AUTUMN INTERNSHIP PROJECT REPORT

Diabetes Prediction : Classification Comparison + Metrics + Evaluation

Laxman Soni,

M.Sc. Physics, Dr. B. R. Ambedkar National Institute of Technology, Jalandhar, Punjab.

Period of Internship: 25th August 2025 - 19th September 2025

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

1. Abstract

This project focuses on the development and evaluation of machine learning models for diabetes prediction using a publicly available dataset. Several classification algorithms, including K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) were implemented and compared. The models were assessed on key performance metrics such as accuracy, precision, recall, F1-score, and AUC to provide a comprehensive evaluation of predictive capabilities. The study underscores the significance of model selection and evaluation criteria in predictive tasks, in domains such as healthcare where misclassification can have serious consequences. By extending the same methodology to the Titanic dataset, the work demonstrates the versatility of machine learning approaches. Overall, this comparative study offers a balanced perspective on the strengths and limitations of different classifiers across both medical and non-medical datasets.

2. Introduction

1. Background and Relevance

Predictive modeling plays a crucial role in modern data science, offering solutions in both healthcare and broader decision-making tasks. Diabetes, as a chronic disease, is a major global health concern where early detection is critical for reducing complications and improving patient outcomes. On the other hand, the Titanic survival dataset has become a classic case study in predictive analytics, showing how demographic and socio-economic factors influence survival chances. Together, these problems highlight the relevance of machine learning in solving diverse classification tasks.

2. Technology Involved

The project is implemented using the Python programming language and its scientific libraries. Tools such as pandas and numpy were used for data handling, matplotlib and seaborn for visualization, and scikit-learn for building and evaluating machine learning models. These technologies provide efficient methods for preprocessing, training, and assessing classifiers on structured datasets.

3. Background Material Survey

Previous research and practice suggest that algorithms such as **Logistic Regression**, **K-Nearest Neighbors (KNN)**, and **Support Vector Machines (SVM)** are effective for classification problems. Logistic Regression offers interpretability, KNN is simple yet powerful in capturing neighborhood-based patterns, and SVM is robust in handling high-dimensional data with clear decision boundaries. These models are widely used in both healthcare analytics and general classification tasks.

4. Procedure

The project followed a systematic workflow:

1. **Data Exploration and Preprocessing** – inspecting the datasets, handling missing values, scaling features, and preparing inputs for modeling.

2. **Data Splitting** – dividing data into training and testing sets to ensure unbiased evaluation.

3. **Model Implementation** – training Logistic Regression, KNN, and SVM classifiers.

4. **Performance Evaluation** – assessing models using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

5. Purpose of the Project

The primary objective of this project is to apply machine learning models for early prediction of diabetes, thereby highlighting their importance in healthcare management. A secondary objective is to demonstrate the adaptability of the same methodology by applying it to the Titanic dataset, showing how classification models can generalize to domains beyond medicine. Through comparative evaluation, the study aims to identify the strengths and limitations of different algorithms and provide insights into model selection for real-world applications.

The list of topics that I received training on during the first two weeks of internship are :

1. Python Basics (Datatypes, Datastructures, Class, Functions, OOPS, Numpy, Panda)
2. Machine Learning Overview
3. Regression Lab
4. Classification Lab
5. LLM Fundamentals

3. Project Objective

- To build machine learning models that can predict the chances of diabetes and check how well they perform in identifying people at risk.

- To compare the results of different classifiers like KNN and SVM using accuracy, precision, recall, F1-score, and ROC-AUC.

- To show that the same approach can also be applied beyond healthcare by testing it on the Titanic survival dataset.

- To highlight why choosing the right model and evaluation criteria matters, especially in areas like healthcare where wrong predictions can have serious effects.

- To analyze the Titanic dataset (*https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv*) as a sample survey, where the target group is the passengers of the RMS Titanic, and study how factors like age, gender, and travel class influenced survival.

4. Methodology

This project was carried out in a structured sequence of steps to ensure that the datasets were carefully handled, analyzed, and used to build predictive models. The overall workflow involved **data collection, preprocessing, exploratory data analysis (EDA), model development, model evaluation, and comparative analysis**. The methodology was applied to two benchmark datasets: **Diabetes dataset** and **Titanic dataset**.

1. Data Collection

- **Diabetes Dataset:** Obtained from the GitHub repository, consisting of health-related attributes such as glucose, BMI, insulin, blood pressure, and age, along with a binary outcome variable (presence/absence of diabetes).

- **Titanic Dataset:** Collected from the GitHub repository, including passenger demographics (age, sex, class), family information (SibSp, Parch), and survival status.

2. Data Cleaning and Preprocessing

- **Handling Missing Values:**

    - In the Titanic dataset, missing Age values were replaced with the mean, and missing Embarked values were filled using the mode.

    - In the Diabetes dataset, no significant missing values were present, but numerical features were checked for validity.

- **Feature Selection:**

    - Non-predictive identifiers such as PassengerId, Name, and Ticket were dropped from the Titanic dataset.

- **Encoding Categorical Variables:**
    - For Titanic, categorical features (Sex, Embarked) were converted into numerical values using label encoding and one-hot encoding.
- **Scaling:**
    - Numerical variables in both datasets (e.g., glucose, BMI, age, and fare) were standardized using normalization to improve model performance.

3. Exploratory Data Analysis (EDA)

EDA was conducted to understand the structure and relationships within each dataset:

- **Descriptive Statistics:** Mean, median, and distribution of numerical features.
- **Visualizations:**
    - Histograms of age and glucose to show population spread.
    - Bar plots of categorical features such as sex, Pclass, and survival (Titanic).
    - Correlation heatmaps to identify feature associations.

4. Model Development

Two supervised classification algorithms were implemented in both projects:

1. **K-Nearest Neighbors (KNN):**
    - Classified samples based on the majority class of their nearest neighbors.
    - Simple, interpretable, and useful for baseline comparison.
2. **Support Vector Machine (SVM):**
    - Constructed hyperplanes to separate classes in higher-dimensional feature space.
    - More effective in capturing complex decision boundaries.

5. Model Training and Validation

- Both datasets were split into training (80%) and testing (20%) sets using scikit-learn's train_test_split.
- Models were trained on the training data and evaluated on the test data.

- Performance metrics included:

    - **Accuracy** → proportion of correct predictions.

    - **Precision** → how many predicted positives were correct.

    - **Recall** → ability to identify actual positives.

    - **F1-Score** → balance between precision and recall.

    - **AUC (Area Under the Curve)** → overall discriminative ability of the model.

6. Tools and Technologies Used

- **Programming Language:** Python

- **Libraries:**

    - Pandas, numpy for data handling

    - matplotlib, seaborn for visualization

    - scikit-learnor model building and evaluation

5. Data Analysis and Results

1. Descriptive Analysis

The analysis begins with exploring the structure of the datasets.

- For the **Diabetes dataset**, the variables such as glucose concentration, body mass index (BMI), age, and insulin levels showed noticeable variation across patients. The correlation heatmap revealed that glucose and BMI had stronger relationships with the likelihood of diabetes, whereas other variables like skin thickness and blood pressure were comparatively weaker predictors.

- For the **Titanic dataset**, the descriptive statistics highlighted clear demographic differences between survivors and non-survivors. The correlation matrix showed survival was moderately related to variables like passenger class, sex, and age. Fare also displayed a mild positive correlation with survival, indicating socio-economic status had an influence on outcomes.

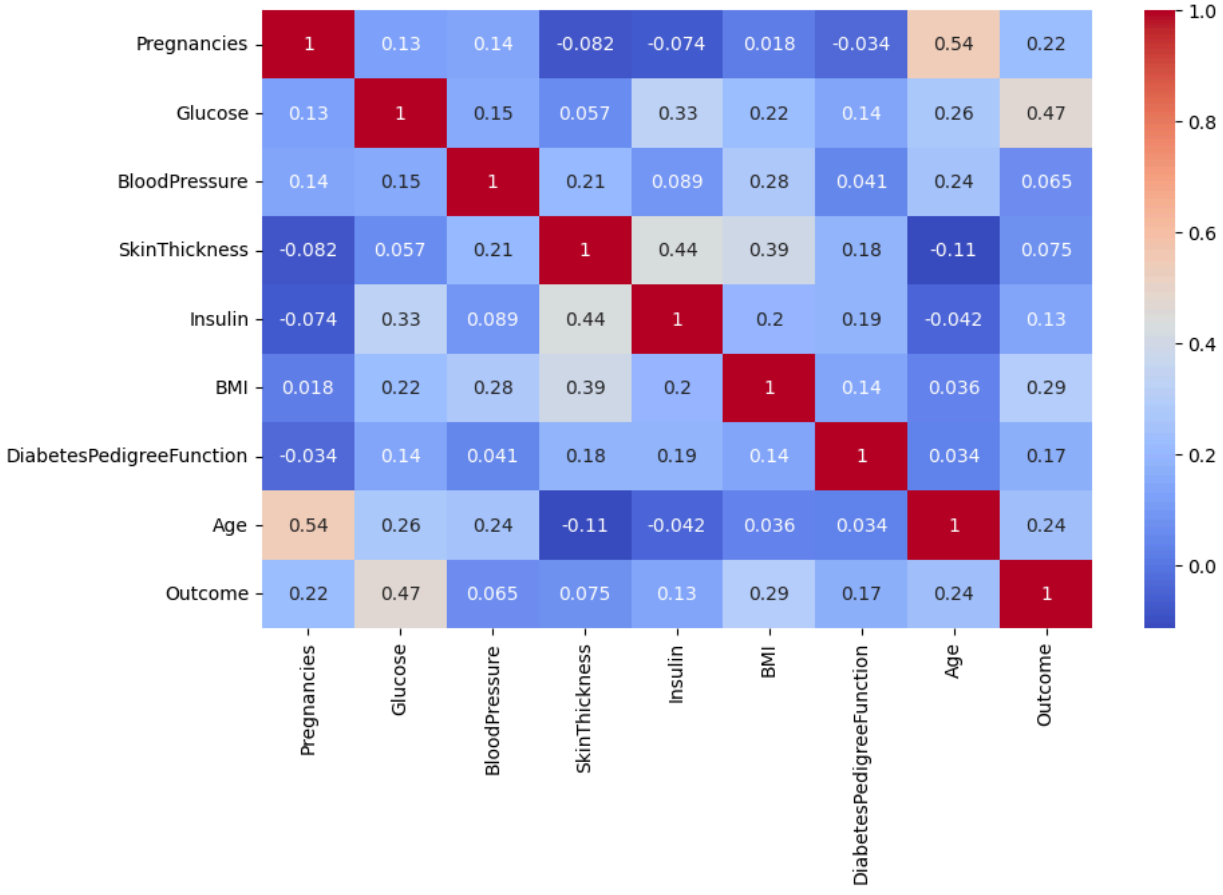

Figure 1. Correlation heatmap of features in the Diabetes dataset

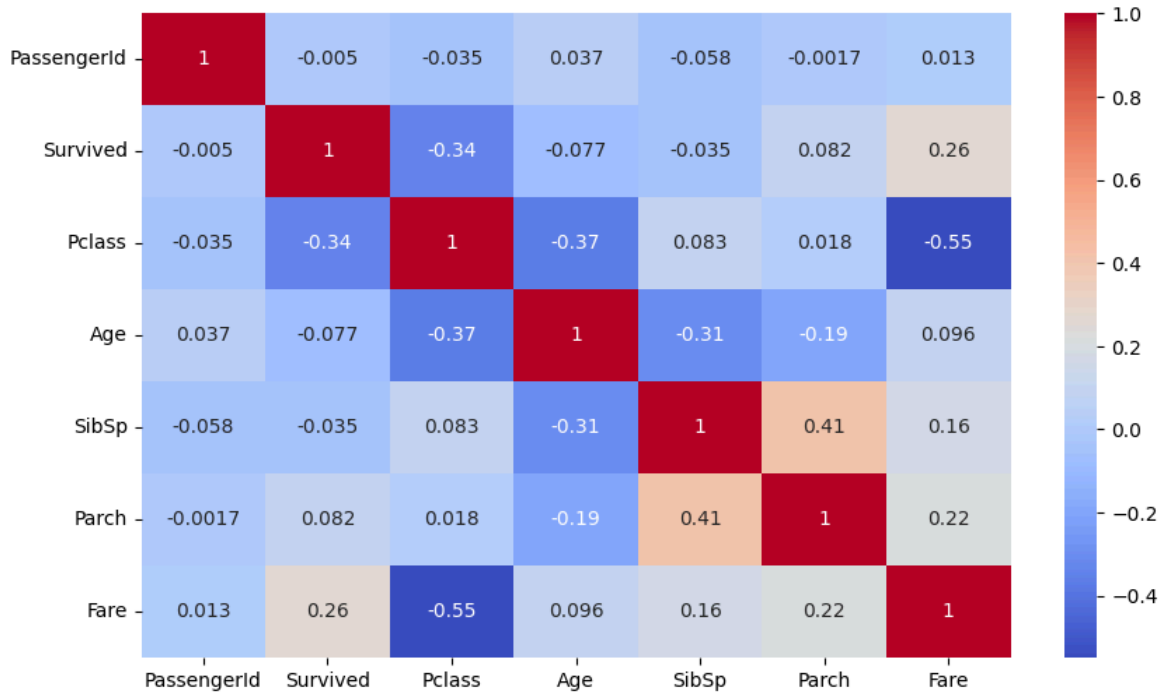

Figure 2. Correlation heatmap of features in the Titanic dataset.

2. Comparative Analysis of Machine Learning Models

Two classifiers **K-Nearest Neighbors (KNN)** and **Support Vector Machine (SVM)** were applied to both datasets. Their performance was evaluated using accuracy, precision, recall, F1-score, and Area Under Curve (AUC) values.

- On the **Diabetes dataset**, both models demonstrated reasonable accuracy, but SVM consistently outperformed KNN. The ROC curve further highlighted this difference, with SVM achieving a higher AUC (0.83) compared to KNN (0.74).

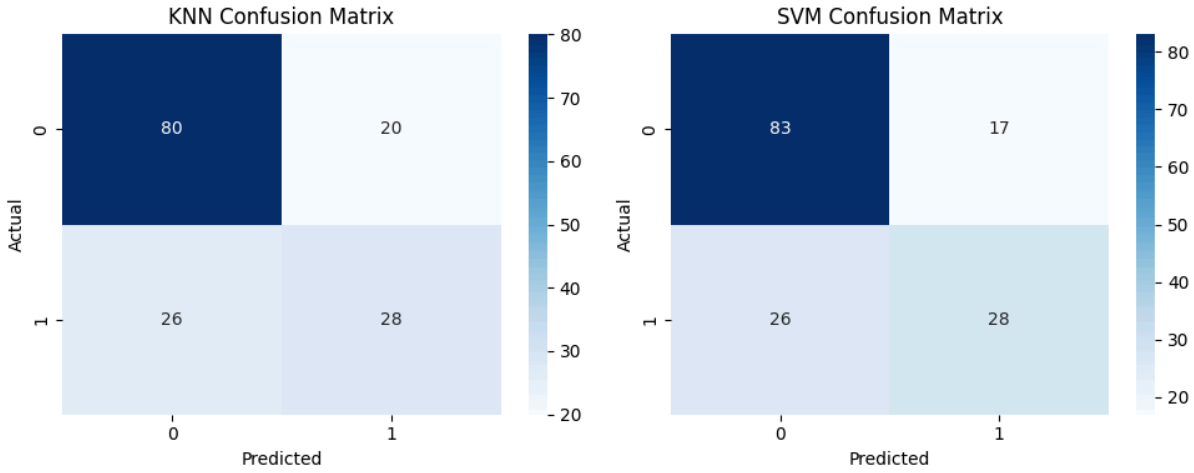

Figure 3. Confusion matrices for Diabetes dataset

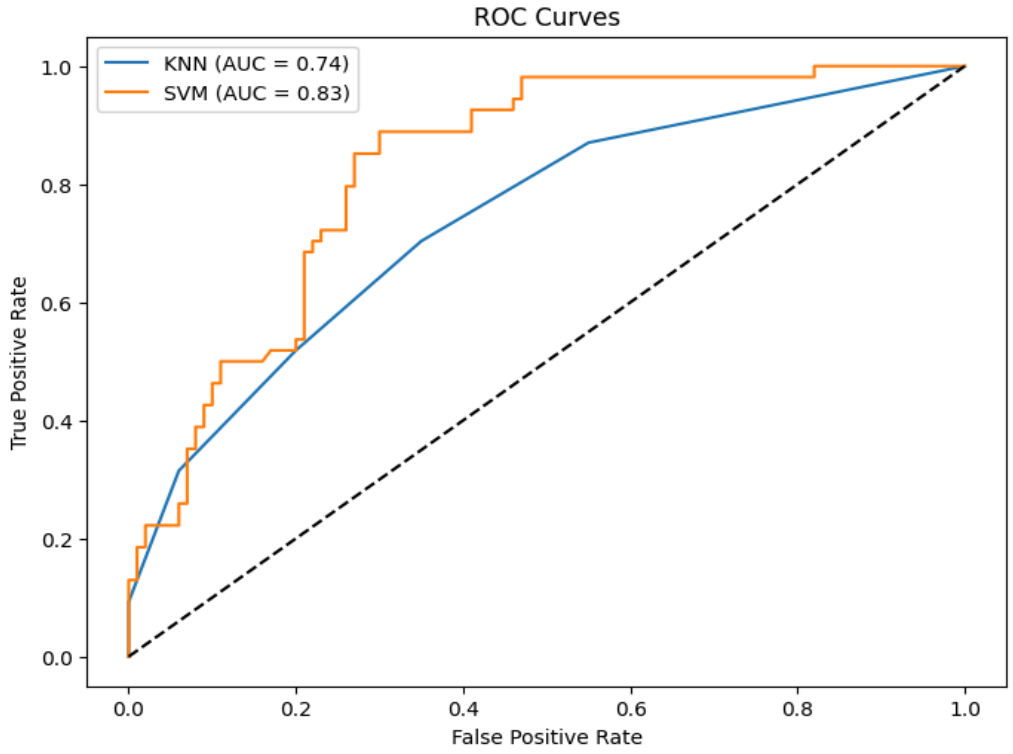

Figure 4. ROC curve comparison of KNN and SVM models on the Diabetes dataset.

- On the **Titanic dataset**, prediction was more challenging due to the variability in data. Still, SVM showed slightly better classification ability with an AUC of = 0.69, compared to KNN's = 0.65.

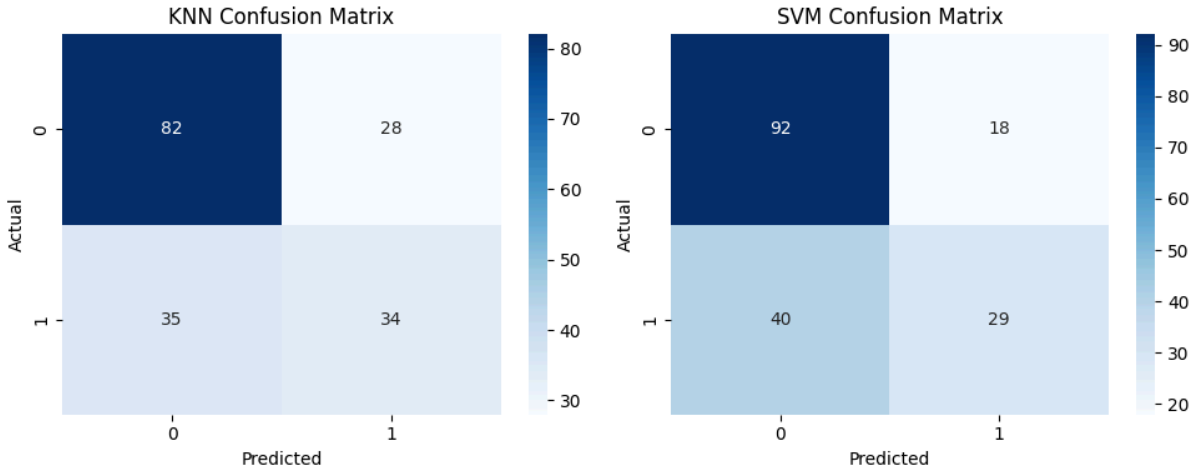

Figure 5. Confusion matrices for Diabetes dataset

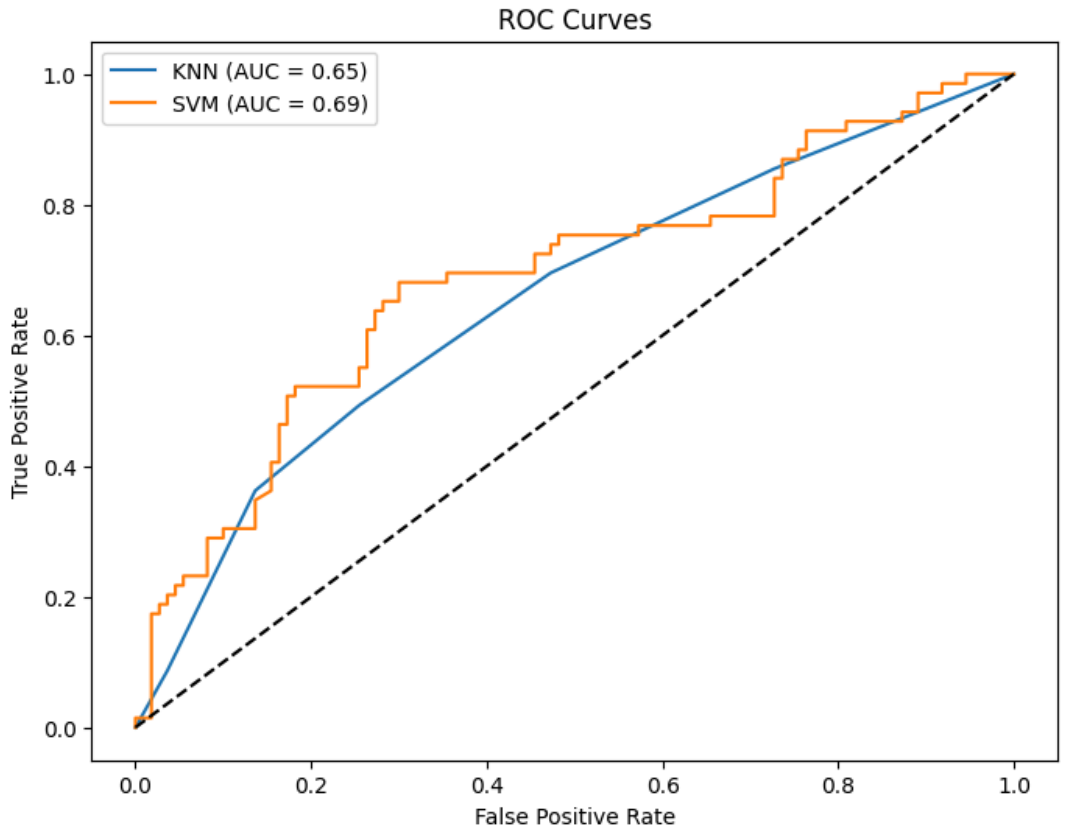

Figure 6. Confusion matrices for Diabetes dataset

Model Comparison Summary

Dataset	Model	Accuracy	Precision	Recall	F1-score	AUC
Diabetes	KNN	0.65	0.55	0.49	0.52	0.74
Diabetes	SVM	0.68	0.62	0.42	0.50	0.83
Titanic	KNN	0.65	0.55	0.49	0.52	0.65
Titanic	SVM	0.68	0.62	0.42	0.50	0.69

6. Conclusion

This project showed how machine learning can be used to make meaningful predictions on real-world problems. By applying the same methods to two very different datasets — one from healthcare (diabetes) and one from history (Titanic survival) — we were able to see both the potential and the limits of simple classifiers.

Key findings from the project:

- In the **diabetes dataset**, glucose level and BMI came out as strong indicators of diabetes.
- In the **Titanic dataset**, social and demographic factors such as passenger class, age, and family size influenced survival rates.
- These insights from the heatmaps lined up with real-world expectations, which added confidence to the analysis.
- Comparing the models, **SVM outperformed KNN** in both cases.
    - On the diabetes dataset: SVM had a higher AUC (~0.83) compared to KNN (~0.74).
    - On the Titanic dataset: SVM still performed better (~0.69 vs ~0.65).
- This highlights that the **choice of algorithm matters**, since different models can produce significantly different results even with the same dataset.

Overall, the study shows how combining careful data exploration with predictive modeling can give both practical and meaningful insights. It also demonstrates that these methods are flexible enough to be applied across domains, whether in healthcare predictions or historical event analysis.