2024-2025 Fermat Institute Education Proposal

Xuanqi Wei[1, 2], Nick Pestov[2], Rustam Atahodja[1], Cristian Antonuccio[3], and Khalid Edris[1]

[1] *Art & Science, University of Toronto*
[2] *Computer Science, University of Toronto*
[3] *Computer Science, University of Bologna*

June 2024

Contents

1 About us

At Fermat Math Institute, our values and educational principles are the cornerstone of our identity and the driving force behind our mission. They guide our actions, shape our culture, and influence the experiences we create for our students, educators, and partners. These values and principles reflect our commitment to fostering personal, social, and academic growth among our learners. Recognizing the diverse challenges that students face, we aim to create an environment that supports confidence building, engages and motivates learners, promotes progress and achievement, inspires creativity, and unleashes the full potential of every individual. These guiding beliefs are not just ideals we strive towards but principles we live by every day to ensure we make a positive impact in the field of education and in the communities we serve.

1.1 Our Values

Our values are the key beliefs and principles that guide the actions of our team and the experiences we offer to our students and partners. They define the way we work and strive to create a positive impact in the field of education.

Integrity

- We uphold the highest standards of integrity and transparency in all our activities.
- We are committed to fostering an environment of trust and respect among our team, students, and partners.
- We believe in making socially responsible decisions that benefit our community and the broader society.

Inclusivity

- We celebrate diversity and inclusion in our outlook, our people, and our programs.
- We are dedicated to providing equal access to educational opportunities for all learners, regardless of background or circumstance.
- We strive to create a welcoming and supportive environment where everyone can thrive.

Relevance

- We engage closely with educators, students, and industry professionals to ensure our programs are relevant and practical.
- We understand the challenges of teaching and learning, and we aim to provide real-world solutions and support.
- We are committed to continuous improvement based on feedback and real-life experiences.

Partnership

- We believe in the power of collaboration and work alongside our partners to achieve shared goals.
- We foster strong relationships locally, nationally, and internationally to enhance the educational experiences we offer.
- We encourage teamwork and the sharing of knowledge within our community.

Innovation

- We are dedicated to staying at the forefront of educational practices and technology.
- We are passionate about exploring new and effective ways of teaching and learning.
- We aim to prepare our students for future challenges by equipping them with the skills and knowledge they need to succeed.

1.2 Our Education Principles

Our educational principles underpin our approach to fostering the personal, social, and academic abilities of our learners.
Our starting point is to recognize that many learners experience barriers to learning.

Building Confidence

- Our programs support students and educators in developing the confidence, self-esteem, and resilience to enjoy and persist in learning.
- We aim to help learners believe in themselves and the value of their education.

Engaging and Motivating

- We are dedicated to providing active, relevant, and accessible learning experiences.
- We strive to develop curious, engaged, and motivated learners.

Promoting Progress and Achievement

- We motivate and engage learners by making their progress visible and celebrating their genuine achievements frequently.

Inspiring Creativity

- We encourage students and educators to be creative in their learning and ambitious in their roles at home, at work, and in their communities.

Releasing Potential

- At our best, we equip students and educators to fully utilize their talents and abilities as a positive force in the world.

2 Computer Science Projects

2.1 Project 1: Using Machine Learning and Satellite Imagery to Predict Real Estate Growth. This Project Won Multiple Awards at Georgia Tech's Hacklytics 2024.

2.1.1 Difficulty 8/10

2.1.2 Keywords

Machine Learning, API, Python, Winning Project, Computer Vision, Satellite Imagery, Finance, Convolutional Neural Network

2.1.3 Project Description

This project was built by one of our members in Georgia Tech's prestigious Hacklytics 2024 Data-Science Hackathon. It uses satellite imagery information in combination with datasets for quality of living to feed a Machine-Learning model to predict future American house prices. This project was **one of four** to win multiple awards out of over 1200 participants, including **Third Best Financial-Technology Project** and **Third in Best Use of Intel Developer Cloud**.

2.1.4 Background Objectives

This project aims to create a real-estate prediction algorithm using ResNet-50's Convolutional Neural Network to identify changes in satellite imagery of local regions via Google Earth paired with dataset information on statistics like crime rate, cost, etc. to accurately predict future house price costs in any American city/town. This project will be built using the programming language Python. The goal is to accurately predict future pricing in any arbitrary region of the United States.

2.1.5 Project Schedule

- **Day 1: (1 hour)**
  - Project kickoff meeting
    - Outline project goals and deliverables.
    - Project Overview.
    - Set up project management tools.
- **Day 2: (2 hours)**
  - Setting Up Satellite Imagery
    - Connect to the API.
    - Create the code to pinpoint the coordinates of a region.
    - Create the code to collect images from the satellite to feed them into ResNet-50 algorithm.

- **Day 3: (2 hours)**
  - Collecting Dataset Information
    - Set up datasets that contain information on the quality of life of regions based on coordinates and link them to the project.
- **Day 4: (1 hour)**
  - Create an Algorithm to Predict Future Trends from Dataset information
    - Create a machine-learning algorithm to predict future trends from dataset information
- **Day 5: (2 hours)**
  - Combine ResNet-50 Satelite Program with Dataset Machine-Learning Al to make Predictions
    - Create a larger algorithm that takes input from the two previous algorithms to create a final output
- **Optional**
  - Quick Sessions on how Each Algorithm Works and How to Code Can Be Added
    - Learn how ResNet-50 works
    - Understand Basics of Utilizing Libraries in Python
    - Machine Learning Algorithm broken down

2.1.6 Achievements and Benefits

Completing the "Using Machine Learning and Satellite Imagery to Predict Real Estate Growth" project offers numerous significant achievements and benefits that can greatly enhance a client's professional profile and skill set. It has also gained successful public recognition from esteemed institutions such as the Georgia Institute of Technology, one of the top three technology schools in the United States, and renowned companies like Intel.

Advanced Technical Skills

- **Computer Vision and Machine Learning:** Clients will gain hands-on experience with advanced concepts in computer vision and machine learning, specifically in implementing a convolutional neural network for detecting changes in landscape.
- **Python and TensorFlow Understanding:** By working with the most common Programming Language used for the Artificial Intelligence space coupled with the Tensorflow library used for these applications, clients will have important experience under their portfolio showcasing highly sought-after skills.
- **Libraries and API:** Clients will develop practical knowledge in working with Libraries and API's which are used by all advanced programmers and significantly aid in understanding overall software development.
- **Unique and a Winning Idea:** Due to the uniqueness of this project (no one else does it), it will stand out in anyone's portfolio, whether it be for financial or software positions. It is also very applicable and utilizes many sought-after skills, ensuring a strong, yet very unique project, as demonstrated by the awards it won.

2.2 Project 2: Development of a Face Recognition System Using OpenCV's Local Binary Patterns Histogram (LBPH) Algorithm and K-Fold Cross-Validation on Raspberry Pi Platform

2.2.1 Difficulty: 8/10

2.2.2 Keywords

Machine Learning, Face Recognition, Local Binary Patterns Histogram (LBPH), OpenCV, K-Fold Cross-Validation, Computer Vision, Embedded Systems, Internet of Things (IoT), Image Preprocessing, Model Optimization, Computer Hardware

2.2.3 Project Description

Face recognition technology is a critical area of research in computer vision and artificial intelligence. It involves analyzing and identifying facial features from images to verify and recognize individuals. With the advancement of embedded systems and the Internet of Things (IoT), developing low-cost, efficient face recognition systems on platforms like Raspberry Pi has gained significant interest.
REQUIRED: Raspberry Pi Mini-Computer

2.2.4 Background Objectives

This project aims to develop a robust and efficient face recognition system using the Local Binary Patterns Histogram (LBPH) algorithm available in the OpenCV library. The system will be implemented on the Raspberry Pi platform, leveraging K-Fold Cross-Validation to evaluate and optimize the model's performance. The goal is to ensure high accuracy and reliability in various environmental conditions.

2.2.5 Project Schedule

Week 1: Data Collection and Initial Setup (10 hours)

- **Day 1: (2 hours)**
  - Project kickoff meeting
    - Outline project goals and deliverables.
    - Assign roles and responsibilities.
    - Set up project management tools.
- **Day 2: (3 hours)**
  - Data collection and organization
    - Collect facial images from various sources (online databases, personal collection).
    - Ensure diversity in lighting, angles, and backgrounds.
    - Organize the collected data into training and testing sets.
- **Day 3: (2 hours)**
  - Hardware setup
    - Set up Raspberry Pi and install necessary libraries (OpenCV, Python).
    - Test the camera module for capturing images.
- **Day 4: (3 hours)**
  - Initial preprocessing of images
    - Convert images to grayscale.
    - Normalize pixel values.
    - Apply histogram equalization.

Week 2: Model Training and Validation (10 hours)

- **Day 5: (2 hours)**
  - Feature extraction using LBPH
    - Implement LBPH algorithm to extract features from preprocessed images.
- **Day 6: (3 hours)**
  - K-Fold Cross-Validation setup
    - Split the dataset into K subsets.
    - Implement K-Fold Cross-Validation to evaluate the model.
- **Day 7: (2 hours)**
  - Parameter tuning
    - Experiment with different LBPH parameters (radius, neighbors).

    - Select the best parameters based on validation results.

- **Day 8: (3 hours)**
  - Model training and optimization
    - Train the model using the optimal parameters.
    - Validate the model to ensure it meets performance criteria.

Week 3: System Integration and Testing (10 hours)

- **Day 9: (2 hours)**
  - Code integration on Raspberry Pi
    - Integrate the trained model into the Raspberry Pi system.
    - Ensure the system can capture live video feed.
- **Day 10: (3 hours)**
  - Real-time face recognition implementation
    - Develop code for real-time face recognition using the camera module.
    - Test the system for real-time performance.
- **Day 11: (2 hours)**
  - User interface design
    - Create a simple and intuitive user interface for displaying recognition results.
    - Implement feedback and system status indicators.
- **Day 12: (3 hours)**
  - Performance evaluation and optimization
    - Evaluate the system's performance in terms of accuracy, response time, and robustness.
    - Optimize the system based on testing results.

Week 4: Final Testing, Documentation, and Presentation (10 hours)

- **Day 13: (2 hours)**
  - Comprehensive testing under various conditions
    - Test the system in different lighting and background scenarios.
    - Record the results and identify any remaining issues.
- **Day 14: (3 hours)**
  - Final optimization and bug fixing
    - Address any issues identified during comprehensive testing.
    - Optimize the code for better performance and reliability.

- **Day 15: (2 hours)**
  - Documentation preparation
    - Document the project, including system design, implementation details, and test results.
    - Prepare user manuals and technical documentation.
- **Day 16: (3 hours)**
  - Final presentation preparation
    - Prepare a presentation summarizing the project goals, methodology, and outcomes.
    - Create slides and practice the presentation.
- **Day 17: (1 hour)**
  - Project review and wrap-up
    - Review the entire project with the team.
    - Discuss lessons learned and potential future improvements.

2.2.6 Achievements and Benefits

Completing the "Development of a Face Recognition System Using OpenCV's Local Binary Patterns Histogram (LBPH) Algorithm and K-Fold Cross-Validation on Raspberry Pi Platform" project offers several significant achievements and benefits that can greatly enhance a client's professional profile and skill set.

Advanced Technical Skills

- **Computer Vision and Machine Learning:** Clients will gain hands-on experience with advanced concepts in computer vision and machine learning, specifically in implementing and optimizing the LBPH algorithm for face recognition.
- **OpenCV Proficiency:** By working extensively with OpenCV, clients will become proficient in one of the most widely used libraries for computer vision tasks, a valuable skill in many tech-related fields.
- **Raspberry Pi Expertise:** Clients will develop practical knowledge in working with Raspberry Pi, a popular platform for IoT and embedded systems projects, enhancing their ability to build and deploy hardware-software integrated systems.

Project Management and Development

- **End-to-End Project Experience:** Clients will experience the full lifecycle of a project, from initial planning and data collection to implementation, testing, and optimization, providing a comprehensive understanding of project management.
- **Problem-Solving Skills:** The project will challenge clients to solve complex problems, such as optimizing the model's performance and ensuring real-time functionality, thereby sharpening their analytical and troubleshooting skills.

Professional Growth

- **Resume Enhancement:** Successfully completing this project and gaining expertise in cutting-edge technologies will significantly boost a client's resume, making them more attractive to potential employers in fields like AI, computer vision, and IoT.
- **Portfolio Development:** The project can serve as a standout piece in a client's portfolio, demonstrating their ability to handle advanced technical challenges and deliver functional solutions.

In summary, clients who complete this project will emerge with a robust set of skills and experiences that not only enhance their technical capabilities but also significantly bolster their professional profiles, leading to greater career opportunities and advancement.

2.3 Project 3: Dog Breed Identification using Convolutional Neural Networks (CNNs)

2.3.1 Difficulty: 7/10

2.3.2 Key Words

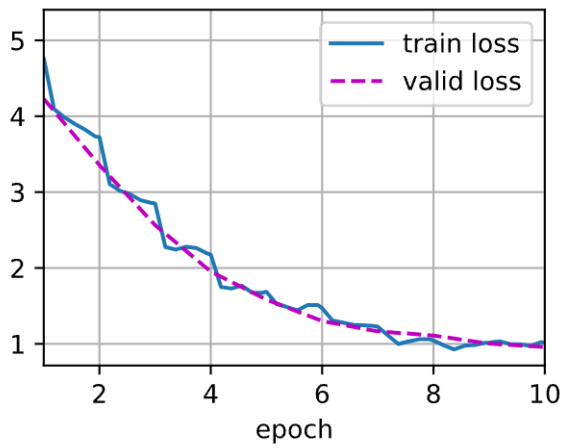
Machine Learning, Convolutional Neural Networks (CNNs), Image Classification, Dog Breed Identification, Deep Learning, Data Preprocessing, Kaggle Competition

2.3.3 Project Description

Dog breed identification is a challenging task in computer vision due to the large number of breeds and the subtle differences between them. This project leverages Convolutional Neural Networks (CNNs) to classify dog breeds from images. Using a deep learning approach, the project aims to accurately identify the breed of a dog in a given image by training a CNN model on a large dataset of labeled dog images.

2.3.4 Background Information

The primary objective of this project is to develop a robust image classification model for identifying dog breeds by applying CNNs to the Kaggle Dog Breed Identification dataset. CNNs are well-suited for image classification tasks due to their ability to automatically learn and extract features from raw pixel data. This project will involve data preprocessing, model training, validation, and fine-tuning to achieve high accuracy in breed identification. This approach aims to provide insights into the effectiveness of deep learning for fine-grained image classification tasks.



(a) Convolutional Neural Network

(b) Dog Breeds

Figure 1: Dog Breed Identification using CNNs

2.3.5 Project Schedule

Week 1: Data Collection and Initial Setup (10 hours)

- **Day 1: (2 hours)**
  - Project kickoff meeting
    - Outline project goals and deliverables.
    - Assign roles and responsibilities.
    - Set up project management tools.
- **Day 2: (3 hours)**
  - Data acquisition and organization
    - Download and preprocess the Kaggle Dog Breed Identification dataset.
    - Ensure data quality and completeness.
    - Organize the data into suitable formats for analysis.
- **Day 3: (2 hours)**
  - Environment setup
    - Set up computational environment and necessary libraries (e.g., TensorFlow, Keras, OpenCV).
    - Verify compatibility with data analysis tools.
- **Day 4: (3 hours)**
  - Initial exploratory data analysis
    - Perform exploratory analysis on the data set.
    - Visualize initial findings to understand data distribution and challenges.

Week 2: Model Implementation and Validation (10 hours)

- **Day 5: (2 hours)**
  - CNN model setup
    - Implement a basic CNN architecture for initial modeling of dog breed classification.
- **Day 6: (3 hours)**
  - Model training
    - Train the CNN model on the training dataset.
    - Monitor training process and adjust hyperparameters as necessary.
- **Day 7: (2 hours)**
  - Model validation

    - Validate the trained model using the validation dataset.
    - Evaluate the model's performance using metrics such as accuracy and F1-score.

- **Day 8: (3 hours)**
  - Model fine-tuning
    - Fine-tune the CNN model for better accuracy.
    - Apply techniques like data augmentation and dropout for improved generalization.

Week 3: Analysis and Interpretation (10 hours)

- **Day 9: (2 hours)**
  - Detailed analysis of model performance
    - Analyze the model's predictions and identify areas for improvement.
- **Day 10: (3 hours)**
  - Cross-validation with external datasets
    - Compare the model's performance with other known benchmarks.
    - Validate findings with additional data if available.
- **Day 11: (2 hours)**
  - Interpretation of results
    - Interpret the results in the context of image classification and deep learning applications.
- **Day 12: (3 hours)**
  - Preparation of analysis report
    - Document the analysis and findings.
    - Prepare a comprehensive report detailing the results.

Week 4: Final Testing, Documentation, and Presentation (10 hours)

- **Day 13: (2 hours)**
  - Comprehensive testing of the model
    - Test the model under different scenarios and datasets.
- **Day 14: (3 hours)**
  - Final model optimization
    - Optimize the model for better accuracy and efficiency.
- **Day 15: (2 hours)**

  - Documentation preparation
    * Document the project, including model details, implementation, and analysis results.
    * Prepare user manuals and technical documentation.

- **Day 16: (3 hours)**
  - Final presentation preparation
    * Prepare a presentation summarizing the project goals, methodology, and outcomes.
    * Create slides and practice the presentation.

- **Day 17: (1 hour)**
  - Project review and wrap-up
    * Review the entire project with the team.
    * Discuss lessons learned and potential future improvements.

2.3.6 Achievements and Benefits

Completing the "Dog Breed Identification using Convolutional Neural Networks (CNNs)" project offers several significant achievements and benefits that can greatly enhance a client's professional profile and skill set.

Advanced Technical Skills

- **Deep Learning and Neural Networks:** Clients will gain hands-on experience with advanced concepts in deep learning and neural networks, specifically in implementing and optimizing CNNs.
- **Data Analysis Proficiency:** By working extensively with large image datasets, clients will become proficient in data preprocessing, analysis, and visualization, valuable skills in many scientific and technical fields.
- **Computer Vision Expertise:** Clients will develop practical knowledge in image classification and computer vision, enhancing their ability to contribute to projects

3 Data Science & Statistics Projects

3.1 Project 2: Utilizing Gaussian Mixture Models and Markov Chain Monte Carlo (MCMC) Methods for the Identification of Stellar Streams/Dwarf Galaxies in the Dark Energy Spectroscopic Instrument Survey Early Data Release

3.1.1 Difficulty: 9/10

3.1.2 Key Words

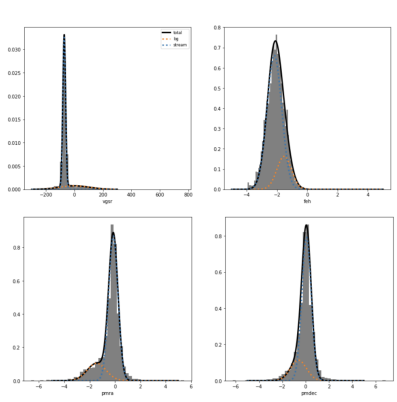
Machine Learning, Gaussian Mixture Models (GMM), Markov Chain Monte Carlo (MCMC), Stellar Streams, Dwarf Galaxies, Statistical Modeling, Astrophysical Data Analysis, Data Preprocessing

3.1.3 Project Description

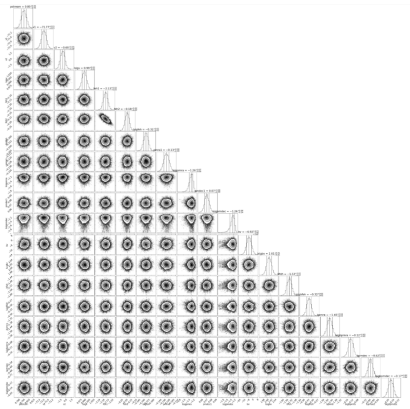
The identification and analysis of stellar streams and dwarf galaxies are crucial for understanding the structure and formation history of the Milky Way. This project leverages Gaussian Mixture Models (GMM) and Markov Chain Monte Carlo (MCMC) methods to identify and characterize these celestial objects within the Dark Energy Spectroscopic Instrument (DESI) Survey's early data release. The integration of these statistical techniques aims to enhance the precision and reliability of detecting stellar streams and dwarf galaxies.

3.1.4 Background Information

The primary objective of this project is to develop a robust framework for identifying stellar streams and dwarf galaxies by applying GMM and MCMC methods to the DESI survey data. GMM will be used to model the distribution of stars, allowing for the distinction between different stellar populations. MCMC methods will be employed to sample from the posterior distribution, refining the parameters of the GMM and improving the accuracy of the identification process. This approach aims to provide insights into the formation and evolution of the Milky Way and its satellite systems.



(a) Gaussian Mixture Model



(b) Posterior Distribution

Figure 2: Identification on Ursa Minor based on DESI

3.1.5 Project Schedule

Week 1: Data Collection and Initial Setup (10 hours)

- **Day 1: (2 hours)**
  - Project kickoff meeting
    - Outline project goals and deliverables.
    - Assign roles and responsibilities.
    - Set up project management tools.
- **Day 2: (3 hours)**
  - Data acquisition and organization
    - Download and preprocess DESI early data release.
    - Ensure data quality and completeness.
    - Organize the data into suitable formats for analysis.
- **Day 3: (2 hours)**
  - Environment setup
    - Set up computational environment and necessary libraries (e.g., NumPy, SciPy, Astropy).
    - Verify compatibility with data analysis tools.
- **Day 4: (3 hours)**
  - Initial exploratory data analysis
    - Perform exploratory analysis on the data set.
    - Visualize initial findings to identify potential stellar streams and dwarf galaxies.

Week 2: Model Implementation and Validation (10 hours)

- **Day 5: (2 hours)**
  - Gaussian Mixture Model setup
    - Implement GMM for initial modeling of star distributions.
- **Day 6: (3 hours)**
  - Parameter estimation using MCMC
    - Implement MCMC methods for parameter estimation.
    - Define prior distributions and likelihood functions.
- **Day 7: (2 hours)**
  - Model validation
    - Validate the GMM and MCMC model using synthetic data.

    * Refine the model based on validation results.

- **Day 8: (3 hours)**
  - Application to DESI data
    * Apply the validated model to the DESI early data release.
    * Identify and characterize stellar streams and dwarf galaxies.

Week 3: Analysis and Interpretation (10 hours)

- **Day 9: (2 hours)**
  - Detailed analysis of identified structures
    * Analyze the properties of identified stellar streams and dwarf galaxies.
- **Day 10: (3 hours)**
  - Cross-validation with external datasets
    * Compare identified structures with known catalogs.
    * Validate findings with external data.
- **Day 11: (2 hours)**
  - Interpretation of results
    * Interpret the results in the context of galactic formation and evolution.
- **Day 12: (3 hours)**
  - Preparation of analysis report
    * Document the analysis and findings.
    * Prepare a comprehensive report detailing the results.

Week 4: Final Testing, Documentation, and Presentation (10 hours)

- **Day 13: (2 hours)**
  - Comprehensive testing of the model
    * Test the model under different scenarios and datasets.
- **Day 14: (3 hours)**
  - Final model optimization
    * Optimize the model for better accuracy and efficiency.
- **Day 15: (2 hours)**
  - Documentation preparation
    * Document the project, including model details, implementation, and analysis results.

    - Prepare user manuals and technical documentation.

- **Day 16: (3 hours)**
  - Final presentation preparation
    - Prepare a presentation summarizing the project goals, methodology, and outcomes.
    - Create slides and practice the presentation.
- **Day 17: (1 hour)**
  - Project review and wrap-up
    - Review the entire project with the team.
    - Discuss lessons learned and potential future improvements.

3.1.6 Achievements and Benefits

Completing the "Utilizing Gaussian Mixture Models and Markov Chain Monte Carlo (MCMC) Methods for the Identification of Stellar Streams/Dwarf Galaxies in the Dark Energy Spectroscopic Instrument Survey Early Data Release" project offers several significant achievements and benefits that can greatly enhance a client's professional profile and skill set.

Advanced Technical Skills

- **Statistical Modeling and Machine Learning:** Clients will gain hands-on experience with advanced concepts in statistical modeling and machine learning, specifically in implementing and optimizing GMM and MCMC methods.
- **Data Analysis Proficiency:** By working extensively with large astronomical datasets, clients will become proficient in data preprocessing, analysis, and visualization, valuable skills in many scientific and technical fields.
- **Astrophysics Expertise:** Clients will develop practical knowledge in astrophysical data analysis, enhancing their ability to contribute to research in galactic dynamics and structure.

Project Management and Development

- **End-to-End Project Experience:** Clients will experience the full lifecycle of a project, from initial planning and data collection to implementation, testing, and optimization, providing a comprehensive understanding of project management.
- **Problem-Solving Skills:** The project will challenge clients to solve complex problems, such as optimizing the model's performance and ensuring accurate identification of stellar structures, thereby sharpening their analytical and troubleshooting skills.

Professional Growth

- **Resume Enhancement:** Successfully completing this project and gaining expertise in cutting-edge statistical and computational techniques will significantly boost a client's resume, making them more attractive to potential employers in fields like AI, data science, and astrophysics.
- **Portfolio Development:** The project can serve as a standout piece in a client's portfolio, demonstrating their ability to handle advanced technical challenges and deliver functional solutions.

In summary, clients who complete this project will emerge with a robust set of skills and experiences that not only enhance their technical capabilities but also significantly bolster their professional profiles, leading to greater career opportunities and advancement.

3.2 Project 3: Unbiased Metallicity (Iron-Abundances) Distribution Analysis for Stellar Streams/Dwarf Galaxies with DESI Data

3.2.1 Difficulty: 10/10

3.2.2 Key Words

Machine Learning, Gaussian Mixture Models (GMM), Metallicity Analysis, Stellar Streams, Distribution Analysis, Statistical Modeling, Astrophysical Data Analysis, Galactic Archaeology

3.2.3 Project Description

This project focuses on analyzing the metallicity distribution of stellar streams and dwarf galaxies using data from the Dark Energy Spectroscopic Instrument (DESI). By leveraging Gaussian Mixture Models (GMM), the project aims to provide an unbiased analysis of the metallicity distribution. The approach involves comparing this model to one that utilizes iron abundances, replicating methods previously applied in studies such as Draco et al.

3.2.4 Background Information

The primary goal of this project is to develop an unbiased framework for analyzing the metallicity distribution of stellar streams and dwarf galaxies within the DESI data. GMM will be employed to model the distribution of stars, facilitating the distinction between different stellar populations without relying on iron abundances. This methodology will be compared with models that include iron abundances to evaluate the effectiveness and bias of each approach. The analysis will be based on the methodology used in the study of the dwarf galaxy.

3.2.5 Project Schedule

Week 1: Data Collection and Initial Setup (10 hours)

- **Day 1: (2 hours)**
  - Project kickoff meeting
    - Outline project goals and deliverables.
    - Assign roles and responsibilities.
    - Set up project management tools.
- **Day 2: (3 hours)**
  - Data acquisition and organization
    - Download and preprocess DESI data.
    - Ensure data quality and completeness.
    - Organize the data into suitable formats for analysis.
- **Day 3: (2 hours)**
  - Environment setup
    - Set up computational environment and necessary libraries (e.g., NumPy, SciPy, Astropy).
    - Verify compatibility with data analysis tools.
- **Day 4: (3 hours)**
  - Initial exploratory data analysis
    - Perform exploratory analysis on the data set.
    - Visualize initial findings to identify potential stellar streams and dwarf galaxies.

Week 2: Model Implementation and Validation (10 hours)

- **Day 5: (2 hours)**
  - Gaussian Mixture Model setup
    - Implement GMM for initial modeling of star distributions.
- **Day 6: (3 hours)**
  - Parameter estimation
    - Estimate parameters for the GMM.
    - Define prior distributions and likelihood functions.
- **Day 7: (2 hours)**
  - Model validation
    - Validate the GMM model using synthetic data.

    * Refine the model based on validation results.

- **Day 8: (3 hours)**
  - Application to DESI data
    * Apply the validated model to the DESI data.
    * Identify and characterize stellar streams and dwarf galaxies.

Week 3: Analysis and Interpretation (10 hours)

- **Day 9: (2 hours)**
  - Detailed analysis of identified structures
    * Analyze the properties of identified stellar streams and dwarf galaxies.
- **Day 10: (3 hours)**
  - Comparison with iron-abundance models
    * Compare the results with models that use iron abundances.
    * Evaluate the biases and differences.
- **Day 11: (2 hours)**
  - Interpretation of results
    * Interpret the results in the context of galactic formation and evolution.
- **Day 12: (3 hours)**
  - Preparation of analysis report
    * Document the analysis and findings.
    * Prepare a comprehensive report detailing the results.

Week 4: Final Testing, Documentation, and Presentation (10 hours)

- **Day 13: (2 hours)**
  - Comprehensive testing of the model
    * Test the model under different scenarios and datasets.
- **Day 14: (3 hours)**
  - Final model optimization
    * Optimize the model for better accuracy and efficiency.
- **Day 15: (2 hours)**
  - Documentation preparation
    * Document the project, including model details, implementation, and analysis results.

    - Prepare user manuals and technical documentation.

- **Day 16: (3 hours)**
  - Final presentation preparation
    - Prepare a presentation summarizing the project goals, methodology, and outcomes.
    - Create slides and practice the presentation.
- **Day 17: (1 hour)**
  - Project review and wrap-up
    - Review the entire project with the team.
    - Discuss lessons learned and potential future improvements.

3.2.6 Achievements and Benefits

Completing the "Unbiased Metallicity (Iron-Abundances) Distribution Analysis for Stellar Streams/Dwarf Galaxies with DESI Data" project offers several significant achievements and benefits that can greatly enhance a client's professional profile and skill set.

Advanced Technical Skills

- **Statistical Modeling:** Clients will gain hands-on experience with advanced concepts in statistical modeling, specifically in implementing and optimizing GMM.
- **Data Analysis Proficiency:** By working extensively with large astronomical datasets, clients will become proficient in data preprocessing, analysis, and visualization, valuable skills in many scientific and technical fields.
- **Astrophysics Expertise:** Clients will develop practical knowledge in astrophysical data analysis, enhancing their ability to contribute to research in galactic dynamics and structure.

Project Management and Development

- **End-to-End Project Experience:** Clients will experience the full lifecycle of a project, from initial planning and data collection to implementation, testing, and optimization, providing a comprehensive understanding of project management.
- **Problem-Solving Skills:** The project will challenge clients to solve complex problems, such as optimizing the model's performance and ensuring accurate identification of stellar structures, thereby sharpening their analytical and troubleshooting skills.

Professional Growth

- **Resume Enhancement:** Successfully completing this project and gaining expertise in cutting-edge statistical and computational techniques will significantly boost a client's resume, making them more attractive to potential employers in fields like AI, data science, and astrophysics.
- **Portfolio Development:** The project can serve as a standout piece in a client's portfolio, demonstrating their ability to handle advanced technical challenges and deliver functional solutions.

In summary, clients who complete this project will emerge with a robust set of skills and experiences that not only enhance their technical capabilities but also significantly bolster their professional profiles, leading to greater career opportunities and advancement.

3.3 Project 4: Statistical Exploration of Galactic Distance-Luminosity Correlations Utilizing Comprehensive NASA-Sloan Atlas Data and Hypothesis-Driven Simulations

3.3.1 Difficulty: 6/10

3.3.2 Key Words

Statistical Analysis, Data Cleaning, Data Preprocessing, Python, R, Hypothesis Testing, Data Visualization

3.3.3 Project Description

This project involves a comprehensive statistical analysis of the correlation between galaxy distance and luminosity. By examining data from over 600,000 galaxies in the NASA-Sloan Atlas, the investigation specifically focuses on the variance in luminosity between galaxies closest to and farthest from Earth. The project employs Python and R for data preprocessing and analysis, utilizing advanced techniques to clean and prepare the data, and ggplot in R for detailed data visualization. The statistical analysis is performed using hypothesis testing on 10,000 simulations to identify significant differences in galaxy luminosity.

3.3.4 Background Information

The objective of this project is to investigate the relationship between the distance of galaxies from Earth and their luminosity. By analyzing a vast dataset from the NASA-Sloan Atlas, which includes information on over 600,000 galaxies, the project aims to uncover significant patterns and correlations. Data cleaning and preparation are critical steps in this process, handled using Python and R to remove NA values and create suitable subsets for analysis. Detailed data visualizations are created using ggplot in R to highlight differences in galaxy luminosity. The statistical analysis involves hypothesis testing on 10,000 simulations, revealing a notable 9 magnitude difference in luminosity between distant and nearby galaxies.

3.3.5 Project Schedule

Week 1: Data Collection and Initial Setup (10 hours)

- **Day 1: (2 hours)**
  - Project kickoff meeting
    - Outline project goals and deliverables.
    - Assign roles and responsibilities.
    - Set up project management tools.
- **Day 2: (3 hours)**
  - Data acquisition and organization
    - Download and preprocess NASA-Sloan Atlas data.
    - Ensure data quality and completeness.
    - Organize the data into suitable formats for analysis.
- **Day 3: (2 hours)**
  - Environment setup
    - Set up computational environment and necessary libraries (e.g., NumPy, pandas, ggplot2).
    - Verify compatibility with data analysis tools.
- **Day 4: (3 hours)**
  - Initial exploratory data analysis
    - Perform exploratory analysis on the dataset.
    - Visualize initial findings to identify potential correlations.

Week 2: Data Cleaning and Preparation (10 hours)

- **Day 5: (2 hours)**
  - Data cleaning in Python
    - Remove NA values and outliers.
    - Create subsets for detailed analysis.
- **Day 6: (3 hours)**
  - Data preparation in R
    - Prepare data for visualization.
    - Ensure datasets are compatible with ggplot2.
- **Day 7: (2 hours)**
  - Initial visualizations

    - Create initial plots using ggplot2.
    - Highlight key differences in galaxy luminosity.

- **Day 8: (3 hours)**
  - Refining visualizations
    - Refine and improve visualizations for clarity.
    - Ensure plots effectively communicate findings.

Week 3: Statistical Analysis (10 hours)

- **Day 9: (2 hours)**
  - Hypothesis formulation
    - Formulate hypotheses for luminosity differences.
- **Day 10: (3 hours)**
  - Simulation setup
    - Set up simulations for hypothesis testing.
    - Ensure reproducibility of results.
- **Day 11: (2 hours)**
  - Executing simulations
    - Run 10,000 simulations.
    - Analyze simulation outputs.
- **Day 12: (3 hours)**
  - Analysis and interpretation
    - Interpret the results of hypothesis testing.
    - Confirm the 9 magnitude difference in luminosity.

Week 4: Final Testing, Documentation, and Presentation (10 hours)

- **Day 13: (2 hours)**
  - Comprehensive testing of the analysis
    - Test the analysis under different scenarios.
- **Day 14: (3 hours)**
  - Final optimization
    - Optimize analysis methods for better accuracy and efficiency.
- **Day 15: (2 hours)**
  - Documentation preparation

    * Document the project, including methods, implementation, and results.
    * Prepare user manuals and technical documentation.

- **Day 16: (3 hours)**
  - Final presentation preparation
    * Prepare a presentation summarizing the project goals, methodology, and outcomes.
    * Create slides and practice the presentation.
- **Day 17: (1 hour)**
  - Project review and wrap-up
    * Review the entire project with the team.
    * Discuss lessons learned and potential future improvements.

3.3.6 Achievements and Benefits

Completing the "Statistical Analysis of Galaxy Distance and Luminosity Correlations" project offers several significant achievements and benefits that can greatly enhance a client's professional profile and skill set.

Advanced Technical Skills

- **Statistical Analysis:** Clients will gain hands-on experience with advanced concepts in statistical analysis, specifically in implementing hypothesis testing and simulations.
- **Data Analysis Proficiency:** By working extensively with large astronomical datasets, clients will become proficient in data preprocessing, analysis, and visualization, valuable skills in many scientific and technical fields.
- **Astrophysics Expertise:** Clients will develop practical knowledge in astrophysical data analysis, enhancing their ability to contribute to research in galactic dynamics and structure.

Project Management and Development

- **End-to-End Project Experience:** Clients will experience the full lifecycle of a project, from initial planning and data collection to implementation, testing, and optimization, providing a comprehensive understanding of project management.
- **Problem-Solving Skills:** The project will challenge clients to solve complex problems, such as optimizing analysis methods and ensuring accurate identification of correlations, thereby sharpening their analytical and troubleshooting skills.

Professional Growth

- **Resume Enhancement:** Successfully completing this project and gaining expertise in cutting-edge statistical and computational techniques will significantly boost a client's resume, making them more attractive to potential employers in fields like data science and astrophysics.
- **Portfolio Development:** The project can serve as a standout piece in a client's portfolio, demonstrating their ability to handle advanced technical challenges and deliver functional solutions.

In summary, clients who complete this project will emerge with a robust set of skills and experiences that not only enhance their technical capabilities but also significantly bolster their professional profiles, leading to greater career opportunities and advancement.

3.4 Project 5: Utilizing Gaussian Mixture Models and Markov Chain Monte Carlo (MCMC) Methods for Customer Segmentation in Retail

3.4.1 Difficulty: 7/10

3.4.2 Key Words

Machine Learning, Gaussian Mixture Models (GMM), Markov Chain Monte Carlo (MCMC), Customer Segmentation, Retail Data Analysis, Data Preprocessing, Statistical Modeling, E-commerce

3.4.3 Project Description

Customer segmentation is crucial for understanding the diverse needs and behaviors of customers in the retail sector. This project leverages Gaussian Mixture Models (GMM) and Markov Chain Monte Carlo (MCMC) methods to identify and characterize distinct customer segments within an e-commerce dataset. The integration of these statistical techniques aims to enhance the precision and reliability of detecting customer segments, enabling targeted marketing and personalized customer experiences.

3.4.4 Background Information

The primary objective of this project is to develop a robust framework for identifying customer segments by applying GMM and MCMC methods to retail transaction data. GMM will be used to model the distribution of customers based on their purchasing behavior, allowing for the distinction between different customer segments. MCMC methods will be employed to sample from the posterior distribution, refining the parameters of the GMM and improving the accuracy of the segmentation process. This approach aims to provide insights into customer behavior, aiding in strategic decision-making for marketing and customer relationship management.

(a) Gaussian Mixture Model (b) Posterior Distribution

Figure 3: Customer Segmentation Analysis

3.4.5 Project Schedule

Week 1: Data Collection and Initial Setup (10 hours)

- **Day 1: (2 hours)**
  - Project kickoff meeting
    - Outline project goals and deliverables.
    - Assign roles and responsibilities.
    - Set up project management tools.
- **Day 2: (3 hours)**
  - Data acquisition and organization
    - Download and preprocess the retail dataset.
    - Ensure data quality and completeness.
    - Organize the data into suitable formats for analysis.
- **Day 3: (2 hours)**
  - Environment setup
    - Set up computational environment and necessary libraries (e.g., NumPy, SciPy, Pandas).
    - Verify compatibility with data analysis tools.
- **Day 4: (3 hours)**
  - Initial exploratory data analysis
    - Perform exploratory analysis on the data set.
    - Visualize initial findings to identify potential customer segments.

Week 2: Model Implementation and Validation (10 hours)

- **Day 5: (2 hours)**
  - Gaussian Mixture Model setup
    - Implement GMM for initial modeling of customer distributions.
- **Day 6: (3 hours)**
  - Parameter estimation using MCMC
    - Implement MCMC methods for parameter estimation.
    - Define prior distributions and likelihood functions.
- **Day 7: (2 hours)**
  - Model validation
    - Validate the GMM and MCMC model using synthetic data.

    - Refine the model based on validation results.

- **Day 8: (3 hours)**
  - Application to retail data
    - Apply the validated model to the retail dataset.
    - Identify and characterize customer segments.

Week 3: Analysis and Interpretation (10 hours)

- **Day 9: (2 hours)**
  - Detailed analysis of identified segments
    - Analyze the properties of identified customer segments.
- **Day 10: (3 hours)**
  - Cross-validation with external datasets
    - Compare identified segments with known benchmarks.
    - Validate findings with external data.
- **Day 11: (2 hours)**
  - Interpretation of results
    - Interpret the results in the context of customer behavior and business strategy.
- **Day 12: (3 hours)**
  - Preparation of analysis report
    - Document the analysis and findings.
    - Prepare a comprehensive report detailing the results.

Week 4: Final Testing, Documentation, and Presentation (10 hours)

- **Day 13: (2 hours)**
  - Comprehensive testing of the model
    - Test the model under different scenarios and datasets.
- **Day 14: (3 hours)**
  - Final model optimization
    - Optimize the model for better accuracy and efficiency.
- **Day 15: (2 hours)**
  - Documentation preparation
    - Document the project, including model details, implementation, and analysis results.

    - Prepare user manuals and technical documentation.

- **Day 16: (3 hours)**
  - Final presentation preparation
    - Prepare a presentation summarizing the project goals, methodology, and outcomes.
    - Create slides and practice the presentation.
- **Day 17: (1 hour)**
  - Project review and wrap-up
    - Review the entire project with the team.
    - Discuss lessons learned and potential future improvements.

3.4.6 Achievements and Benefits

Completing the "Customer Segmentation in Retail using Gaussian Mixture Models and Markov Chain Monte Carlo (MCMC) Methods" project offers several significant achievements and benefits that can greatly enhance a client's professional profile and skill set.
Advanced Technical Skills

- **Statistical Modeling and Machine Learning:** Clients will gain hands-on experience with advanced concepts in statistical modeling and machine learning, specifically in implementing and optimizing GMM and MCMC methods.
- **Data Analysis Proficiency:** By working extensively with large retail datasets, clients will become proficient in data preprocessing, analysis, and visualization, valuable skills in many scientific and technical fields.
- **Business Intelligence Expertise:** Clients will develop practical knowledge in customer segmentation and retail data analysis, enhancing their ability to contribute to strategic business decisions.

4 Mathematics Specialized Curriculum

4.1 Introduction

Our math curriculum is designed for high school students to excel in university entrance exams like the SAT and ACT, and to gain the necessary knowledge for university courses up to the third year.

4.2 About the Creators

The creators of our curriculum designed it to ensure success in the challenging early years of undergraduate university. Based on our personal experience, we observed that a significant portion of students drop out in their first year (around a 10% drop-out rate for general programs at the University of Toronto) due to the challenging transition from high school to university. One of our instructors, who maintained a 97% high school average and a perfect grade in high school calculus, received only 51% in UofT's first-year class MAT137 (Calculus with Proofs) due to the difference in curriculum teaching. Staying in the chosen program can also be difficult because of the new environment, with even more students completing their studies in an alternative program. All these experiences inspired us to build a curriculum that bridges the gap between high school and university, addressing the shortcomings of other programs that focus solely on high school success!

Our curriculum is created by highly successful competitive mathematics students, who ranked among the **top 100 globally** in contests like the **University of Waterloo's** prestigious **Euclid contest**, and young, highly-rated teachers with over **6 years of experience** in various teaching environments, including top Canadian extracurricular math schools and 1-on-1 tutoring. This combination of expertise ensures that important and high-level mathematical concepts are taught in a digestible and easy-to-understand manner, with topics flowing logically from one to the next. Challenging concepts from computer science, economics, and statistics

related to computation are also integrated throughout the curriculum. We believe that any difficult concept can be taught with the right help, and that is exactly what we provide!

4.3 Classes Overview

We recommend starting our program for SAT I in the first year of high school, as this is when we begin teaching the accelerated curriculum. Joining in later years will require additional tutoring sessions to catch up unless the student passes the curriculum assessment. The curriculum is designed to be taught in an easy and logical progression, with each class building on the previous one, eventually covering topics found in third-year university courses. We expect students to dedicate 2 hours to class time weekly and 1 hour for homework to succeed in our classes.

4.3.1 SAT Series Classes (14 Weeks Each)

We offer three classes—SAT I, II, and III—designed to prepare students for university exams such as the SAT and ACT, applicable to American, British, and Canadian universities. Our aim is to have students ready for these exams by the middle of their second year of high school, providing plenty of time to retake the exams if needed and achieve top scores.

Additionally, these three classes emphasize building a strong foundation in mathematics essential for advanced studies, with a focus on Algebra, Geometry, Basic Probability and Statistics, and Trigonometry. By the end of these courses, students will be roughly 1.5 years ahead of their peers and equipped with the knowledge needed for success in high school non-Calculus classes. These classes also serve as prerequisites for our University (UNI I, II, III, IV, and V) courses.

4.3.2 UNI Series Classes (14 Weeks Each)

We offer five university-level classes—UNI I through UNI V—designed to ensure success in university mathematics courses. These classes also prepare students for the mathematical components of computer science, statistics, and economics courses. The curriculum covers topics that are integral to second-year university courses, including Linear Algebra, Calculus with Proofs, and Multivariable Calculus.

This curriculum is designed specifically for success in major topics for these University of Toronto mathematics courses:

MAT133 - Calculus and Linear Algebra for Commerce
MAT135 - Calculus I
MAT136 - Calculus II
MAT137 - Calculus with Proofs
MAT138 - Introduction to Proofs
MAT223 - Linear Algebra I

MAT224 - Linear Algebra II
MAT235 - Multivariable Calculus
MAT237 - Multivariable Calculus with Proofs

Concepts covered in our curriculum are also found in the following courses:

STA130 - An Introduction to Statistical Reasoning and Data Science
CSC165 - Mathematical Expression and Reasoning for Computer Science
CSC236 - Introduction to the Theory of Computation
STA237 - Probability, Statistics and Data Analysis I
STA257 - Probability and Statistics I

Although these classes are tailored for University of Toronto courses, the university is considered a benchmark among top global institutions, and other universities offer similar courses with comparable content. For instance, the University of Waterloo's MATH 135 (Algebra for Honours Mathematics) covers the same material as the University of Toronto's MAT135. As a result, this preparation is easily transferable to similar courses at other institutions.

4.4 Classes Offered By Us - In Order of Difficulty

4.4.1 First Year High School / Grade 9 / Freshman (SAT I and II)

SAT I (14 Weeks)
This class is the first in our three-part SAT series, focusing on building mathematical fundamentals and preparing for university entrance exams. By the end of the series, we expect students to be approximately a year ahead of their peers.
Some of the topics covered include:

- Number Sets including Complex Numbers
- Simplifying Algebraic Expressions
- Inequalities
- Exponential and Polynomial Laws
- Similar Triangles and Triangle Angle Theorems

This class is the prerequisite for SAT II and is recommended to be taken in September of the first year of High School.

SAT II (14 Weeks)
This class is the second of three SAT series classes and is a continuation of SAT I. Some topics that are covered:

- Working with Circles
- Trigonometric Identities

- Standard Deviation in Statistics
- Working with Shapes
- Permutations and Combinations
- SAT I and II review

This class is recommended to be taken in January of the first year of High School, following the immediate completion of SAT I.

4.4.2 Second Year High School / Grade 10 / Sophomore (SAT III and UNI I)

SAT III (14 Weeks)
This class is the final SAT series class, that focuses on building the mathematical fundamentals and university entrance exam preparation. By the end of the two classes, we expect students to be about 1.5 years ahead of their peers. Some topics that are covered:

- Quadratic Functions
- Polynomial and Radical Functions
- Modulus Theorems
- Compound Angle Formulas in Trigonometry
- Working with Logarithms

This class is the prerequisite to UNI I and is recommended to be taken in September of the second year of High School.

UNI I (14 Weeks)
This is the first class of the University (UNI) series and begins the university-level exploration into mathematics, specifically Linear Algebra and Calculus (denoted "CALC I" in the poster). Some topics that are covered:

- Mathematical Notation
- Introduction to Differentiation
- Introduction to Integration
- Working with Vectors
- Working with Matrices
- SAT II and UNI I review

This class is recommended to be taken in January of the second year of High School, following the immediate completion of SAT III.

4.4.3 Third Year High School / Grade 11 / Junior (UNI II and UNI III)

UNI II (14 Weeks):
This class is the second of the University series. It focuses on introducing topics found in proofs, differentiation found in higher-level calculus (denoted "CALC II" in the poster) and the mathematical aspects of Computer Science. From this point on, students will be about 2 years ahead of their peers, following the completion of the two classes. Some topics that are covered:

- Propositional and Predicate Logic
- Set Theory
- Proof Techniques
- Series
- Differentiation Rules and Origins of Them

This class is recommended to be taken in September of the third year of High School, following the completion of UNI I.

UNI III (14 Weeks):
This is the third class of the University (UNI) series and builds on Differentiation in UNI II ("CALC II" in the poster), explores higher-level Integration and Statistics. Some topics that are covered include:

- Maclaurin series
- Integration Techniques
- Trapezium Rule
- Working with Discrete and Continuous Data in Statistics
- Conditional Probability
- UNI II and UNI III review

This class is recommended to be taken in January of the third year of High School, following the immediate completion of UNI II.

4.4.4 Fourth Year High School / Grade 12 / Senior (UNI IV and UNI V)

UNI IV (14 Weeks):
This class is the fourth of the University series. It dives into higher-level Calculus concepts, denoted as "CALC III" in the poster. These Calculus topics revolve primarily around various differentiation theorems that are explored in more advanced university proof-based courses. Some topics that are covered:

- Deep Dive into Limits and Definitions
- Continuity vs Differentiation
- Derivatives of Inverse Functions
- Mean Value Theorem

This class is recommended to be taken in September of the fourth year of High School, following the completion of UNI III.

UNI V (14 Weeks):
This is the final class that is offered in our math curriculum. Here we finalize all the key points we believe are important for success in University technical courses. This course wraps up integration and statistics concepts. Some topics that are covered:

- Riemann-Sums
- Reduction Formulas
- Improper Integrals
- de Moivre's theorem
- Discrete Random Variable Distributions
- Review of Curriculum

Following the completion of this curriculum, students will have the tools necessary to succeed up until their third year in University. For further learning, we offer tutoring where we continue more in-depth exploration and contest-style mathematics.

5 Tutoring

5.1 Description

Our tutoring classes are entirely customizable and we offer a couple of different packages based on your needs.

5.2 General Tutoring

Currently, our full-time tutors from the University of Toronto specialize in university-level mathematics, computer science, statistics, and economics, comfortably teaching these subjects up to the third-year university level. We also offer tutoring in high school sciences, including biology, chemistry, physics, and English. If you need help with a specific subject not listed here, we can arrange specialized tutoring services through our extensive network.

5.3 Mentorship Program

We also offer a **mentorship** program that connects you with current students from specific institutions and subjects you are interested in. These mentors not only tutor you but also provide additional services like resume assistance and project selection guidance. For instance, if you need a mentor from MIT Aerospace, we can find one for you!