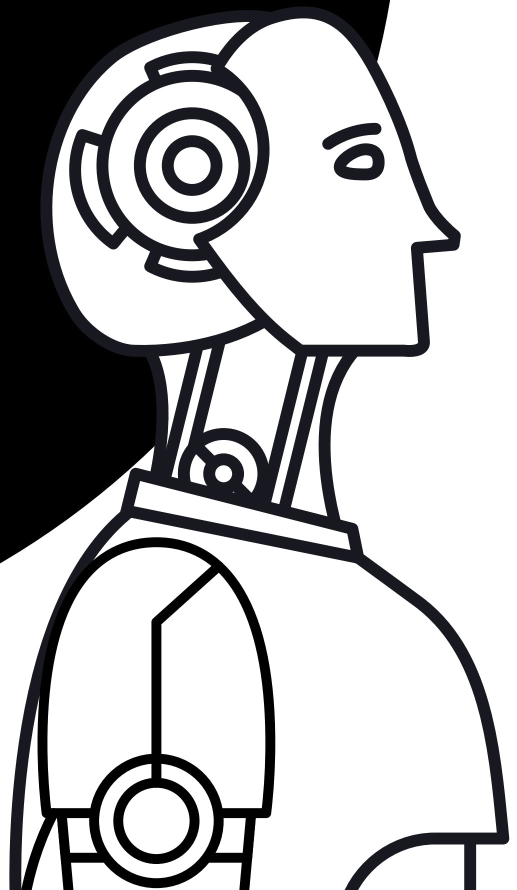
SE 3007: INTRODUCTION TO MACHINE LEARNİNG

AI IMPACT ON JOBS: PREDICTING JOB RISK & FUTURE SKILLS



presented by:

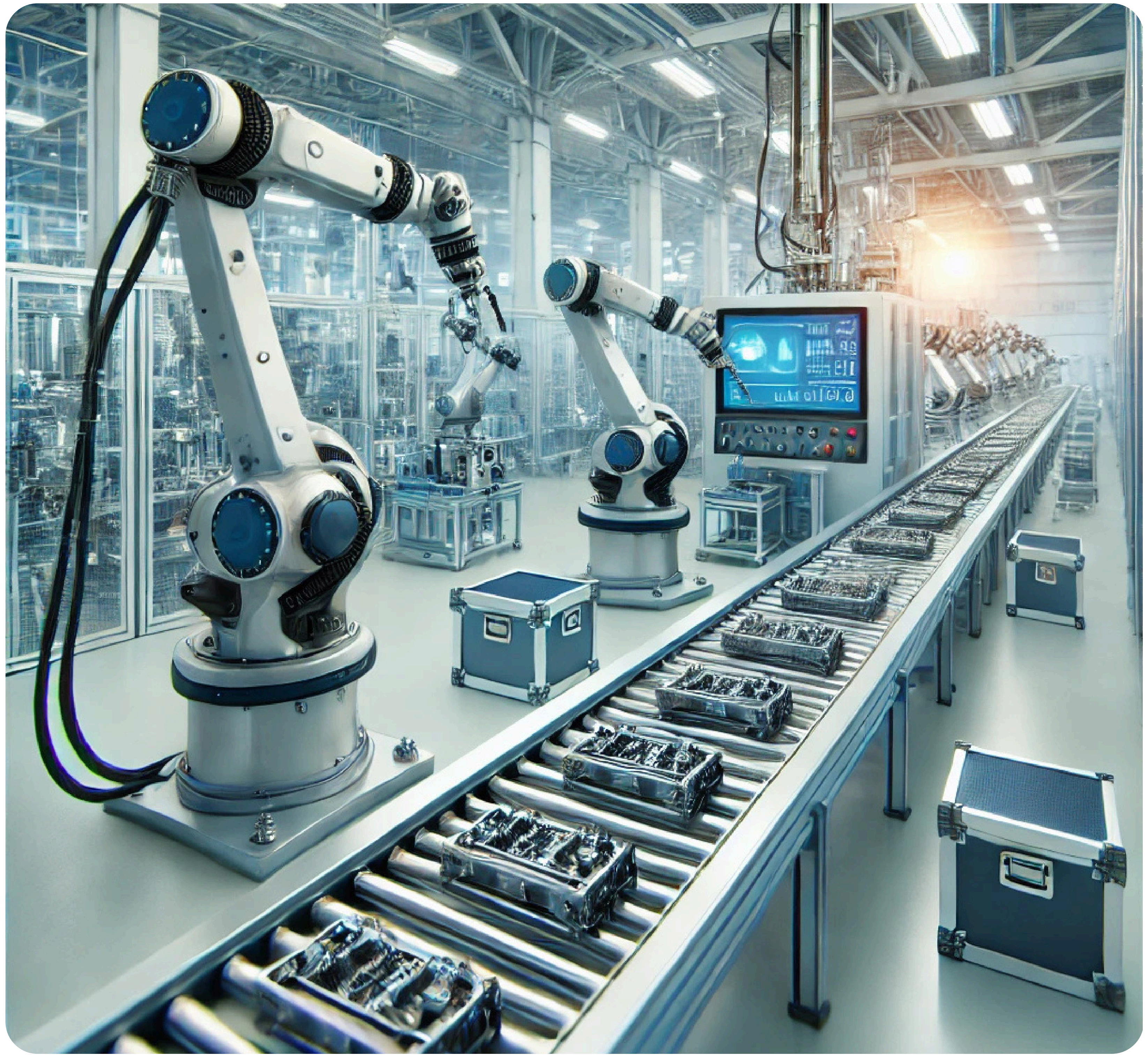
Ömer Faruk Kaba
Ahmet Sefa Yıldırım
Yusuf Şaban Tosuncuk

WHAT THIS STUDY DOES?

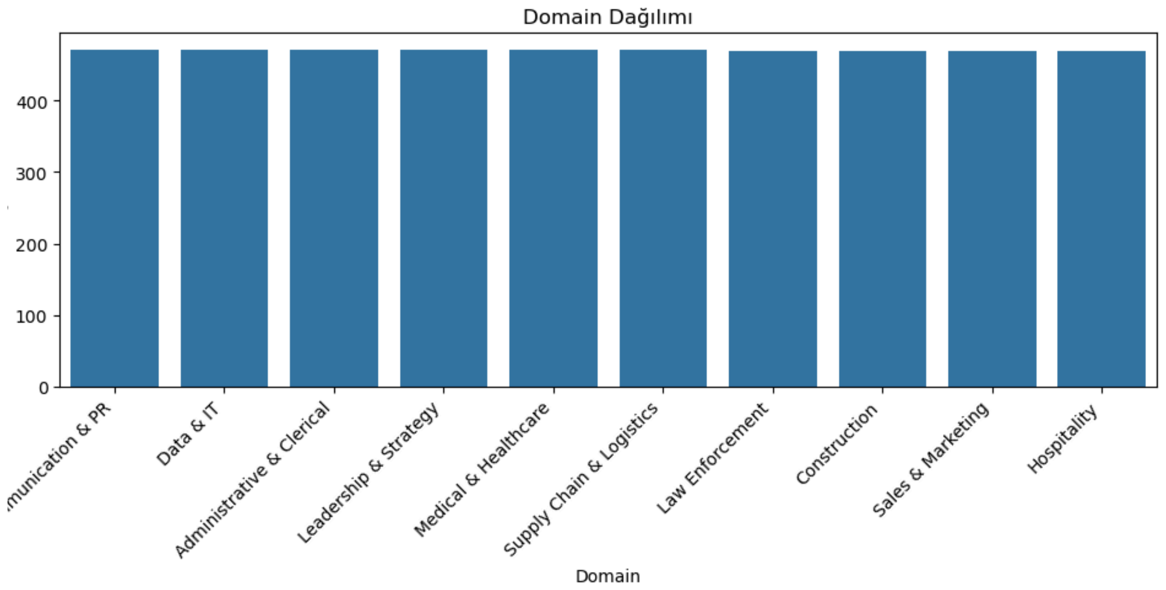
- Predicts the automation risk of occupations using machine learning
- Classifies jobs into High / Medium / Low AI vulnerability
- Suggests alternative low-risk jobs based on skills
- Builds a sector-level AI Resilience Index
- Provides a data-driven framework for future workforce decisions

WHY IT MATTERS?

AI will transform millions of jobs, understanding risk today helps prepare for tomorrow.



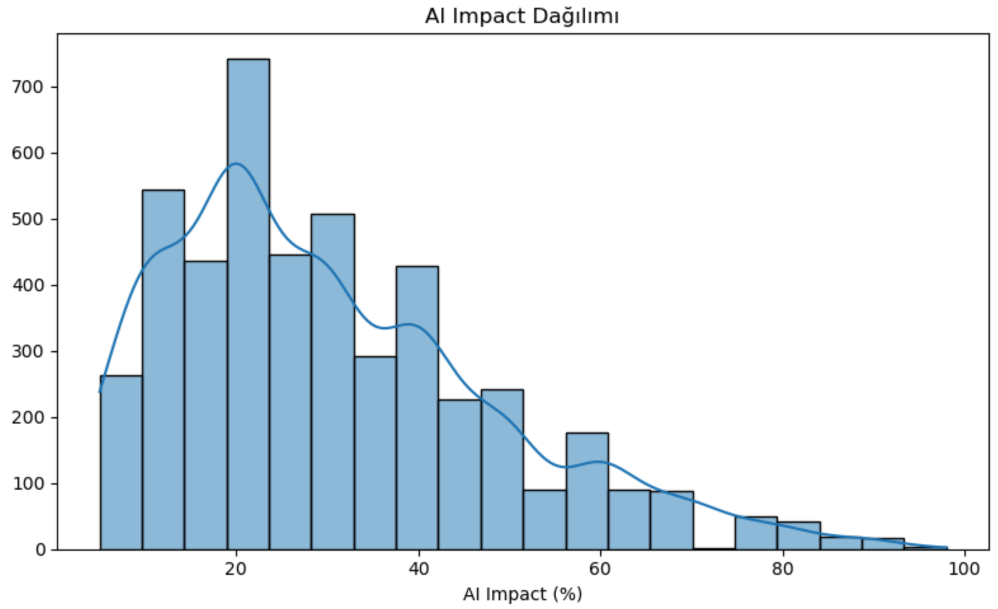
EXPLORATORY DATA ANALYSİS



01

DOMAIN DISTRIBUTION

The dataset is evenly distributed across all domains. The average AI Impact values are very similar across domains, indicating a balanced and synthetic dataset structure.

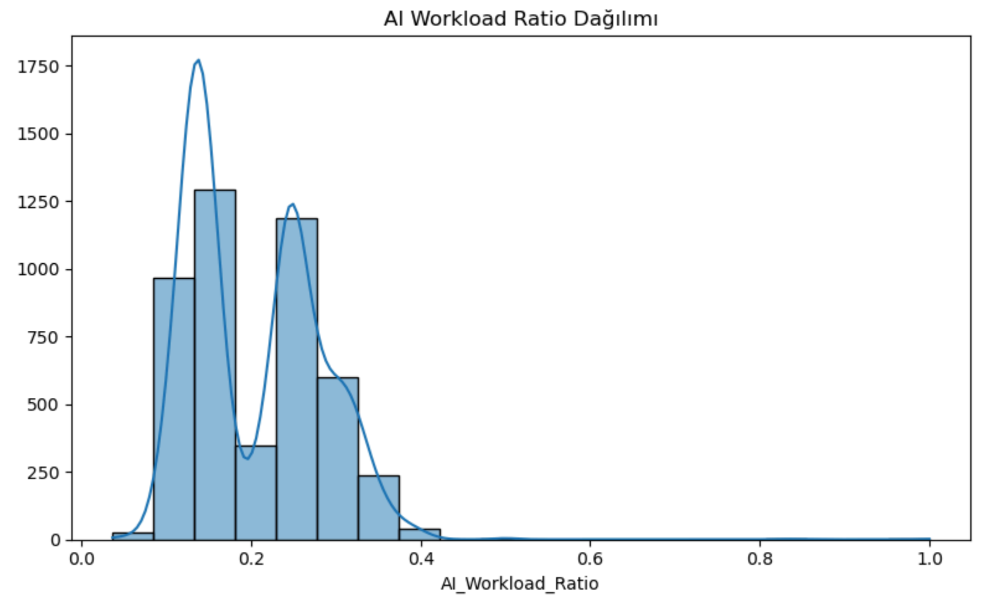


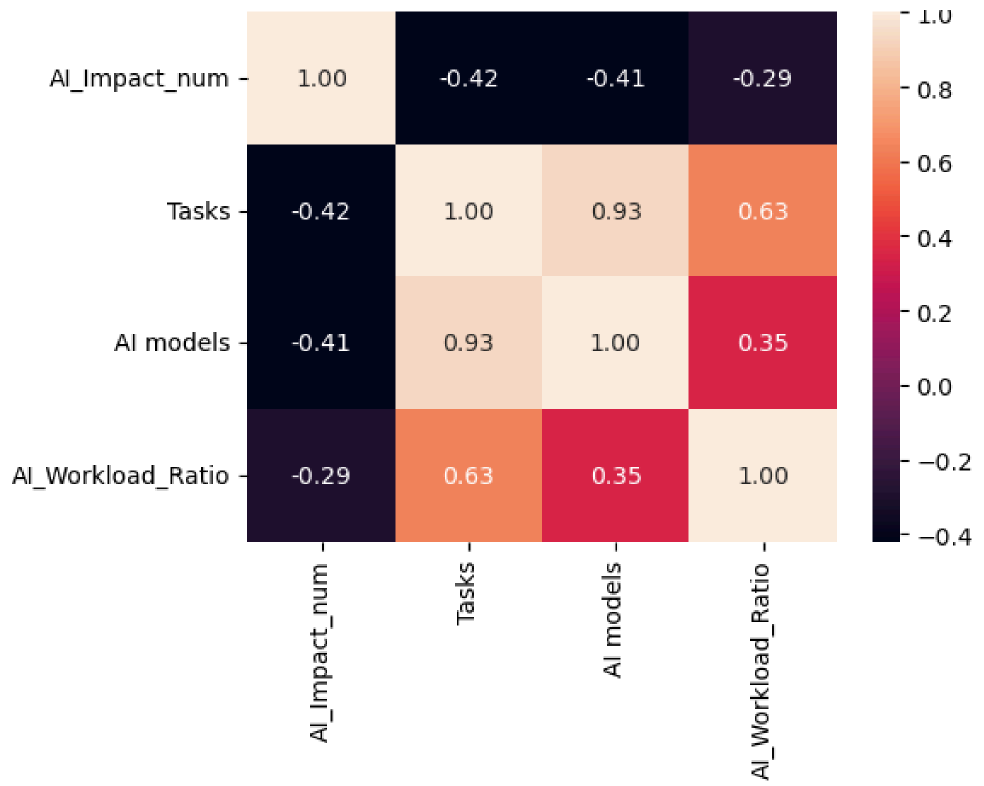
02

AI IMPACT DISTRIBUTION

The AI Impact distribution is right-skewed, meaning low and medium risk jobs are more common, while highly automatable jobs are fewer

EXPLORATORY DATA ANALYSİS





03

AI WORKLOAD RATIO

AI Workload Ratio is mostly concentrated between 10%–30%; meaning most jobs contain a moderate amount of automatable tasks.

04

CORRELATION MATRIX

There is a strong positive correlation between Tasks and AI models. AI Impact shows a moderate negative correlation with these variables.

AI RISK PREDICTION MODEL - REGRESSION

```
numeric_transformer = Pipeline(steps=[
    ("scaler", StandardScaler())
])

categorical_transformer = Pipeline(steps=[
    ("onehot", OneHotEncoder(handle_unknown='ignore'))
])

preprocessor = ColumnTransformer(
    transformers=[
        ("num", numeric_transformer, numeric_features),
        ("cat", categorical_transformer, categorical_features),
    ]
)
```

PIPELINE

The preprocessing steps operate through separate pipelines for numeric and categorical features. Numeric data is scaled, while the Domain feature is one-hot encoded and combined through a ColumnTransformer.

MODEL SETUP

All preprocessing steps and the RandomForestRegressor are merged into a single Pipeline, ensuring consistent training and reproducible results.

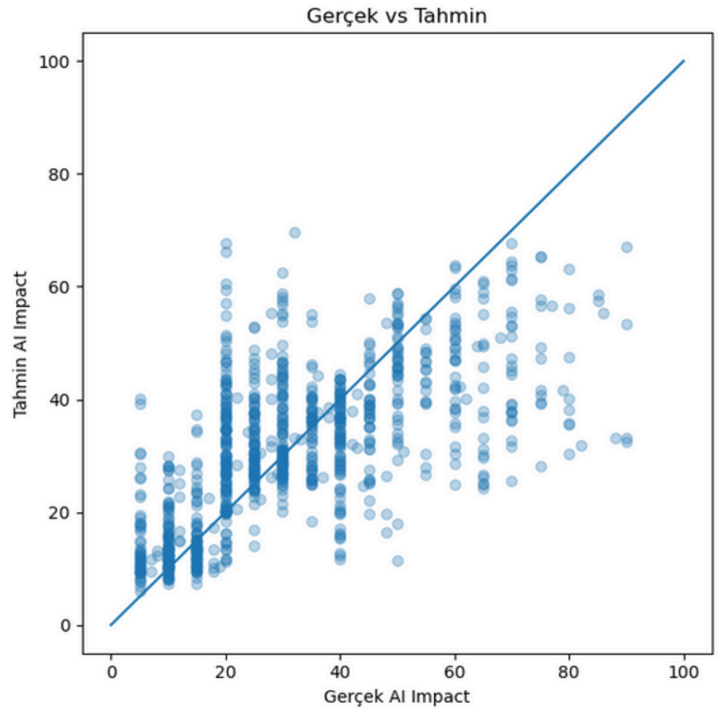
```
model = RandomForestRegressor(
    n_estimators=300,
    random_state=42,
    n_jobs=-1
)

pipeline = Pipeline(steps=[
    ("preprocessor", preprocessor),
    ("model", model)
])
```

AI RISK PREDICTION MODEL - REGRESSION

```
MAE  : 9.987
RMSE : 14.162
R^2  : 0.439
```

MODEL PERFORMANCE

The model achieves MAE ≈ 9.99 and RMSE ≈ 14.16, indicating moderate prediction error. $R^2 \approx 0.439$ means the model explains about 44% of the variance.



MODEL VISUALIZATION

The scatter plot shows that predictions follow the overall trend. Many points align near the 45° line, indicating strong consistency between predicted and actual AI Impact values.

WHY RANDOM FOREST?

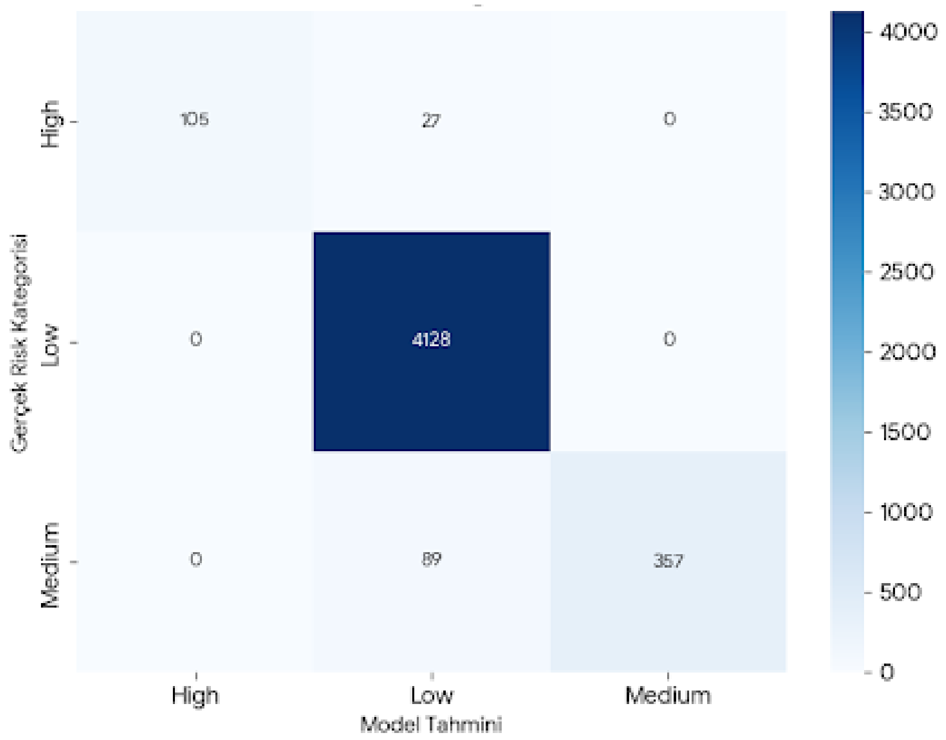
- Captures non-linear relationships
- Robust against noise and outliers
- Works well with mixed data types
- Provides stable and explainable predictions

JOB VULNERABILITY CLASSIFICATION

To classify occupations into Low – Medium – High risk groups based on their AI Impact scores.

STEPS

- Normalized the AI Impact score and converted it into 3 classes.
- Built preprocessing pipeline (Scaler + OneHotEncoder).
- Trained a RandomForestClassifier.
- Evaluated performance using accuracy + confusion matrix.



CONFUSION MATRIX

The model predicts the Low risk class with extremely high accuracy. Although there is minor confusion between High and Medium classes, the overall accuracy (97.54%) is very strong.

JOB VULNERABILITY CLASSIFICATION

Model Output

	Job titiles	Domain	AI Impact	AI_Impact_num	Risk_Category	Model_Prediction
19	Warehouse Worker	Sales & Marketing	90%	90.0	High	High
20	Web Search Evaluator	Hospitality	90%	90.0	High	High
21	Development Manager	Communication & PR	89%	89.0	High	Low
22	Delivery Driver	Data & IT	88%	88.0	High	Low
23	Chief Security Officer	Administrative & Clerical	88%	88.0	High	High
24	File Clerk	Leadership & Strategy	88%	88.0	High	Low
25	Mail Processing	Medical & Healthcare	88%	88.0	High	High
26	Director Of Operations	Supply Chain & Logistics	87%	87.0	High	High
27	irect Support Professional	Law Enforcement	86%	86.0	High	High
28	f Communications Officer	Construction	85%	85.0	High	High
29	Administrative Clerk	Sales & Marketing	85%	85.0	High	High

FUTURE PROOF SKILLS RECOMMENDER SYSTEM

Suggest safer, low-risk job alternatives based on skill similarity.

What Does the Future-Proof Skills Recommender Do?

Goal: For a high-risk job, suggest lower-risk occupations that require similar skills.

INPUT

- User's job (Job Title)
- Skill scores of that job
- Risk category (High / Medium / Low)

SETUP

- Build a skill vector for every job.
- Compute similarity between vectors using cosine similarity.
- Filter jobs with lower AI risk than the original one.
- Return the top 5–10 jobs with the highest similarity as recommendations.

SENTENCE TRANSFORMER

We used the SentenceTransformer (all-MiniLM-L6-v2) model to encode job titles into numerical vectors that preserve semantic meaning. This allows the system to understand that terms like "developer" and "software engineer" are closely related.

FUTURE PROOF SKILLS RECOMMENDER SYSTEM

Suggest safer, low-risk job alternatives based on skill similarity.

CODE LOGIC

1- Each job is converted into a numerical vector:

job→[s1,s2,s3,...]

This vector describes the strengths of that job across different skills.

2- Create a Skill Matrix

We combine all job vectors into a matrix where:

- Rows = jobs
- Columns = skill values
- This forms the search space for similarity.

3- Use Cosine Similarity

We compute similarity between all job vectors using cosine

$$\text{similarity}(A, B) = \frac{A \cdot B}{||A||||B||}$$

A score close to 1 → very similar skills

A score close to 0 → very different skills

5 - Filter by Lower Risk Jobs

When the user selects a job:

- We check its risk level (High / Medium / Low)
- We filter only the jobs with lower AI risk
- (e.g., High → Medium & Low)

This makes the recommendations both safe and relevant.

6 - Rank by Similarity

Among the lower-risk jobs,

we sort them by similarity score in descending order.

Top 5 or Top 10 most similar jobs are selected.

7 - Return Recommendations

The system outputs a list of careers that have:

✔ similar skills

✔ lower AI risk

FUTURE PROOF SKILLS RECOMMENDER SYSTEM

Model Output:

	High_Risk_Job	High_Risk_Domain	Alternative_Job	Alternative_Domain	Similarity
1	Communications Manager	Communication & PR	Channel Marketing Manager	Communication & PR	0.7874
2	Communications Manager	Communication & PR	Senior Marketing Manager	Communication & PR	0.7849
3	Communications Manager	Communication & PR	Business Development Manager	Communication & PR	0.7744
4	Data Collector	Data & IT	Information Architect	Data & IT	0.725
5	Data Collector	Data & IT	Scanner	Data & IT	0.7018
6	Data Collector	Data & IT	Information Security Manager	Data & IT	0.6929
7	Data Entry	Administrative & Clerical	Inside Sales Representative	Administrative & Clerical	0.6736
8	Data Entry	Administrative & Clerical	It Business Analyst	Administrative & Clerical	0.6591
9	Data Entry	Administrative & Clerical	Insurance Verification Specialist	Administrative & Clerical	0.6571
10	Mail Clerk	Leadership & Strategy	Postal Clerk	Leadership & Strategy	0.8956
11	Mail Clerk	Leadership & Strategy	Grocery Clerk	Leadership & Strategy	0.8583
12	Mail Clerk	Leadership & Strategy	Bakery Clerk	Leadership & Strategy	0.8442
13	Compliance Officer	Medical & Healthcare	Compliance Specialist	Medical & Healthcare	0.9303
14	Compliance Officer	Medical & Healthcare	Credit Officer	Medical & Healthcare	0.8375
15	Compliance Officer	Medical & Healthcare	Legal Officer	Medical & Healthcare	0.8231
16	Chief Executive Officer (CEO)	Supply Chain & Logistics	Chief Engineer	Supply Chain & Logistics	0.8217

SIMILARITY SCORE :

Our recommendation engine uses a hybrid similarity model that combines Sentence Transformer semantic embeddings with cosine similarity on domain and numeric features. The final similarity score is a weighted blend of these three components.

```
hybrid_sim = (
    0.45 * job_sim +
    0.35 * domain_sim +
    0.20 * feat_sim
)
```

SECTOR AI RESILIENCE INDEX

Measuring how resilient each sector is against AI-driven automation.

We created an AI Resilience Index to measure how resistant each sector is to AI-driven automation. The index is calculated from multiple factors such as average AI impact, task complexity, model usage intensity, and AI workload ratio.

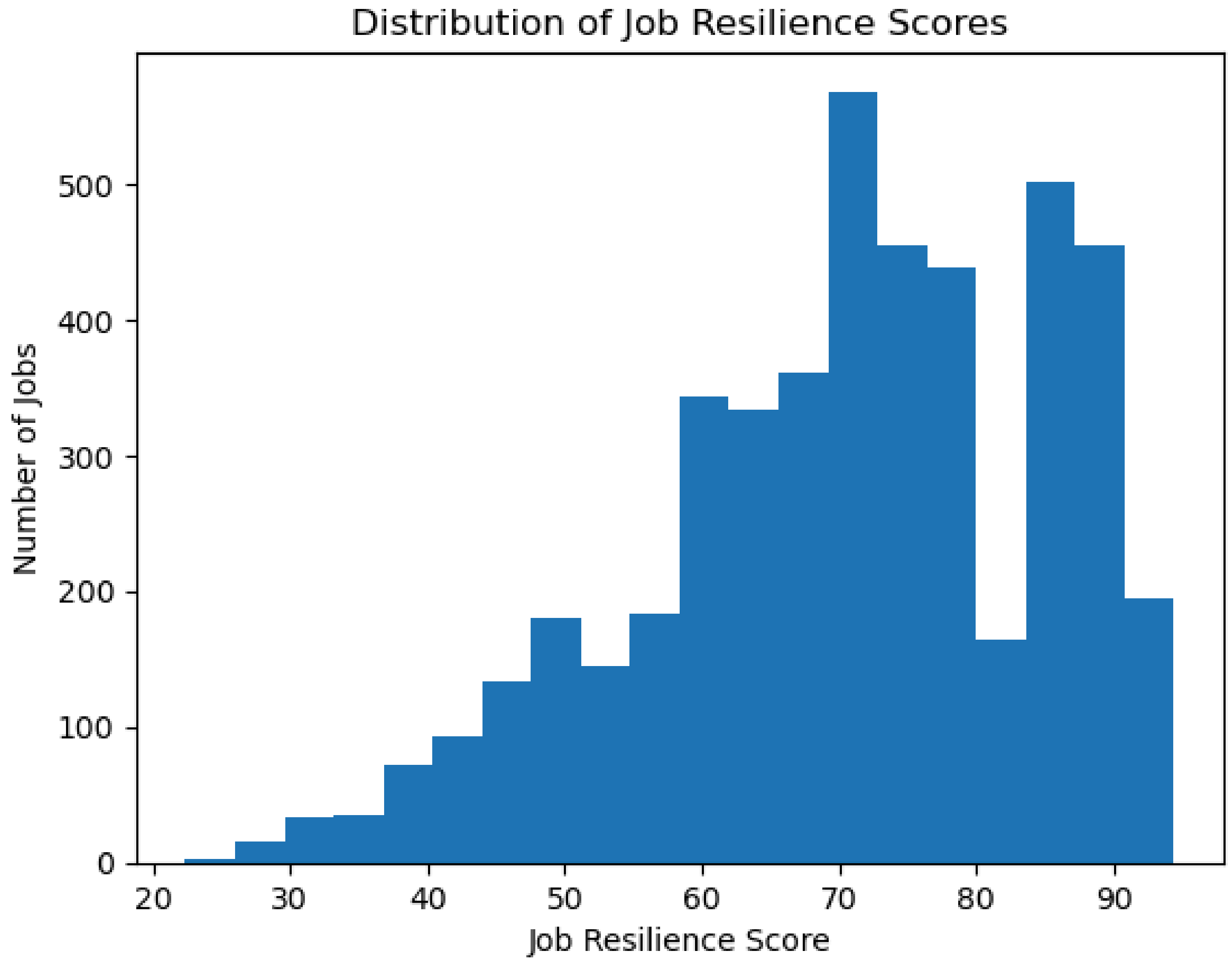
To compute the Sector AI Resilience Index, we aggregated each sector's:

- Average AI Impact
- Average Tasks
- Average AI models
- Average AI_Workload_Ratio

All metrics were normalized, inverted where necessary, and combined into a single resilience score.

Resilience Index:

0.35 × (1 – AIImpactNorm)
+ 0.25 × (1 – AIWorkloadNorm)
+ 0.20 × TasksNorm
+ 0.20 × DomainComplexityNorm



SECTOR AI RESILIENCE INDEX

Measuring how resilient each sector is against AI-driven automation.

Model Output:

	Job titiles	Domain	AI Impact	AI_Impact_num	Model_AI_Impact	Job_Resilience_Score
76	Printing Press Operator	Supply Chain & Logistics	80%	80.0	62.44	37.56
77	Automation Test Engineer	Law Enforcement	80%	80.0	72.596	27.403999999999996
78	Automation Tester	Construction	80%	80.0	72.016	27.983999999999995
79	QA Automation Engineer	Sales & Marketing	80%	80.0	69.552	30.447999999999993
80	Tax Preparer	Hospitality	80%	80.0	29.938	70.062
81	Associate	Communication & PR	79%	79.0	33.694	66.306
82	Chief Learning Officer	Data & IT	79%	79.0	66.328	33.672
83	Web Project Manager	Administrative & Clerical	78%	78.0	60.046	39.954
84	Medical Coding Specialist	Leadership & Strategy	78%	78.0	59.96	40.04

LIMITATIONS

While the models provide meaningful insights, several limitations should be considered:

- **Synthetic Dataset:** The dataset is artificially generated, meaning real-world labour dynamics, economic factors, and nuanced job definitions may not be fully captured.
- **Feature Availability:** The skill columns and workload indicators are limited.
- **Model Interpretability:** RandomForest provides strong predictive performance but limited interpretability compared to linear or SHAP-based approaches.
- **Semantic Dependence:** SentenceTransformer embeddings rely solely on job titles, not full job descriptions, which may reduce semantic accuracy.
- **Static Analysis:** The models assume static AI risk; real-world risk changes over time due to technology adoption and economic shifts.
- **Domain Encoding Simplicity:** One-hot encoding treats domains as independent categories and does not capture relationships between similar sectors.

These limitations highlight the need for more robust, real-world data and more advanced modelling techniques in future work.

CONCLUSION

In this project, we explored how machine learning can be used to understand and forecast the impact of AI-driven automation on jobs and sectors.

We developed four interconnected models:

AI Risk Prediction Model estimated automation risk using numerical and categorical features.

Job Vulnerability Classification grouped occupations into Low, Medium, and High risk.

Future-Proof Skills Recommender System suggested safer alternative careers using a hybrid similarity model combining semantic (SentenceTransformer) and skill-based measures.

Sector AI Resilience Index quantified how resilient each sector is by aggregating job-level risk and workload indicators.

Overall, our results show that machine learning can effectively capture patterns related to task complexity, domain structure, and skill similarity, enabling data-driven insights for workforce planning, career guidance, and policy decision-making.