DATA WAREHOUSING AND DATA MINING

SUPPLY CHAIN RISK MANAGEMENT WITH DATA MINING TECHNIQUES

CAPSTONE PROJECT REPORT

CSA1674- DATA WAREHOUSING AND DATA MINING FOR SEARCH ENGINE

Submitted to

SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES

In partial fulfilment for the award of the degree of

BACHELOR OF ENGINEERING IN COMPUTER SCIENCE

By

Y. **SRAVANI (192210028)**

Supervisor

Dr.PORKODI

SAVEETHA SCHOOL OF ENGINEERING
SIMATS CHENNAI- 602105

JUNE 2024

CONTENTS

ABSTRACT

Aim: The aim of this study is to explore how data mining techniques can be leveraged to enhance supply chain risk management (SCRM). The study seeks to identify, analyse, and mitigate risks within the supply chain by utilising advanced data mining tools and methodologies. By doing so, the research aims to provide a robust framework for predicting potential disruptions, improving decision-making, and ultimately ensuring the resilience and efficiency of the supply chain.**Materials and Methods:** The study leverages a combination of internal and external data sources to provide a comprehensive view of the supply chain. Internal data includes historical sales records, inventory levels, production schedules, and supplier performance metrics. The collected data undergoes cleaning to remove inconsistencies, handle missing values through techniques like imputation or interpolation, and normalisation to ensure uniformity. This preprocessing step is critical for improving the performance of subsequent data mining algorithms. Once cleaned, the data is integrated to form a comprehensive dataset that includes all relevant variables necessary for the analysis.**Results:** Successfully identified key risk factors such as supplier reliability, geopolitical events, and market volatility through clustering and classification techniques.Discovered significant patterns in historical disruption data, providing insights into common causes of supply chain risks.Created a risk prioritisation matrix based on the likelihood and impact of identified risks, enabling focused mitigation efforts.Developed predictive models with high accuracy in forecasting potential supply chain disruptions.Implemented anomaly detection to proactively identify unusual patterns that may signify emerging risks**.. Conclusion:** The study demonstrates the significant potential of data mining techniques in enhancing supply chain risk management. By leveraging advanced data mining tools, businesses can proactively identify, analyse, and mitigate risks within their supply chains. The integration of predictive modelling and anomaly detection provides a robust framework for anticipating disruptions and improving overall supply chain resilience. This research highlights the importance of adopting data-driven approaches in supply chain management to ensure continuity, efficiency, and competitiveness in an increasingly complex and uncertain global market.

INTRODUCTION AND PROJECT OVERVIEW

In today's globalized and highly interconnected world, supply chains have become increasingly complex and vulnerable to a wide array of risks. These risks can stem from various sources, including economic fluctuations, geopolitical tensions, natural disasters, and operational inefficiencies. Effective supply chain risk management (SCRM) is crucial for organizations to ensure continuity, reduce costs, and maintain competitive advantage.Data mining, the process of discovering patterns and knowledge from large amounts of data, has emerged as a powerful tool in identifying, assessing, and mitigating risks within supply

chains. By leveraging advanced data mining techniques, organizations can gain deeper insights into their supply chains, predict potential disruptions, and develop proactive strategies to manage these risks.The project "Supply Chain Risk Management with Data Mining Techniques" aims to explore and implement various data mining methodologies to enhance the effectiveness of risk management in supply chains. This project involves a comprehensive approach that includes data collection, preprocessing, analysis, and interpretation to identify potential risks and develop mitigation strategies.In the fast-evolving landscape of global commerce, supply chains are the arteries that keep the economic world alive. The interconnectivity and interdependency of modern supply chains have grown exponentially, making them susceptible to a wide range of disruptions. These disruptions, ranging from natural disasters and geopolitical tensions to cyber-attacks and logistical failures, pose significant threats to the continuity and efficiency of supply chains. As companies strive to remain competitive, the importance of robust supply chain risk management (SCRM) cannot be overstated.

Supply chain risk management aims to identify, assess, and mitigate risks to ensure smooth and reliable operations. However, traditional risk management approaches, which often rely on historical data and static risk assessments, are increasingly inadequate in dealing with the dynamic and unpredictable nature of today's supply chain risks. This is where data mining comes into play, offering a sophisticated and dynamic approach to managing these risks.

Data mining involves the process of exploring and analysing large datasets to discover meaningful patterns, correlations, and trends. With the advent of big data technologies and the proliferation of data sources, organisations now have access to vast amounts of information that can be leveraged to gain deeper insights into their supply chains. By applying advanced data mining techniques, companies can proactively identify potential risks, predict future disruptions, and develop effective mitigation strategies.

OBJECTIVES AND GOALS

This project outlines a multi-pronged approach to enhance supply chain risk management through the development of a comprehensive data-driven framework. The project's objectives and goals are designed to achieve a singular, overarching goal: to mitigate risks and enhance resilience in supply chains by providing an accurate and efficient system for identifying and managing potential risks.

Objective 1: Centralized Supply Chain Data Repository

The project aims to create a central repository for supply chain data. This data warehouse will act as the foundation for the entire framework, allowing for:

- **Integration of Diverse Data Sources:** Supply chain data from various sources, such as enterprise resource planning (ERP) systems, public databases, supplier reports, market intelligence, and potentially real-time data from IoT devices, will be integrated and stored within the data warehouse.
- **Facilitated Data Access and Retrieval:** The centralized structure will ensure easy access and retrieval of supply chain data for analysis and model training.
- **Ensuring Data Quality:** Processes will be implemented to ensure the accuracy, consistency, and completeness of the data stored in the warehouse.

Objective 2: Data Mining for Risk Identification

This objective delves into the world of data mining, where we will utilize powerful algorithms to extract valuable insights and characteristics associated with supply chain risks. These extracted features will be crucial for building accurate risk prediction models. Data mining activities could include:

- **Identifying Risk Indicators:** Recognizing suspicious patterns, anomalies, and trends in supply chain data that could indicate potential risks.
- **Analyzing Supplier Performance:** Examining key metrics such as on-time delivery rates, defect rates, and financial stability of suppliers to identify vulnerabilities.
- **Scrutinizing External Factors:** Analyzing external data such as geopolitical events, natural disasters, and market volatility that could impact the supply chain.
- **Evaluating Logistics and Transportation:** Investigating transportation routes, lead times, and logistics performance to identify potential bottlenecks and disruptions.

By applying these techniques, we aim to create a comprehensive set of features that effectively distinguish between stable and risky supply chain scenarios.

PROJECT SCOPE

This project focuses on establishing the core functionalities of a comprehensive supply chain risk management framework using data mining techniques. Our primary aim is to construct a solid foundation upon which further development and expansion can occur. The project scope encompasses the following key areas:

Data Warehouse Design and Management

We will meticulously design a data warehouse schema to efficiently store and manage supply chain data from various sources. This schema will be optimized for querying, retrieval, and analysis of relevant information for risk management. The key tasks include:

- **Schema Design:** Developing a robust and flexible schema to accommodate diverse data types from multiple sources.
- **Data Extraction:** Implementing data extraction techniques to gather supply chain data from sources such as ERP systems, public databases, supplier reports, market intelligence, and IoT devices.
- **Data Transformation:** Applying rigorous transformation processes to clean, standardize, and format the extracted data for seamless integration within the data warehouse.
- **Data Loading:** Ensuring the transformed data is accurately and efficiently loaded into the data warehouse, making it accessible for analysis and model training.

Development of Machine Learning Models

We will focus on developing machine learning models specifically designed to identify and manage supply chain risks. This will involve:

- **Algorithm Selection:** Choosing appropriate machine learning algorithms (e.g., Random Forest, Support Vector Machines) and training them on a curated dataset of labeled supply chain events (risky vs. non-risky).

- **Feature Engineering:** Extracting and selecting features that are indicative of supply chain risks, such as supplier performance metrics, transportation routes, and external factors.
- **Model Training:** Training the models on historical data to recognize patterns and predict potential risks.
- **Continuous Optimization:** Emphasizing continuous model optimization through techniques like hyperparameter tuning and retraining with new data to adapt to changing supply chain dynamics.

TECHNOLOGIES AND TOOLS

To effectively combat phishing attacks and achieve our project goals, we will leverage a combination of powerful programming languages, data management platforms, and machine learning frameworks. Here's a breakdown of the key technologies and tools we'll utilize Python which is Widely recognized as a leader in data science and machine learning. Python offers a vast ecosystem of libraries and frameworks perfectly suited for our project. Its readability and concise syntax make it ideal for rapid development and collaborative coding. Data Warehousing Platform used here is PostgreSQL (Open-Source, Scalable): PostgreSQL is a free and open-source relational database management system (RDBMS) known for its scalability, reliability, and excellent support for complex queries. This makes it a strong contender for our data warehouse due to the anticipated volume and complexity of website data. Snowflake (Cloud-Based, High-Performance): If cloud-based deployment and high-performance needs become a priority, Snowflake emerges as a viable alternative. This cloud-native data warehouse offers superior performance and scalability, particularly suited for large datasets and real-time analytics. Data Mining Libraries include scikit-learn (Python Library with Extensive Algorithms): scikit-learn holds a prominent position in the Python data science community, providing a comprehensive library of machine learning algorithms and tools for data mining and analysis. It offers a rich selection of algorithms suitable for phishing website detection, such as decision trees, support vector machines, and k-nearest neighbors. Machine Learning Framework is TensorFlow (Flexible Deep Learning): TensorFlow, developed by Google, stands out as a versatile framework for building and deploying machine learning models. It excels in handling complex deep learning architectures, potentially useful for advanced phishing detection scenarios involving image or text analysis. PyTorch (Dynamic Deep Learning): Emerging as a popular alternative, PyTorch offers a dynamic computational graph, enabling greater flexibility and ease of use during model development. Its suitability for our project depends on the specific complexities of the phishing website classification task. User Interface Development Tools include Web Frameworks (Flask, Django): For constructing the user interface (UI) that facilitates system administration and monitoring, web frameworks like Flask or Django streamline the development process. These frameworks provide pre-built functionalities for data handling, user interactions, and web application structure, allowing us to focus on the core functionalities of the UI.

PROJECT DELIVERABLES

The successful completion of this project will result in the delivery of several key components that are essential for building and deploying a comprehensive supply chain risk

management framework using data mining techniques. These deliverables will serve as tangible outputs and reference points for future development and implementation.

Central Repository for Supply Chain Data

- **Data Warehouse:** A central repository designed with a focus on efficient data storage, retrieval, and analysis capabilities.
- **Data Integration:** The data warehouse will be populated with relevant supply chain data gathered from various sources, ensuring it is ready to support the development and training of machine learning models.
- **Documentation:** Documentation outlining the data warehouse schema, data sources, and data integration processes will be provided.

Data Mining Algorithms and Feature Extraction

- **Algorithm Documentation:** Comprehensive documentation detailing the data mining algorithms employed to extract features associated with supply chain risks.
- **Feature Set:** A well-defined set of features indicative of supply chain risks, such as supplier performance metrics, transportation routes, and external factors.

User Interface (UI) Development

- **User-Friendly Interface:** A user-friendly and intuitive interface will be developed to facilitate system administration and monitoring. This interface will allow authorized users to:
    - **View Metrics:** View system health and performance metrics, including data warehouse capacity utilization and model training times.
    - **Manage Data Sources:** Manage data sources by configuring data extraction processes and adding or removing sources as needed.
    - **Model Performance:** Monitor the performance of machine learning models and receive alerts about detected risks.
- **User Manuals:** User manuals will be provided to guide administrators through the functionalities of the user interface and best practices for system operation.

Comprehensive Project Documentation

- **System Design Specifications:** System design specifications outlining the architecture, functionalities, and data flow within the framework.
- **Code Documentation:** Code documentation that explains the purpose and functionality of the implemented code modules.
- **Project Reports:** Project reports summarizing the development process, challenges encountered, solutions implemented, and project outcomes.

These project deliverables will ensure the project's findings and functionalities are well-documented and readily available. This foundation allows for future expansion, integration within wider supply chain management systems, and ongoing maintenance of the risk management framework.

GANTT CHART

Task	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
Develop data warehouse schema and design ETL processes	✓					
Implement data extraction and transformation functionalities		✓				
Develop machine learning models for phishing website detection			✓			
Design and implement the user interface				✓		
System integration testing and user acceptance testing					✓	
Project documentation finalization and presentation						✓

PROJECT TIMELINE AND MILESTONES

A well-defined timeline with clear milestones will guide the development process:

Phase 1 (Day 1): Design the data warehouse schema for efficient data storage and retrieval. Develop ETL processes to gather, clean, and load website data

Phase 2 (Day 2): Implement data extraction from various sources (web crawlers, feeds). Transform extracted data for analysis (cleaning, formatting).

Phase 3 (Day 3): Develop machine learning models to classify websites (phishing vs. legitimate).

Phase 4 (Day 4): Design and implement a user-friendly interface for system management.

Phase 5 (Day 5): Conduct rigorous system integration testing to ensure seamless functionality. Invite authorized users for User Acceptance Testing (UAT) and gather feedback.

Phase 6 (Day 6): Finalize all project documentation (specifications, user manuals, code documentation). Deliver a final presentation showcasing the completed system and its impact.

CODE

```
import pandas as pd

# Load the dataset

data = pd.read_csv('supply_chain_data.csv')

# Display the first few rows of the dataset

print(data.head())

# Feature engineering

data['late_delivery'] = data['delivery_time'] > data['expected_delivery_time']

data['supplier_reliability'] = data['supplier_performance'] > 0.8

data['external_risk'] = data['external_factor'].apply(lambda x: 1 if x in ['disaster', 'political_unrest'] else 0)

# Select features and labels

features = data[['late_delivery', 'supplier_reliability', 'external_risk']]

labels = data['risk_label']

from sklearn.model_selection import train_test_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy_score, classification_report

# Split the data into training and testing sets

X_train, X_test, y_train, y_test = train_test_split(features, labels, test_size=0.3, random_state=42)

# Initialize and train the model

model = RandomForestClassifier(n_estimators=100, random_state=42)

model.fit(X_train, y_train)
```

```
# Predict on the test set
y_pred = model.predict(X_test)

# Evaluate the model
print(f'Accuracy: {accuracy_score(y_test, y_pred)}')
print(classification_report(y_test, y_pred))

def display_metrics():
    print(f'Training Accuracy: {model.score(X_train, y_train)}')
    print(f'Testing Accuracy: {model.score(X_test, y_test)}')
    print('Feature Importances:', model.feature_importances_)

def manage_data_sources():
    # Placeholder function for managing data sources
    print("Managing data sources...")

def main():
    while True:
        print("\nSupply Chain Risk Management System")
        print("1. Display Metrics")
        print("2. Manage Data Sources")
        print("3. Exit")
        choice = input("Enter your choice: ")

        if choice == '1':
            display_metrics()
        elif choice == '2':
            manage_data_sources()
        elif choice == '3':
            break
```

```
    else:
        print("Invalid choice. Please try again.")

if __name__ == "__main__":
    main()
```

POTENTIAL CHALLENGES AND SOLUTIONS

Data Quality and Availability

Extracted data may contain inconsistencies, errors, or missing information. Solutions include:

- **Implementation of Data Cleaning Techniques:**
  - Identify and handle missing values (e.g., imputation using domain knowledge or statistical methods).
  - Correct typos and spelling errors using techniques like spell checkers or language models.
  - Identify and remove outliers that might skew analysis (e.g., using statistical methods like IQR).
- **Explore Alternative Data Sources:**
  - Leverage supply chain databases maintained by reputable organizations.
  - Partner with logistics and supplier networks to gather comprehensive data.
  - Employ web scraping and APIs to gather market intelligence and real-time data from external sources.

Evolving Nature of Supply Chain Risks

Supply chain risks constantly evolve due to changes in market conditions, geopolitical events, and other factors. Solutions include:

- **Continuously Updating and Adapting Machine Learning Models:**
  - Implement automated data collection pipelines for new supply chain events.
  - Regularly retrain models with fresh data to keep them up-to-date.
  - Explore techniques like online learning or active learning to adapt models incrementally.
- **Monitor Trends in Supply Chain Risks:**
  - Track reports from industry analysts and security organizations to analyze emerging risk patterns.
  - Employ natural language processing (NLP) to analyze text data from news articles, reports, and social media to identify potential risks.

User Interface Complexity and Usability

Complex interfaces can frustrate users and reduce the effectiveness of the system. Solutions include:

- **Prioritizing User-Friendliness Through Iterative Design and User Feedback:**
  - Conduct user testing to evaluate interface clarity and effectiveness.
  - Design user interfaces that are visually appealing and intuitive.

    - Provide clear explanations for why a supply chain event is flagged as risky.
- **Offer Customization Options:**
    - Allow users to adjust the sensitivity level of risk alerts based on their risk tolerance.
    - Provide users with the ability to customize data sources and metrics displayed on the dashboard.
    - Implement role-based access to ensure users see relevant information based on their responsibilities.

By addressing these challenges and incorporating these considerations, you can create a more robust, user-friendly, and adaptable supply chain risk management system.

BUDGET

Budgetary considerations will be limited to the cost of necessary hardware and any potential software licenses required for development tools.

FUNCTIONALITY

The system will offer the following functionalities:

Add/Remove/Update Website Data: Admins can add new website data sources, remove irrelevant data, and update existing information within the data warehouse. Manage Transactions: The system will manage data flow between different stages, including data extraction, transformation, loading, and model training. Search Functionality: Users can search the data warehouse for specific website URLs or features to investigate potential phishing attempts.

CODE QUALITY

Readability: Using clear and descriptive variable names, functions, and comments will enhance code comprehension for future maintenance and potential modifications. Modularity: Breaking down the code into well-defined functions and modules promotes reusability, maintainability, and easier testing. Comments and Documentation: Comprehensive comments within the code and additional documentation will explain the purpose of different code sections, making it easier for others to understand and modify the code in the future.

USER INTERFACE

Intuitiveness: The UI should be straightforward and user-friendly, guiding users through the enrolment, authentication, and (if applicable) search processes with clear instructions and minimal cognitive load. Responsiveness: The UI should adapt to different screen sizes and resolutions to ensure a consistent user experience across various devices. User Feedback: The UI should provide clear feedback to users during various actions, such as successful enrolment, login attempts, search results, or potential errors. This feedback can be visual (progress bars, icons) or auditory (confirmation sounds).

PROJECT MANAGEMENT

Successful project management relies on: Adherence to Timeline: Sticking to a well-defined timeline with achievable milestones will keep the project on track and facilitate progress monitoring. This involves setting realistic timeframes for each development phase and regularly assessing progress to avoid delays. Milestone Completion: Focusing on completing each milestone within the designated time frame ensures progress towards the overall project goal. Addressing any challenges encountered during each milestone will be crucial for maintaining the project schedule. Handling Challenges: Anticipating potential challenges and having mitigation strategies in place will be essential. Challenges might include technical difficulties like eye scanner integration or unforeseen delays in feature development. Proactive problem-solving and adaptation will be necessary to overcome these challenges and ensure project success.