

# **AI-Powered Career Guidance:** Enhancing Student Workforce Readiness through Machine Learning and Large Language Models

Sherif Abdelhamid, Jude Roberts

Virginia Military Institute





# INTRODUCTION

- Workforce readiness is a growing challenge, particularly for underserved populations.
- Manual resume reviews are inefficient and biased.
- There is a clear need for scalable, equitable, and personalized guidance solutions.

# MOTIVATION



Students often lack structured support to map skills to job opportunities.



Traditional advising systems struggle with volume and personalization.



AI offers a transformative pathway to automate and scale career guidance.

# RESEARCH GOALS

1

Design a hybrid AI system for resume classification and career guidance feedback.

2

Leverage both machine learning and large language models.

3

Address institutional needs for personalization and modularity.

## DATASET & PREPROCESSING



962 resumes manually  
labeled with job categories.



Text Preprocessing.



TF-IDF and Word2Vec  
tested for feature extraction.



TF-IDF selected based on  
higher model performance.





# MODEL SELECTION

- Models tested:
  - **Support Vector Machine (SVM)**
  - Random Forest
  - Logistic Regression
  - K-Nearest Neighbors (KNN)
  - Ensemble (Voting Classifier)
- Evaluated using Accuracy, Precision, Recall, and F1-score.
- Training/testing split: 75/25.

# Text Preprocessing

- Applied NLP techniques to clean resume content:
  - Tokenization
  - Stopword removal
  - Stemming and lemmatization
- Term frequency-inverse document frequency (TF-IDF) vectorization: a method for converting text into numerical vectors.

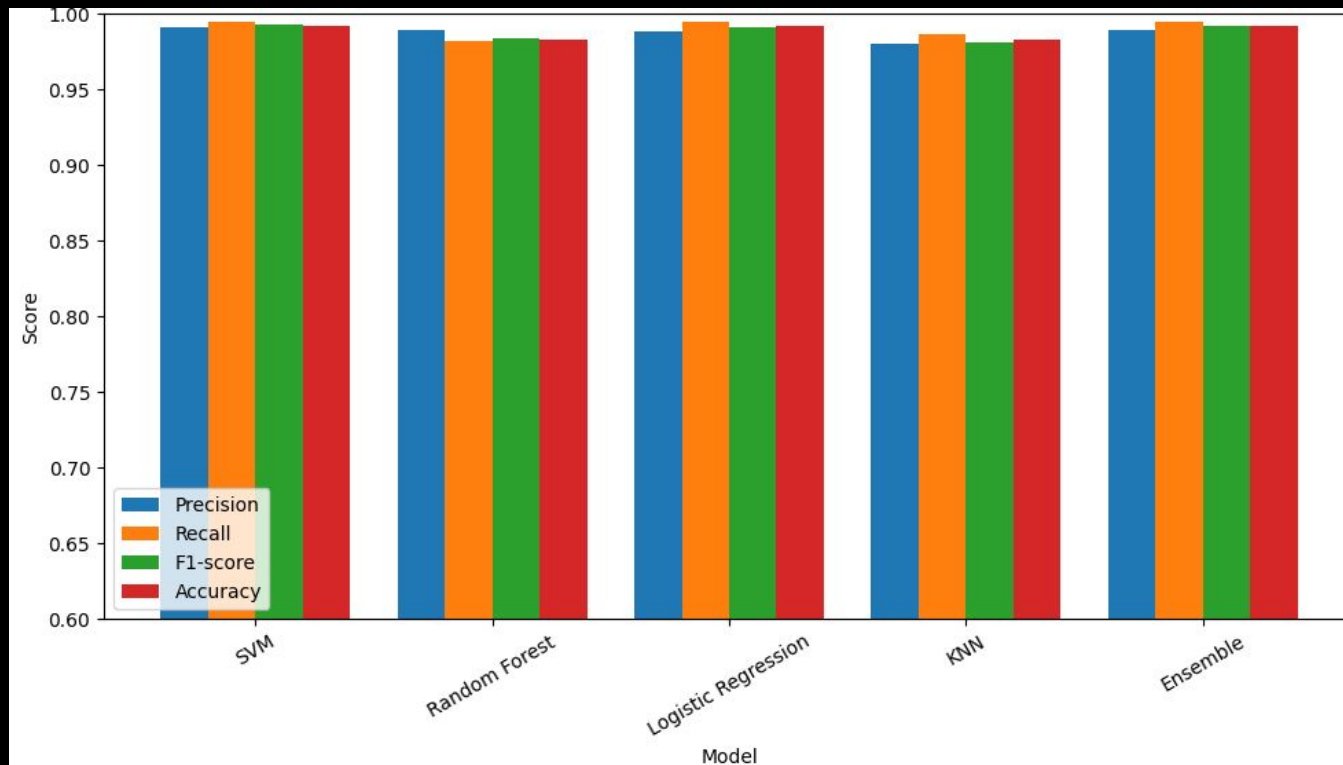
$$TF(t) = \frac{\text{Number of times term } t \text{ appears in the document}}{\text{Total number of terms in the document}}$$

$$IDF(t) = \log \left( \frac{\text{Total number of documents}}{\text{Number of documents containing term } t} \right) + 1$$

$$TF - IDF(t) = TF(t) \times IDF(t)$$

# TF-IDF RESULTS

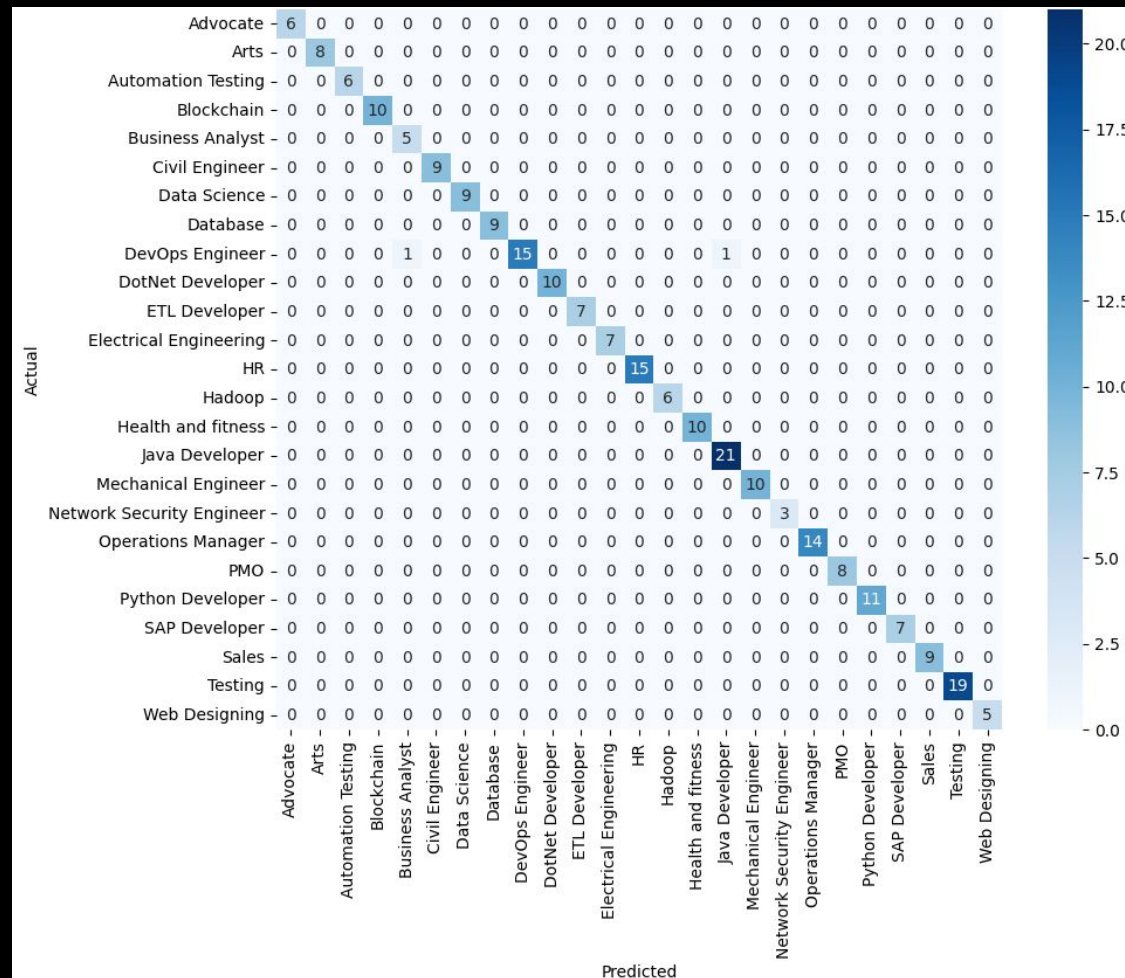
- SVM achieved the highest accuracy: 99.17%.
- All metrics (Precision, Recall, F1-score) above 99%.





# TF-IDF RESULTS

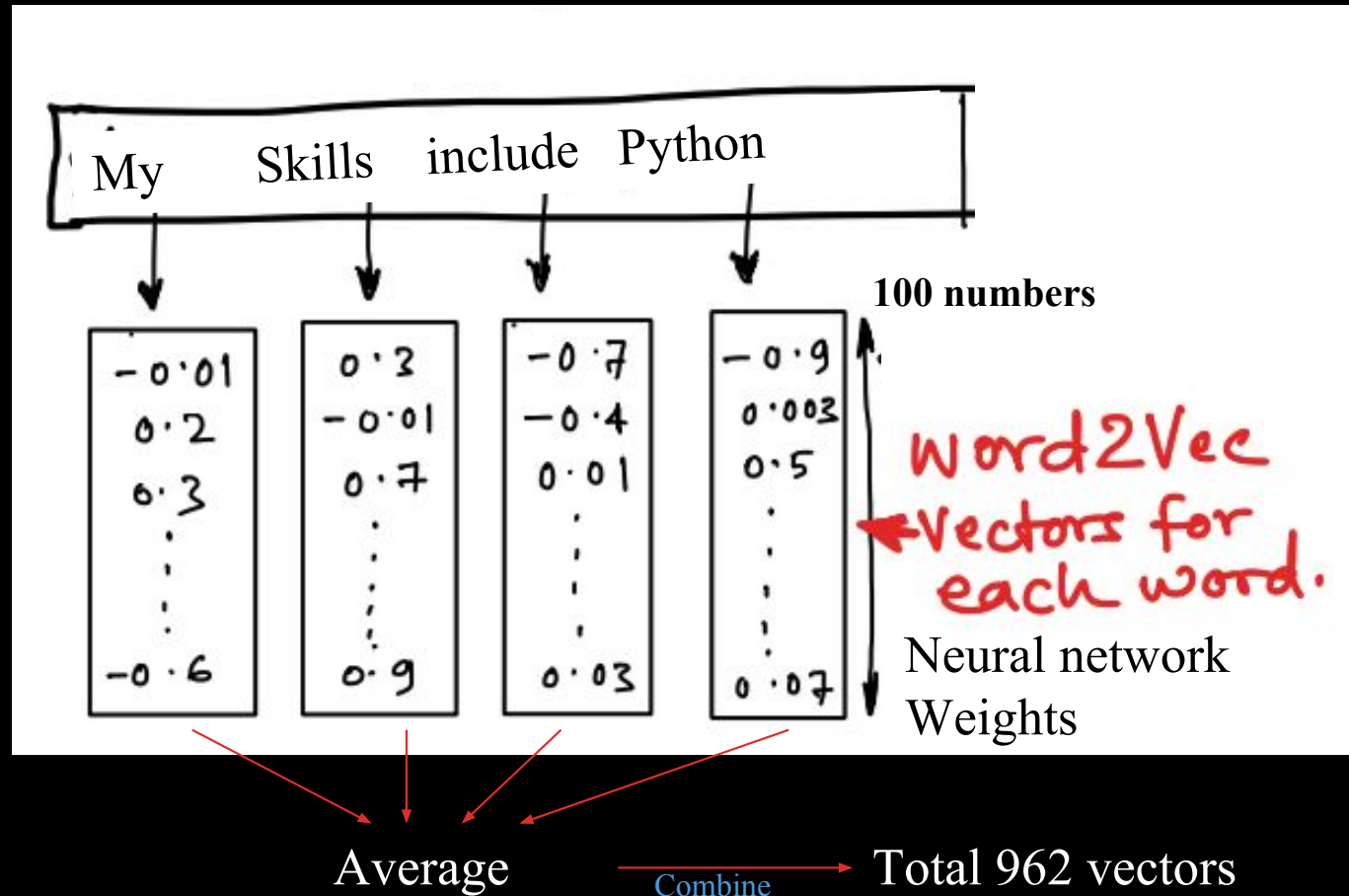
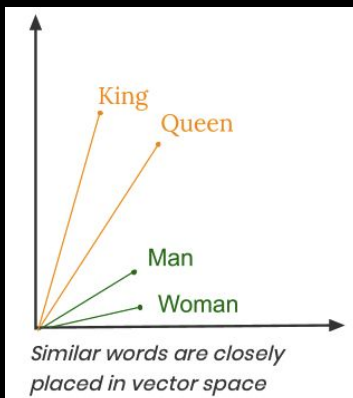
- **Confusion matrix shows near-perfect classification.**



# Text Preprocessing

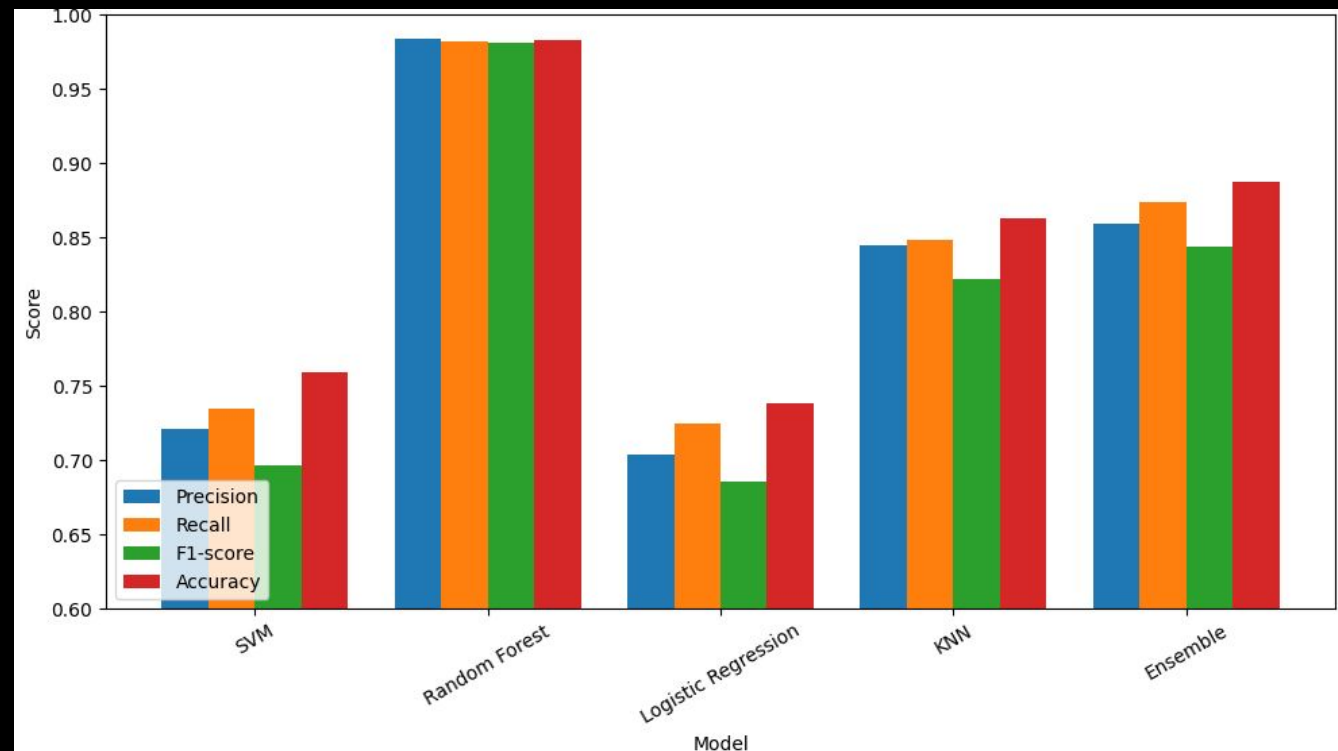
- Word2Vec

CBOW(Continuous bag of words) predicts the probability of a word to occur given the 5 words surrounding it.



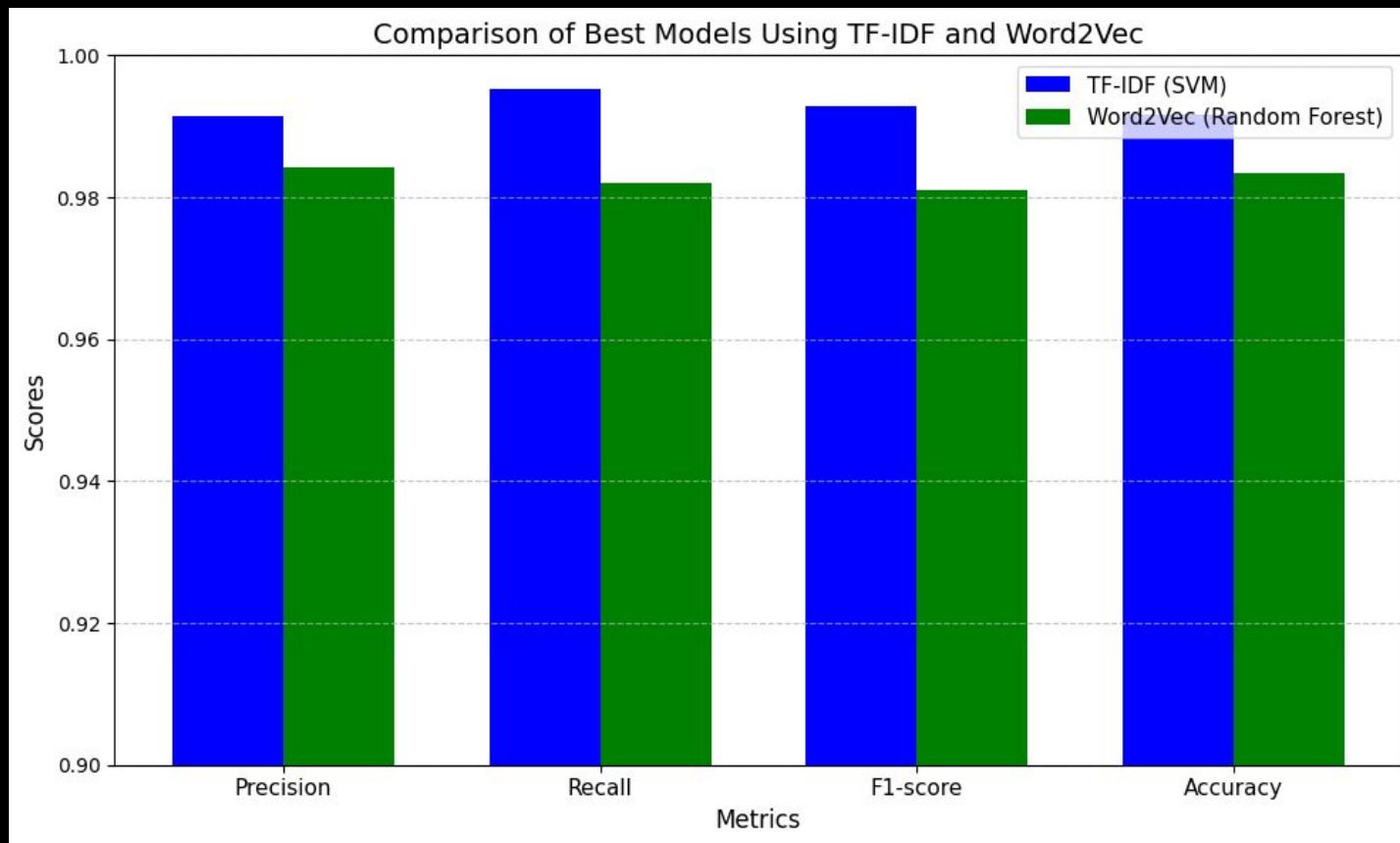
# WORD2VEC RESULTS

- Random Forest performed best with 98.34% accuracy.
- SVM and Logistic Regression underperformed.
- TF-IDF chosen due to consistent superiority across models.



# FINAL MODEL CHOICE

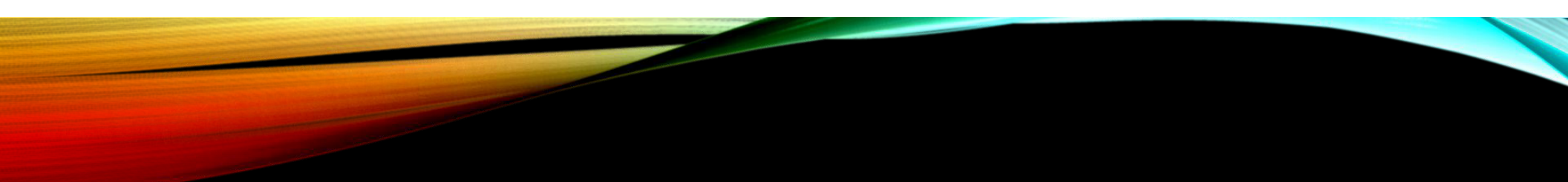
- SVM + TF-IDF adopted as the final classification pipeline.





# GPT-4 INTEGRATION

- Two GPT-4 instances integrated:
  1. General instance: Trained on broad workforce and industry knowledge.
  2. Institution-specific instance: Trained on internal courses, policies, and academic programs.



# PERSONALIZED GUIDANCE

- GPT-4 provides:
  - Skills gap analysis
  - Recommended courses or certifications
  - Learning pathways aligned with resume classification
- Feedback is contextual and adaptive.



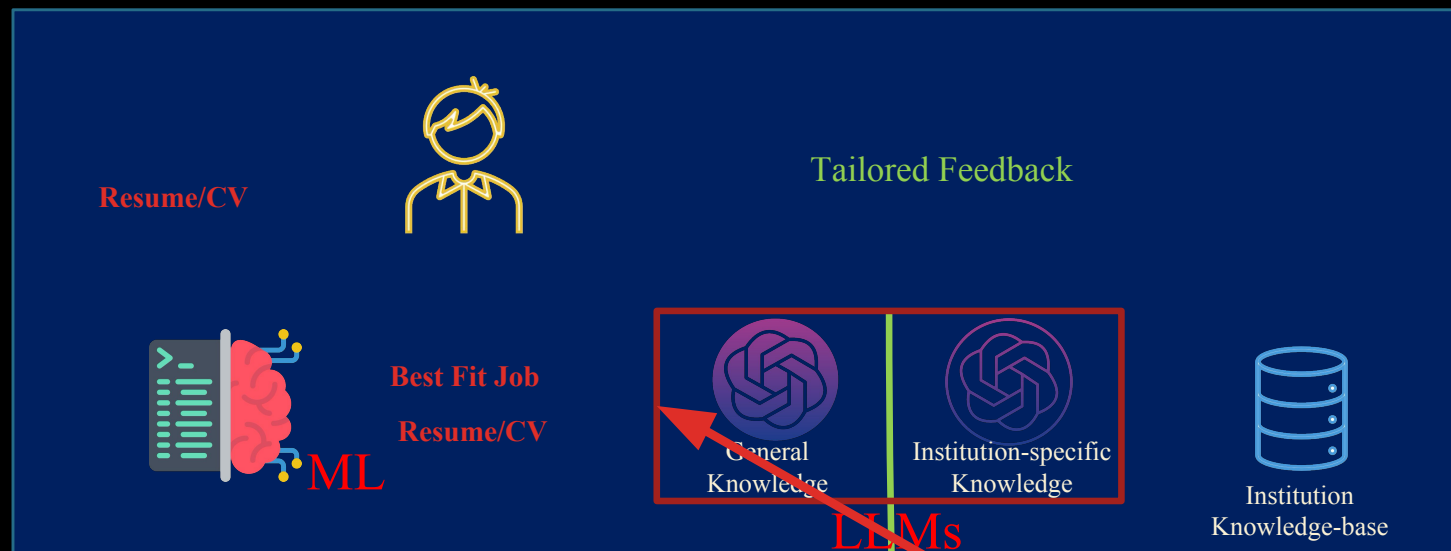


# SYSTEM OVERVIEW

- CareerCompass is a two-part system:
  - 1. Resume Classifier: Predicts job category from resume.
  - 2. GPT-4 Guidance: Provides targeted, contextual feedback.
- Input: Resume → Output: Role, Skills Gap, Learning Path.

# SYSTEM ARCHITECTURE

- Three-layer structure:
  - User Interface: Upload, feedback visualization.
  - Classification Engine: SVM model with TF-IDF.
  - Guidance Engine: GPT-4 General + GPT-4 Institution-specific modules.



# USER INTERFACE

CareerCompass

My Profile

Upload Resume

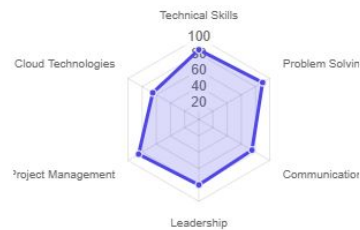
Upload Transcript

## Best Career Match

Software Developer

As a Software Developer, you'll design and build applications, collaborate with teams, and solve complex problems. This role aligns with your strong technical background and problem-solving abilities. Key responsibilities include coding, testing, and maintaining software systems.

## Skills Analysis



### Technical Skills

Strong proficiency in programming languages (Python, Java) and database management. Excellent problem-solving abilities with room for growth in cloud technologies.

### Communication

Good written communication skills. Could benefit from more experience in public speaking and technical presentations.

### Leadership

Strong project management capabilities. Shows initiative in team settings with potential for growth in conflict resolution.

## Recommended Certifications & Training

### AWS Certification

AWS Certified Solutions Architect - Associate

6 months

Online

[View Certification Details →](#)

### Professional Training

Agile Project Management (Scrum) Certification

3 months

Online/Hybrid

[View Certification Details →](#)

### Technical Certification

CompTIA Security+ Certification

4 months

Online

[View Certification Details →](#)

# USER INTERFACE

## Career Pathways

### Technical Leadership Path



### Product Development Path



### Cloud Architecture Path





# USER INTERFACE

## Recommended VMI Courses

### CIS 222 - Database Management

Introduction to database systems and SQL programming. Essential for software development careers.

Credits: 3 Spring Semester

### CIS 430 - Programming Languages

Study of programming language concepts and paradigms. Strengthens software development fundamentals.

Credits: 3 Fall Semester

### BU 330 - Management Information Systems

Integration of business and technical knowledge. Perfect for tech leadership roles.

Credits: 3 Spring Semester

## Course Selection Tips

- Courses are recommended based on your career path and current skill analysis
- Check prerequisites and course availability in SIS before registration
- Consult with your academic advisor for optimal course sequencing



# EVALUATION & RESULTS

- Resume classification accuracy: 99.17%
- GPT-4 responses reviewed for relevance and precision.
- System validated for reliability, scalability, and educational value.





# SOCIETAL IMPACT

- Supports equitable access to career guidance.
- Helps map pathways to success.
- Reduces human bias and advisor workload.



# BEYOND WORKFORCE READINESS

- System supports:
  - Lifelong learning
  - Career pivots
  - Reskilling and upskilling for evolving job markets.



# FUTURE WORK

- Expand resume dataset diversity.
- Integrate real-time labor market data.
- Enhance UI with resume editor and user feedback loop.

## CONCLUSION & ACKNOWLEDGMENTS



SVM + TF-IDF + GPT-4 offers scalable, personalized career guidance.



Enables data-driven, inclusive, and flexible career planning.



Research is funded by the Commonwealth Cyber Initiative (CCI).

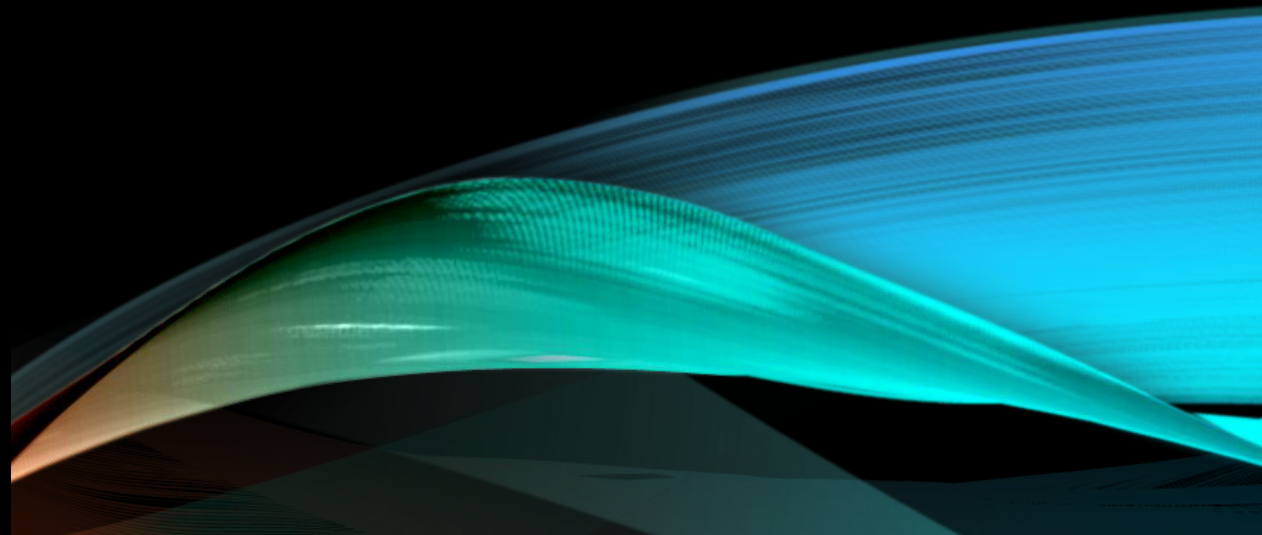


Contact: [abdelhamidse@vmi.edu](mailto:abdelhamidse@vmi.edu)



THANK YOU

Questions Please



# Model Training and Evaluation

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$F_1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{HammingLoss}(x_i, y_i) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{\text{xor}(x_i, y_i)}{|L|},$$





## ML MODELS EVALUATED (1/5)

- K-Nearest Neighbors (KNN):
  - A non-parametric algorithm that classifies data points based on the majority class of their nearest neighbors in the feature space.
  - Strengths: Simple and interpretable; effective for smaller datasets.
  - Weaknesses: Computationally expensive for larger datasets; sensitive to irrelevant features.
  - Why Used: Evaluates performance of proximity-based classification in project classification.



## ML MODELS EVALUATED (2/5)

- Decision Tree:
  - A tree-structured model that splits data into branches based on feature values to make decisions.
  - Strengths: Easy to visualize and interpret; handles categorical and numerical data.
  - Weaknesses: Prone to overfitting; less effective with imbalanced data.
  - Why Used: Provides interpretable decision-making for understanding project classification rules.



## ML MODELS EVALUATED (3/5)

- Random Forest:
  - An ensemble learning method that builds multiple decision trees and merges their outputs for better accuracy and stability.
  - Strengths: High accuracy; robust to overfitting and noise.
  - Weaknesses: Computationally intensive; less interpretable than simpler models.
  - Why Used: Demonstrates the effectiveness of ensemble learning with the hybrid feature set.



## ML MODELS EVALUATED (4/5)

- SVM:
  - It works well when there's a clear separation between groups.
  - It is a powerful and efficient algorithm for high-dimensional spaces, particularly useful for complex classification problems.
  - It's great for small and high-dimensional data (like texts or images).
  - It can be **slow** for very large datasets.



## ML MODELS EVALUATED (5/5)

- Logistic Regression:
  - Standard logistic regression handles binary classification
  - For more than two classes, we use strategies like:
  - One-vs-Rest (OvR): Train one classifier per class
  - Multinomial Logistic Regression: Uses the softmax function
  - Outputs a probability distribution over all classes
  - The class with the highest probability is selected as the prediction