroscience/articles/10.3389/fnagi.2021.747359/full

Pros	Cons
● Can model complex relationships ● Adaptable to various types of data (images, text, etc.) ● Good at finding patterns in noisy data ● Works well with unstructured data ● Can improve performance with more data and layers	● Requires large amounts of data ● Computationally expensive and time-consuming ● Prone to overfitting without proper regularization ● Difficult to interpret ● Sensitive to hyperparameters and architecture choice

Watanabe, Hirohisa, et al. “Characteristics of neural network changes in normal aging and early dementia.” Frontiers in Aging Neuroscience, vol. 13, 22 Nov. 2021, https://doi.org/10.3389/fnagi.2021.747359.

Here is my classification report for Neural Netowork Connections:



```
              precision    recall  f1-score   support

           0       1.00      0.00      0.01       283
           1       0.34      1.00      0.51       147

    accuracy                           0.34       430
   macro avg       0.67      0.50      0.26       430
weighted avg       0.78      0.34      0.18       430
```

This was probably the worst result for my classification report. Neural networks seem to require more hyperparameter tuning and are probably not suitable for this dataset, so I didn’t bother researching this. I know that neural networks are supposed to be for larger, more complex datasets. Neural networks are not the best option for this dataset.

Week 15: 12/04 - 12/10

Final Thoughts

After evaluating the performance of the models, here’s the summary: SVM performed impressively, achieving strong classification results. Gradient Boosting, XGBoost, and RandomForest demonstrated excellent performance with high accuracy, showing that ensemble methods are well-suited for this dataset. Random Forest offered reliable results with minimal tuning required. On the other hand, Logistic Regression and Naive Bayes did not perform as well, with their results lagging behind the other models. Neural networks performed the worst as it was not suitable for this dataset. Based on these findings, SVM, Gradient Boosting, XGBoost, and Random Forest are the best-performing models for evaluating this dataset, with XGBoost being the strongest contender for deployment for this Alzheimer’s dataset.

Bibliography

Alzheimer’s Disease Fact Sheet | National Institute on Aging, www.nia.nih.gov/health/alzheimers-and-dementia/alzheimers-disease-fact-sheet.

Chubb, J., Cowling, P. & Reed, D. Speeding up to keep up: exploring the use of AI in the research process. AI & Soc 37, 1439–1457 (2022). https://doi.org/10.1007/s00146-021-01259-0

Hotz, Nick. “What Is CRISP DM?” Data Science PM, 9 Dec. 2024, www.datascience-pm.com/crisp-dm-2/.

Kharoua, El Rabie. (2024, June). 🧠 Alzheimer's Disease Dataset 🧠. Retrieved from https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset/data.

Kolte, Pranjlee, et al. “Early Alzheimer’s Detection Using Random Forest Algorithm.” 2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT), 25 May 2023, pp. 1–5, https://doi.org/10.1109/iconscept57958.2023.10170234.

Kriegeskorte, N., Douglas, P.K. Cognitive computational neuroscience. Nat Neurosci 21, 1148–1160 (2018). https://doi.org/10.1038/s41593-018-0210-5

Lazarova, Sophia, et al. “Detection of alzheimer’s disease using logistic regression and clock drawing errors.” *Brain Sciences*, vol. 13, no. 8, 29 July 2023, p. 1139, https://doi.org/10.3390/brainsci13081139.

Shojaie, Mehdi, et al. “A transfer learning approach based on gradient boosting machine for diagnosis of alzheimer’s disease.” Frontiers in Aging Neuroscience, vol. 14, 5 Oct. 2022, https://doi.org/10.3389/fnagi.2022.966883.

Vichianin, Yudthaphon, et al. “Accuracy of support-vector machines for diagnosis of alzheimer’s disease, using volume of brain obtained by Structural MRI at Siriraj Hospital.” *Frontiers in Neurology,* vol. 12, 10 May 2021, https://doi.org/10.3389/fneur.2021.640696.

Watanabe, Hirohisa, et al. “Characteristics of neural network changes in normal aging and early dementia.” Frontiers in Aging Neuroscience, vol. 13, 22 Nov. 2021, https://doi.org/10.3389/fnagi.2021.747359.

Wei, Wei, et al. “The application of naive Bayes model averaging to predict alzheimer’s disease from genome-wide data.” *Journal of the American Medical Informatics Association*, vol. 18, no. 4, July 2011, pp. 370–375, https://doi.org/10.1136/amiajnl-2011-000101.

“What Is Artificial Intelligence (AI)? | Google Cloud.” Google, Google, cloud.google.com/learn/what-is-artificial-intelligence.

Yi, Fuliang, et al. “XGBoost-Shap-based interpretable diagnostic framework for alzheimer’s disease.” *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, 25 July 2023, https://doi.org/10.1186/s12911-023-02238-9.